CASP16 meeting, December 1-4, 2024, Punta Cana

Improving protein structure prediction with optimized sequences

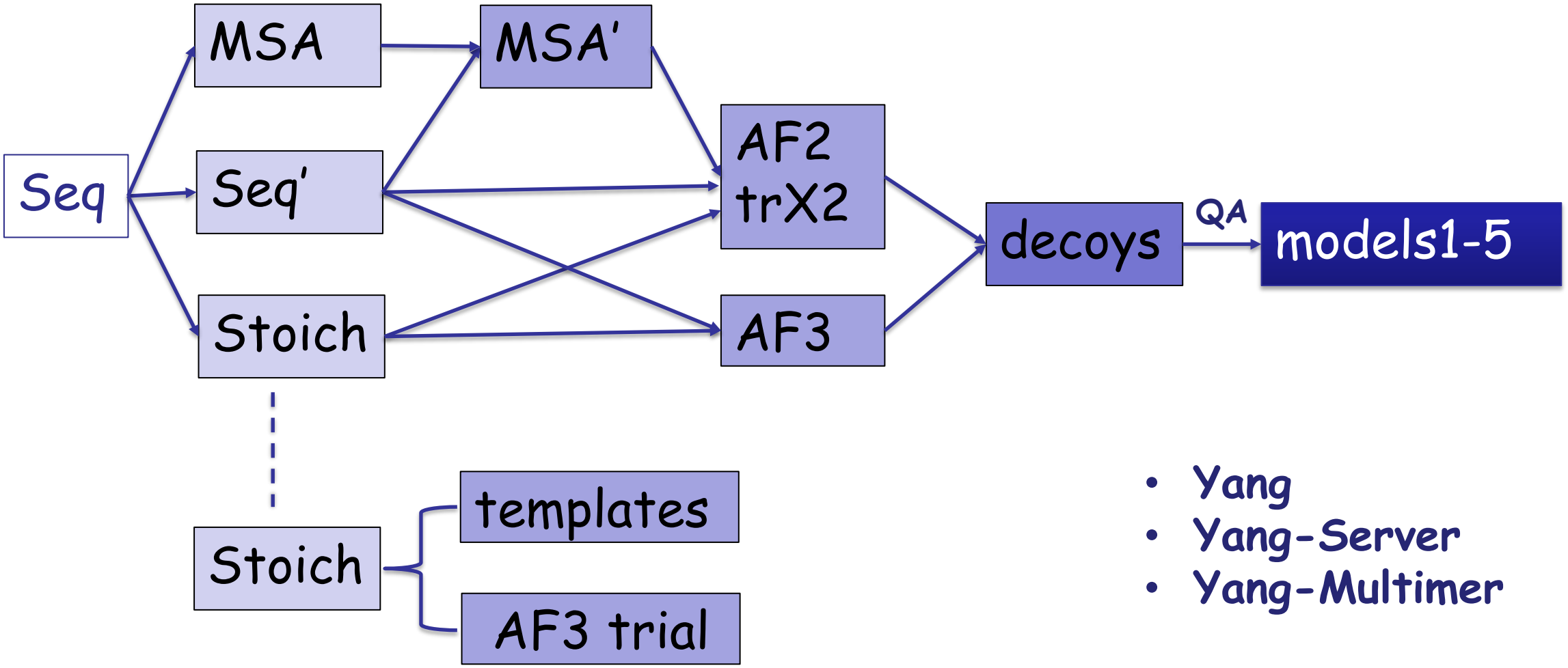
Jianyi Yang

Shandong University

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

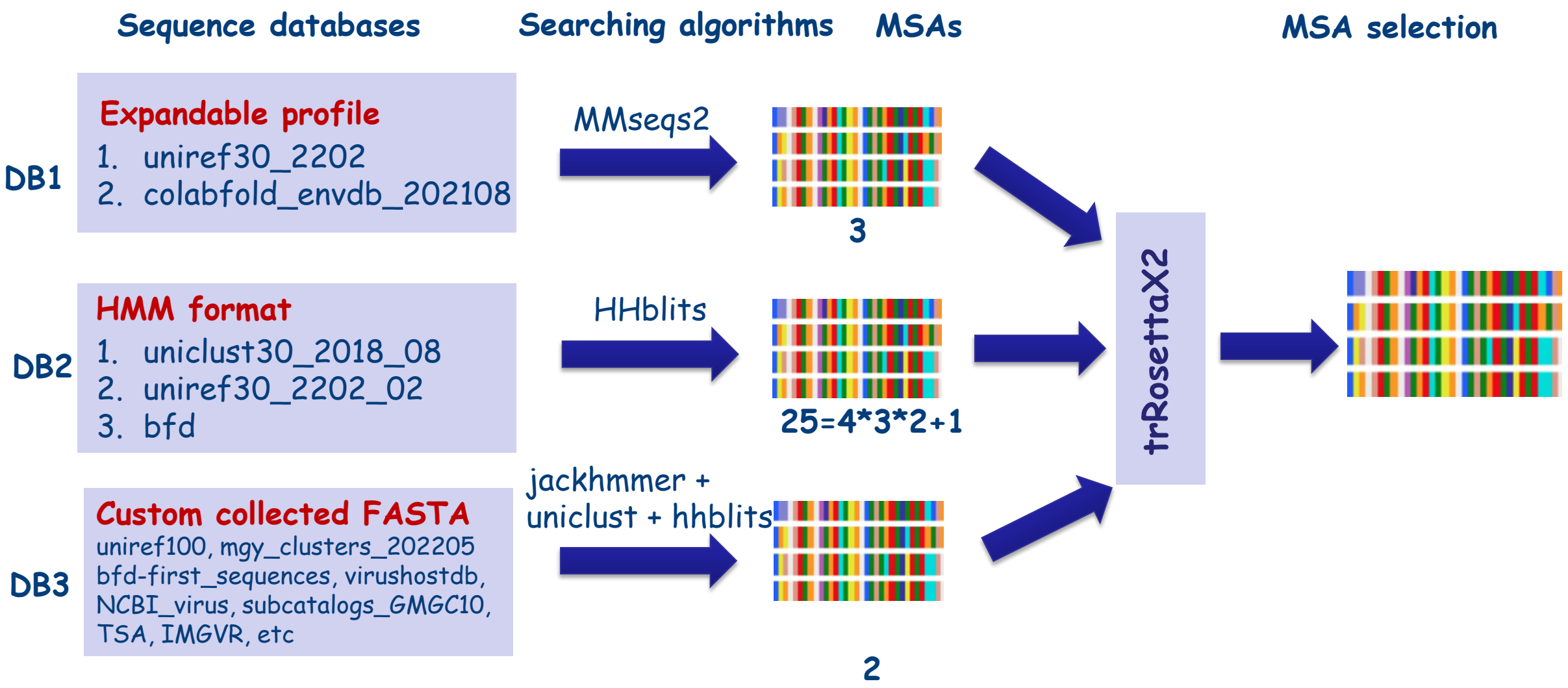
CONTENTS

CASP16 structure prediction strategy

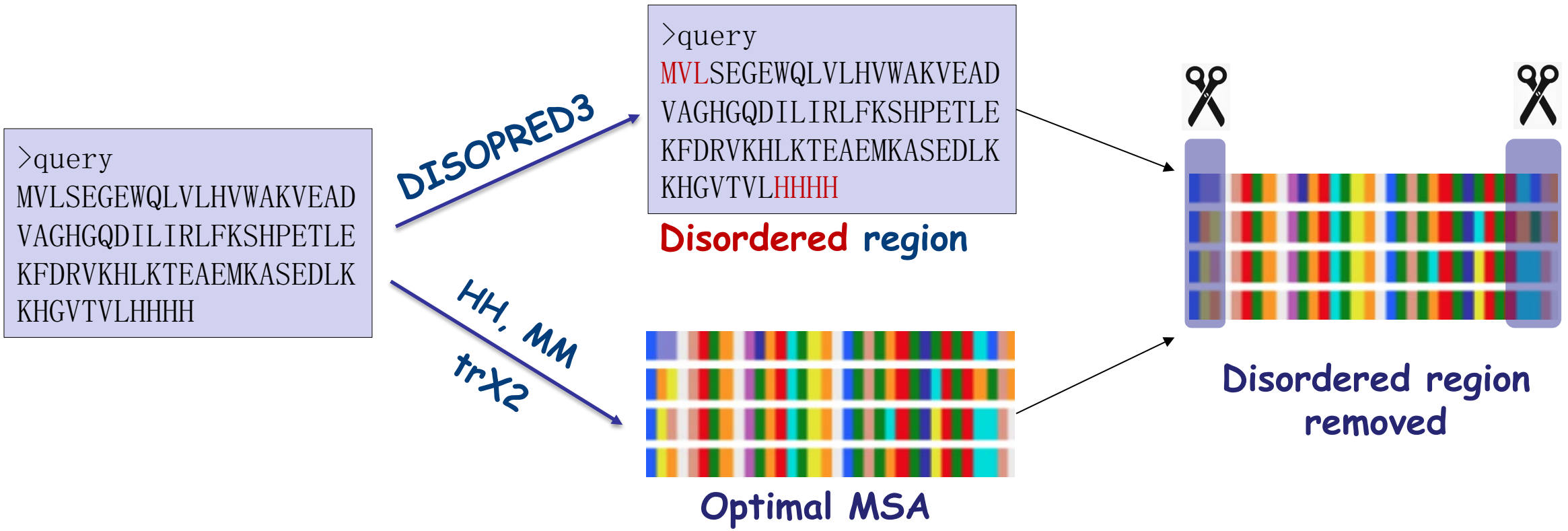


- Yang
- Yang-Server
- Yang-Multimer

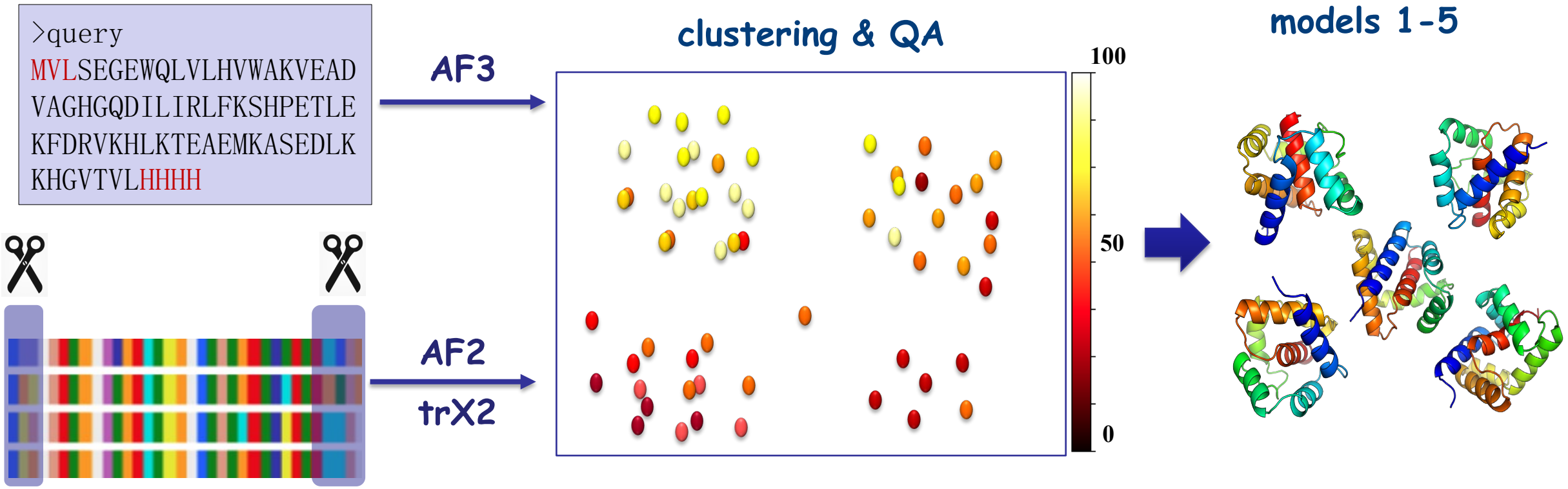
MSA generation & selection



Optimized sequence and MSA



Monomer structure prediction



Note: Monomer model for multimeric targets are from multimeric modeling, unless the multimeric model is in low confidence

