CASP 13
Assembly assessment

Riviera Maya, Dec 2018

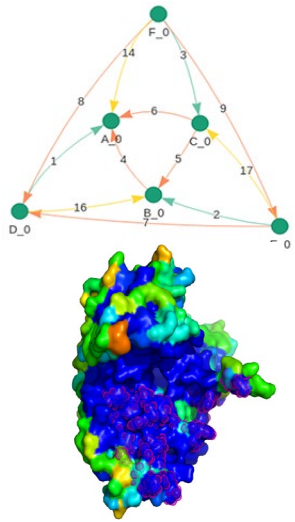
Jose Duarte, Dmytro Guzenko
RCSB Protein Data Bank, UC San Diego

Biological assembly of targets

- The *Ground Truth* is not always 100% clear when talking about biological assemblies of crystal structures (still most structures, 7 EM out of 42 targets)
- Assessors did not always have the structures at time of assignment
- Most of the times authors did not provide experimental evidence for the quaternary structure

Biological assembly of targets: assignment

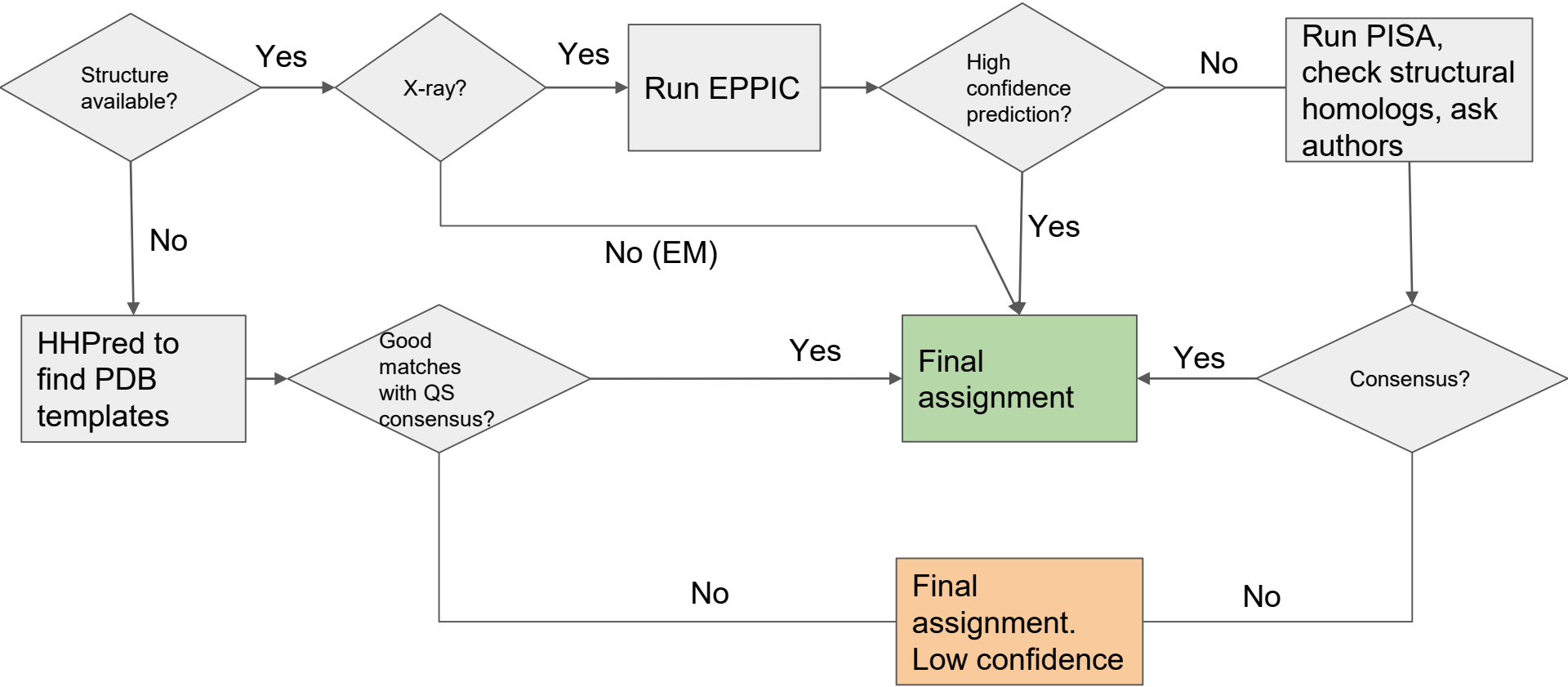
- EPPIC[1] used as main method to find most likely bioassemblies (when structure available)
  - Evaluation of all possible assemblies in crystal. Predictions include confidence values
  - Scoring based on evolutionary conservation of interfaces



