Assessment of **EMA**

(Evaluation of Model Accuracy)

in CASP13

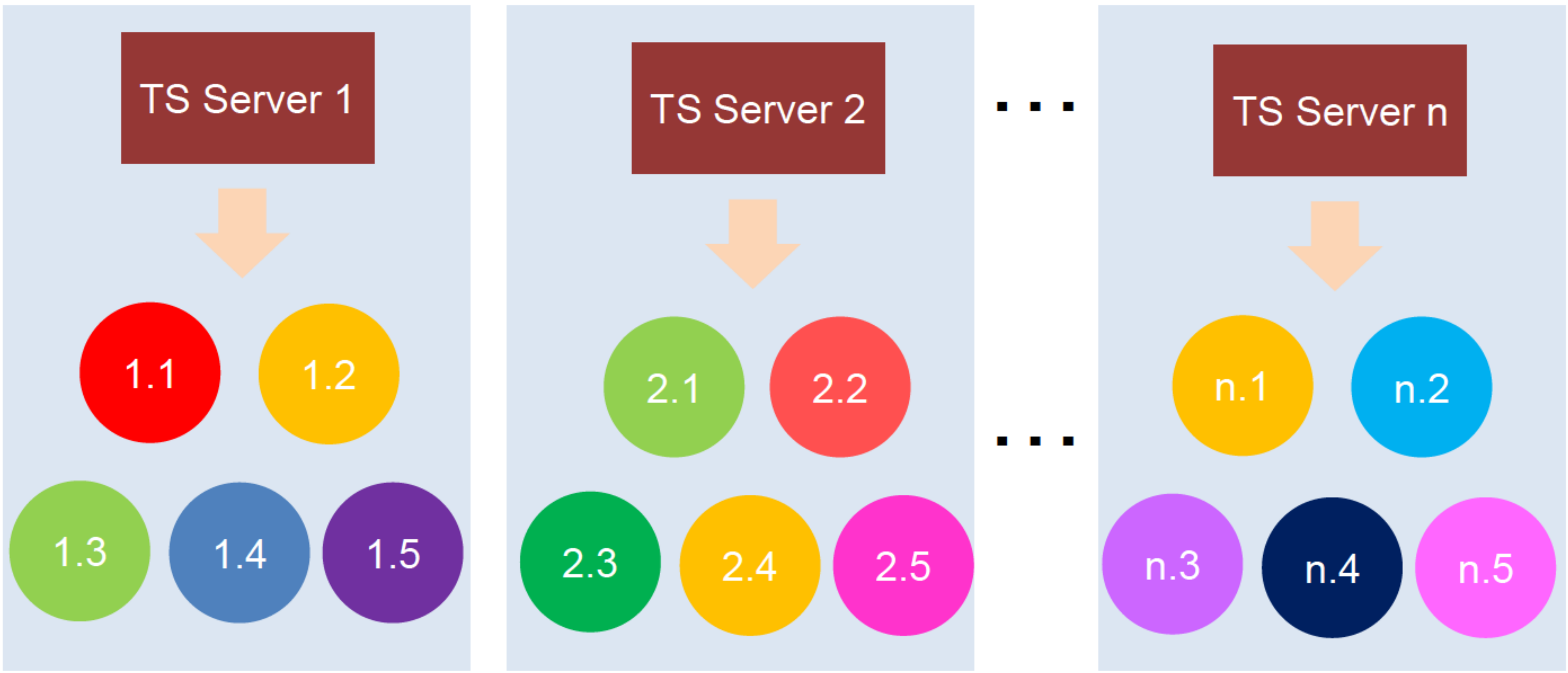
Minkyung Baek, Jonghun Won, and Chaok Seok
Department of Chemistry, Seoul National University

Bohdan Monastyrskyy and Andriy Kryshtafovych
Genome Center, University of California, Davis

John Moult, Krzysztof Fidelis, Torsten Schwede

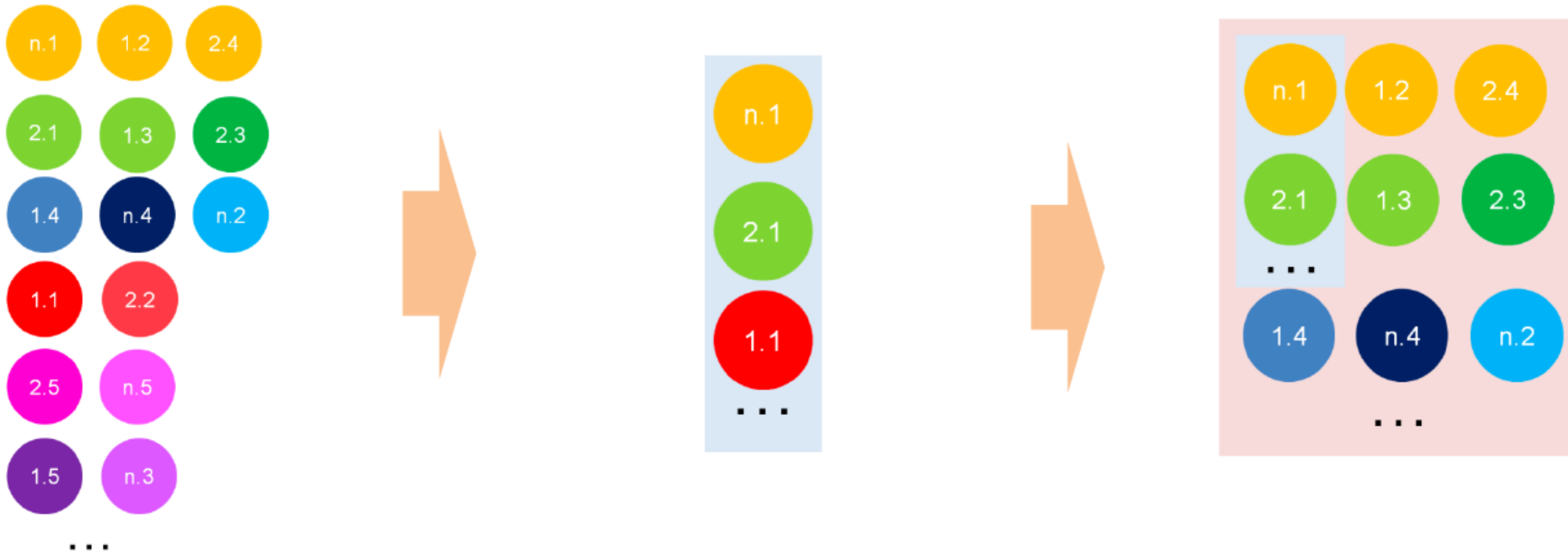
QA (Quality Assessment) of 3D models generated by TS servers

