Assessment of **EMA** in CASP14

(Evaluation of Model Accuracy)



Jonghun Won, Sohee Kwon, and Chaok Seok

Department of Chemistry, Seoul National University

Andriy Kryshtafovych

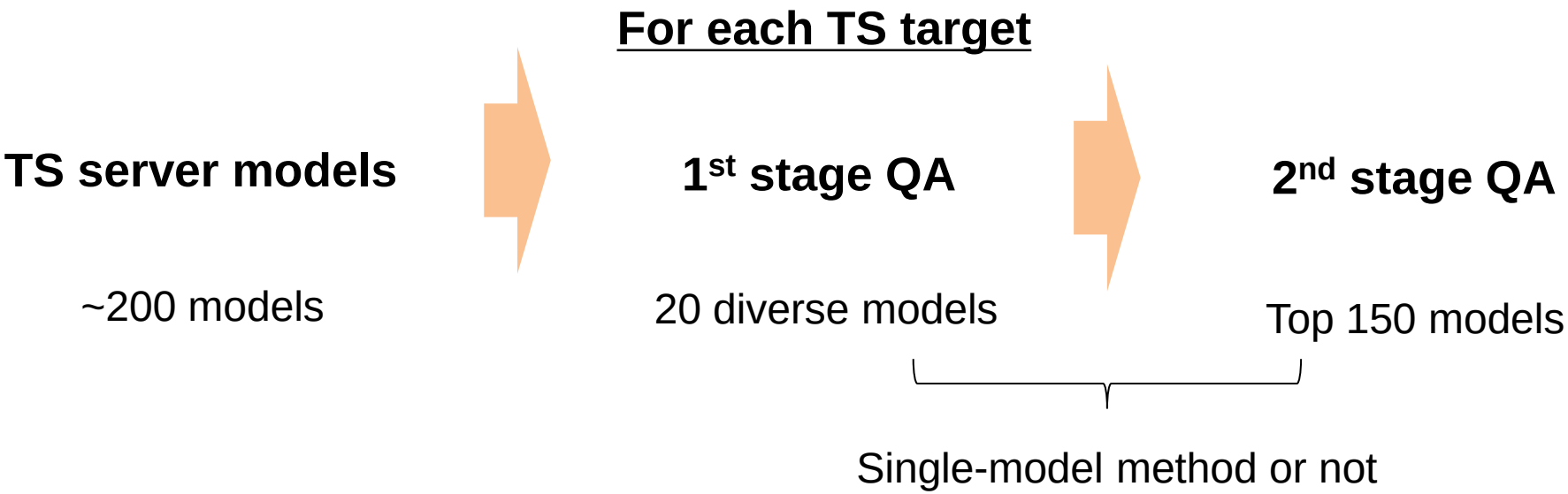
Genome Center, University of California, Davis

CASP Organizers

What useful things can QA do after AlphaFold2?

QA participants did not have a chance to score AlphaFold2 models in CASP14.
(except for a bit of post-analysis for CASP-COVID)

EMA/QA (Quality Assessment) of 3D models generated by TS servers



Global and Local QA:

- Global QA score (0~1) for each model (e.g. GDT-TS or LDDT)
- Local QA score (Å) for each residue of each model (distance deviation upon superposition)

Group Statistics

- Global QA
  - # groups = 73 (51 in CASP13)
  - g005, g044, g428: submitted only ~10% targets
  - g082: ~15%; g398: ~55%

- Local QA
  - # groups = 38 (29 in CASP13)
  - g082: ~15%

Target statistics

- Global QA targets (whole seq): 71 targets (79 in casp13)
  - N(GDT-TS≥40) ≥ 1 (out of 150): 58 targets
  - N(LDDT≥40) ≥ 1 (out of 150): 64 targets
  - (removed after manual inspection: T1048, T1062, T1072s1, T1070, T1080)

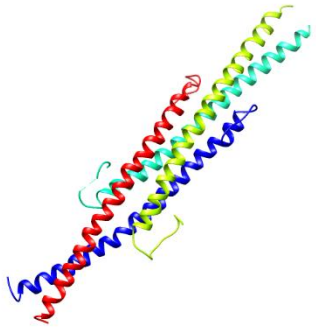
- Local QA targets (EU-wise): 94 EUs (108 in casp13)
  - N(GDT-TS≥40) ≥ 10 (out of 150): 90 EUs
  - (removed after manual inspection: T1070-D1, T1080-D1)

Removed targets after manual inspection

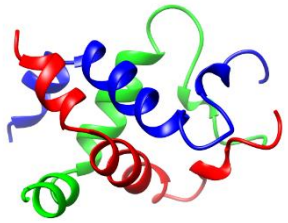
Oligomeric targets whose monomers show no core structures
(QA of oligomeric targets → CAPRI QA)

Removed from both global and local QA

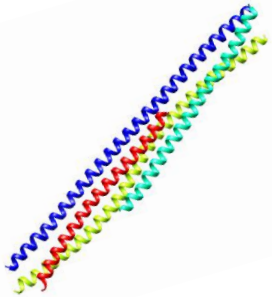
T1048 (A_4)



T1062 (A_3)

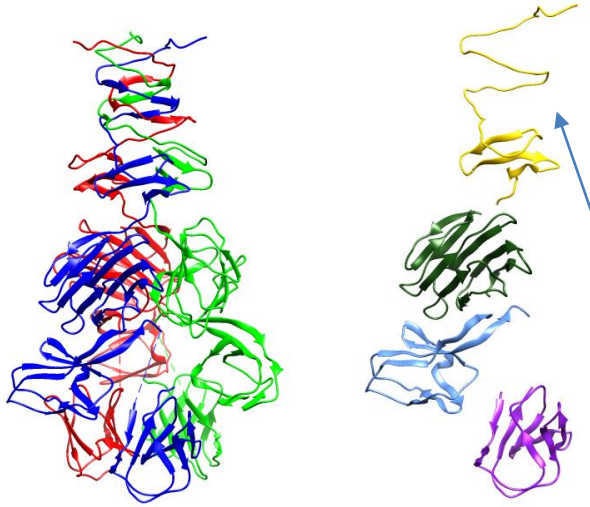


T1072s1 (A_2B_2)

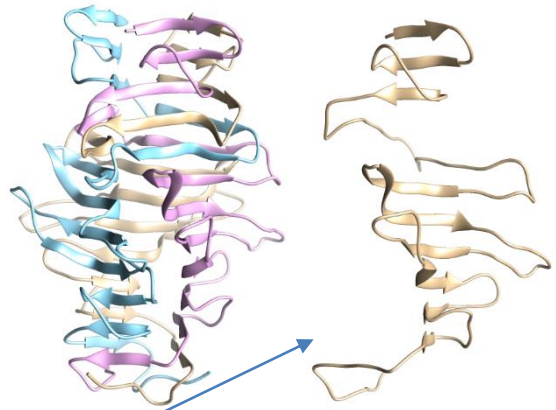


Removed from global QA

T1070 (A_3)



T1080 (A_3)