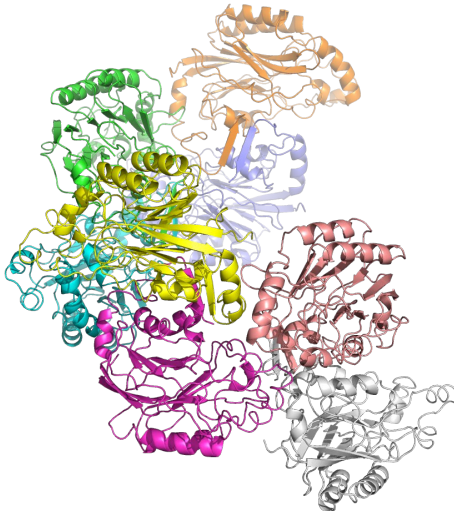
- Other methods used: PISA, structural homologs

[1] Bliven et al, PLoS Comp Biol, 2018 http://eppic-web.org

Biological assembly of targets: assignment



Some difficult cases

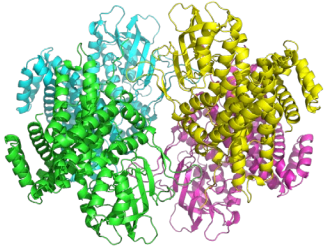


- T0995 (assignment A8)
  - Helical assembly: stoichiometry assignment subjective.
  - We decided A8. But A2 or A4 would have been reasonable choices too
- T0966 (assignment A2)
  - Large dimeric interface 1700 $Å^2$ not very well packed.
  - Bad EPPIC scores (indicating monomer). PISA says monomer
  - A subdomain covering only a small region from the full length protein
  - No experimental evidence provided. Kept A2 from authors assignment.
- T1018 (assignment A2)
  - EPPIC: dimer medium confidence. PISA: dimer.
  - Structural homologs both monomers and dimers (a feature of that family, following literature)
- T0985
  - Released as A1, structure wasn't available
  - Clear A2 (once structure became available)
  - Excluded from our assessment, even though some groups submitted good predictions

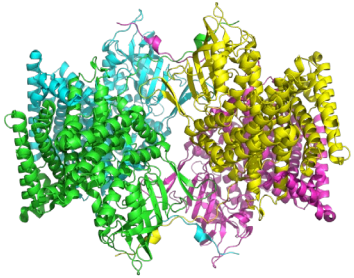
Target difficulty: easy

- Templates with the same quaternary structure can be detected by sequence similarity (HHPred)

T0961o (A4)



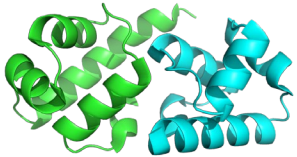
4Y9J



H0974 (A1B1)

CASP:
Image redacted

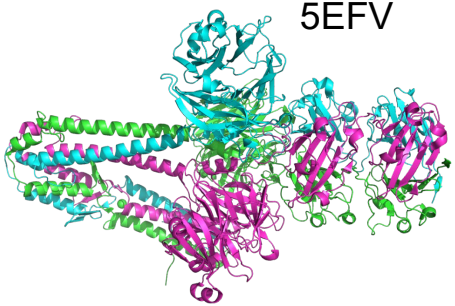
1Y7Y (C2 homodimer)



T0977o (A3)

CASP:
Image redacted

5EFV



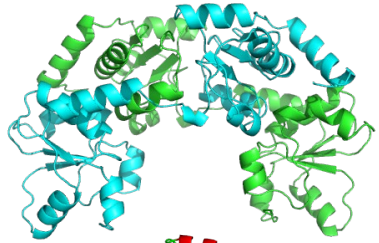
Target difficulty: medium

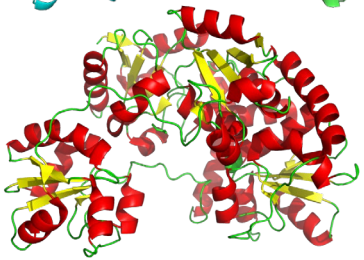
- No assembly template is easily found, (partial) templates for subunits, (partial) interface templates are available

T0999 (A2): all domains and most of the interfaces available, but fragmented. The structure needs to be puzzled together

CASP:
Image redacted

T0976 (A2): the best assembly template is a monomer. Possible domain swap.

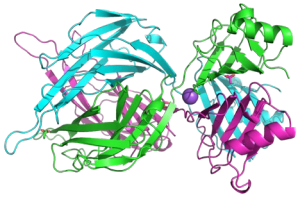




1YT8

T0981 (A3): assembly template for ⅓ of the structure, individual domains for the rest.