Multimer structure prediction

>A
MVLSEGEWQLVLHVWA
KVEADVAGHGQDILIR
LDRVKHLKTEAEMKAS
EDLKKHGVTVLHHHH

>B
SQLTTRVWTVAIDGYR
TIVVEGETGIVAINSF
GTPSQTFGDKPVVAVV
ASIDHLDHT

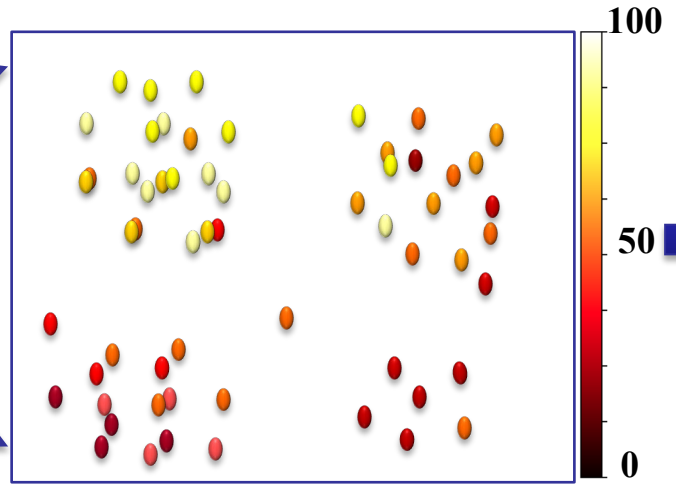
>A
MVLSEGEWQLVLHVWA
KVEADVAGHGQDILIR
LDRVKHLKTEAEMKAS
EDLKKHGVTVLHHHH

>B
SQLTTRVWTVAIDGYR
TIVVEGETGIVAINSF
GTPSQTFGDKPVVAVV
ASIDHLDHT

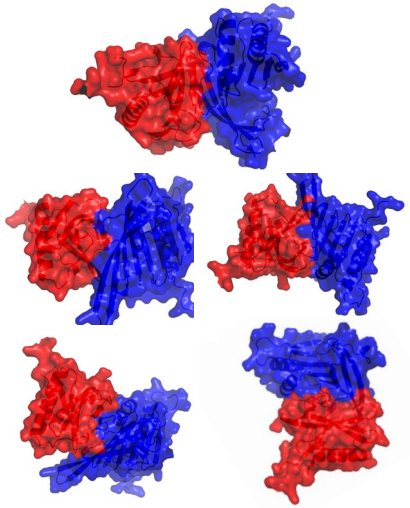
AFM

AF3

clustering & QA

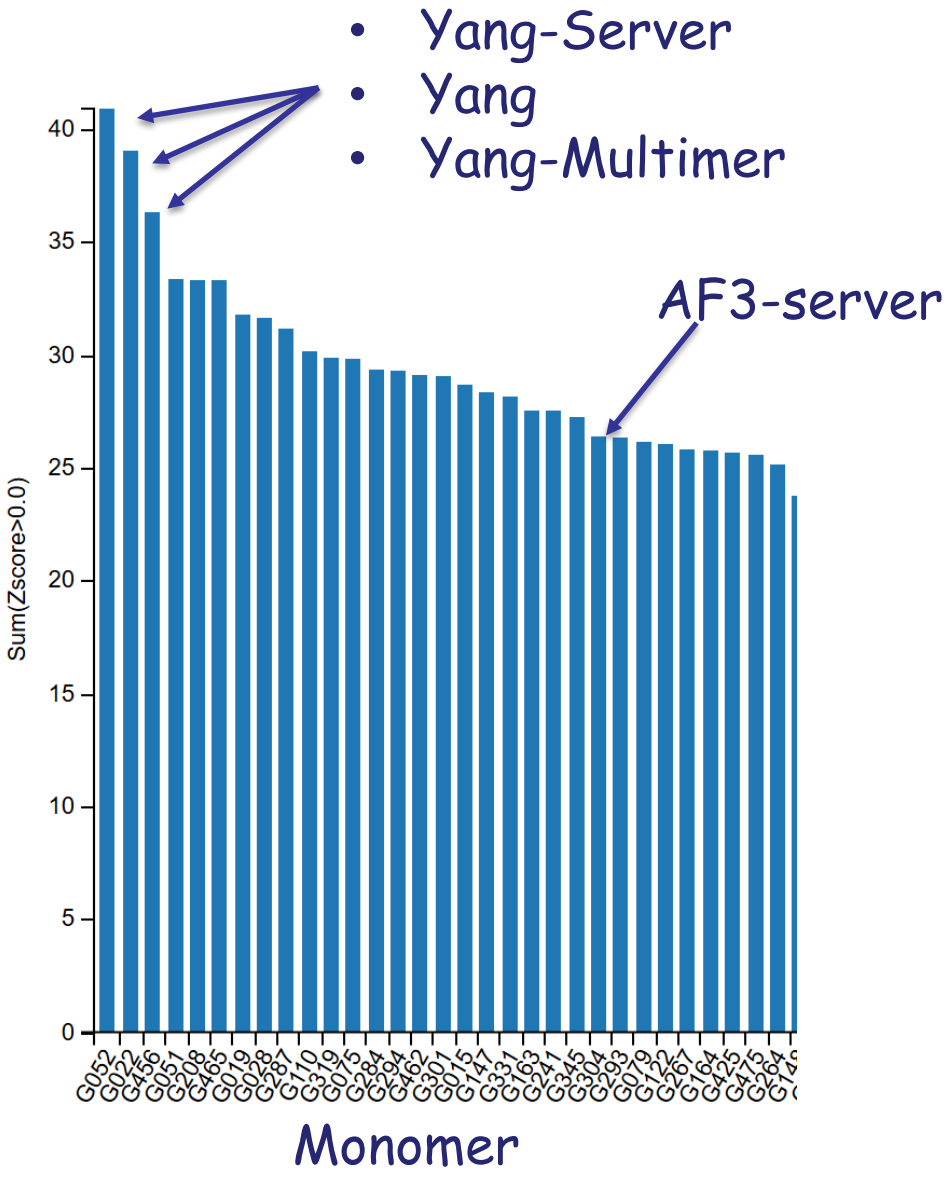


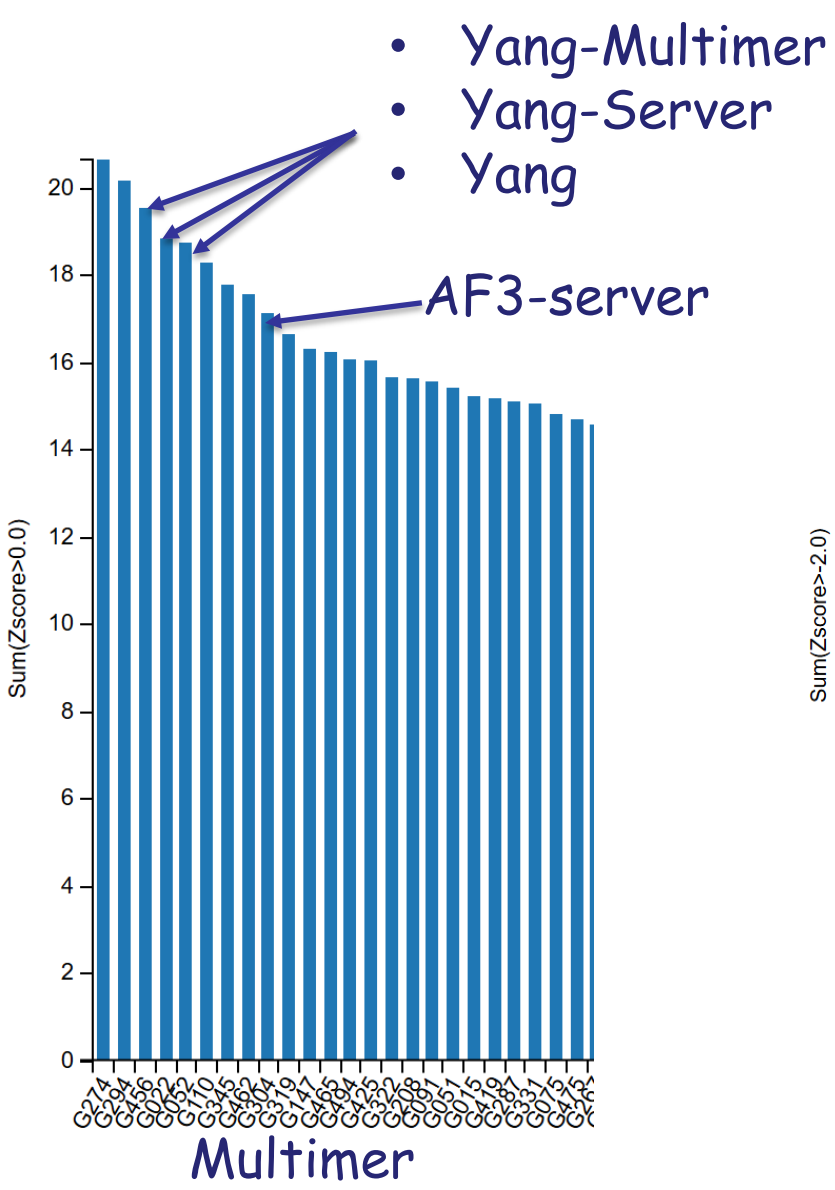
models 1-5

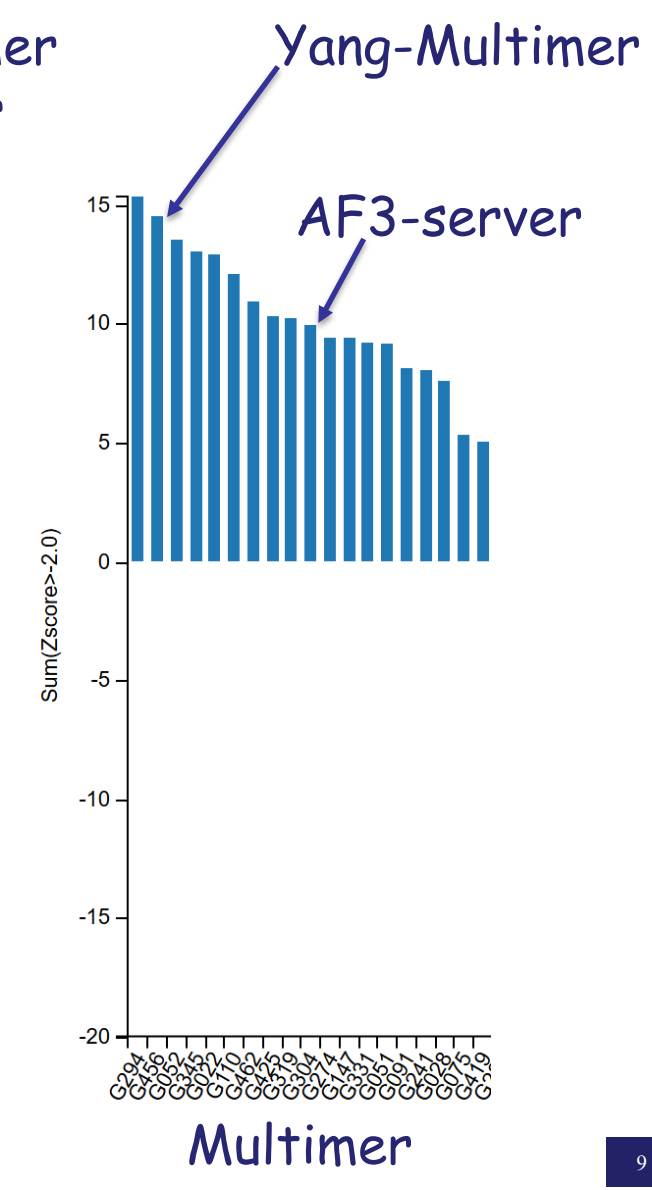


CONTENTS

Official ranking

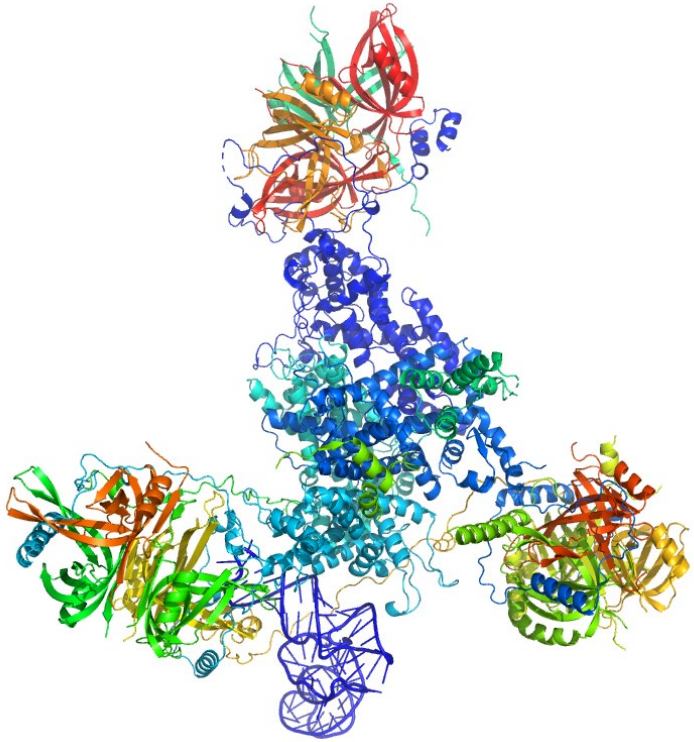




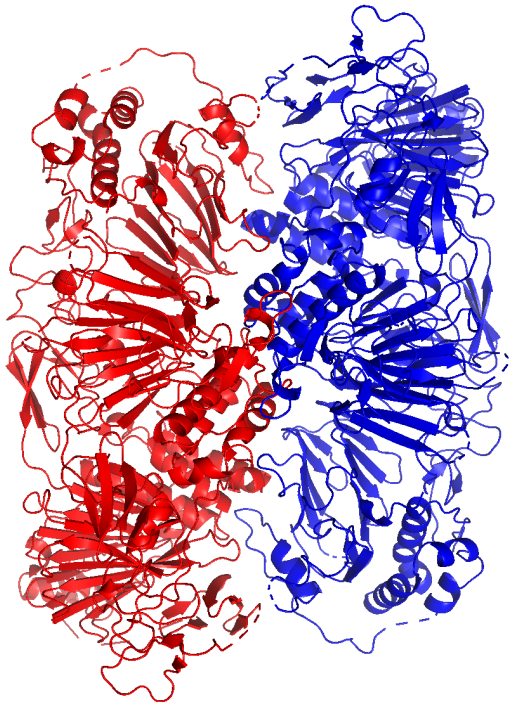


Case studies

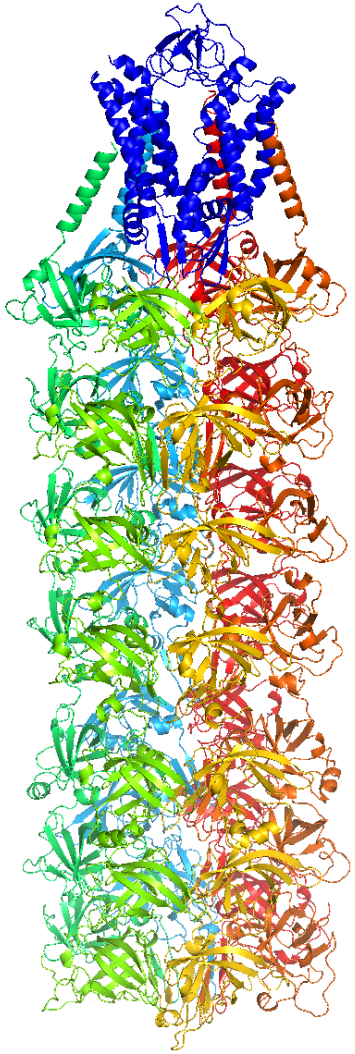
M1271



T1269 (Filament)