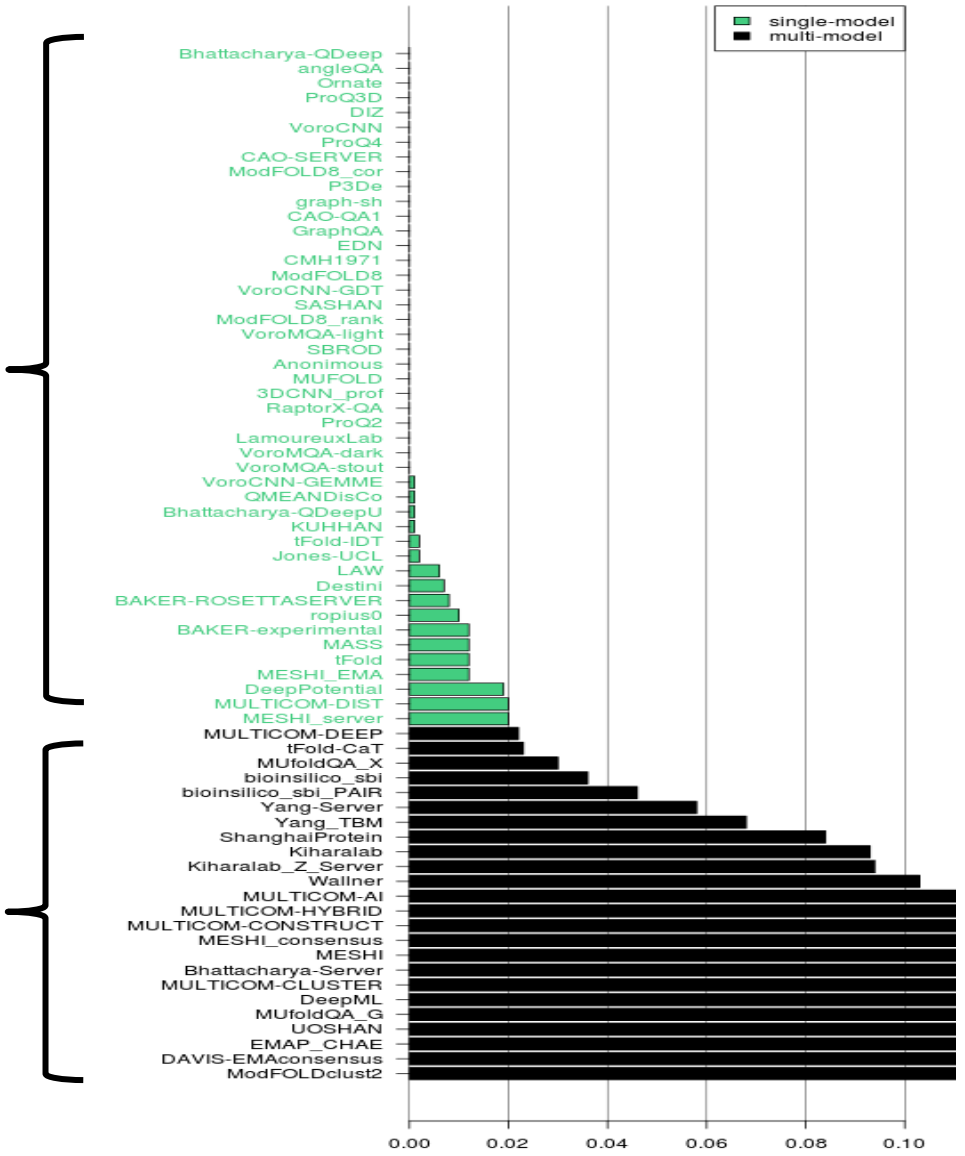


Removed from local QA (D1)
by the criterion N(GDT-TS≥40) ≥ 10 (out of 150)

Difference between stage 1 and stage 2

Single-model methods
(CASP-independent
performance)

Multi-model methods
(Performance
in non-CASP situations
can be different)



Please contact me if you think that your method is misclassified.

How can QA contribute?

Scoring models after structure prediction

Global QA to select the best models
Local QA to identify inaccurately/accurately modeled regions
(with biomedical applications in mind)

Scoring models for better structure prediction

Global QA to guide conformational sampling during iterative structure prediction
Local QA to detect inaccurately modeled regions to improve (e.g. by refinement)

Ranking global QA results (1/2)

Structure quality of top 1 model selected by QA

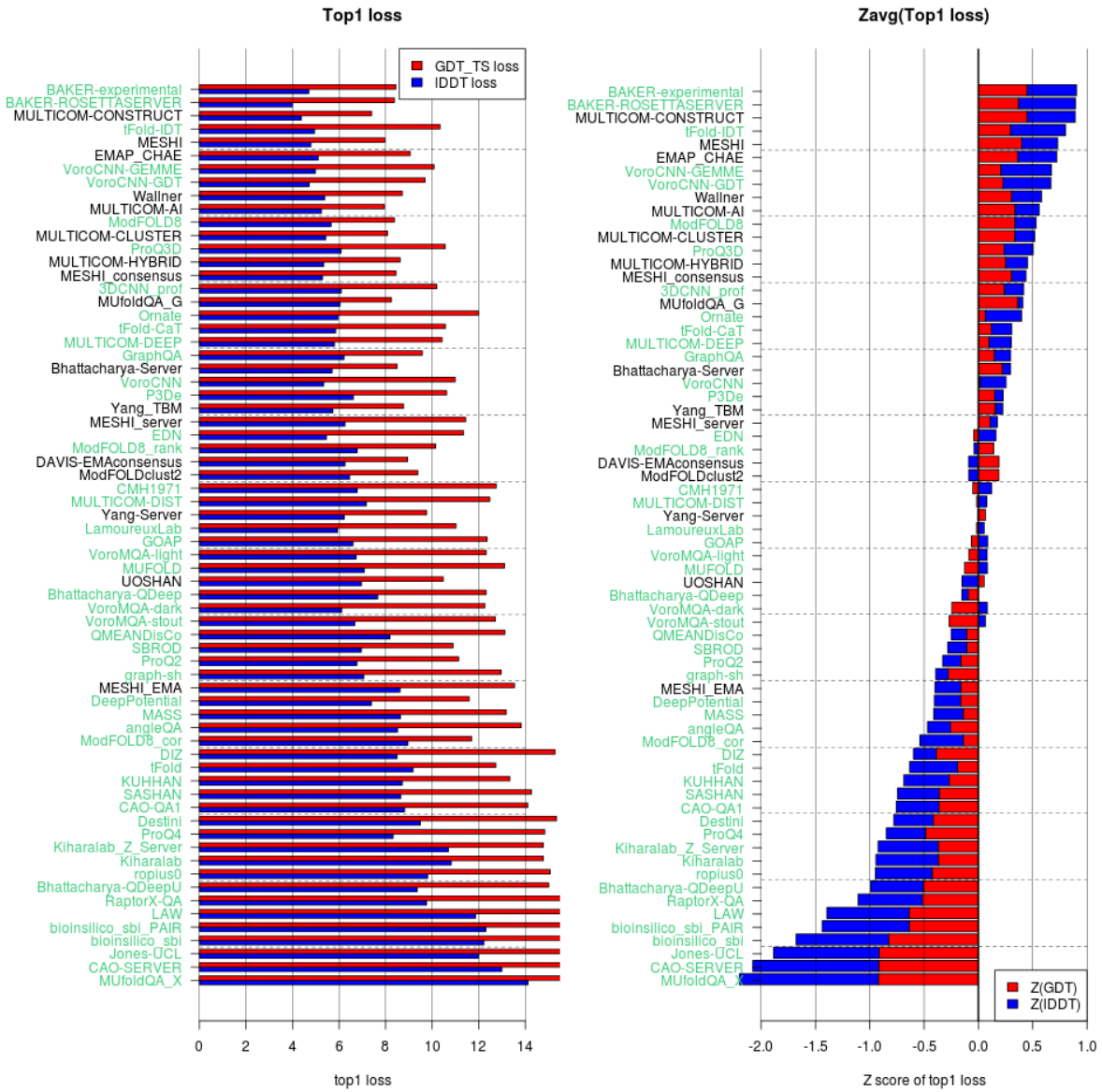
(Assessment for top 5 models resulted in very similar ranking.)

GDT-TS loss = |(GDT-TS of top 1 model) – (best GDT-TS)|
LDDT loss = |(LDDT of top 1 model) – (best LDDT)|

Global QA ranking by sum of Z-scores for GDT-TS and LDDT

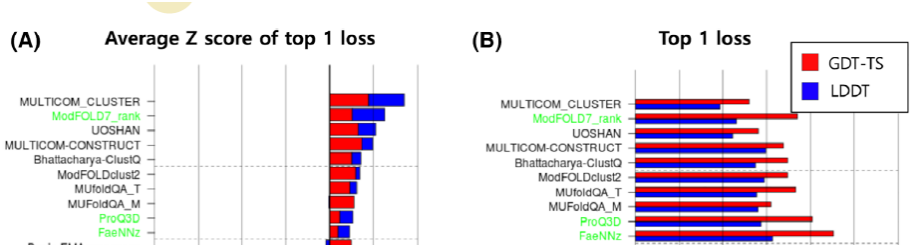
Z-score calculated by the standard CASP procedure with minimum z-score of -2.
Penalty of -2 for un-submitted targets.