For a given TS target



175~185 server models per target

TS server models | **1st stage QA** | **2nd stage QA**



Scoring by Davis-EMAconsensus

20 diverse models

Top 150 models

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

Only those targets w/ maximum GDT-TS > 40 are assessed.

targets: 65

1^{st} stage QA

20 models

QA group.1	s1.1.1	s1.1.2	s1.1.3
QA group.2	s1.2.1	s1.2.2	s1.2.3
⋮	⋮	⋮	⋮
QA group.m	s1.m.1	s1.m.2	s1.m.3

2nd stage QA

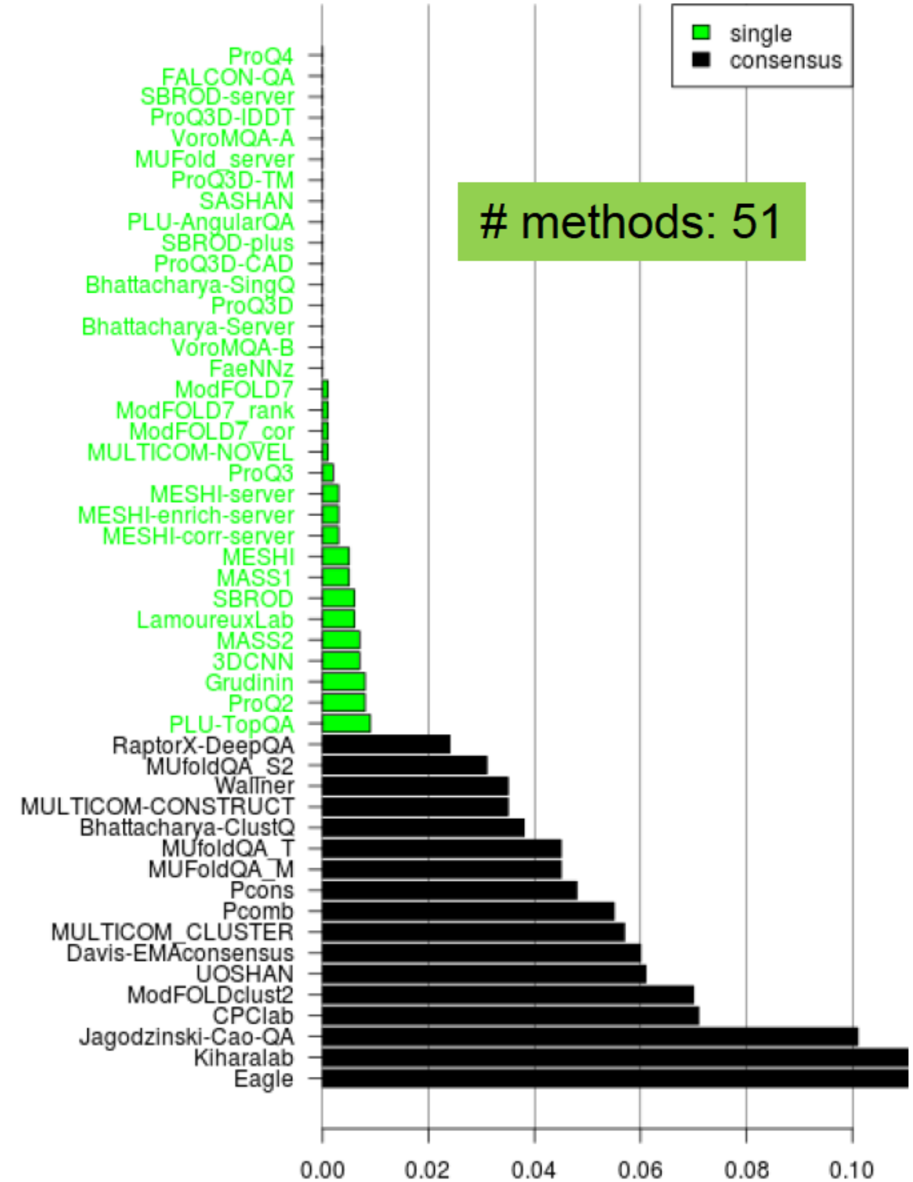
150 models

	1	2	3	4	5	.....
QA group.1	s2.1.1	s2.1.2	s2.1.3	s2.1.4	s2.1.5	
QA group.2	s2.2.1	s2.2.2	s2.2.3	s2.2.4	s2.2.5	
⋮	⋮	⋮			⋮	
QA group.m	s2.m.1	s2.m.2	s2.m.3	s2.m.4	s2.m.5	

Difference between stage 1 and stage 2

Single-model methods
(CASP-independent)

Consensus/clustering methods
(Performance
in non-CASP situations
can be different)



We did not classify quasi-single-model methods as a separate category.

Global QA and Local QA

Scores for global structure accuracy

Single score (0~1) for each of the given server models
(e.g. estimated GDT-TS/LDDT)

Scores for local structure accuracy

Single score (Å) for each residue of each model
(estimated Å deviation upon superposition)

Only 27 out of 51 groups submitted local QA scores.

How can QA contribute to the community?