CASP:
Image redacted

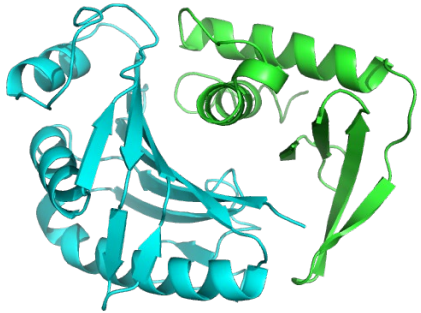


5M9F

Target difficulty: hard

- No or negligible amount of information available on the subunits and assembly

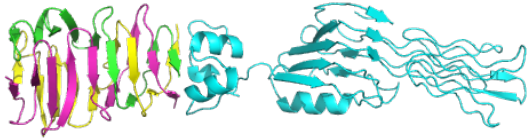
H0986 (A1B1)



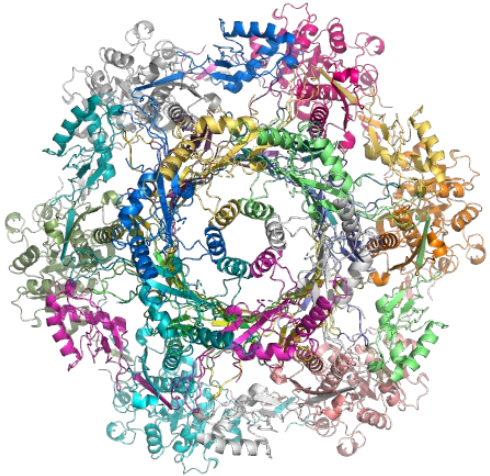
T0989 (A3)

CASP:
Image redacted

H0953 (A3B1)



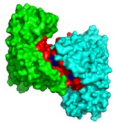
H1021 (A6B6C6): partial templates available, but the total assembly is huge



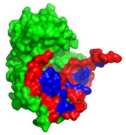
Scores

Interface/local:

Interface Contact Score (F1)

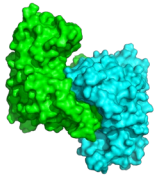


Interface Patch Score (Jaccard)

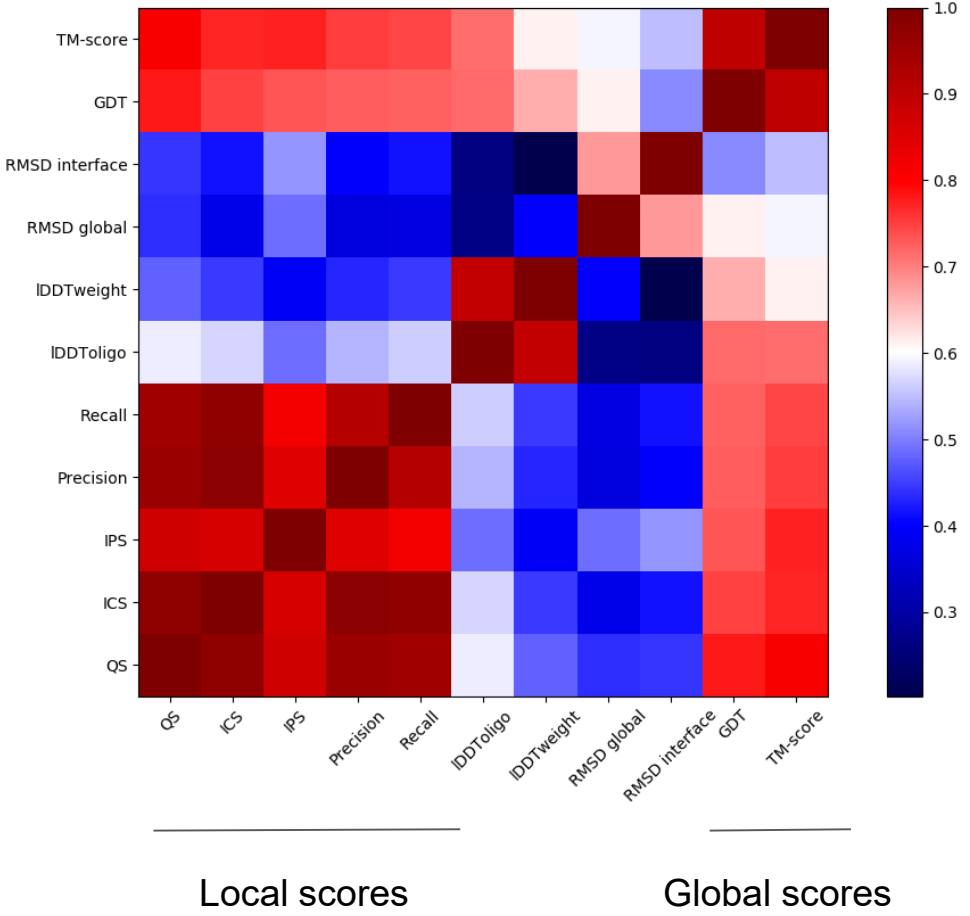


Assembly/global:

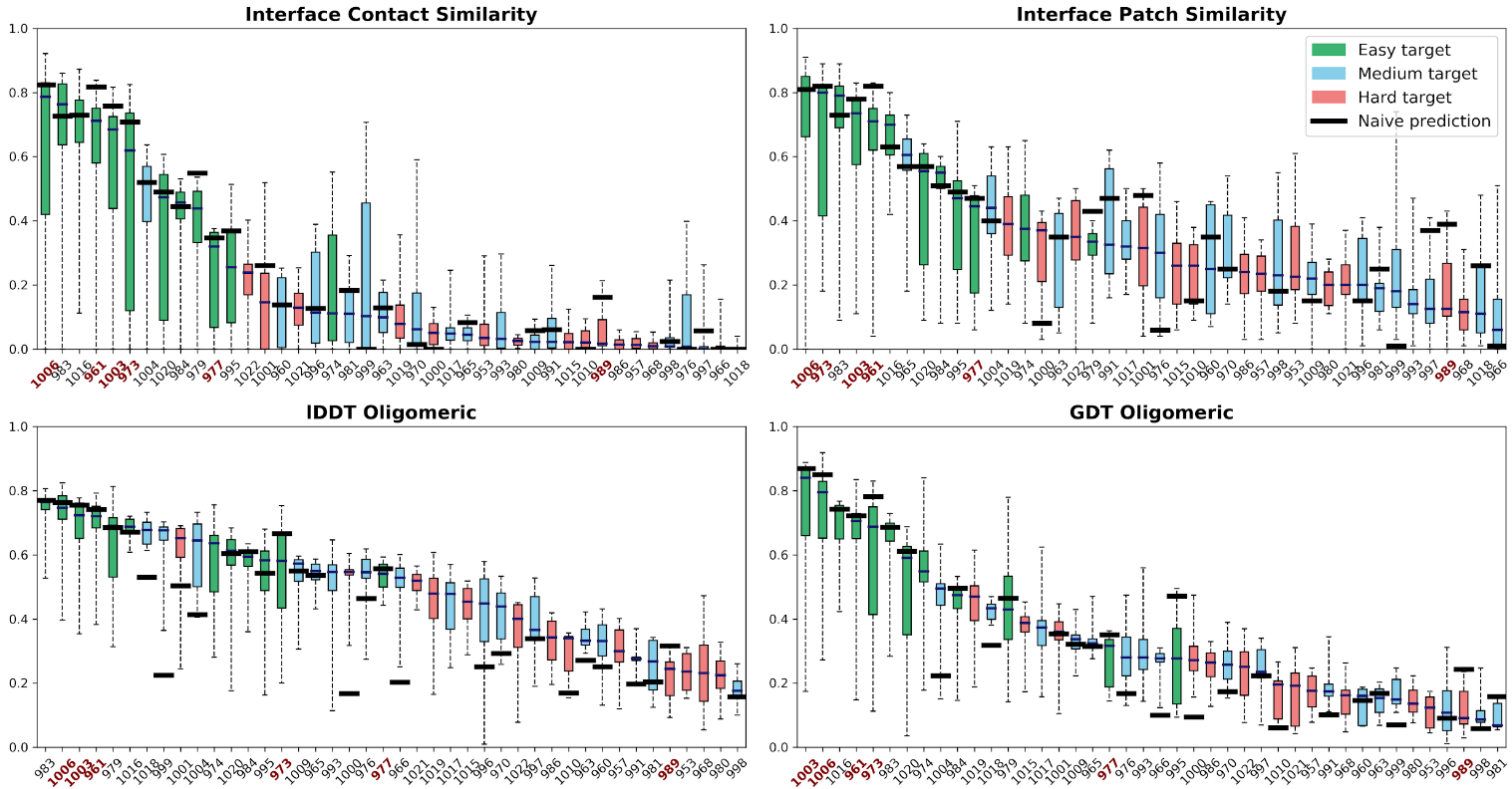
Oligomeric lDDT



Oligomeric GDT



Scores: per target overview



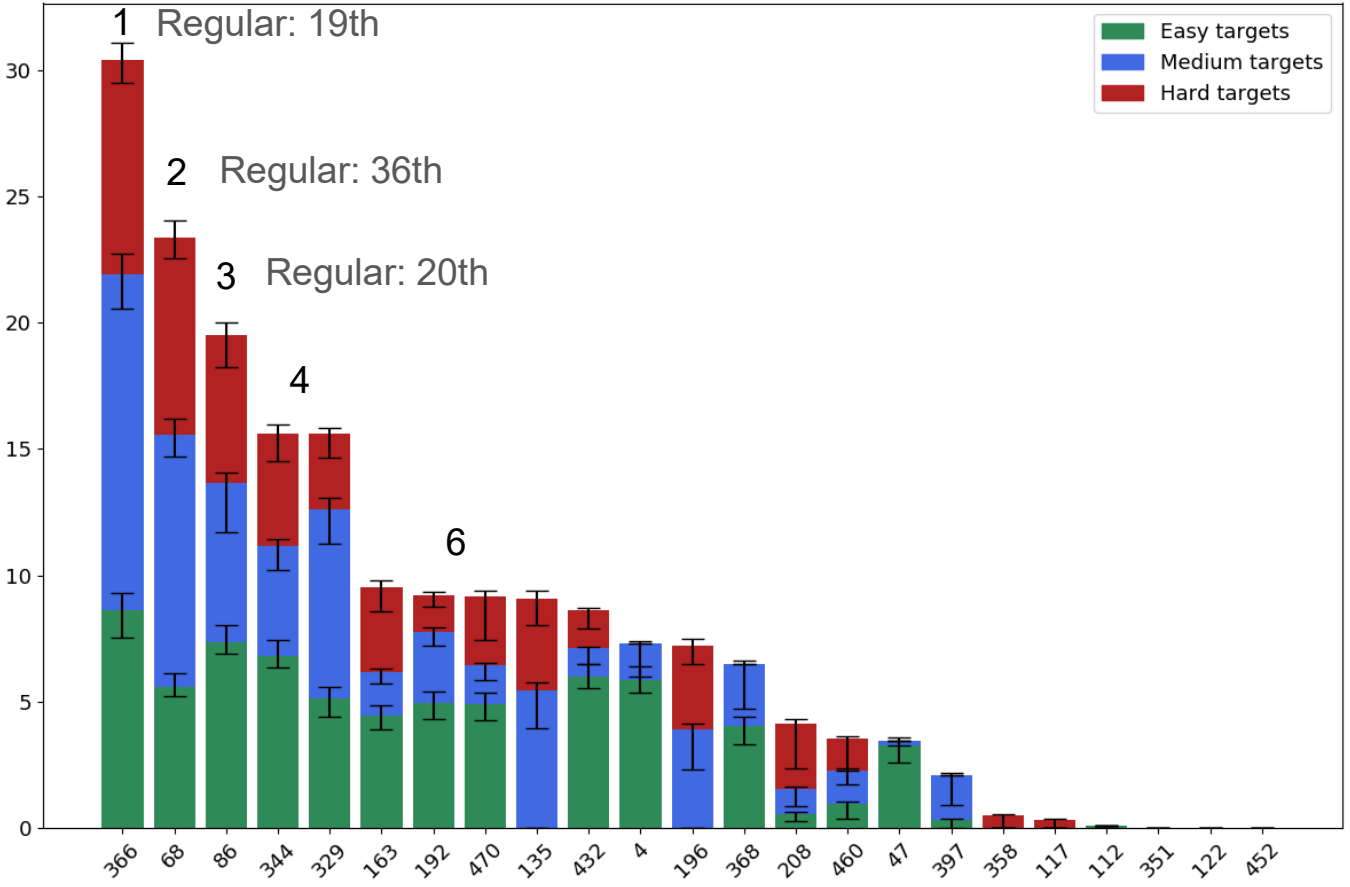
* Naive predictor: Seok-naive_assembly (thanks Seok group!)

Group ranking - methodology

- Interface Patch Score (Jaccard), Interface Contact Score (F1), lDDT (oligo) and GDT (oligo)
  - normalised to Z-scores
  - equal weights
  - Sum(Zi) > 0 only
- Leave-one-out cross-validation performed on the scoring, groups ordered by mean score
  - Scoring by difficulty and stacking the results does not change the overall ranking
  - Excluding targets with poor predictions and small score variance (e.g. H0980, H0968, H0986) does not change the ranking