Global QA results (1/2): Ranking in Top1 loss



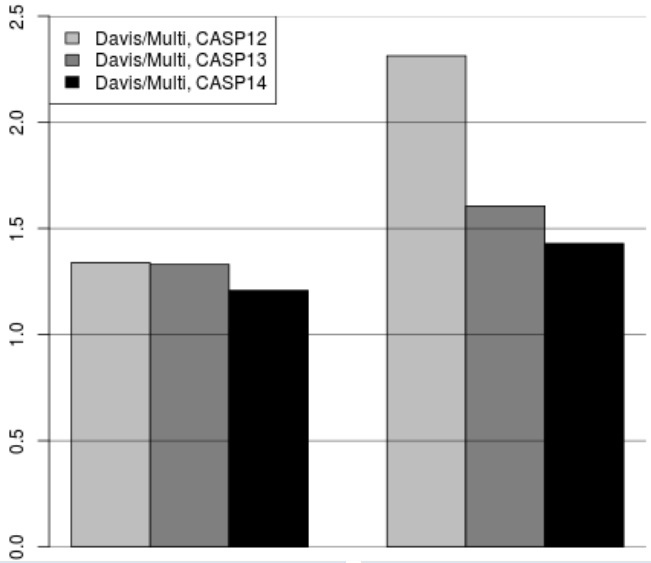
Best **single**-model methods:
BAKER-ROSETTASERVER
BAKER-experimental
tFOLD-IDT

The best **multi**-model method performed worse than the best single-model method unlike in CASP13.



Progress over previous CASPs

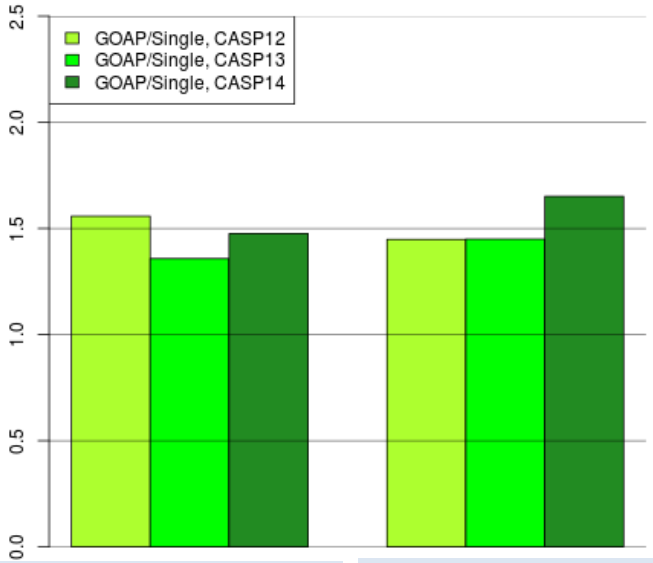
Performance of the best **multi-**model method got worse when scaled by that of **DAVIS**



GDT-TS loss
CASP 12,13, 14

LDDT loss
CASP 12,13, 14

Performance of the best **single**-model method improved when scaled by that of **GOAP**



GDT-TS loss
CASP 12,13, 14

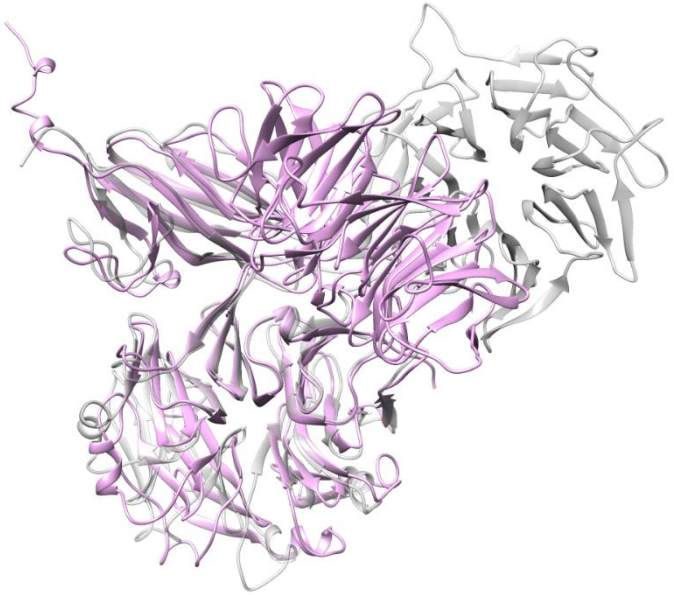
LDDT loss
CASP 12, 13, 14

Davis-EMAconsensus

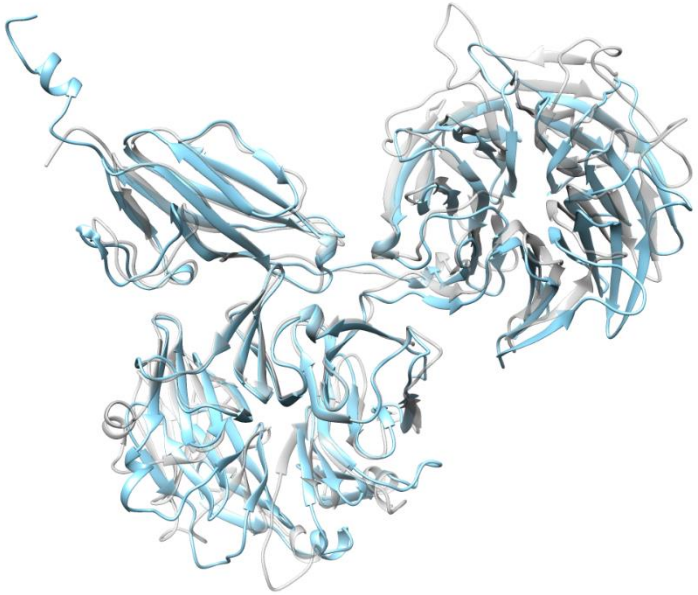
$$\text{score}_i = \left\langle \frac{N_{\text{res, model}}}{N_{\text{res, target}}} (\text{GDT-TS})_{i,\ \text{model}} \right\rangle_{\text{model}}$$

Similar LDDT, lower GDT-TS models were ranked higher (Example 1)

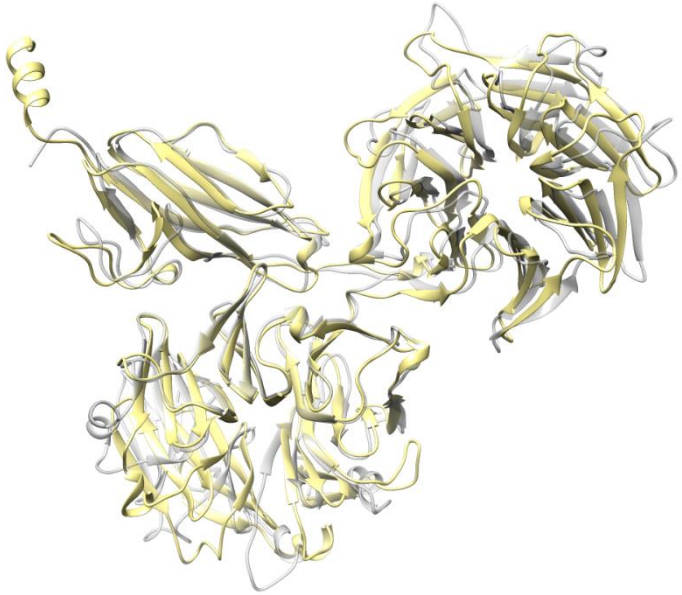
T1050



TS209_1
GDT-TS = 40
LDDT = 68
(Top 1 by 5 QA groups)