Scoring models after structure prediction

Global QA to select final 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 search
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 by QA

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

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

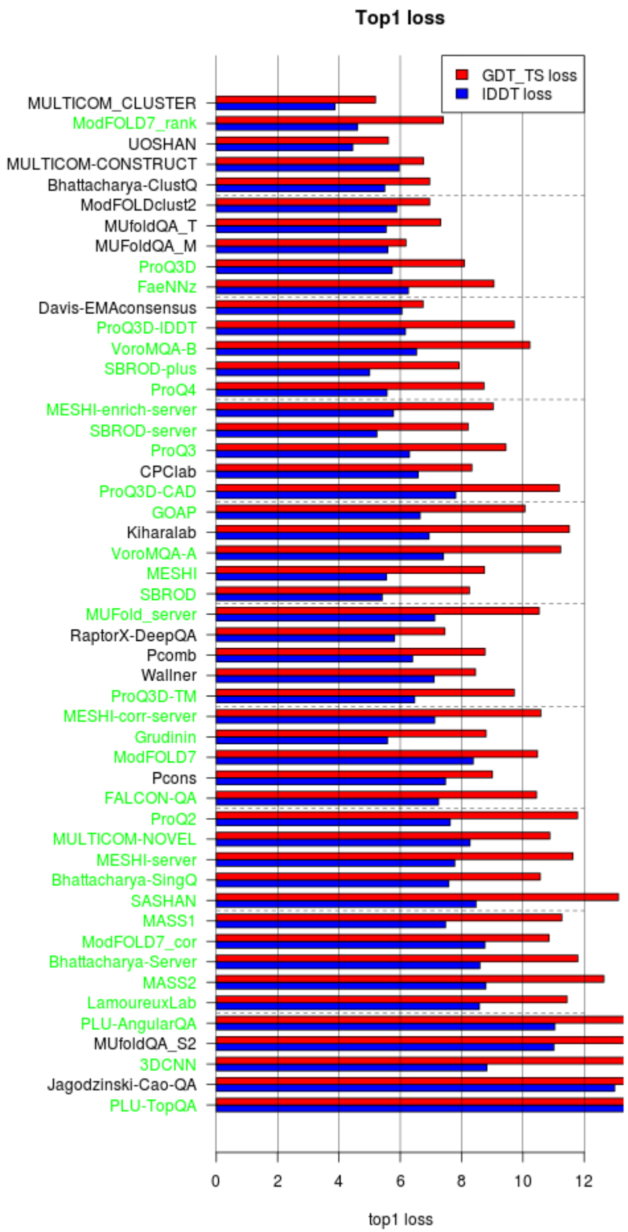
LDDT loss = |(LDDT of top 1 model) – (best LDDT)|

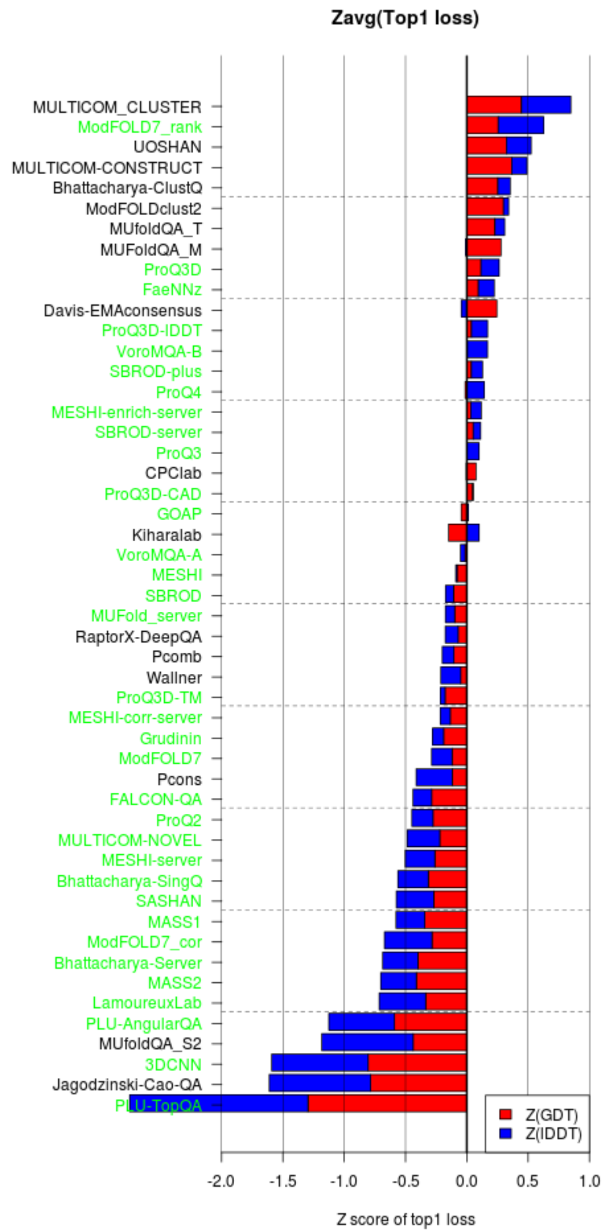
Global QA ranking by sum of Z-scores for GDT-TS and LDDT

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 consensus methods:
- MULTICOM_CLUSTER

Best single-model methods:
- ModFOLD7_rank
- ProQ3D, FaeNNz

- GDT-TS & LDDT scores are correlated.
- Single-model methods tend to do better in LDDT than GDT-TS

Ranking global QA results (2/2)