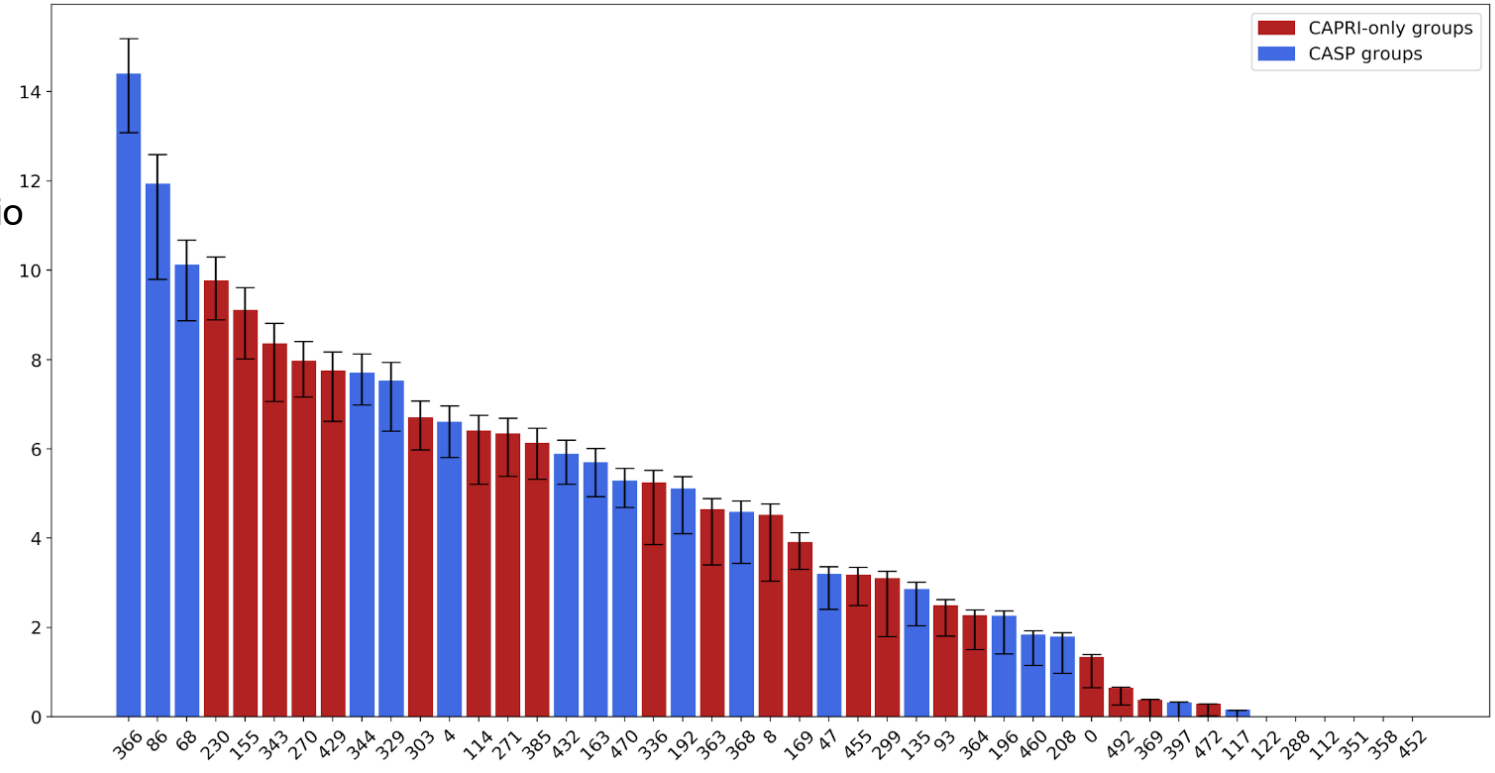
Group ranking - CASP groups, all targets

366 Venclovas
068 Seok
086 Baker
344 Kiharalab
329 D-Haven



Group ranking - CAPRI targets only, all groups

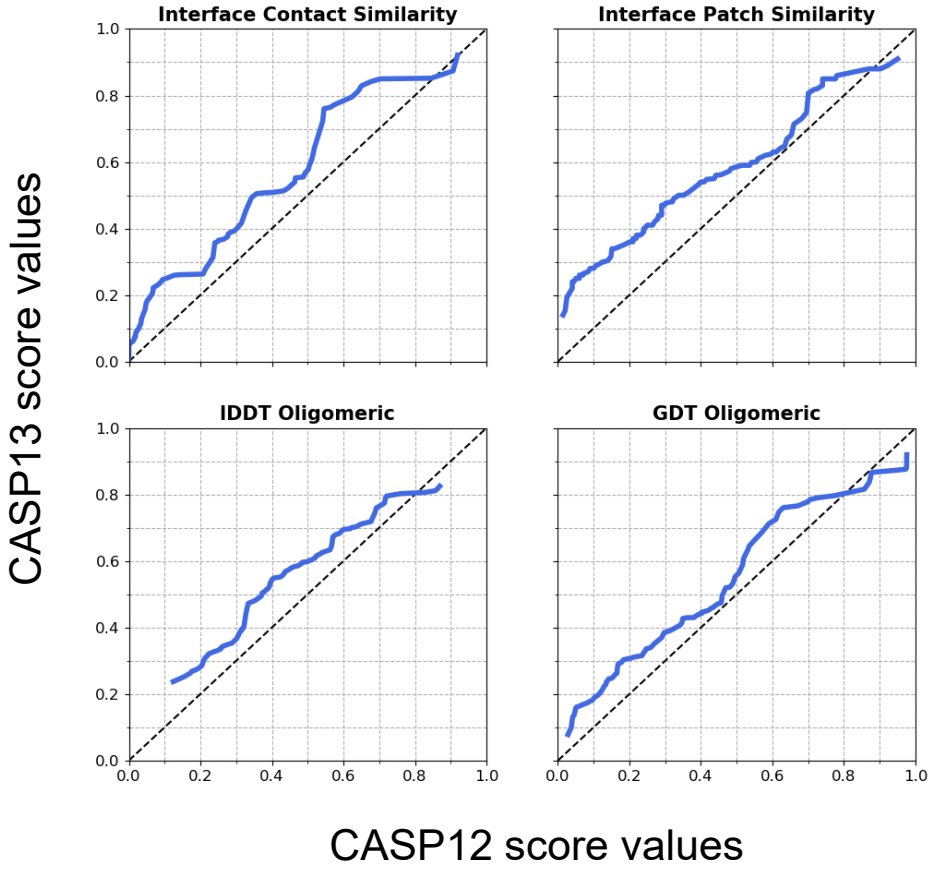
366 Venclovas
086 Baker
068 Seok
230 FernandezRecio
155 ZouTeam