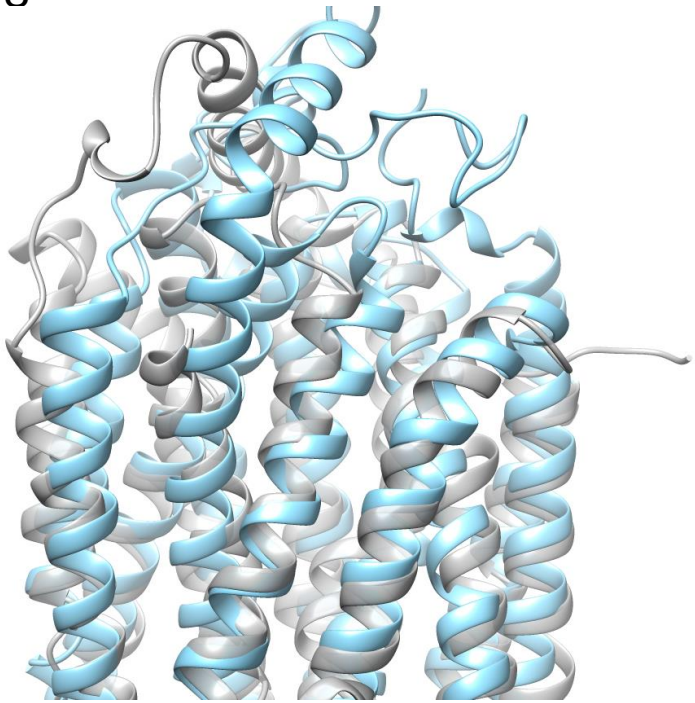
TS487_1
GDT-TS = 56
LDDT = 67
(Top 1 by 13 QA groups)



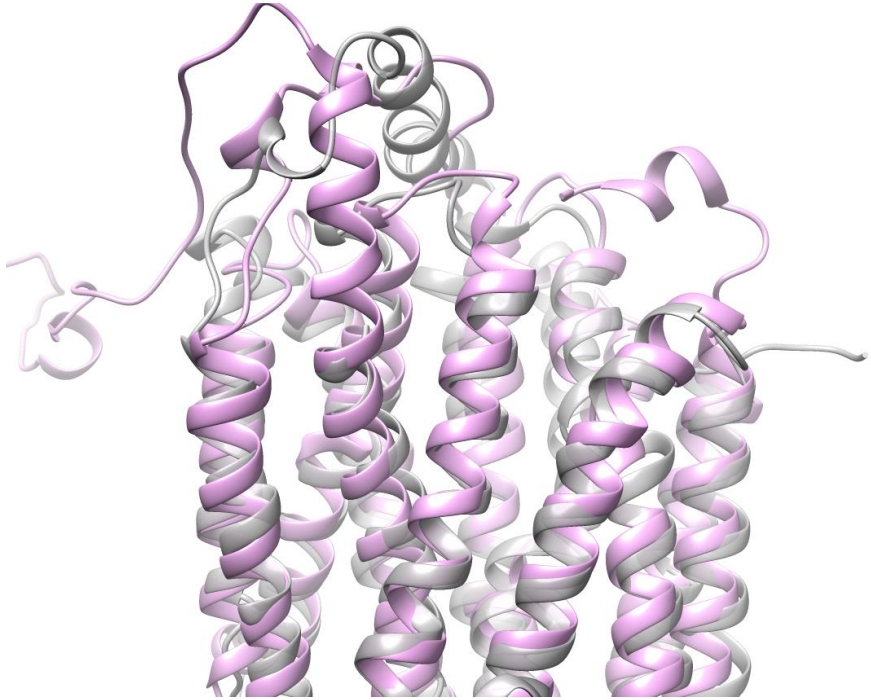
TS487_2
GDT-TS = **65**
LDDT = 67
(not selected as top 1 by any QA)

Similar LDDT, lower GDT-TS models were ranked higher (Example 2)

T1098



TS183_2
GDT-TS = 50
LDDT = 51
(Top 1 by QA247, QA325)

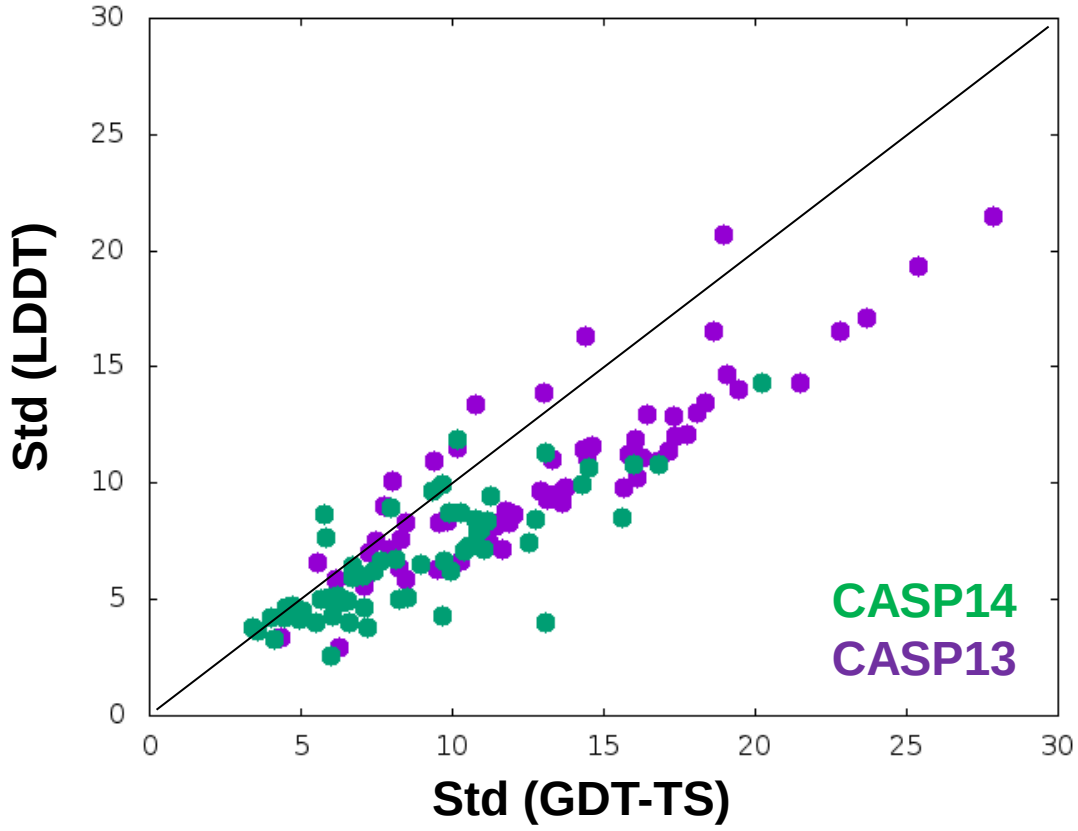


TS075_4
GDT-TS = **57**
LDDT = 52
(not selected as top 1 by any QA)

Hard to find models with large variations in LDDT with similar GDT-TS unlike in CASP13.

Top server models are now more optimized in sidechains for given backbone structures in this CASP.

In addition, server models tend to be closer to each other in this CASP.