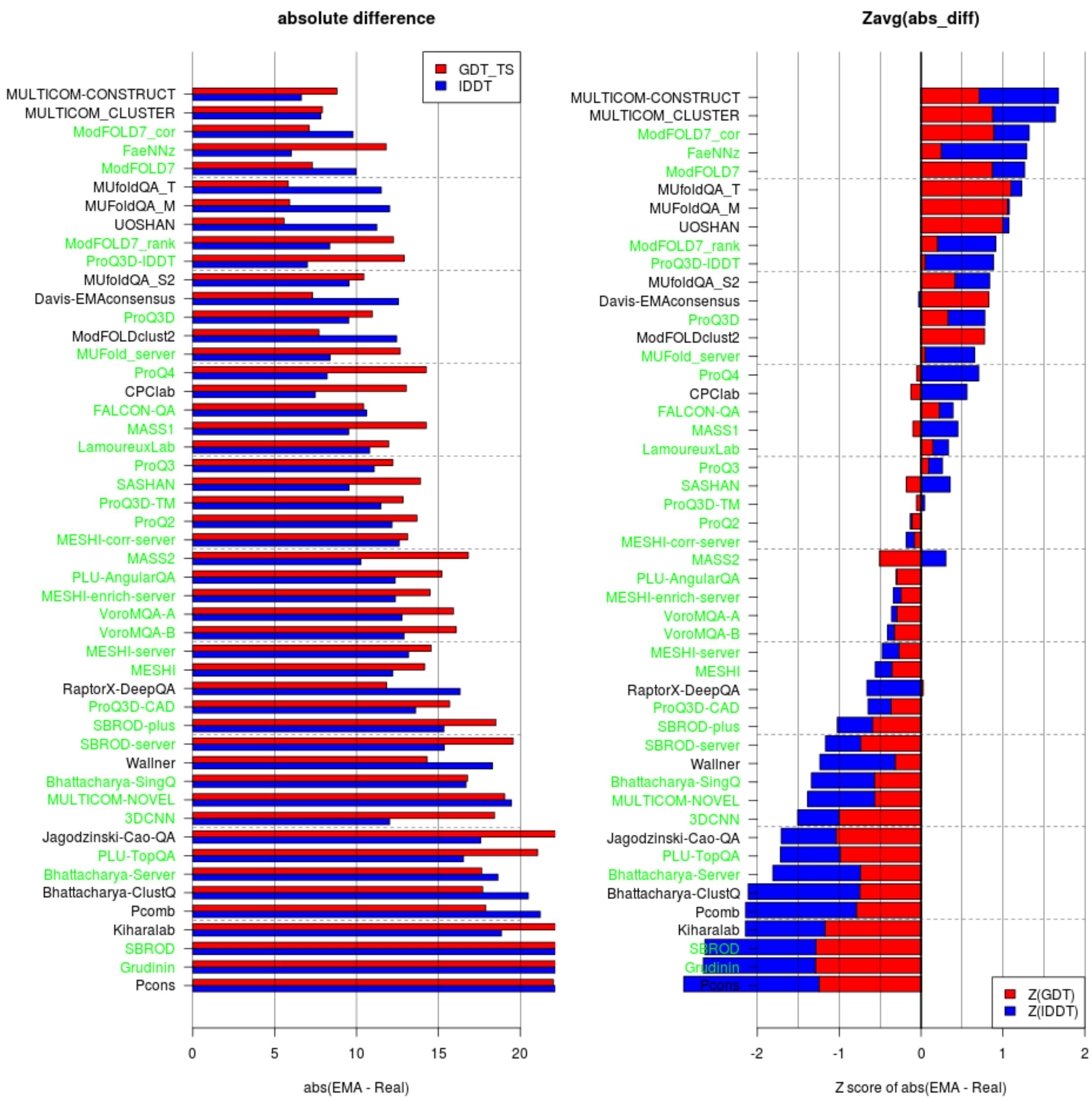
Absolute score

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

(per-model analysis)

Z-score

Global QA results (2/2): Absolute difference



Best absolute LDDT estimation by Δ~6
- FaeNNz
(single-model method)

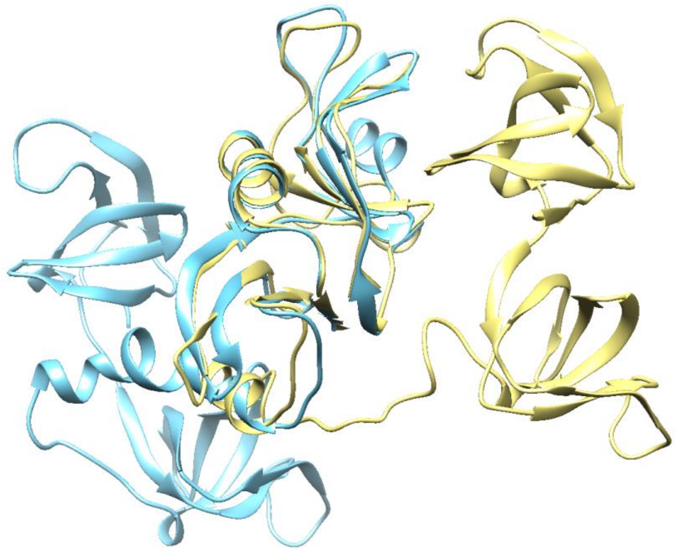
Similar GDT-TS, different LDDT (1/3)

experiment

T1002 (A1)

TS156_2

(42, 61)



TS368_3

(43, 46)

Similar GDT-TS, different LDDT (2/3)

T1004 (A3)

90

CASP:
Images redacted

Similar GDT-TS, different LDDT (3/3)

T0974s2 (A1B1)

CASP:
Images redacted

Similar LDDT, different GDT-TS (1/3)

T0973 (A2)

CASP:
Images redacted

LDDT: Contacts not present in ref structure
are not penalized

Similar LDDT, different GDT-TS (2/3)

T1022s2 (A6B3)

TS368_4

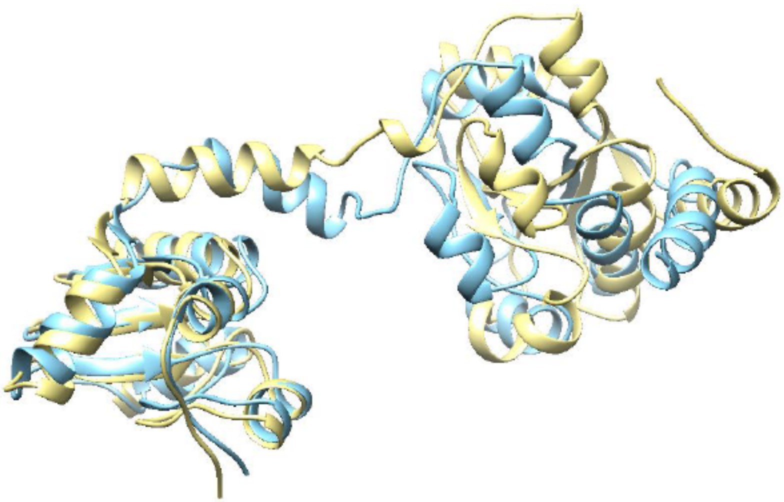
(62, 59)