Comparison with CASP12

- Organisation
  - Oligomeric predictions have their own format
  - No accidental participation by predictors
  - No need to determine if a prediction is meant to be oligomeric
- Participation
  - Almost 5000 models submitted (CASP+CAPRI) vs. 1600 in CASP12
  - 45 groups in CASP13 vs 108 in CASP12
    - Some groups may have participated in this category by accident in CASP12
  - Targets
    - CASP13: 42 regular (12 heteromers), 16 data-assisted.
    - CASP12: 30 regular (8 heteromers), no data-assisted

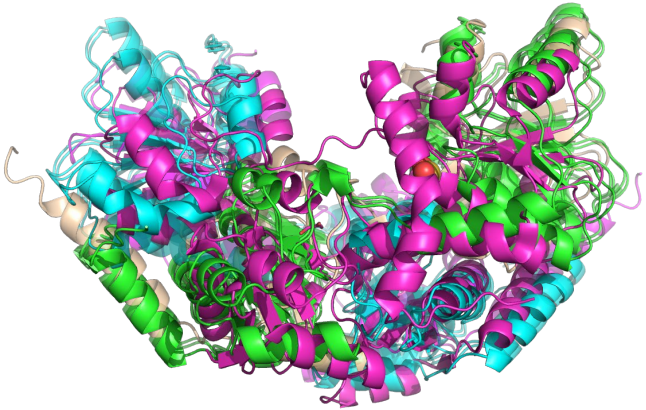
Comparison with CASP12



Improvements across the board!

Prediction highlights: what went well

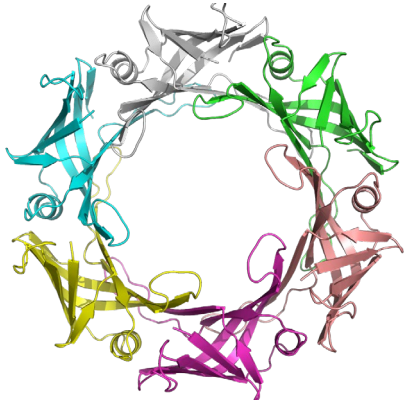
- Target: dimeric
- Each chain: 2 copies of same domain (CATH 3.40.250.10, oxidized rhodanese)
- Plenty of templates for the domain
- Best template: 1YT8 monomeric, with central domain-domain interface very similar to dimeric interface of target
- Best model: 155_4 (CAPRI group). F1 (ICS) 39.8



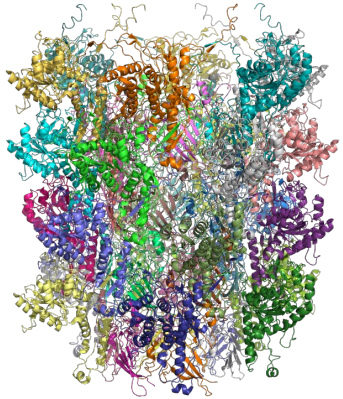
T0976o (A2)

Prediction highlights: what went well

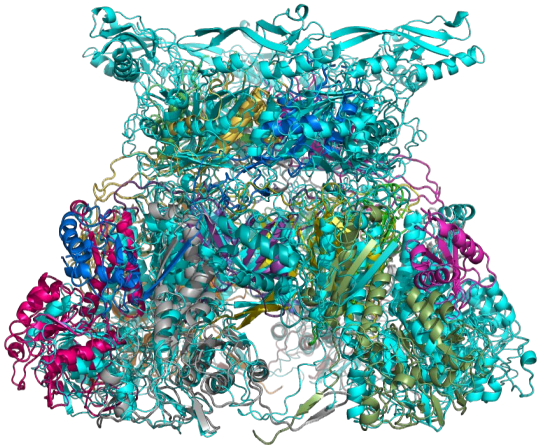
- Huge complex A6B6C6 (798x6 residues)
- Partial templates: 6bdc (A6), 3j9q (6-fold ring with matching B chain and another molecule)
- 068_5: decent global assembly prediction



6bdc (A6)



3j9q (B6D6)

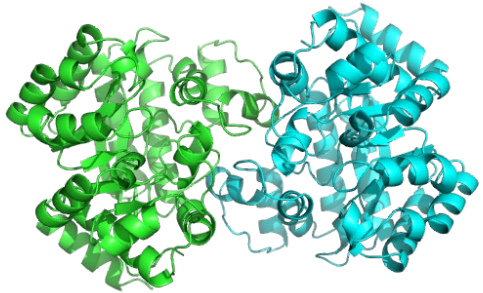


H1021 (A6B6C6)