Distribution of 150 models

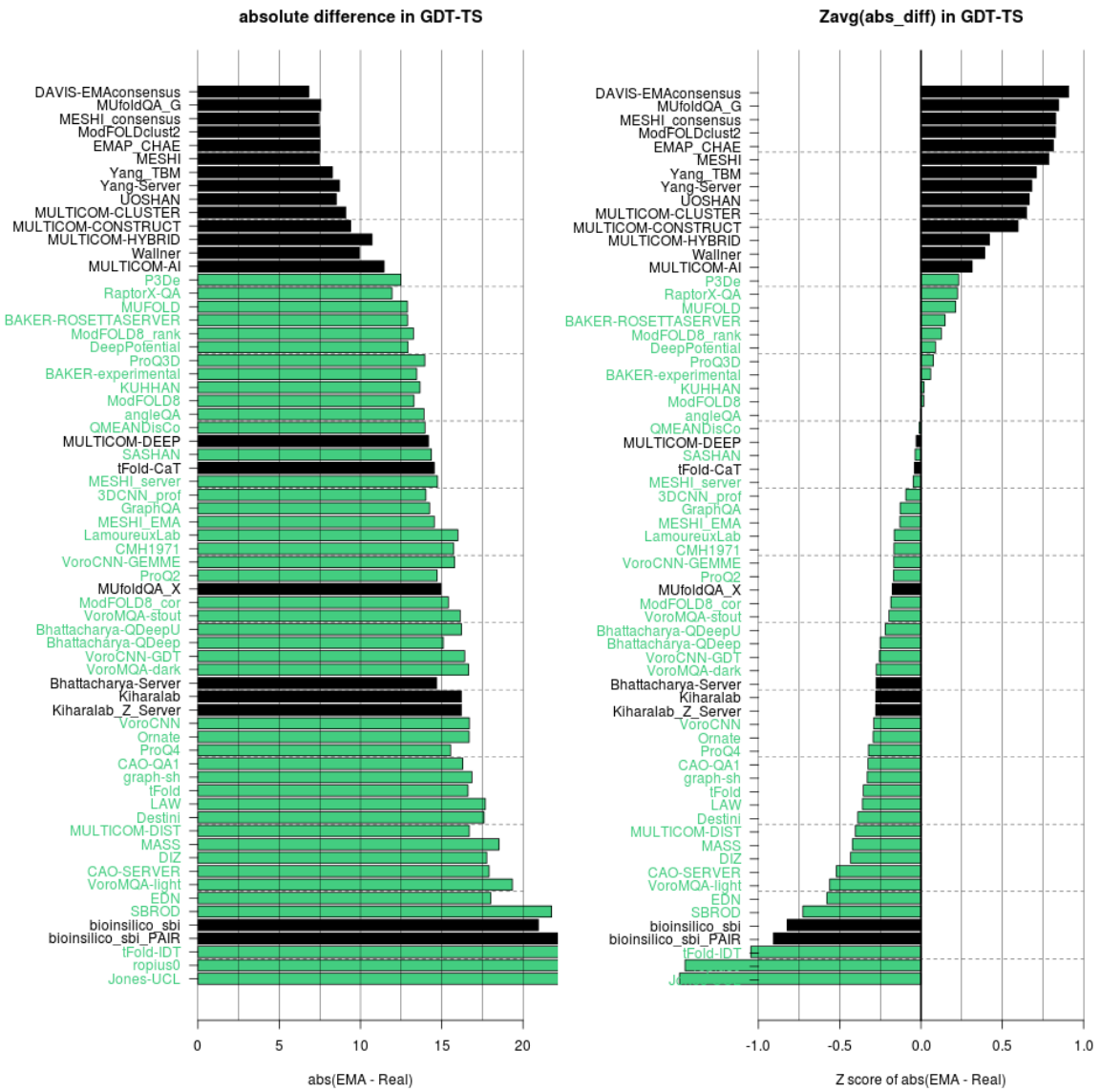
Ranking global QA results (2/2)

Absolute score

GDT-TS difference = |(QA score) – (GDT-TS of model)|
LDDT difference = |(QA score) – (LDDT of model)|

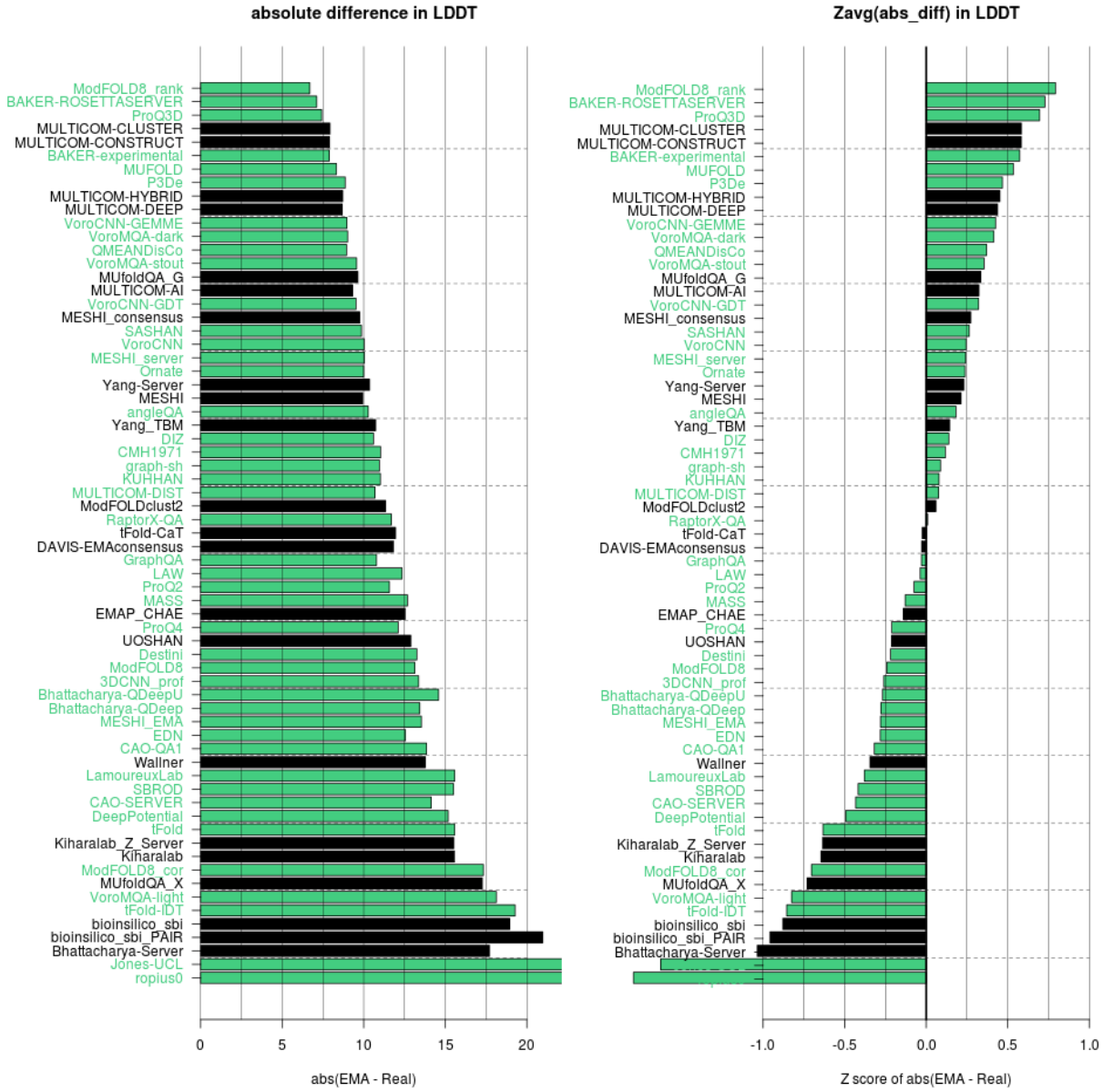
(per-model analysis)

Global QA results (2/2): Absolute GDT-TS difference



- Best absolute GDT-TS estimation by **DAVIS** with Δ=7.5 (6 in casp13)

Global QA results (2/2): Absolute LDDT difference



- Best absolute LDDT estimation by single-model methods with Δ=7 (6 in casp13)
- Best methods:
  ModFOLD8_rank, BAKER-ROSETTASERVER

Ranking local QA results

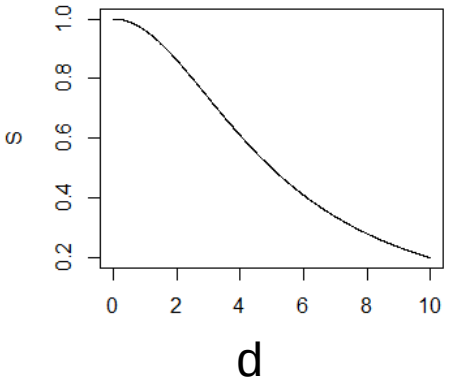
Z-score sum of three measures (ASE, AUC, & ULR-F1)

Model structures GDT-TS > 40 &
Distance deviation calculated after EU-wise LGA superposition.