H0258, H1258

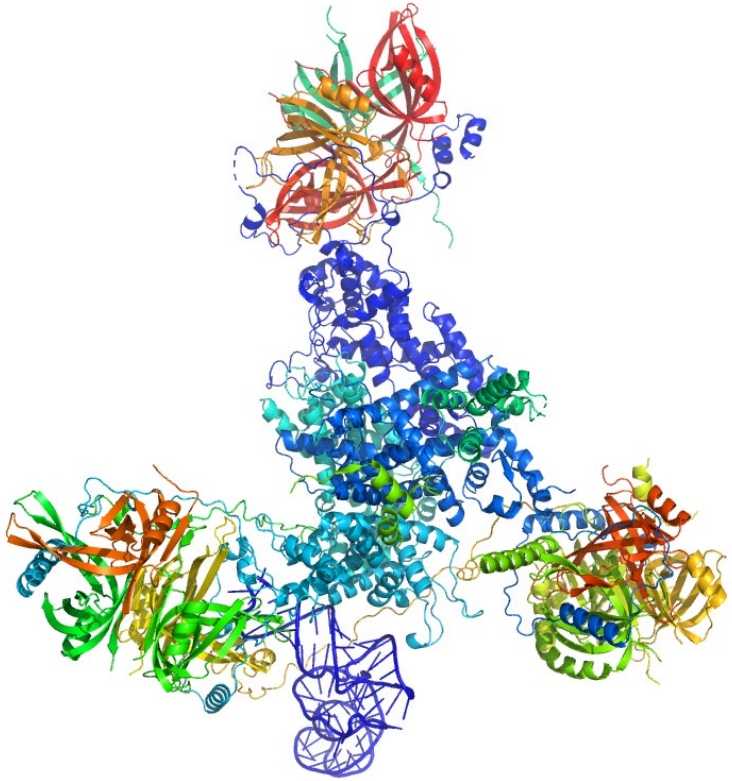


T0257o

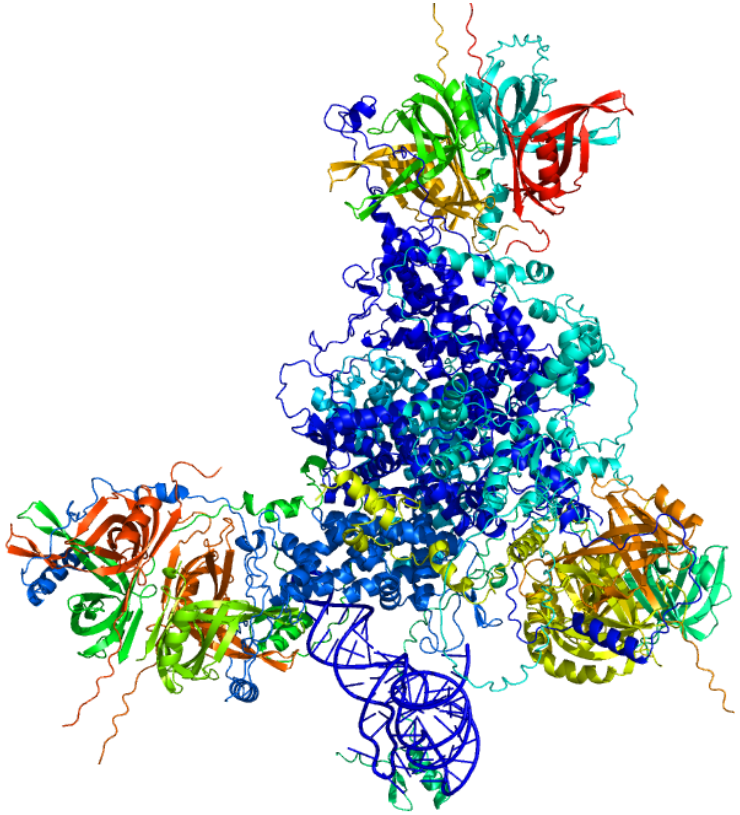
H0227, H1227

What went right?

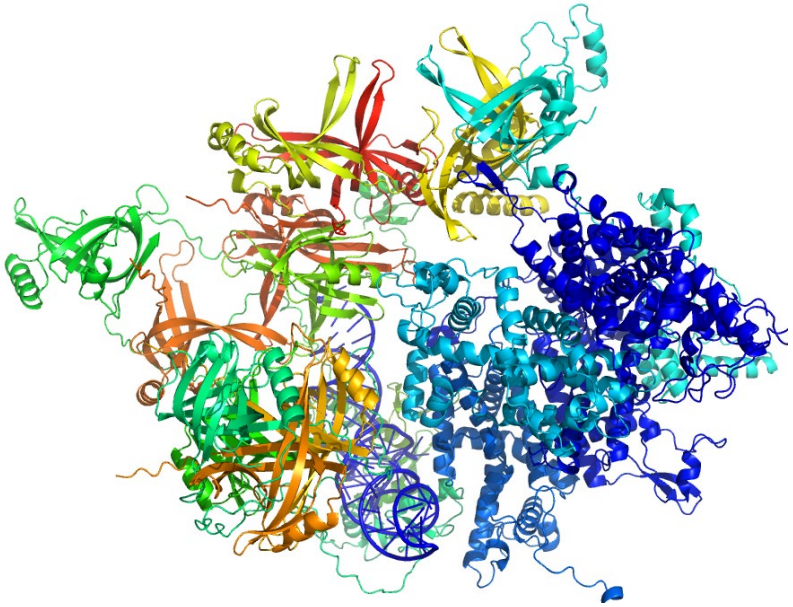
M1271



native

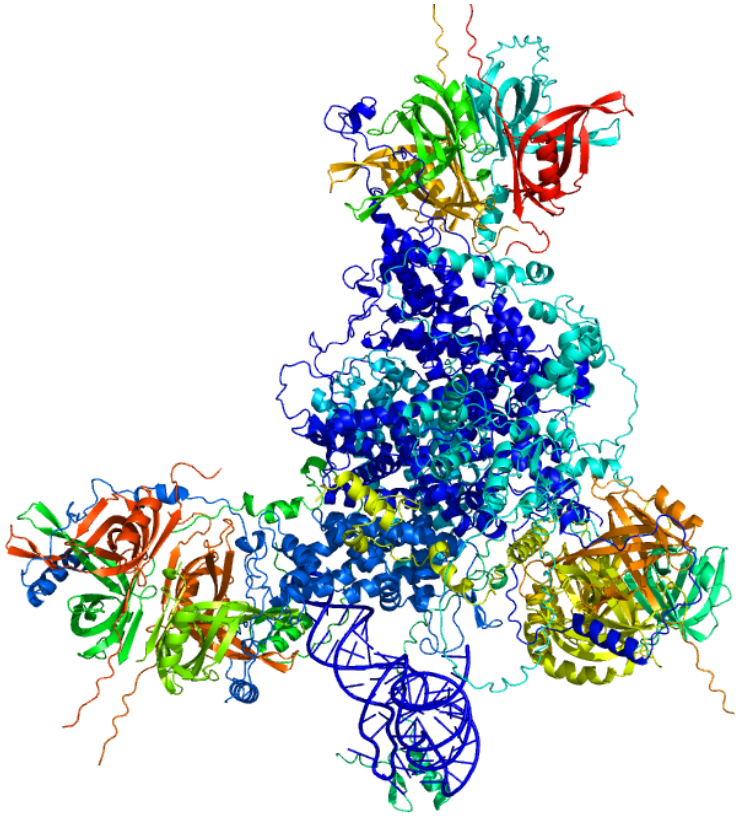


our model based on AF3



AF3-server model
(group 304)

What went right?



our model based on AF3

M1271 5990 tokens >5000

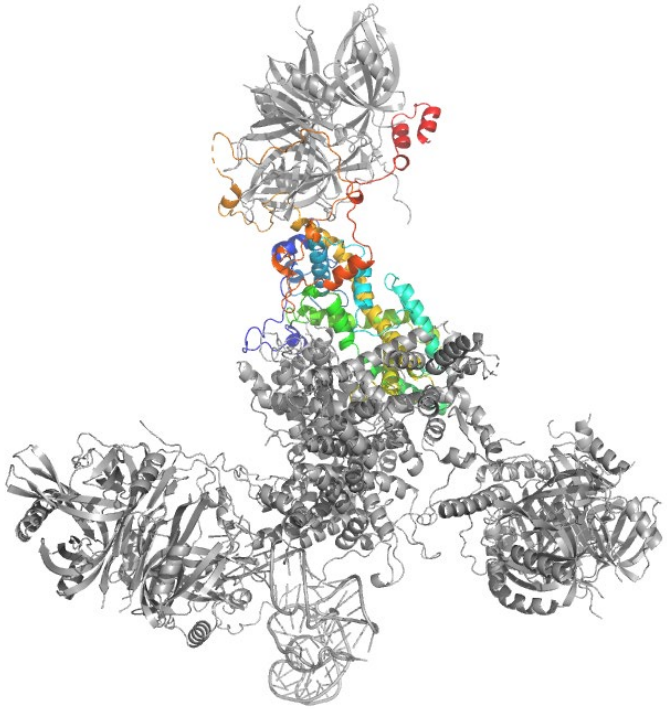
remove disordered regions for subunits:
s1, s2, s4, s5, s6, s7

Number of remaining tokens : ~3700

Number of tokens in native structure: 3276

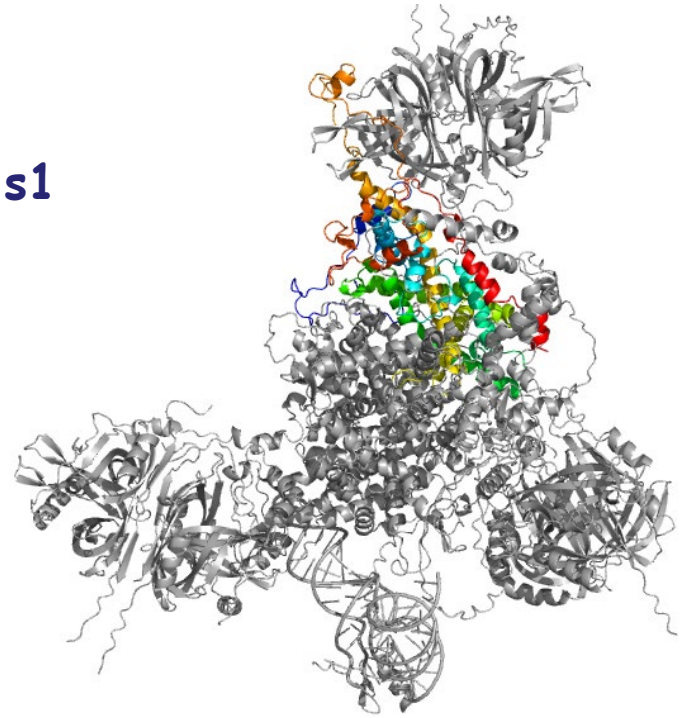
Good models for: s1, s2, s3, s7, s8

What went right?



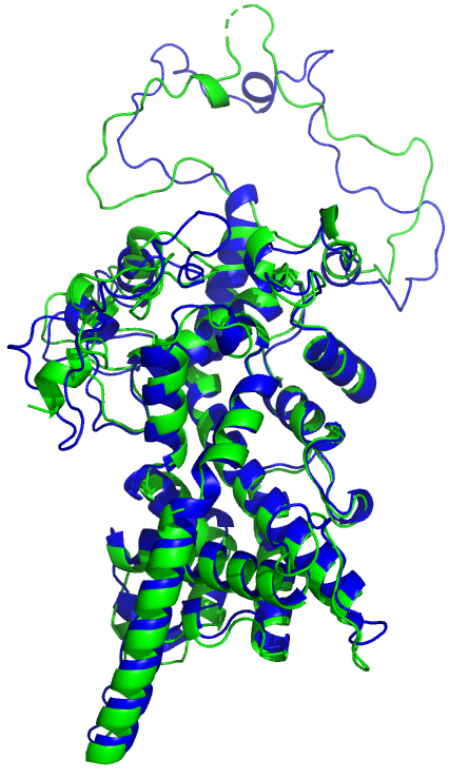
native

s1



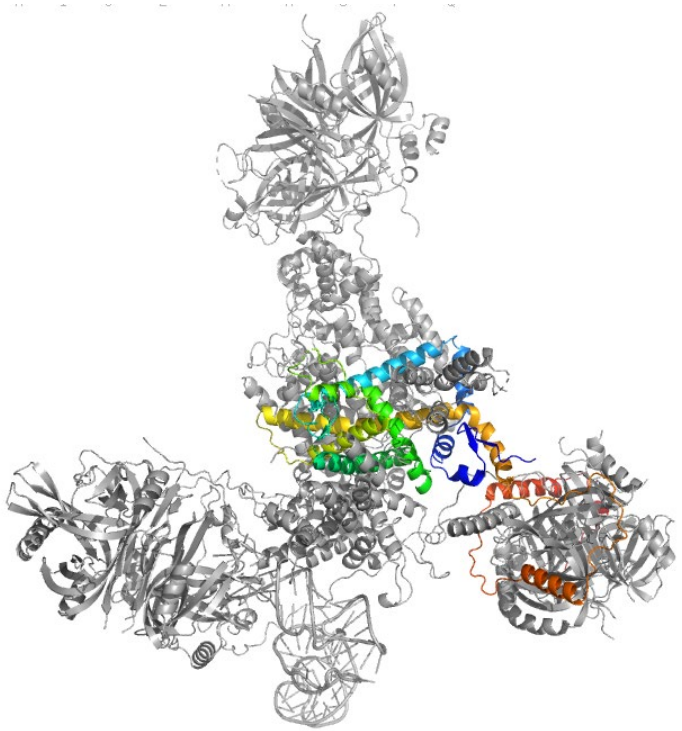
our model

T1271s1-D1, GDT-TS: ~70



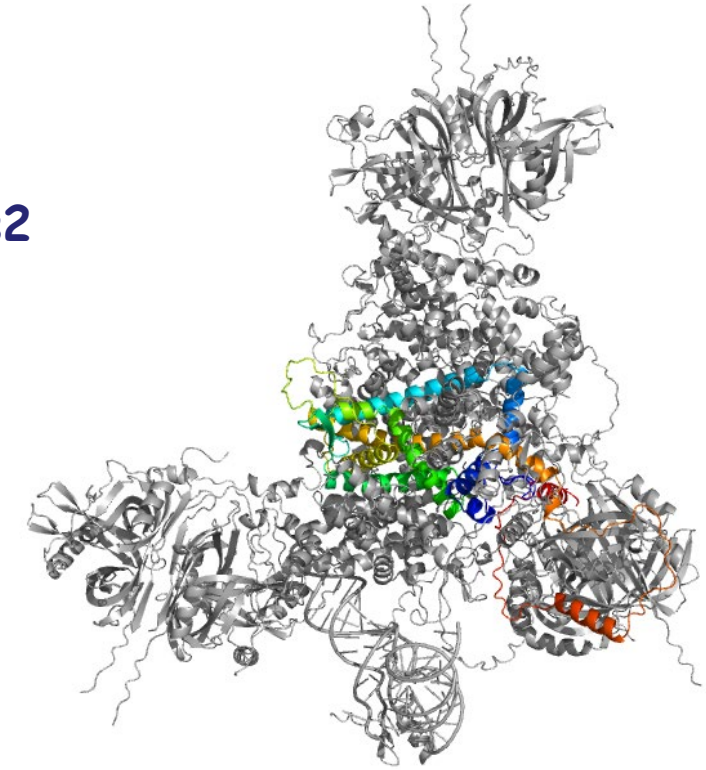
Blue: model
Green: native

What went right?