What did not go so well

- Very good template for monomer
- No templates for assembly or interface
- I.e. pure docking
- But! no good predictions

3RYS



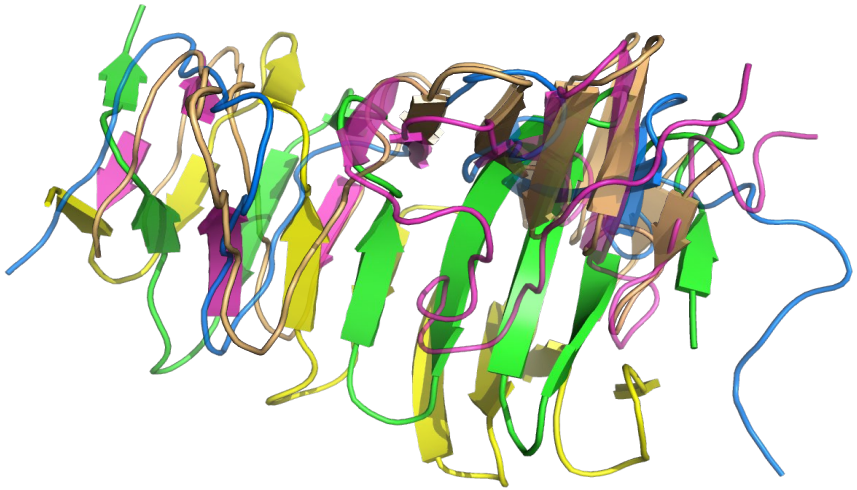
Weak dimer? Crystal contact?

We don’t know!

T1018o (A2)

Quaternary is important for regular modelling

- Bad modelling in C-terminal for almost all regular groups
- Best model (043_1) folds C-terminal in
- Some assembly groups have decent models in the C-terminal (e.g. 086_1)



T0953s1-D1

#	Model	gdt_ts	gdt_ha	gdc_sc	rmsd
1.	T0953s1TS043_1-D1	54.48	42.16	19.74	13.60
2.	T0953s1TS086_1-D1	48.88	27.61	3.64	5.33
3.	T0953s1TS208_1-D1	47.76	34.70	13.39	9.34
4.	T0953s1TS460_1-D1	46.27	31.71	12.39	10.88
5.	T0953s1TS335_1-D1	46.27	32.09	12.69	13.23
6.	T0953s1TS196_1-D1	45.90	30.60	13.54	10.86
7.	T0953s1TS135_1-D1	45.90	30.60	13.54	10.86
8.	T0953s1TS145_1-D1	45.90	30.60	13.54	10.86
9.	T0953s1TS055_1-D1	45.90	30.60	13.54	10.86
10.	T0953s1TS426_1-D1	45.90	30.60	13.54	10.86
11.	T0953s1TS418_1-D1	45.90	30.60	13.54	10.86
12.	T0953s1TS089_1-D1	45.15	30.60	12.06	10.88
13.	T0953s1TS197_1-D1	45.15	30.59	12.95	11.08
14.	T0953s1TS149_1-D1	44.78	31.71	9.69	13.27
15.	T0953s1TS261_1-D1	44.78	29.85	10.95	11.03
16.	T0953s1TS406_1-D1	44.78	29.85	10.95	11.03
17.	T0953s1TS457_1-D1	44.78	29.85	10.95	11.03
18.	T0953s1TS044_1-D1	44.78	29.85	10.95	11.03
19.	T0953s1TS224_1-D1	44.78	30.23	8.50	13.35
20.	T0953s1TS274_1-D1	44.78	30.23	9.69	10.97

Quaternary is important for regular modelling

- Overall bad predictions (best GDT_TS 37.16)
- Homodimer with very large interface (3300 Å^2)
- 366 (best assembly group) is best prediction (GDT_TS 37.16, QCS 69.00). Pretty good from manual inspection.
- Next best prediction (214_1) is good in N-terminal but the helix in C-terminal is folded in

CASP:
Image redacted

T0991-D1

Quaternary is important for regular modelling

There are a few more examples:

- T0998 (mentioned in Multicom’s presentation)
- T0973 (mentioned in Zhang’s group and Seok’s talk). TBM-easy target!
- H0957
- T0981
- T0989 (mentioned in Read’s talk as a problem in refinement)

Question: can quaternary modelling become mainstream? What are the obstacles?

About half of the targets were oligomeric (representative of the PDB)

Acknowledgements

Dmytro Guzenko (see poster also!)

RCSB PDB PROTEIN DATA BANK

UC San Diego

Andriy Kryshtafovych
Bohdan Monastyrskyy

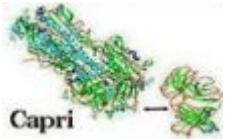
Aleix Lafita

EMBL-EBI

Spencer Bliven

PAUL SCHERRER INSTITUT PSI

Marc Lensink
Shoshana Wodak



Software:
OpenStructure and BioJava developers