TS324_1
(top1 by 10 QA groups)

(40, 55)

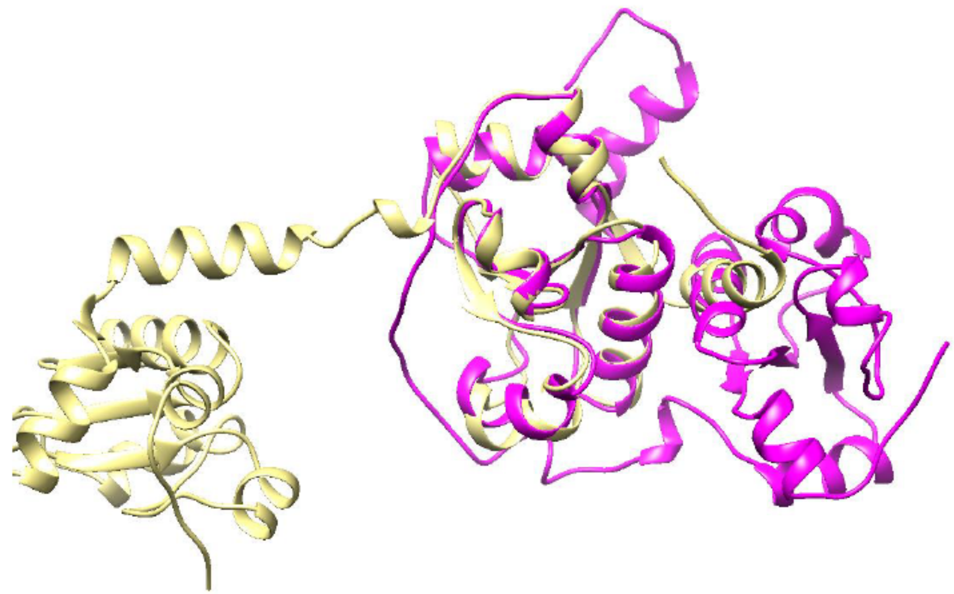
Similar LDDT, different GDT-TS (3/3)

T0976 (A2)



TS145_5

(59, 69)



TS368_3
(top1 by 4 QA groups)

(38, 68)

Issues regarding EMA assessment

- **Multi-EU (Evaluation Unit) targets** (11/65)
  - In cases where EU orientations in models are not well predicted by TS servers, models of higher LDDT are better.

Not much change in ranking when only single-EU targets are considered.

- **Oligomer targets** (**43**/65)
  - Monomer models for oligomer targets were evaluated without the full quaternary structure.
  - Global structures determined by oligomer interactions are not captured by LDDT for monomer.

Ranking local QA results

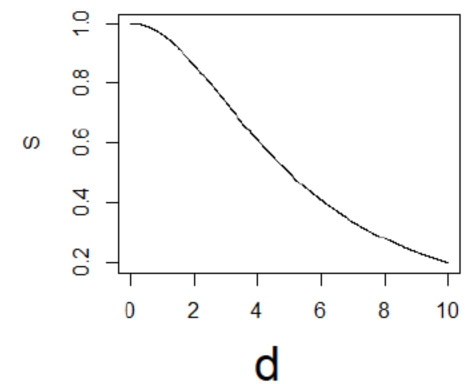
Z-score sum of three measures (ASE, AUC, & ULR F1)

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

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

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

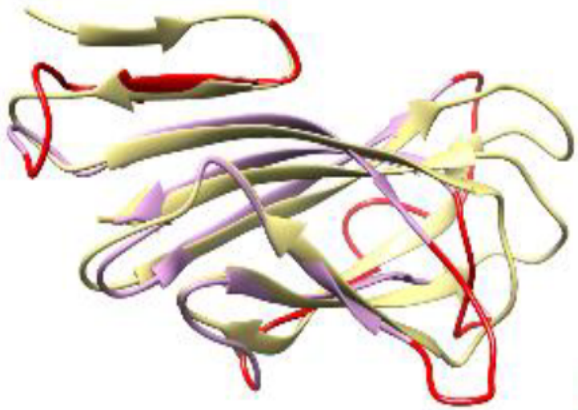
$$S(d) = \frac{1}{1+\left(d/d_0\right)^2} \quad d_0 = 5 \text{ Å}$$



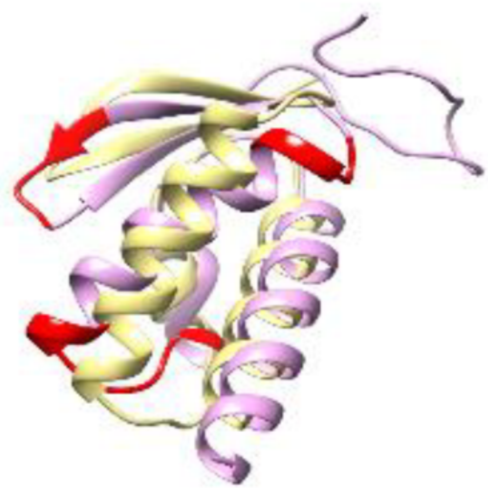
- **AUC-ROC**
  Predictions for Inaccurately/accurately modeled residues (> 3.8 Å) by varying cutoff for each methods