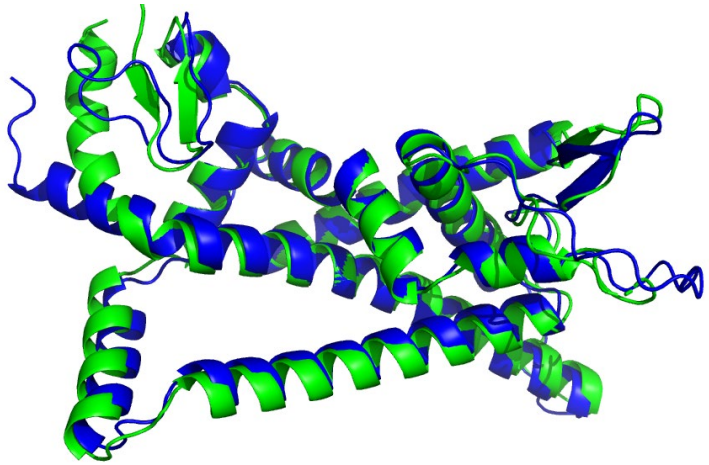
native

s2

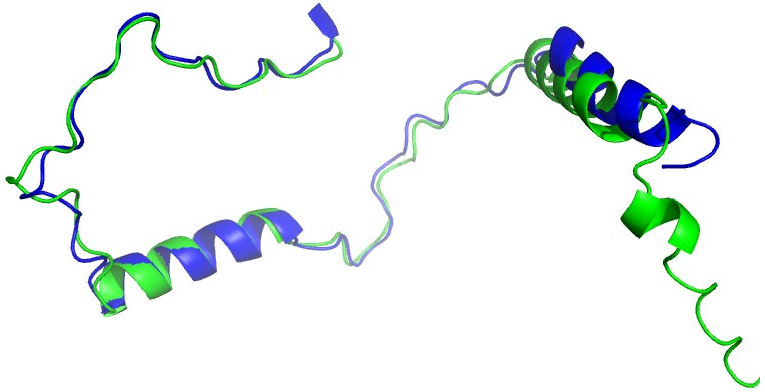


our model

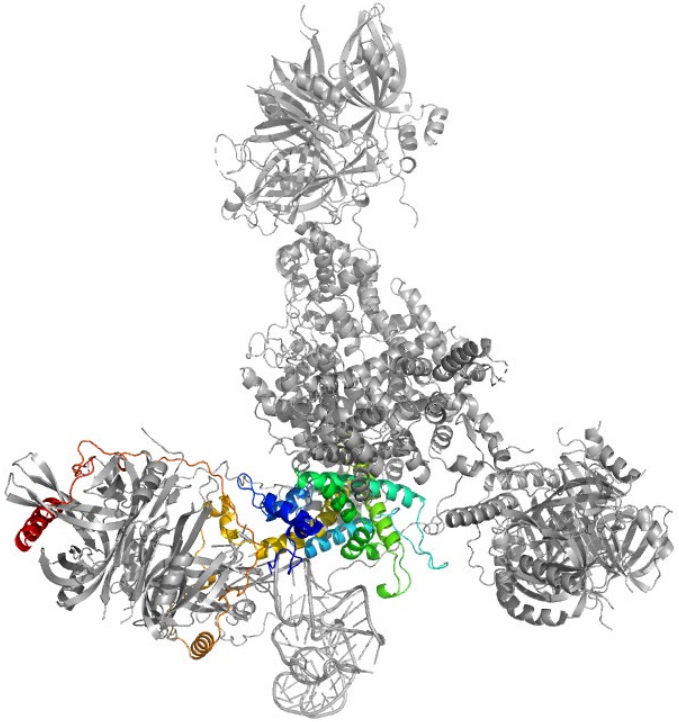
T1271s2-D1, GDT-TS: 85



T1271s2-D2, GDT-TS: 63

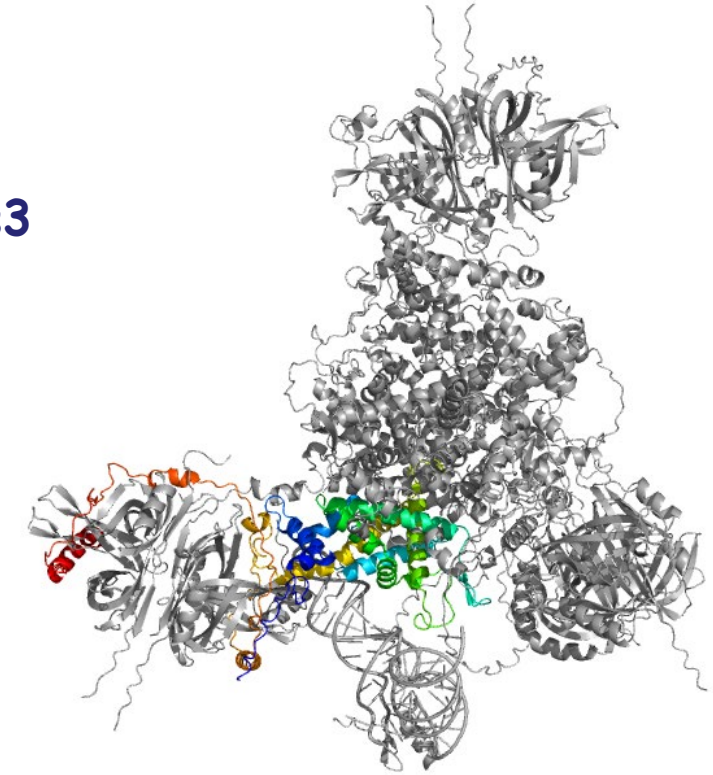


What went right?



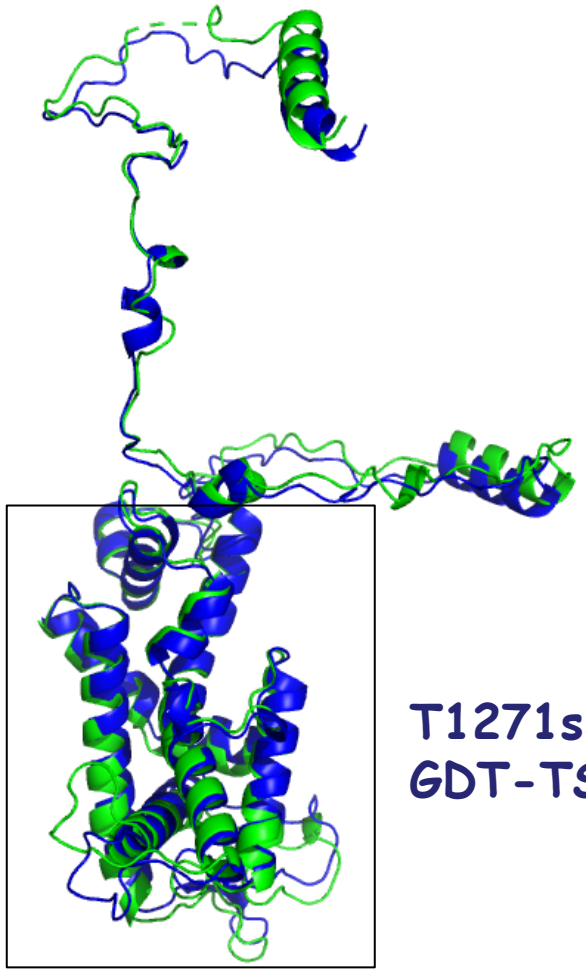
native

s3



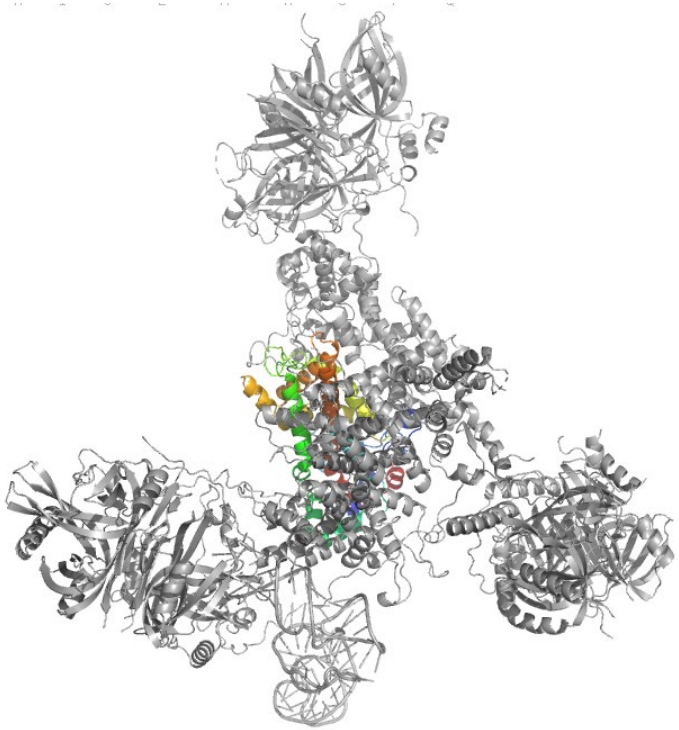
our model

T1271s3, GDT-TS: 79



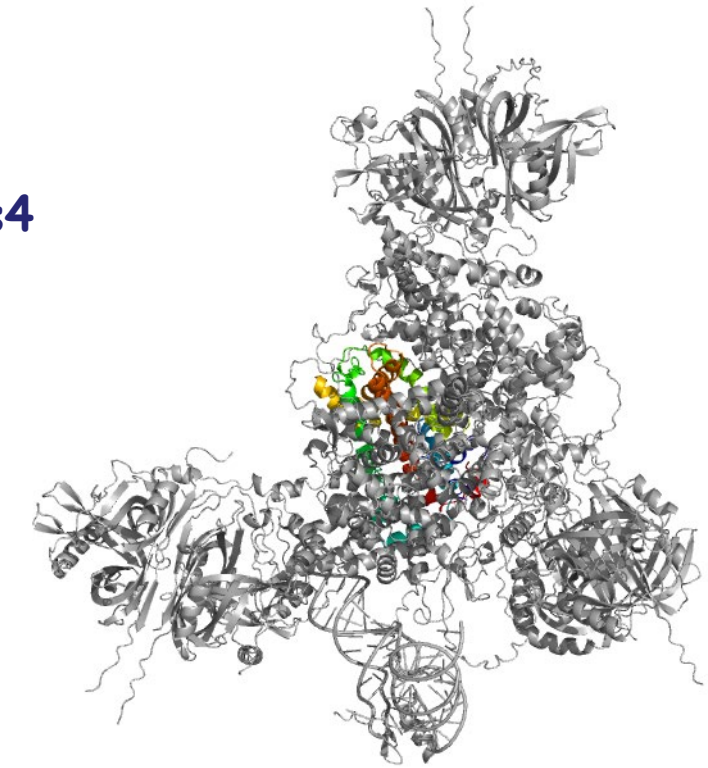
T1271s3-D1
GDT-TS: 91

What went right?



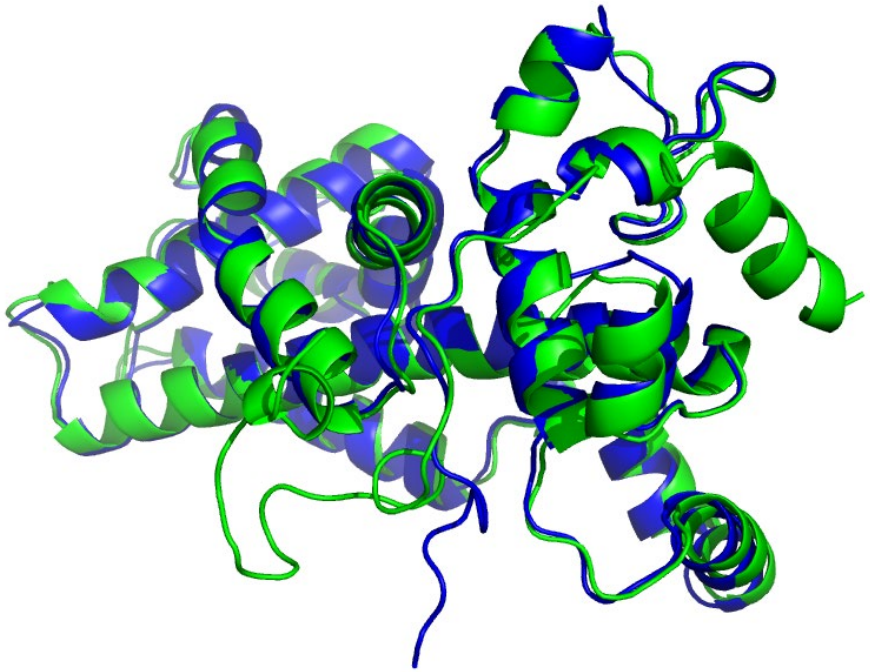
native

s4



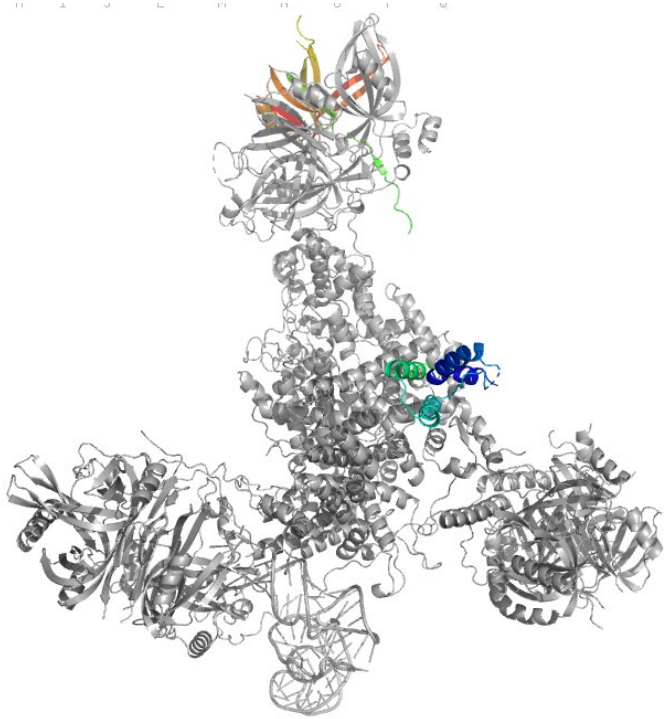
our model

T1271s4-D1, GDT-TS: 81



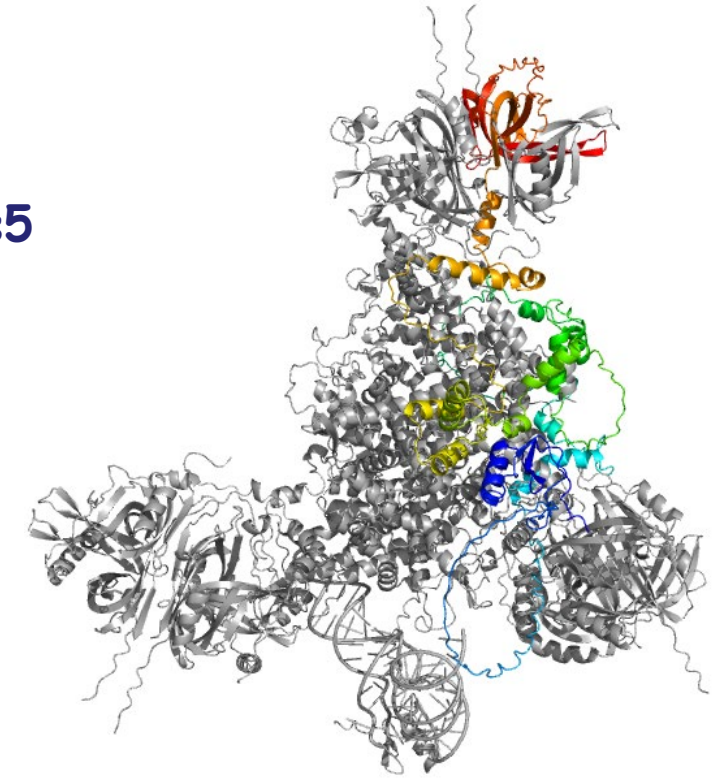
Removed too many residues at the N-term

What went right?