- **ASE**
  Average residue-wise S-score difference

$$\mathrm{ASE} = \left(1 - \frac{1}{N}\sum_{i=1}^{N}\left|S(e_i) - S(d_i)\right|\right) \times 100$$

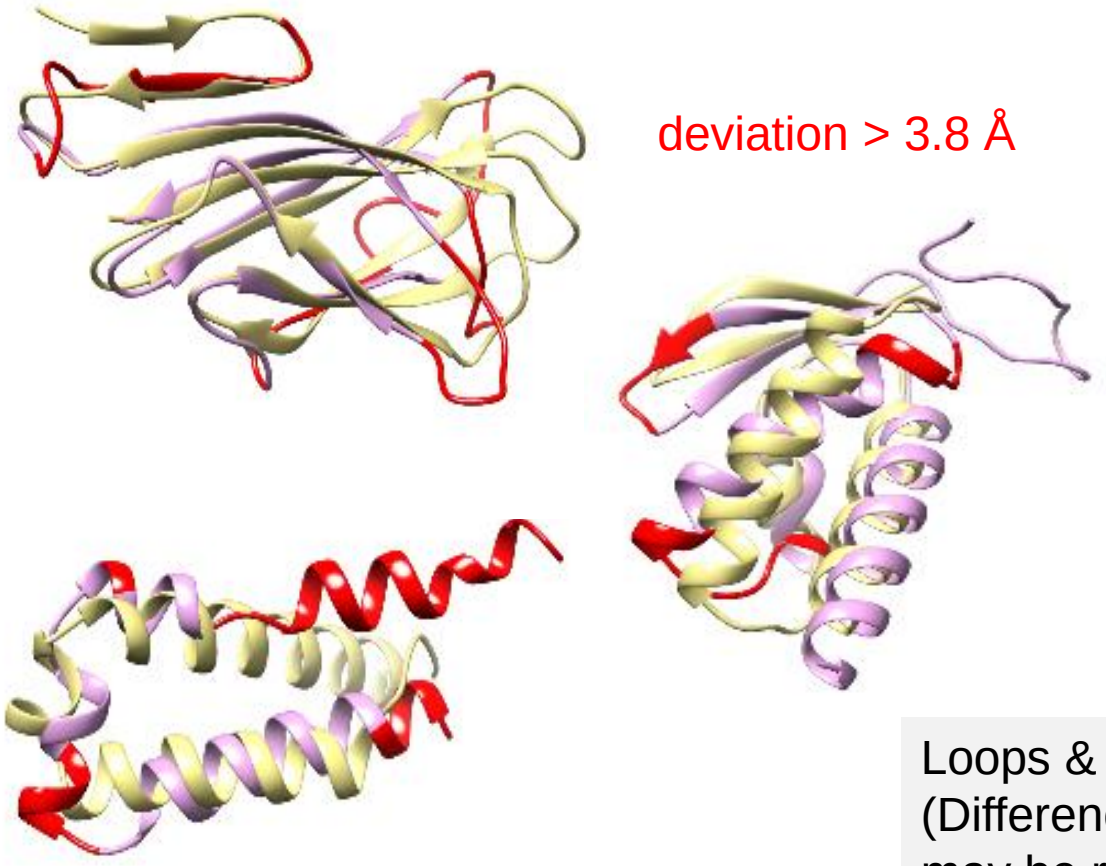
$$S(d) = \frac{1}{1+\left(d/d_0\right)^2} \qquad d_0 = 5 \text{ Å}$$

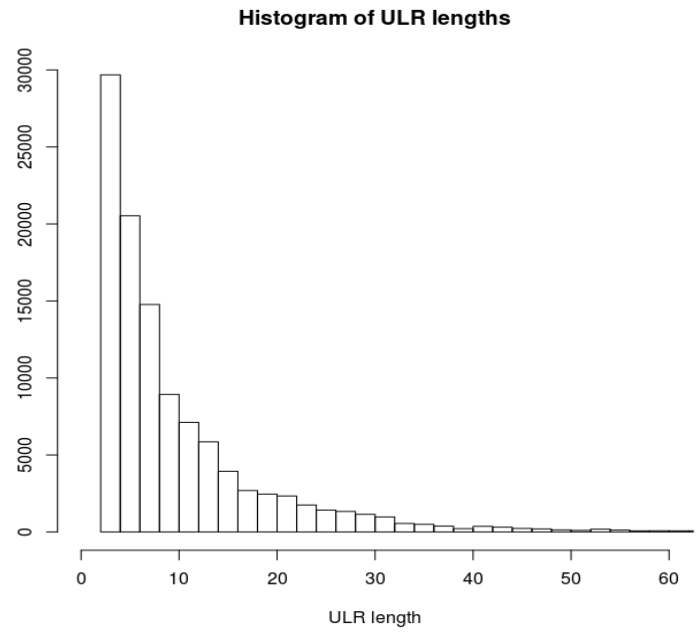


- **AUC-ROC**
  Predictions for Inaccurately/accurately modeled residues (> 3.8 Å) by varying cutoff for each methods

- **ULR-F1**
  Ability to detect inaccurately modeled regions

- **ULR** (unreliable local region):
  A region of sequential residues with distance deviation > 3.8 Å.
  (Single residues sandwiched between ULRs are united
  to neighboring ULRs, Minimum ULR length = 3)





Loops & Termini
(Differences between related proteins,
may be relevant to functional specificity)

- Assessing performance of ULR prediction F1 score
  with tolerance of +2 or -2 residues at each end of ULRs

$$F1 = 2\frac{\text{accuracy} \times \text{coverage}}{\text{accuracy} + \text{coverage}}$$

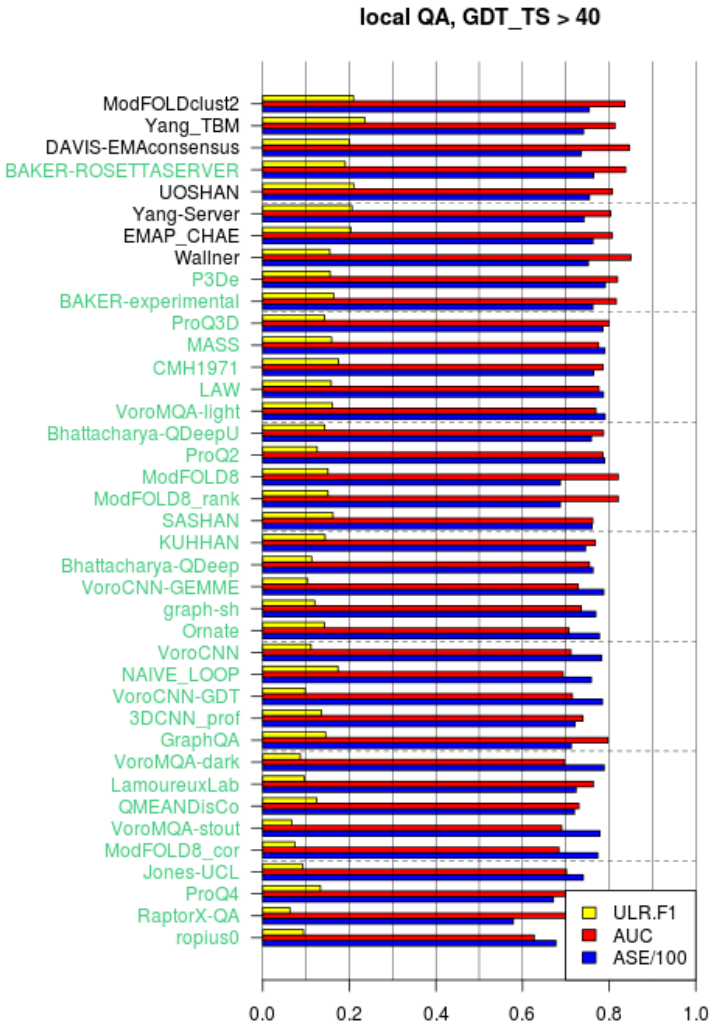
$$\text{accuracy} = \frac{\#\text{ correctly predicted ULR}}{\#\text{ predicted ULR}}$$