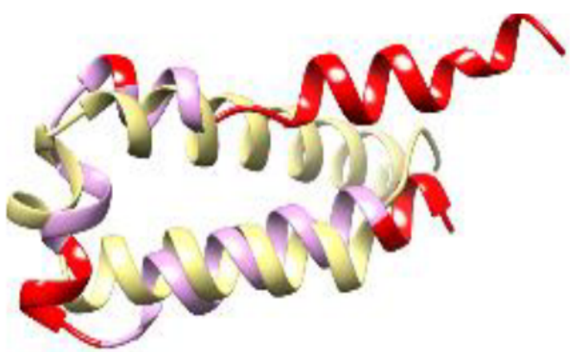
- **ULR F1**
  Ability to detect inaccurately modeled regions

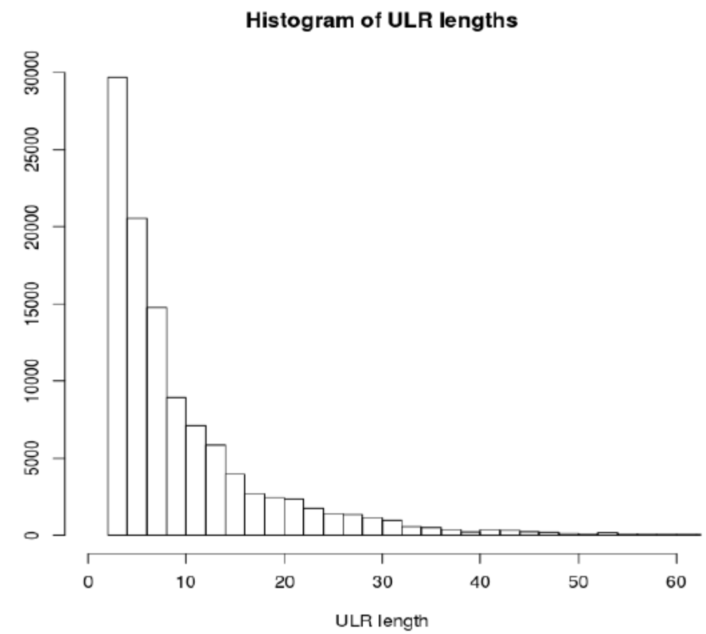
- **ULR** (unreliable local region):
  A region of sequential residues with distance deviation > 3.8 Å.
  (Single residues sandwiched between ULRs are united to neighboring ULRs, Minimum ULR length = 3)



deviation > 3.8 Å







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

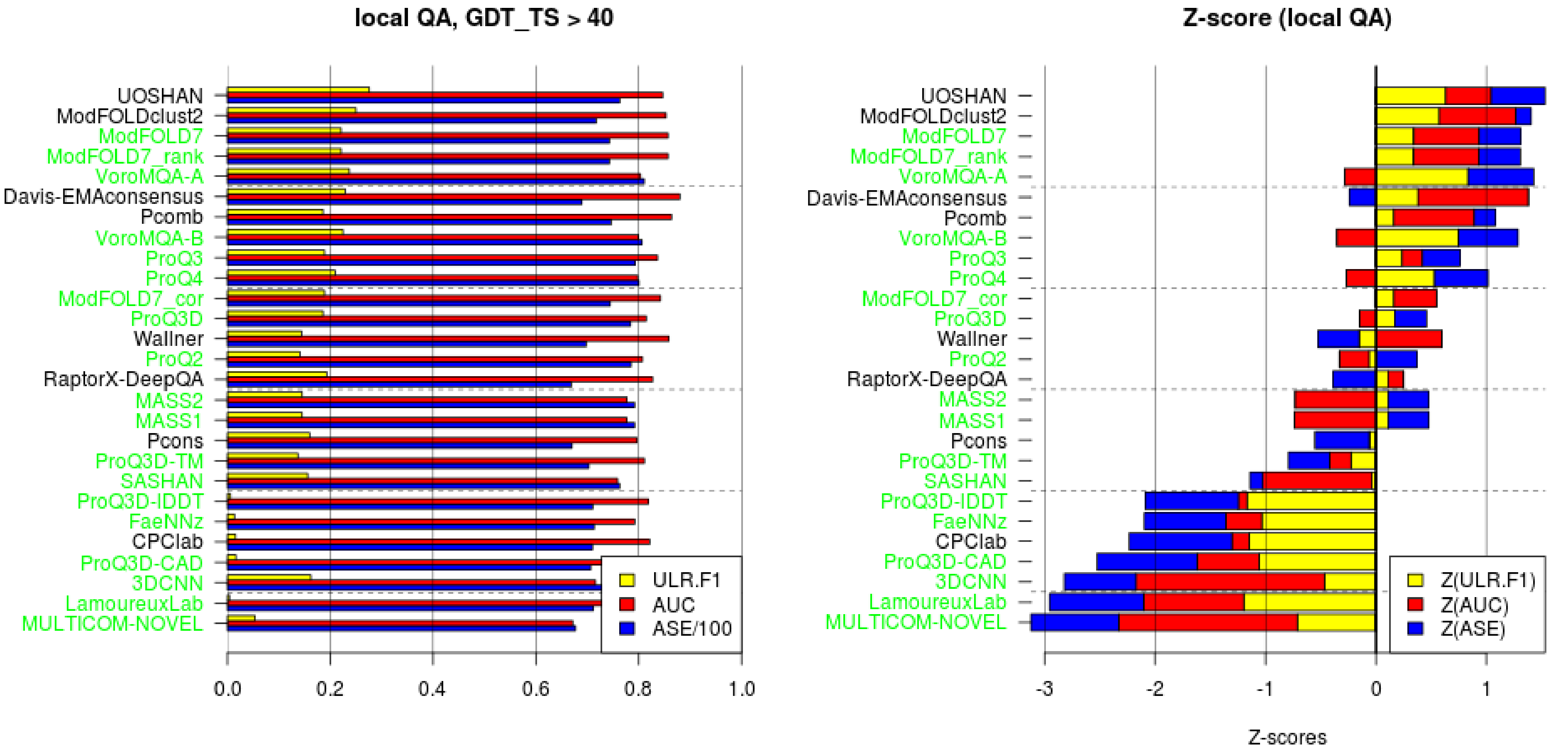
$$\mathrm{F1} = 2\,\frac{\text{accuracy} \times \text{coverage}}{\text{accuracy} + \text{coverage}}$$