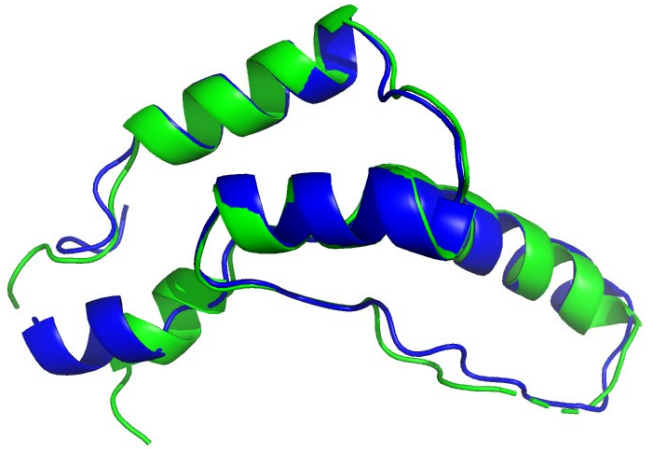
native

s5

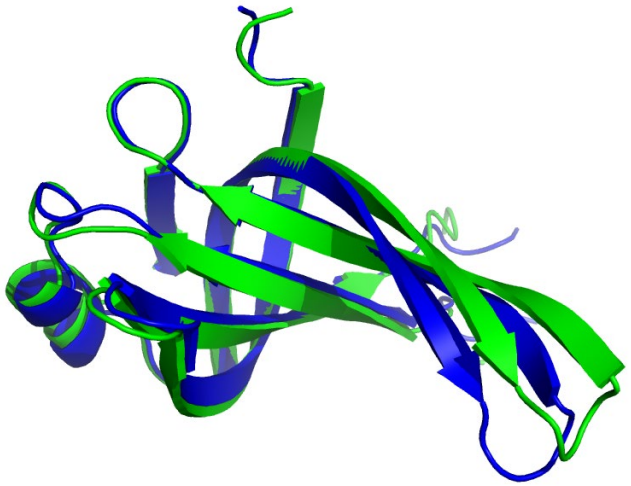


our model

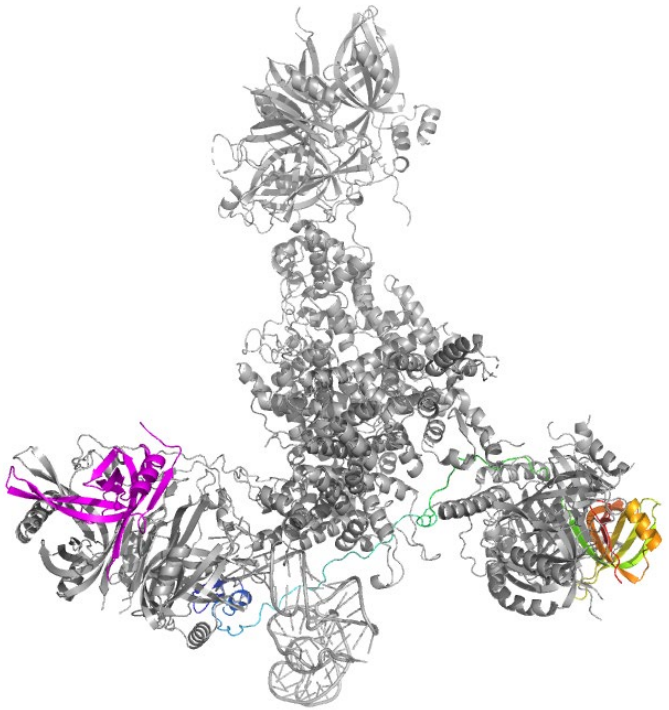
T1271s5-D1, GDT-TS: 85



T1271s5-D2, GDT-TS: 87

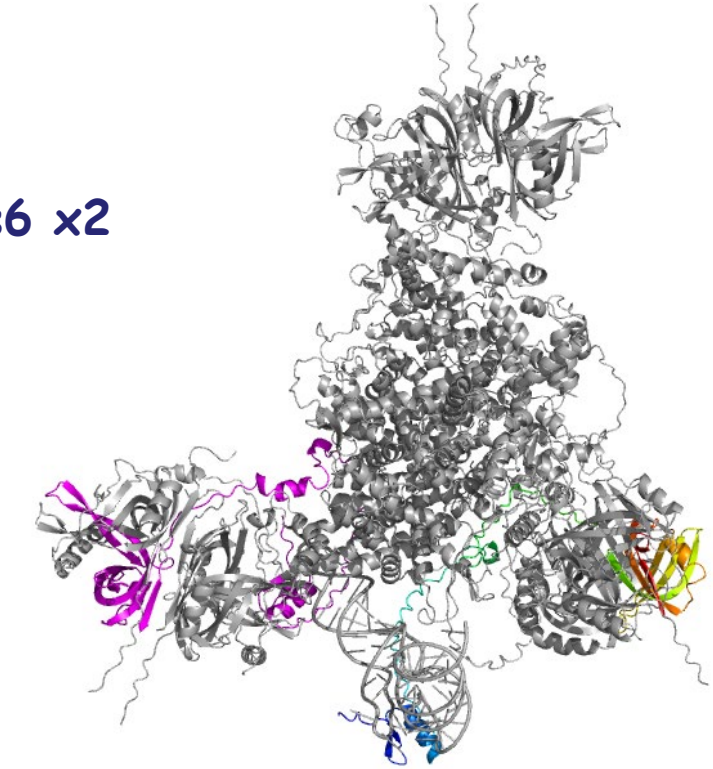


What went right?



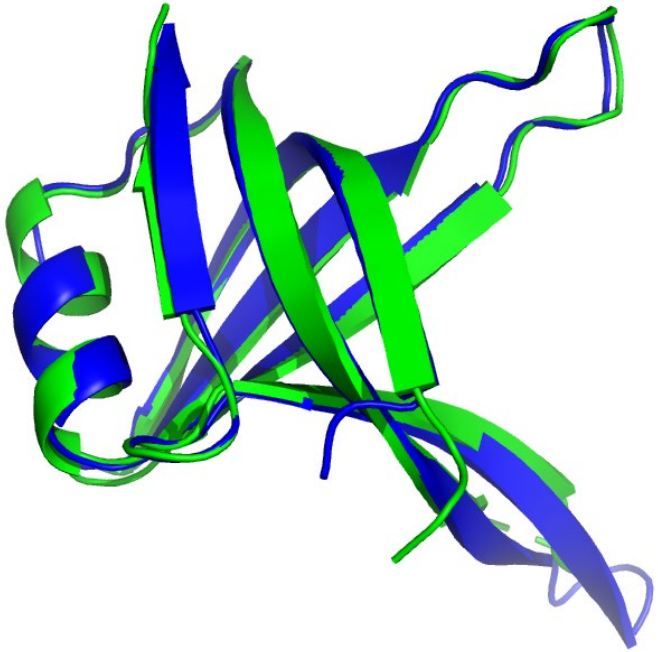
Native, chain L, N

s6 x2



our model

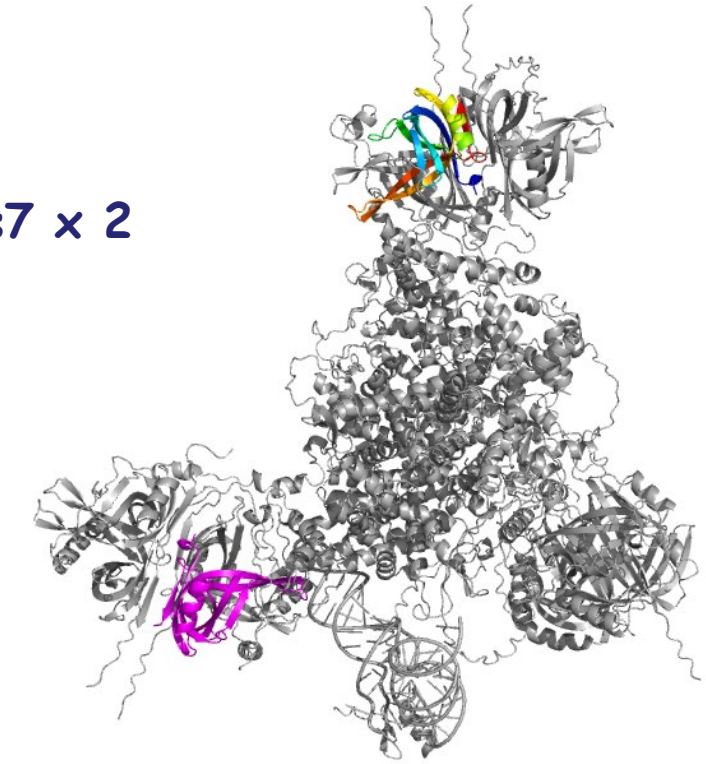
T1271s6-D1, GDT-TS: 95



What went right?

Native, chain G, M

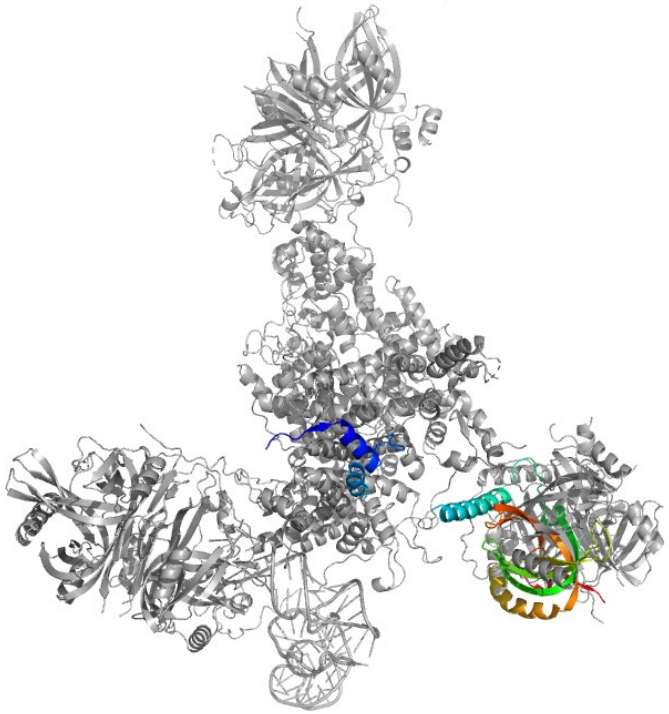
s7 x 2



our model

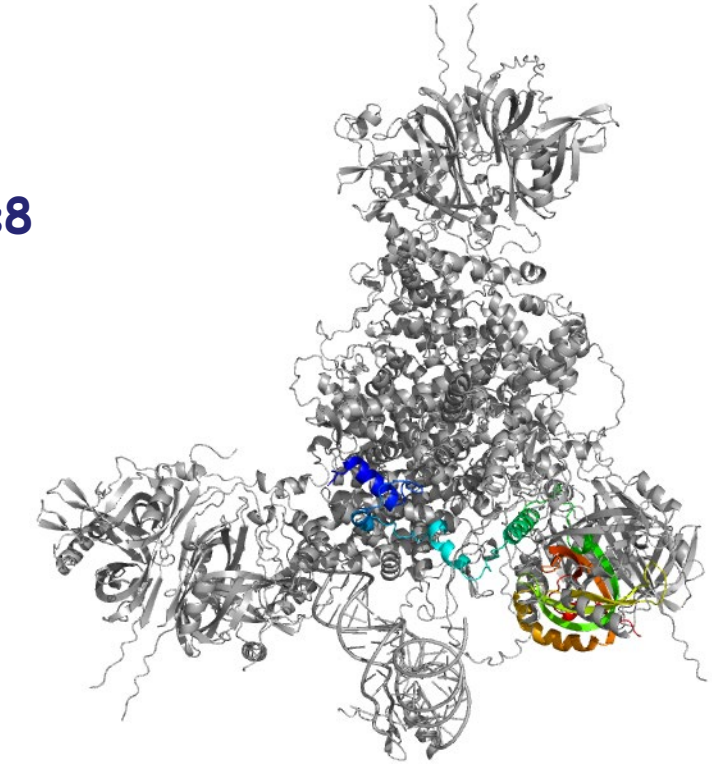
T1271s7-D1, GDT-TS: 93

What went right?