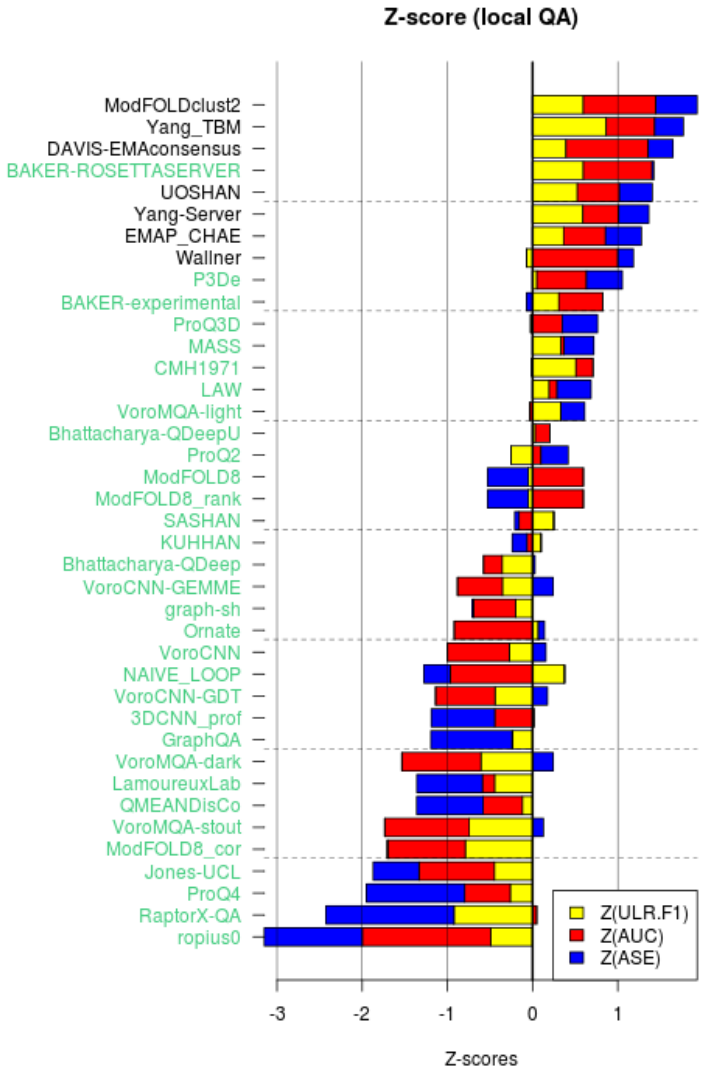
$$\text{coverage} = \frac{\#\text{ correctly predicted ULR}}{\#\text{ actual ULR}}$$

- The best score cutoff to maximize the F1 score was used for each group. (Several groups submitted scores in 0~1 scale)

- **Naïve_Loop method (a reference method for local QA)**
  Amino acid distance (Å) from the closest residue with secondary structure

Local QA ranking





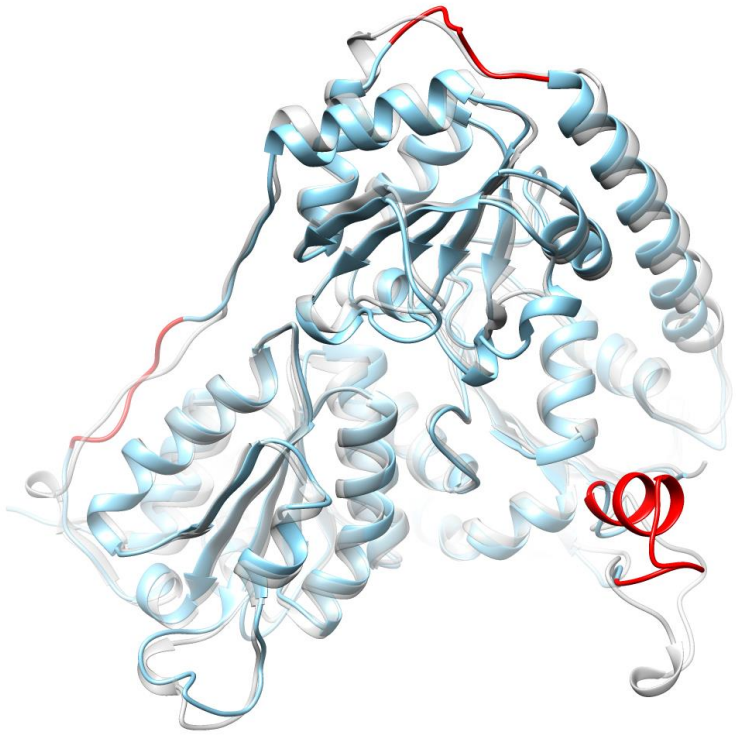
Only two **multi**-model methods did better than **DAVIS**
(Best ULR-F1 of 0.24 by **Yang_TBM**)

Best **single**-model method:
BAKER-ROSETTASERVER
(ULR-F1 = 0.19)

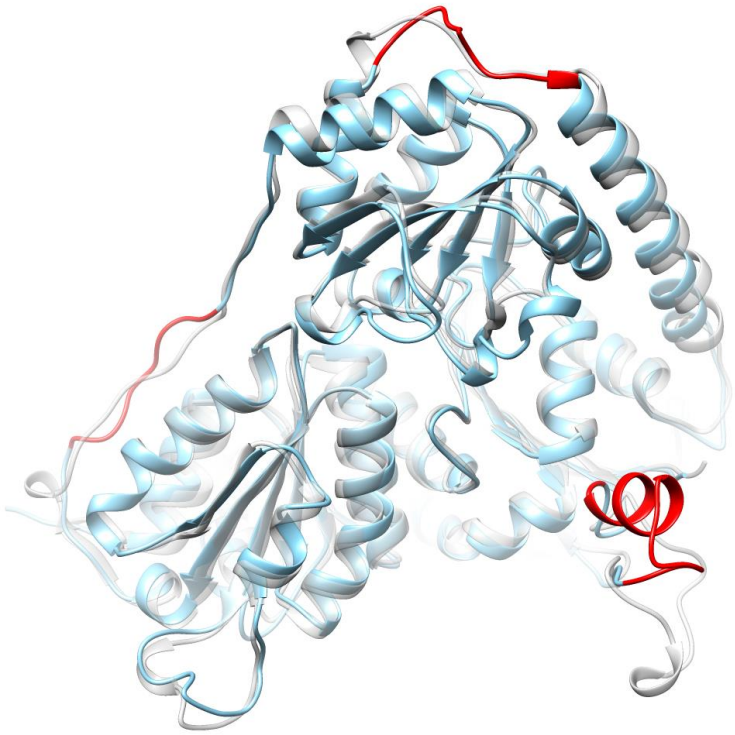
Naïve_Loop:
ULR-F1 = 0.17
ASE = 0.76
AUC = 0.7

Higher performance for multi-model methods implies common structure prediction failures in certain regions (due to less structural or sequence information) for TS-servers.

Example 1: T1076-D1 (almost perfect ULR prediction)