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

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

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

Local QA ranking



Best consensus method: **UOSHAN**

Best single-model method groups:

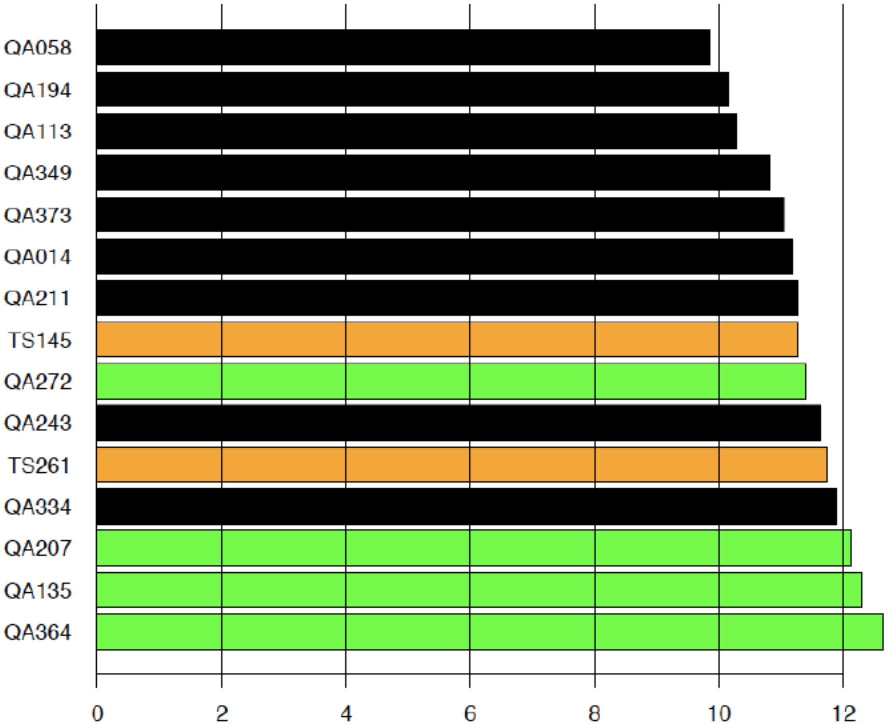
- **ModFOLD7**
- **VoroMQA** (best ULR prediction)

What if EMA methods participated in CASP13 as meta predictors?

(CASP-specific performance)

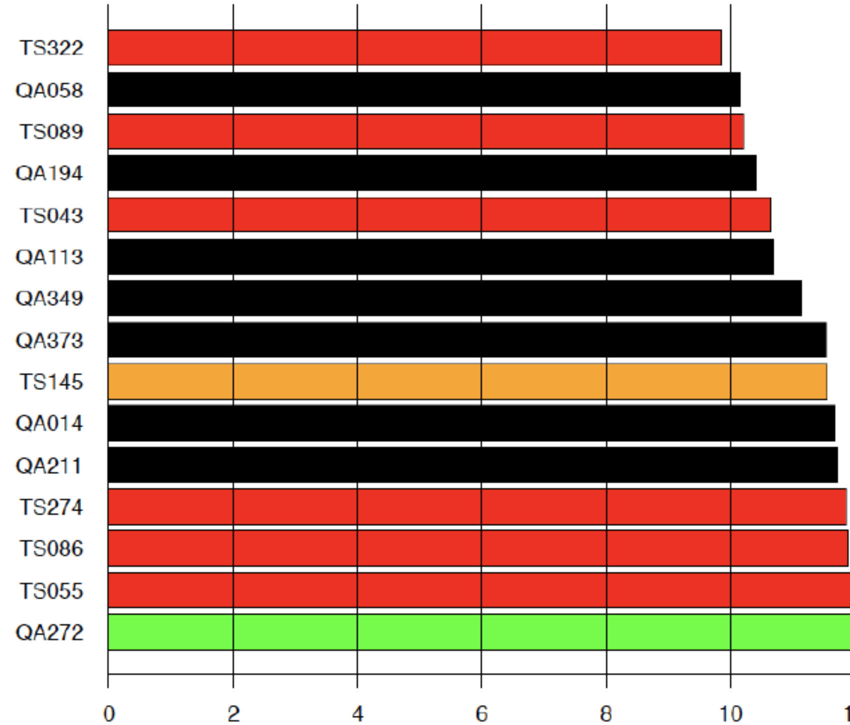
EMA methods perform better than the best TS servers,
but not better than the best TS human groups.
Top TS human groups added some, not great, values beyond consensus.

EMA methods and TS servers on all targets, MDL1



<GDT-TS difference from the best>

EMA methods and all TS groups on human targets, MDL1



<GDT-TS difference from the best>

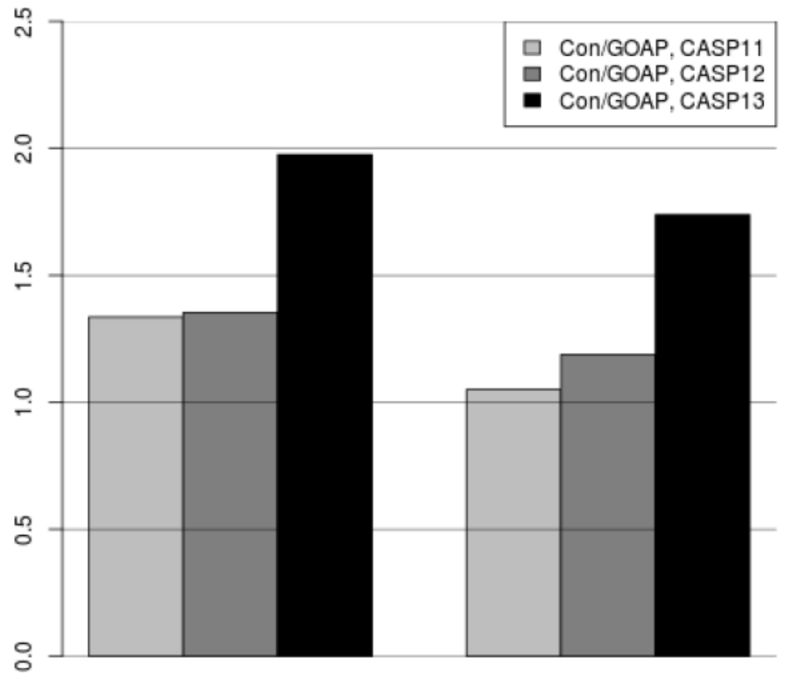
TS server group

TS human group

PROGRESS OVER PREVIOUS CASP?