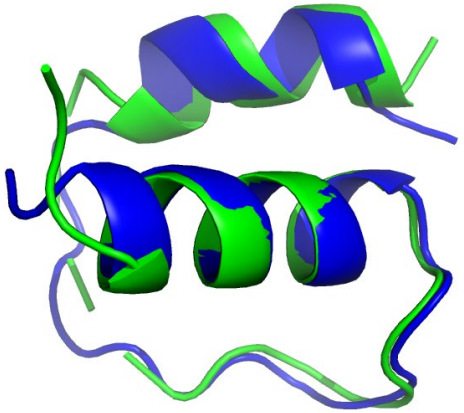
Native, chain H

s8

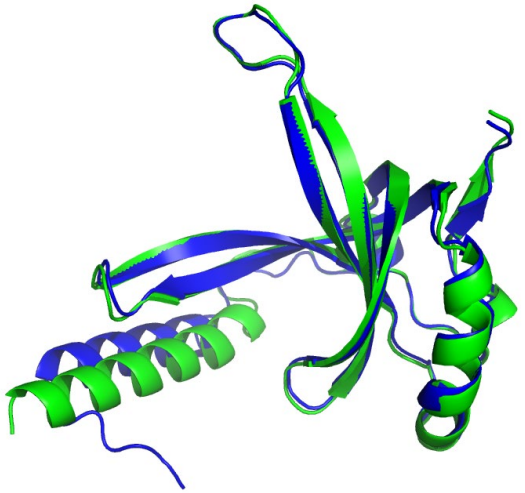


our model

T1271s8-D1, GDT-TS: 84

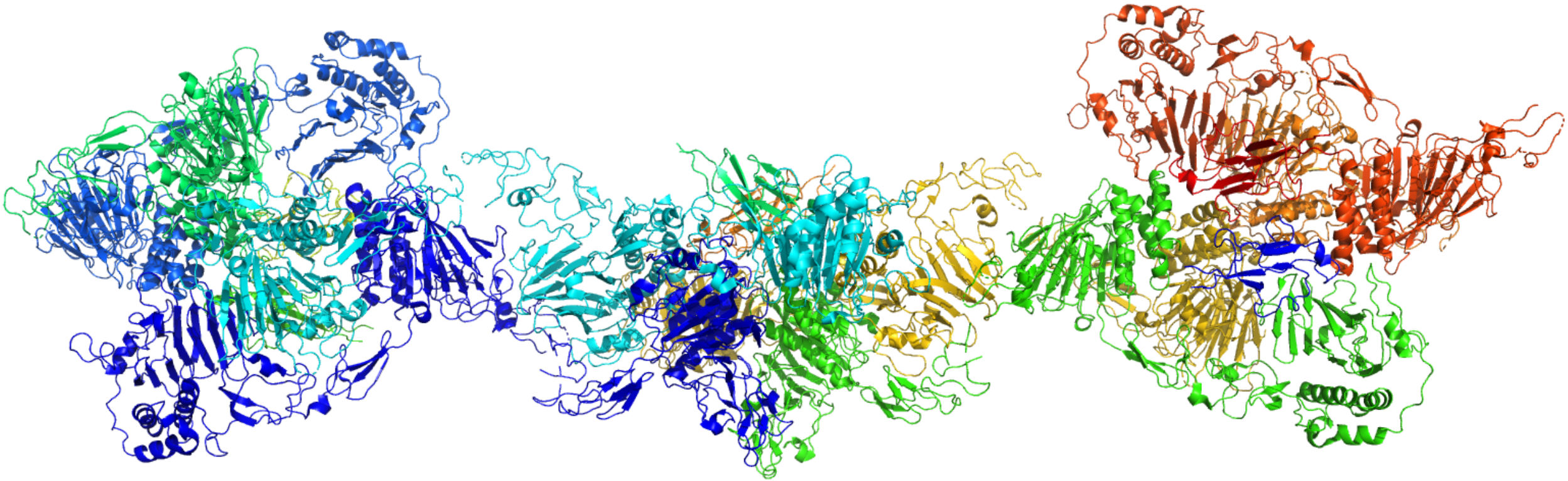


T1271s8-D2, GDT-TS: 87



What went right?

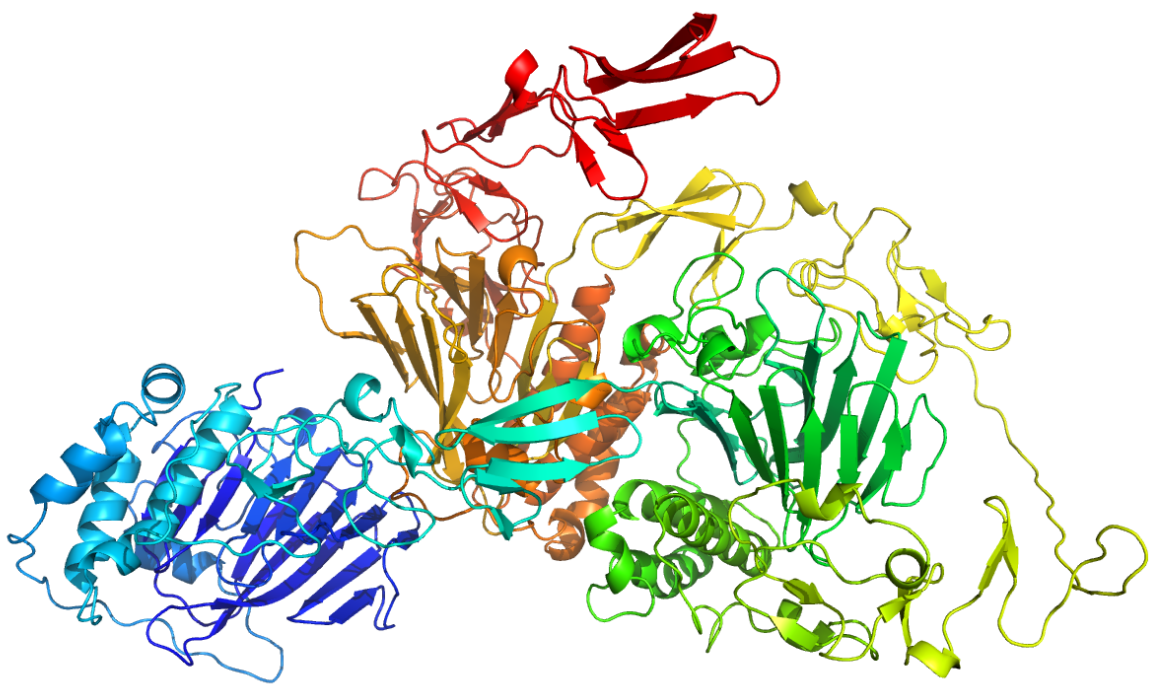
T1269 (Filament) **Multiple templates exist**



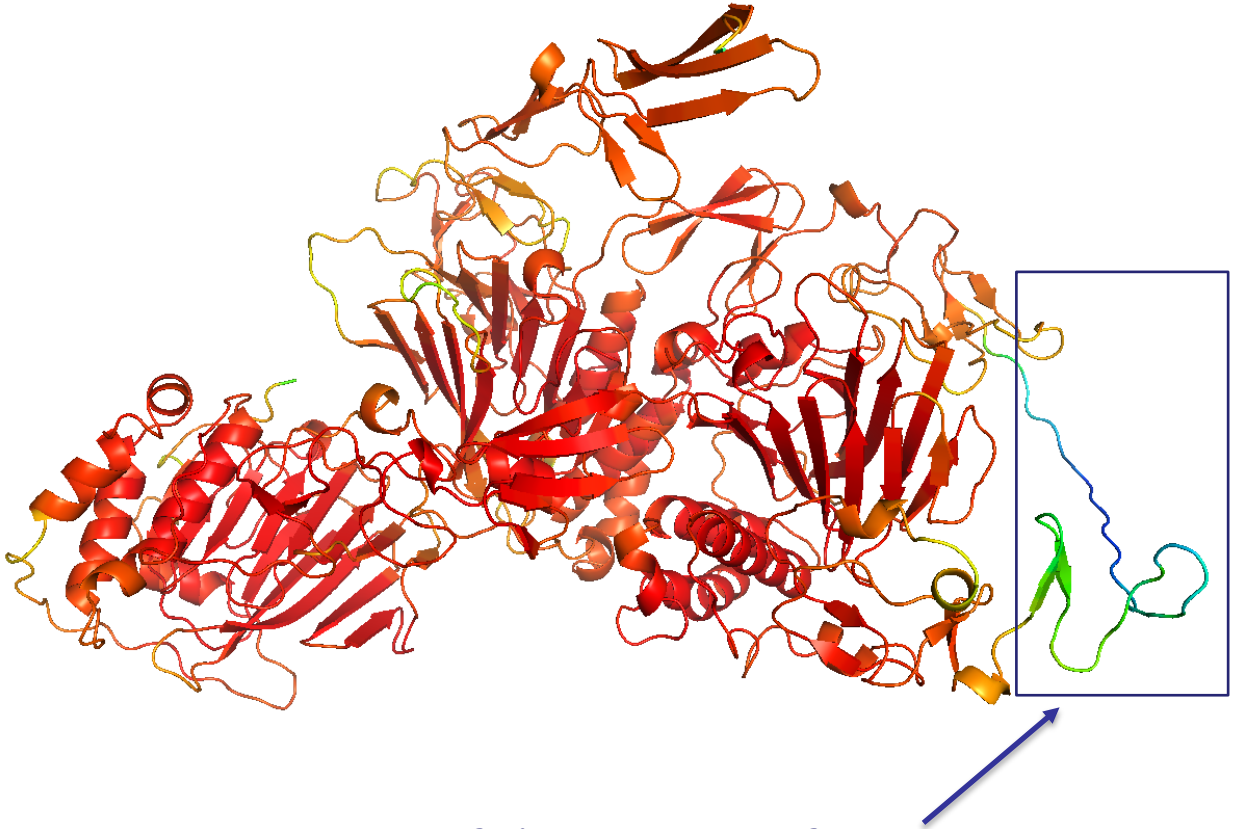
PDB ID: 7A5O (colored by chains)

What went right?

T1269 (1410 AAs) as a monomer

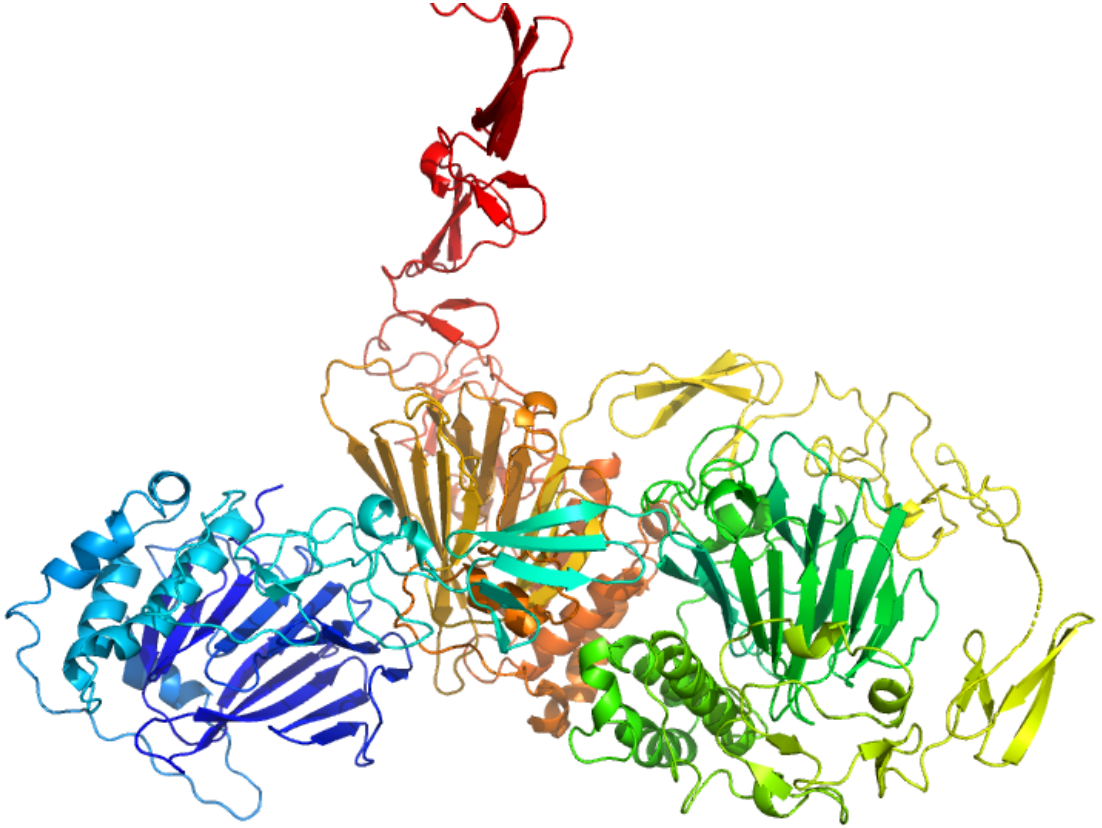


AF3 model for 1-1249
Ranking score: 81

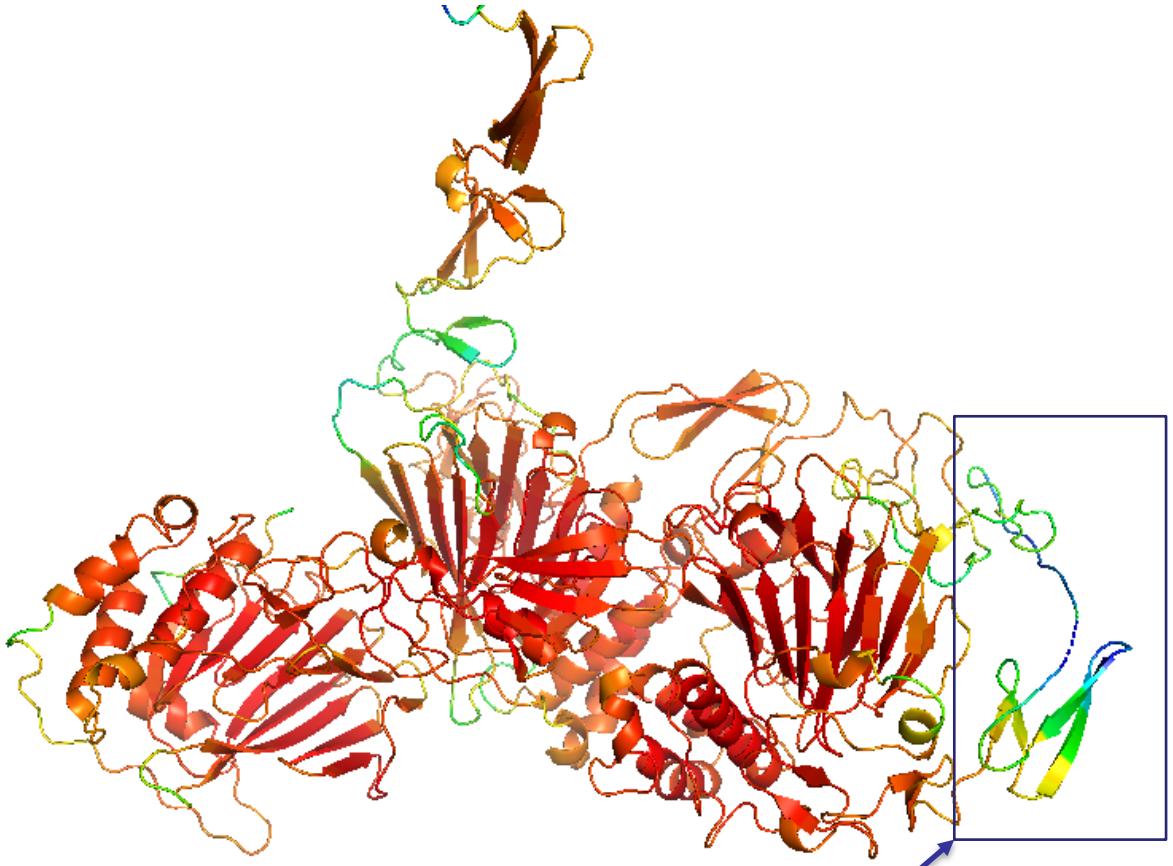


Low confidence score for 688-726

What went right?