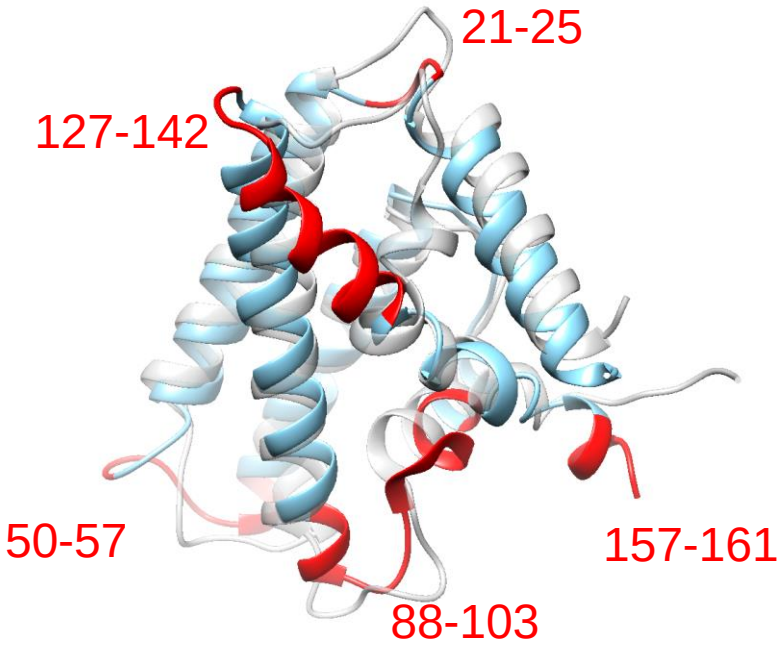


TS198_3
True ULR

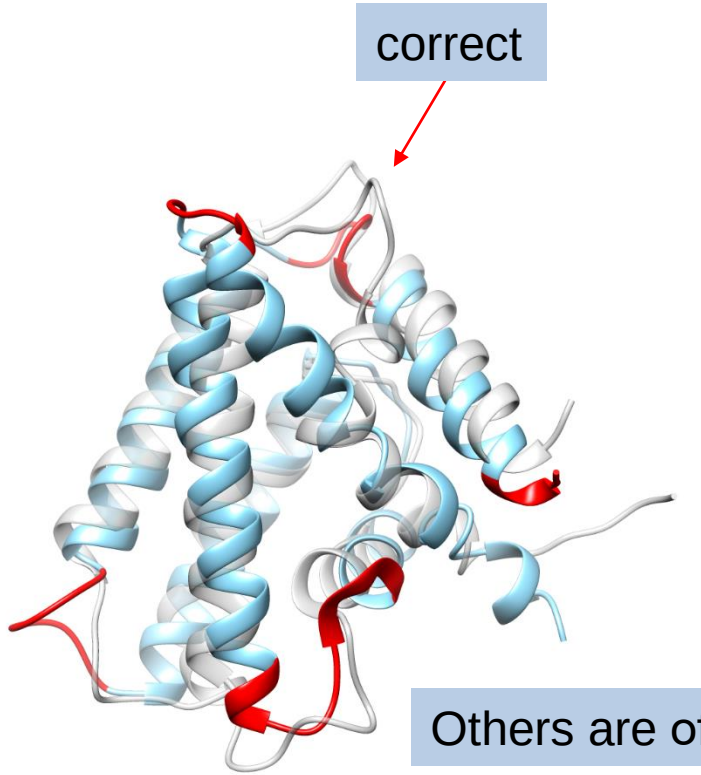


TS198_3
Predicted ULR by BAKER-ROSETTASERVER
F1 = 1.0

Example 2: T1039-D1 (partial ULR prediction)



TS031_1
True ULR



Others are of wrong lengths…

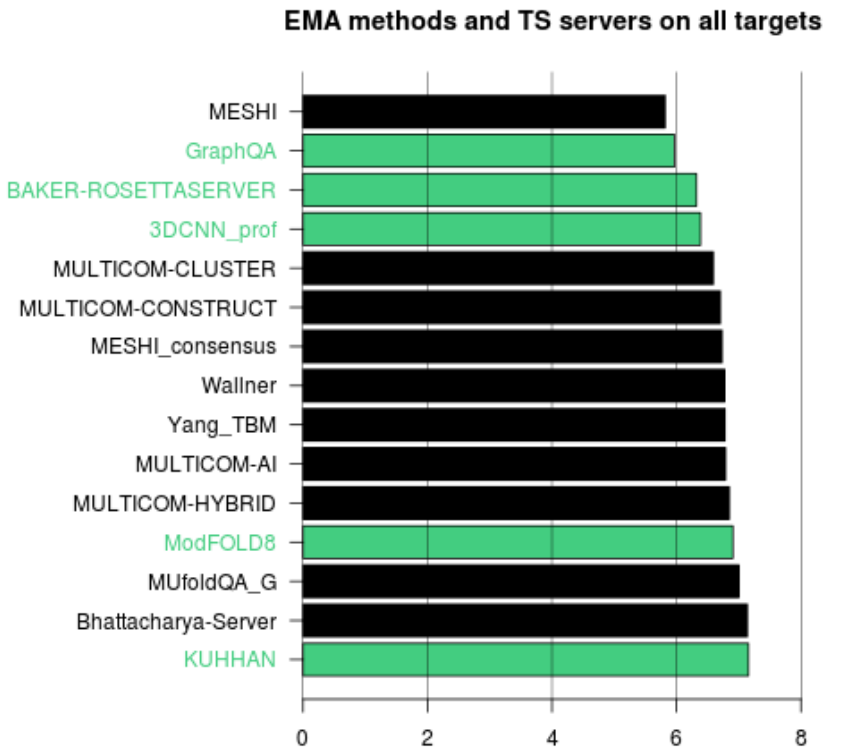
TS031_1
Predicted ULR by BAKER-ROSETTASERVER
F1 = 0.2

(More relaxed criterion→ increased F1, but no ranking change)

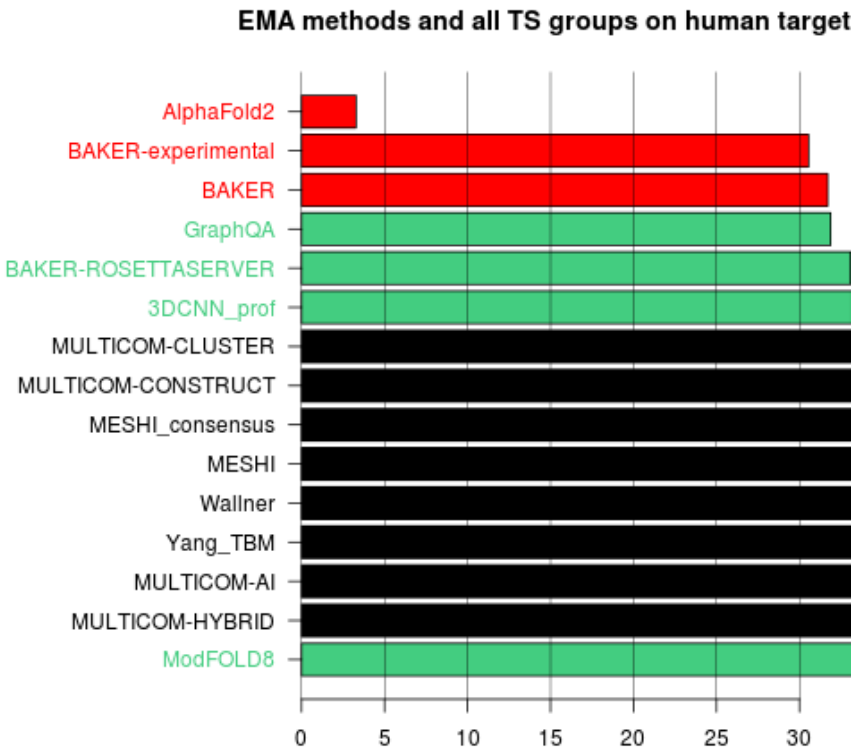
What if EMA methods participated in CASP14 as meta predictors?

(CASP-specific performance)

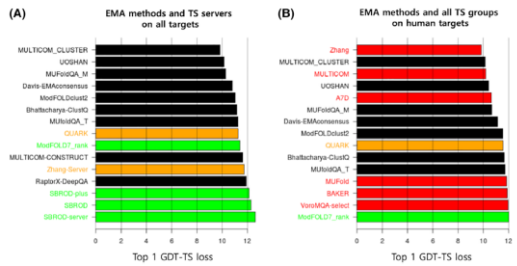
EMA methods perform better than the best TS servers,
but not better than the best TS human groups.
Top TS human groups added some values beyond consensus.



<GDT-TS difference from the best>



<GDT-TS difference from the best>



TS server group

TS human group

Can QA help TS?: Best QA Methods for Top TS-servers

(Multi-model methods were not considered here.)

TS209 may be doing something optimal for it, which is different from QA209.

TS ranked better than QA for

TS209	GDT loss	LDDT loss
TS209	**0.73**	**0.52**
QA209	0.78	0.56
QA167	0.74	0.67

TS324	GDT loss	LDDT loss
TS324	**1.79**	**0.94**
QA263	1.82	1.25
QA209	2.30	0.97

QA ranked better than TS for

TS031	GDT loss	LDDT loss
QA209	1.63	0.71
TS031	**1.26**	**1.47**
QA403	1.95	0.99

TS042	GDT loss	LDDT loss
QA263	1.22	1.06
QA209	2.06	0.85
TS042	**1.97**	**1.05**

TS226	GDT loss	LDDT loss
QA263	1.68	1.20
QA209	2.27	0.83
TS226	**2.16**	**1.66**

Round Table

5’ presentation by each of the following groups:

Nao Hiranuma (Baker group)
(BAKER-ROSETTASERVER, BAKER-experimental)

Liam McGuffin
(ModFOLD8_rank)

Sheng Wang
(tFOLD-lDT)

Lisha Ye (Yang group)
(Yang_TBM)

Questions

How can QA do better in estimating GDT-TS?
[Although low-GDT-TS/high-LDDT matters only for models of intermediate accuracy (50~70)]

How can local accuracy be estimated better?

<Future role of QA with near-perfect structure prediction>

Can QA do something useful for model structures of proteins involving conformational flexibilities or intrinsically disordered proteins?

How does your QA work differently for monomers and oligomers (if you have QA for oligomers)?

Can QA be extended to predict stabilities for monomers and binding affinities for oligomers?

Further visions regarding the role of QA?