AF2 model for 1-1249
Ranking score: 85

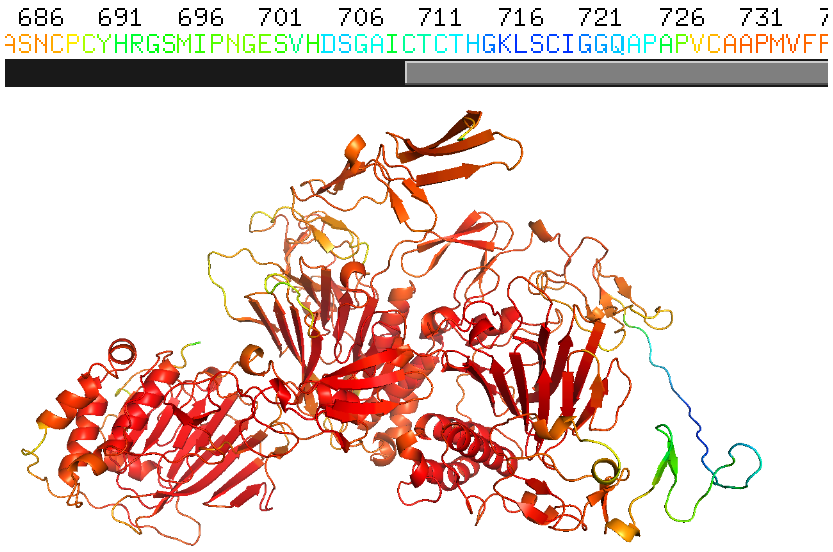


Low confidence score for 688-726

What went right?

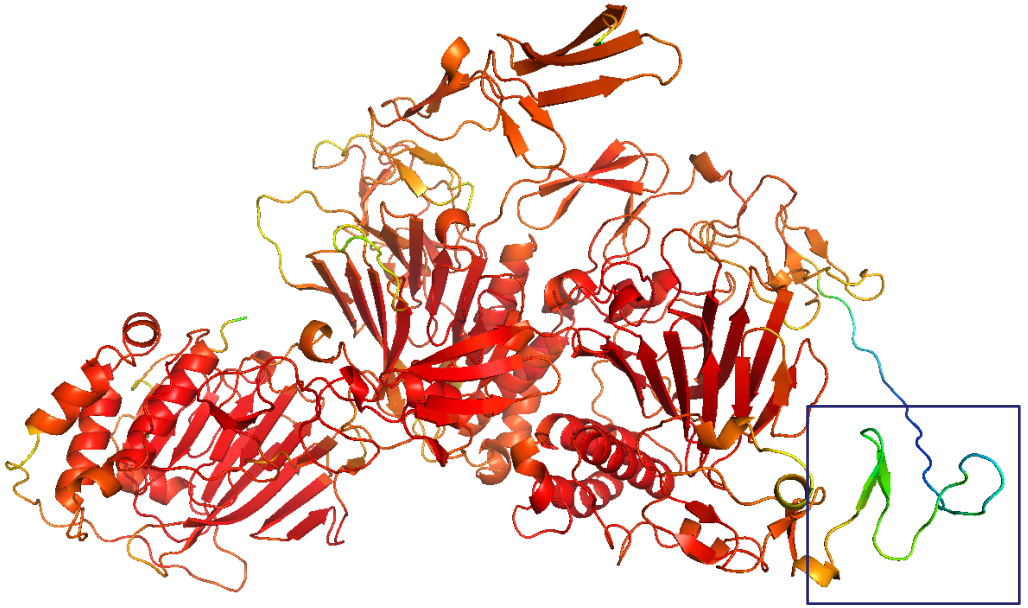
A726

```
Q T1269         711 TCTHGKLSCIGGQAPA-------------------PVCAAPMVFFDCRNATPGDTGAGCQKSCHTLDMTCYSPQCVPGCV  771 (1410)
Q Consensus     711 ~C~~G~~~C~~~~cp~-------------------~~C~~~~~~~~~~C~~~~p~~~g~~C~~tC~~~~~~C~~~~C~~gC~  771 (1410)
                    +|.+|.+.|....+..                   ..|..+.....|........+..|..+|.++...|....|..||.
T Consensus     730 ~c~~~~~~c~~~~~~~~~~~~~~~~~~~~~~~~~~~~~c~~~~~~~~~c~~~~~~~~~~~~c~~tC~~~~~~c~~~~c~~gc~  809 (1646)
T 7KWO_V        730 YCEDGFMHCTMSGVPGSLLPDAVLSSPLSHRSKRSLSCRPPMVKLVCPADNLRAEGLECTKTCQNYDLECMSMGCVSGCL  809 (1646)
T ss_dssp          ---------EEECEE-ETTE------EEEEEEEECCCCCTTSEEECCCSSSSCCCCSTTCCBSSSSSSCCCCSSCCCEEE
T ss_pred          EEcCCEEEEeeCCCCCcccCccccCCccccccccccCCCceEEecCcccccccCCccccCcccCCCCcccccccccCCcc
Confidence        9999999987654321                   123333334445444444455678888988877788888999999
```

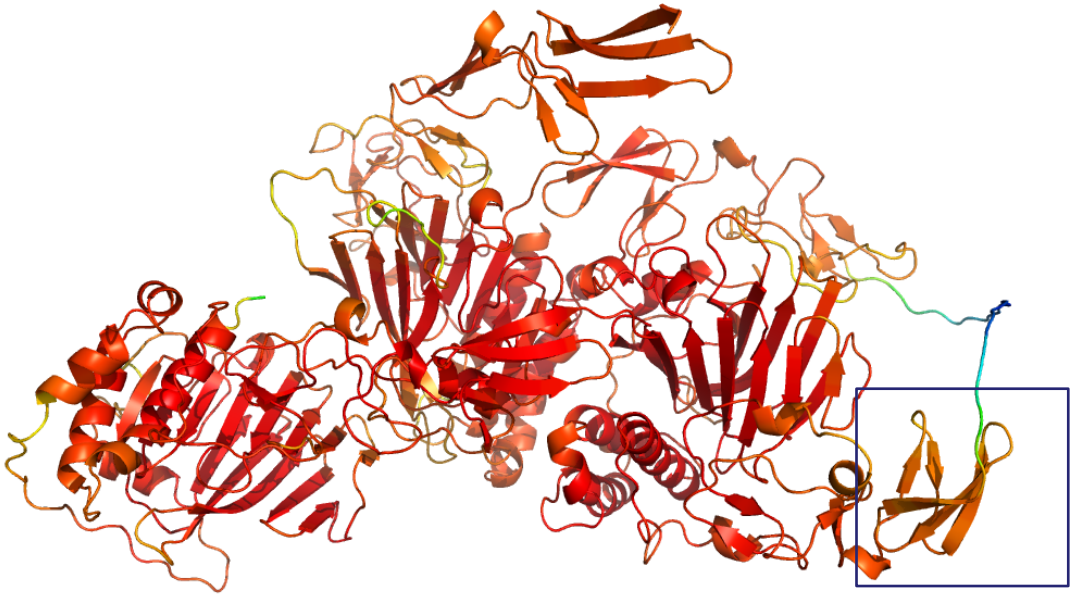


Solution:
Insert an artificial segment (length 19) after position A726

What went right?

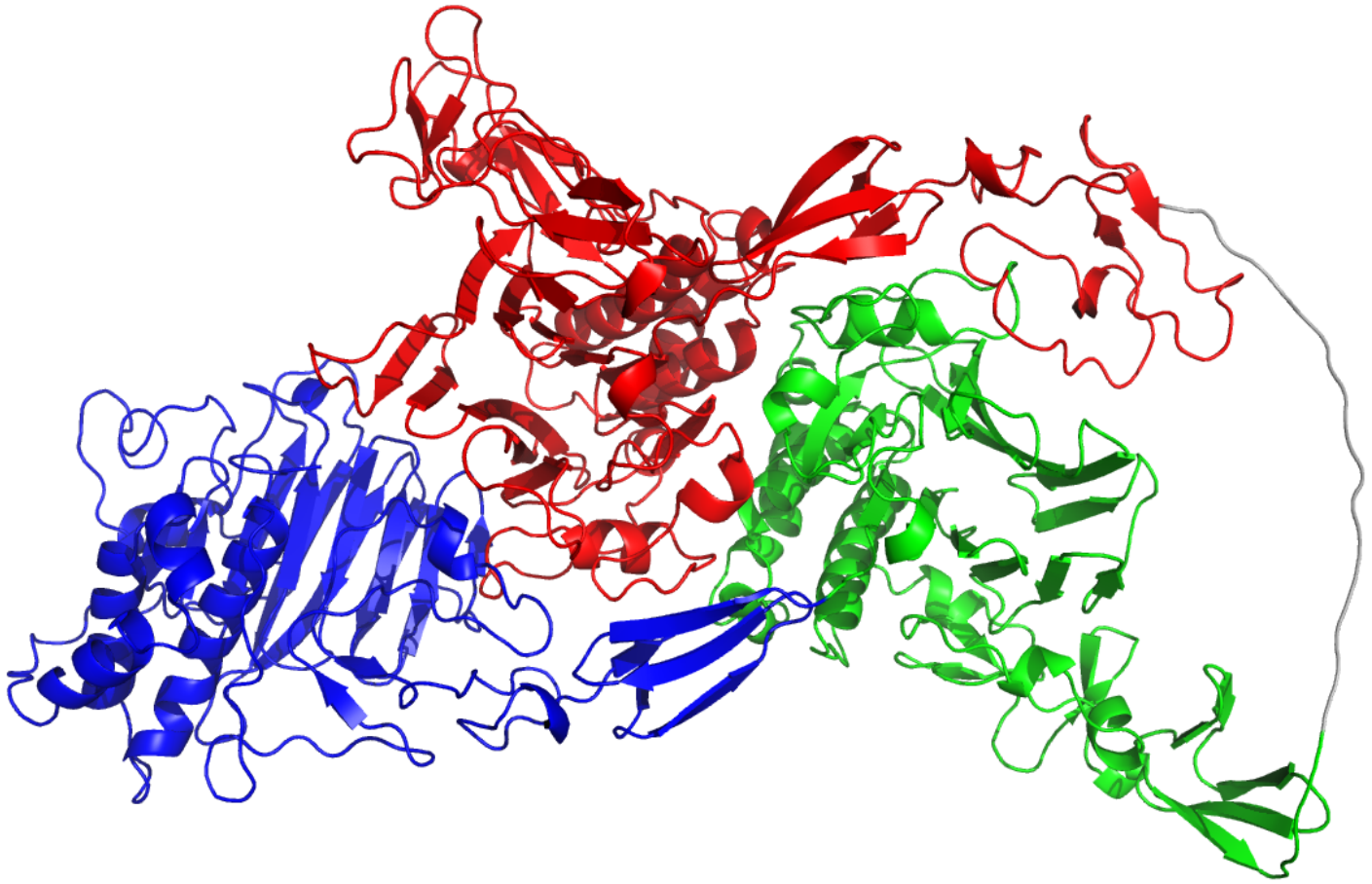


Ranking score: 81



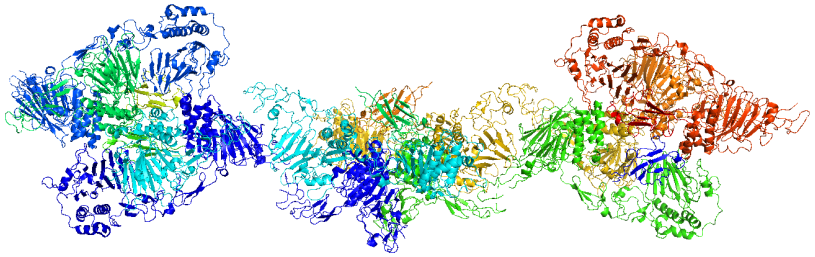
Ranking score: 89

What went right?



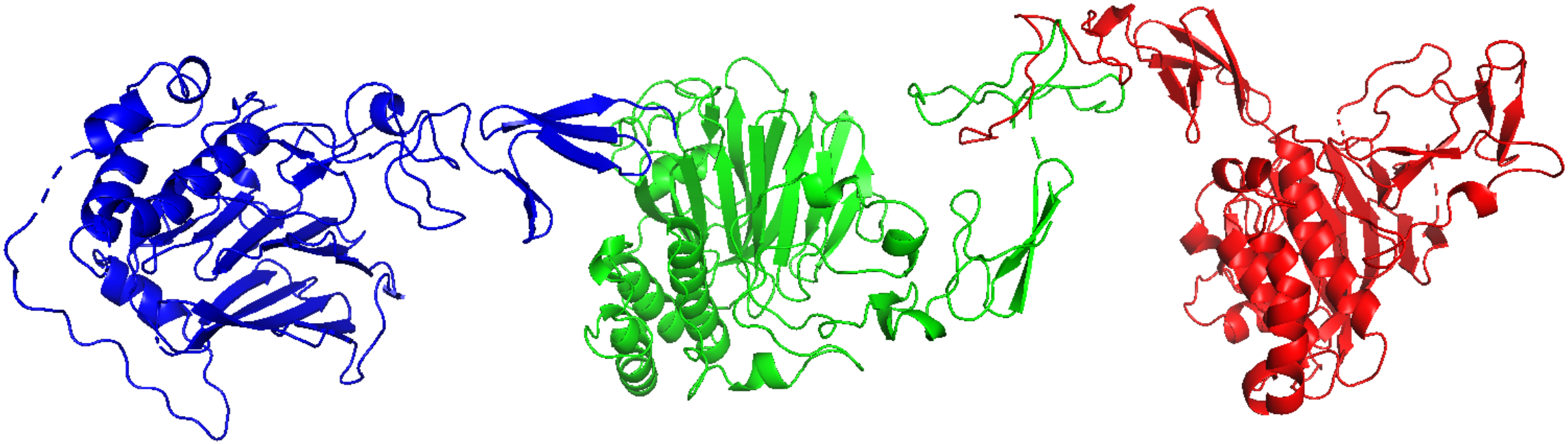
Can be divided into three domains

then assembly them based on the template 7A5O



Remark

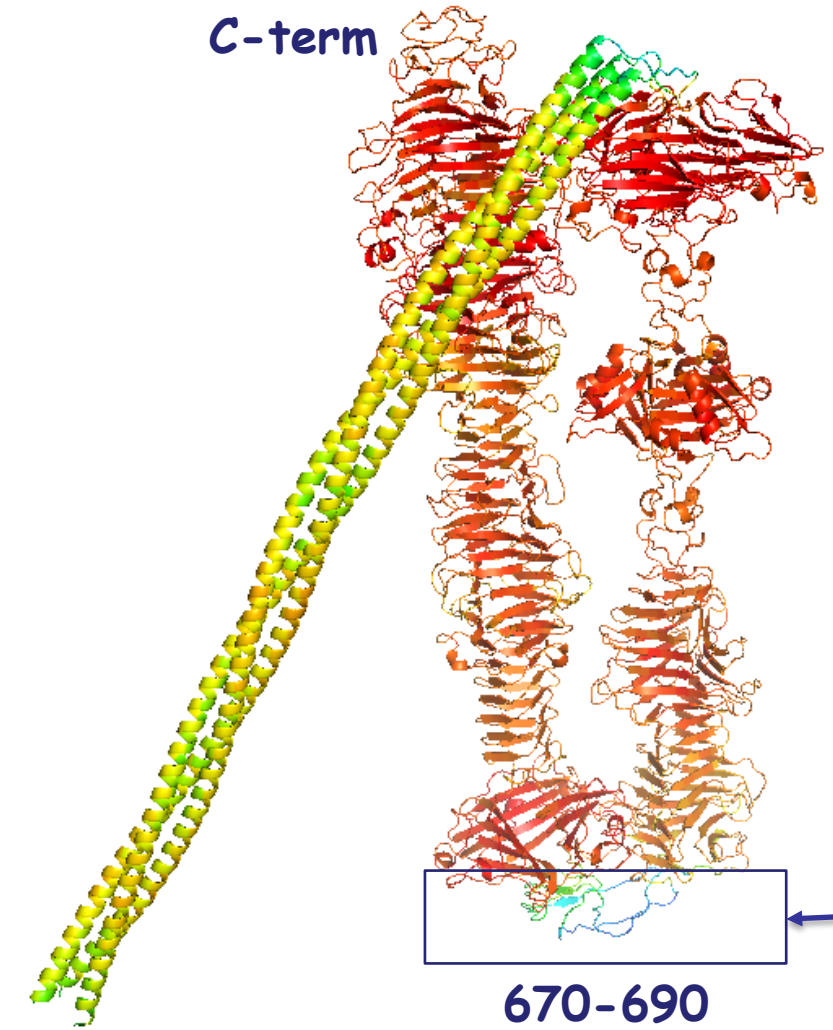
official domain split may be improved for this target



What went right?

ranking score: 56

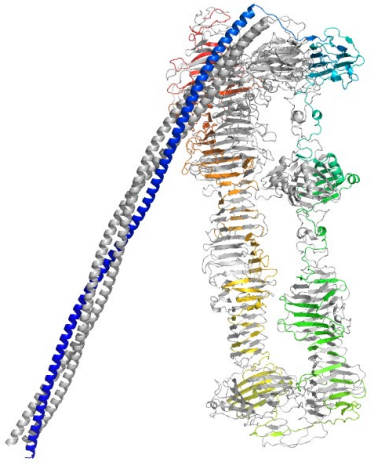
C-term



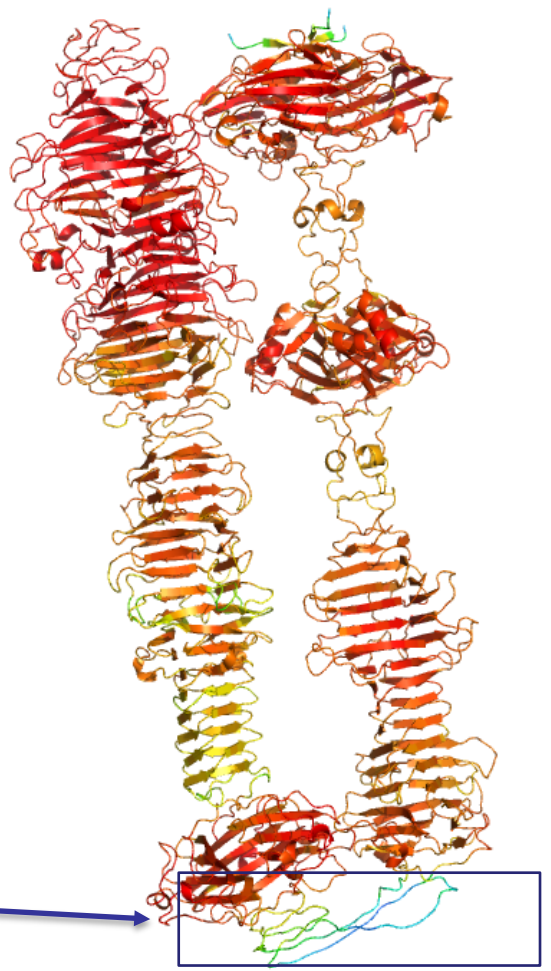
N-term

670-690

AF3 model for T0257o



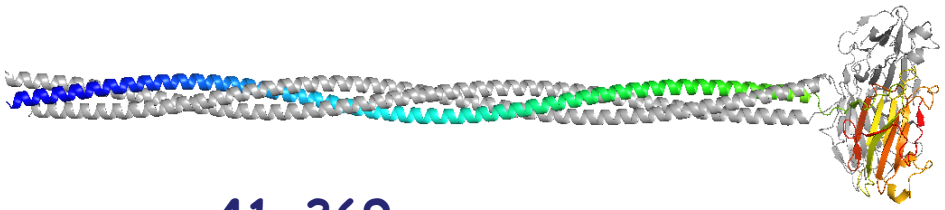
ranking score: 61



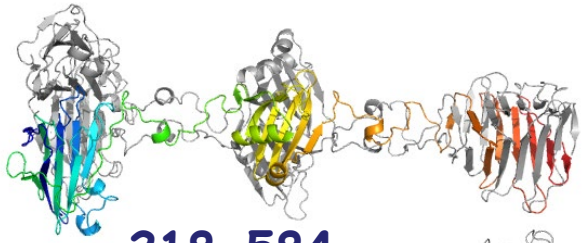
Low confidence

Why?
Wrong co-evolution forms the turn?

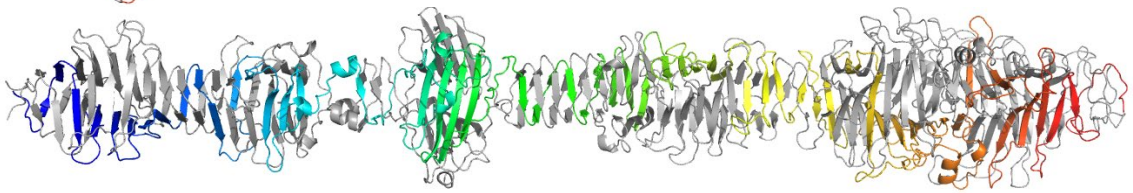
What went right?



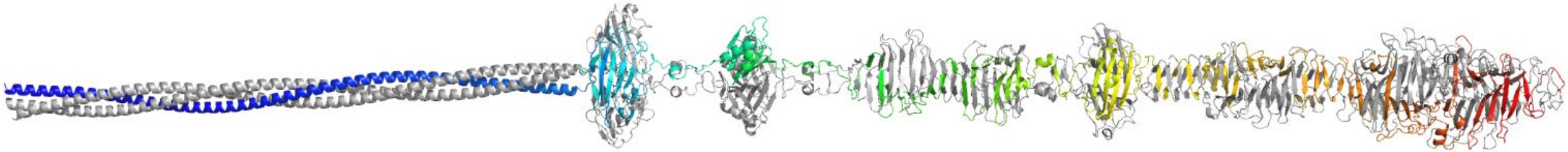
41-369



218-584



501-1263



What went wrong? H0258 phase 0

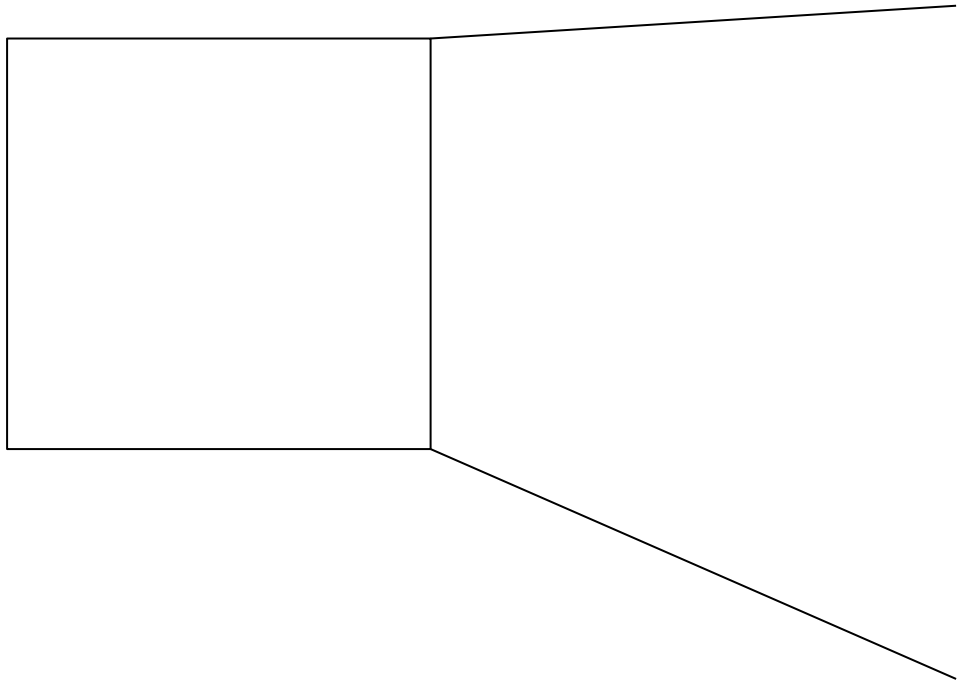
- Wrongly cut the disordered region (893-1014) in chain A, which is however the binding interface against chains B,C

893-1014
disordered in chain A

AF2 model for chain A,
plDDT: ~80

What went wrong? H0258 phase 0

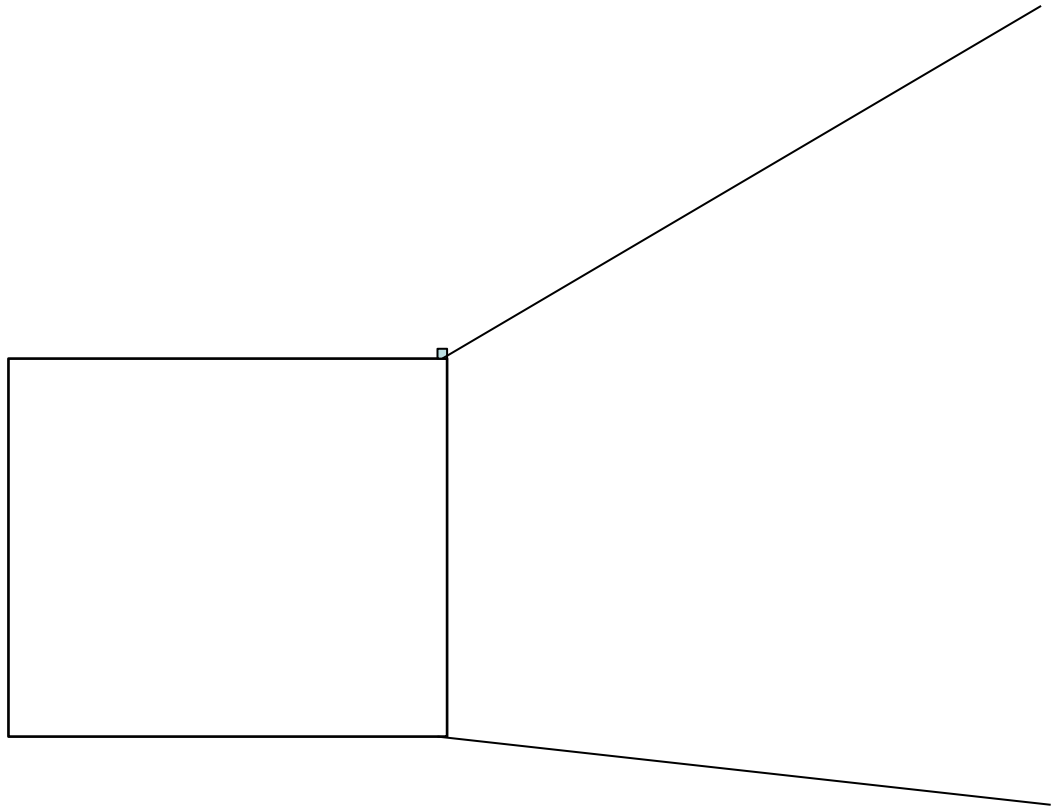
- This mistake makes it impossible to predict correct interface between chain A, and chains B, C



Low ranking score in AF3: ~0.3

What went right? H1258 phase 1 and 2

- AF3 model with full-length sequence has in higher confidence, especially the interface between chain A, and chains B, C



Higher ranking score: ~0.4

What went right? H1258 phase 1 and 2

When modeling interface only

Ranking score: ~0.9

DockQ
A-B: 0.63
A-C: 0.86
B-C: 0.73

Green: native
Blue: model

What went wrong?

H0227: wrong prediction of Stoichiometry led to incorrect complex prediction

A6B6

A1B6

CONTENTS

Conclusion

- Optimized sequences: disorder + stoichiometry
- MSA optimization is less important than in CASP15
- AF3 improves AFM for multimer, but still challenging

Acknowledgments

organizers, assessors

Wenkai Wang