Performance of the best **consensus method** improved in CASP13.

Top1 GDT-TS/LDDT loss for the best consensus method relative to **GOAP**



GDT-TS
CASP 11, 12, 13

LDDT
CASP 11, 12, 13

Performance of **consensus methods** improved because TS servers generated models of more consensus towards higher accuracy.

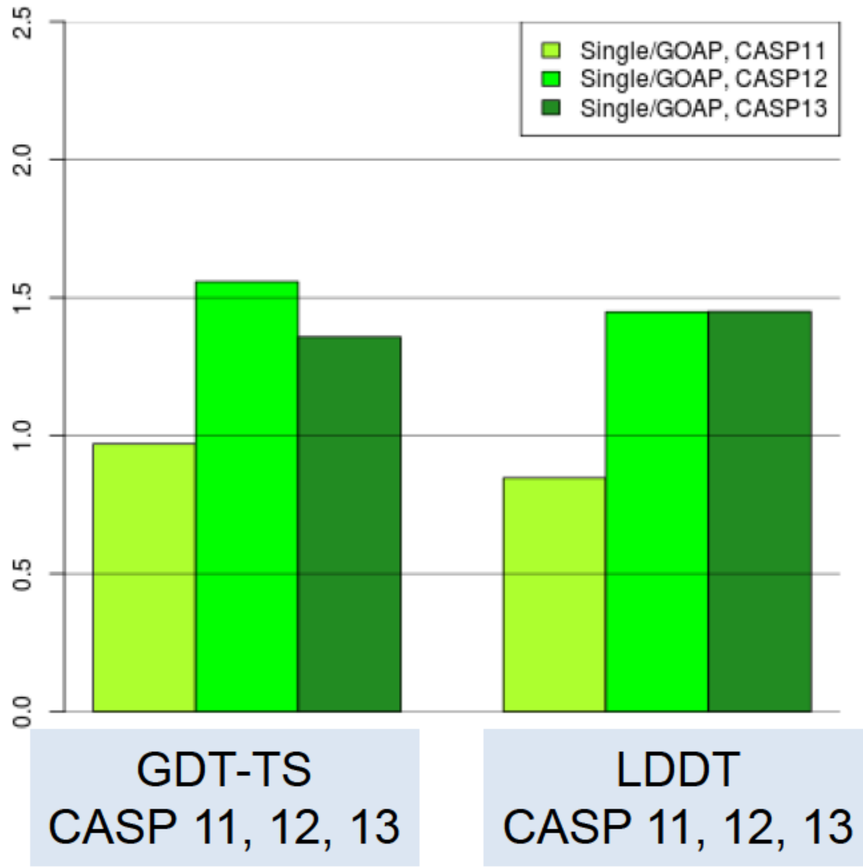
More consensus in CASP13 TS server models.
Average of pairwise GDT-TS for top10 GDT-TS models when GDT-TS of best model > 40: 40 (CASP12) → 59 (CASP13)

More higher-accuracy models for single-EU FM targets in CASP13.
Fraction of FM targets for which GDT-TS of best model > 40:
5/13 (CASP12) → 11/15 (CASP13)

More consensus for FM targets.
Davis-EMAconsensus (pure consensus) won over ProQ3 (a single model method, also tested both in CASP12&13) for higher fraction of FM targets:
1/5 (CASP12) → 8/11 (CASP13)

Performance of **single-model methods** did not improve

Top1 GDT-TS/LDDT loss for the best single-model method relative to GOAP



Single-model methods did particularly worse in CASP13 compared to CASP12 for single-EU FM targets, although consensus methods did significantly better.

Single-model methods tend to score stereochemically correct models highly. In CASP13, more high-accuracy models with poor stereochemistry were generated by TS servers for FM targets

FM target	Davis-EMAconsensus			GOAP			ProQ3		
	model	d(gdt)	molp	model	d(gdt)	molp	model	d(gdt)	molp
T0953s1	149_4	6.0	4.1	085_1	16.8	1.8	261_1	4.1	2.8
T0957s2	324_3	7.7	3.3	402_5	31.0	0.7	261_1	13.1	2.2
T0968s1	498_2	5.9	3.5	368_1	0.0	0.7	368_2	7.8	0.7
T0968s2	498_4	7.8	3.7	407_3	32.8	1.5	368_1	11.7	1.0
T0969	324_4	12.1	3.6	368_5	27.3	1.2	498_5	1.4	$\boxed{3.8}$
T0975	261_2	19.4	3.1	368_1	19.6	1.0	368_1	19.6	1.0
T0980s1	145_1	0.0	3.3	368_1	14.4	1.4	368_1	14.4	1.4
T0986s2	324_5	0.0	3.5	368_1	24.0	1.0	407_1	15.8	1.0
T1001	156_5	17.6	1.0	368_2	0.0	1.1	368_4	1.6	1.2
T1015s1	261_2	2.3	2.4	407_4	27.6	0.5	368_1	5.1	0.7
T1017s2	261_1	3.8	2.9	368_4	12.4	0.9	407_1	29.4	1.2

High-accuracy model could be selected by improving stereochemistry during QA

Was there an advance?

Not really. Single-model methods performed worse than in previous CASPs.

A new challenge for QA

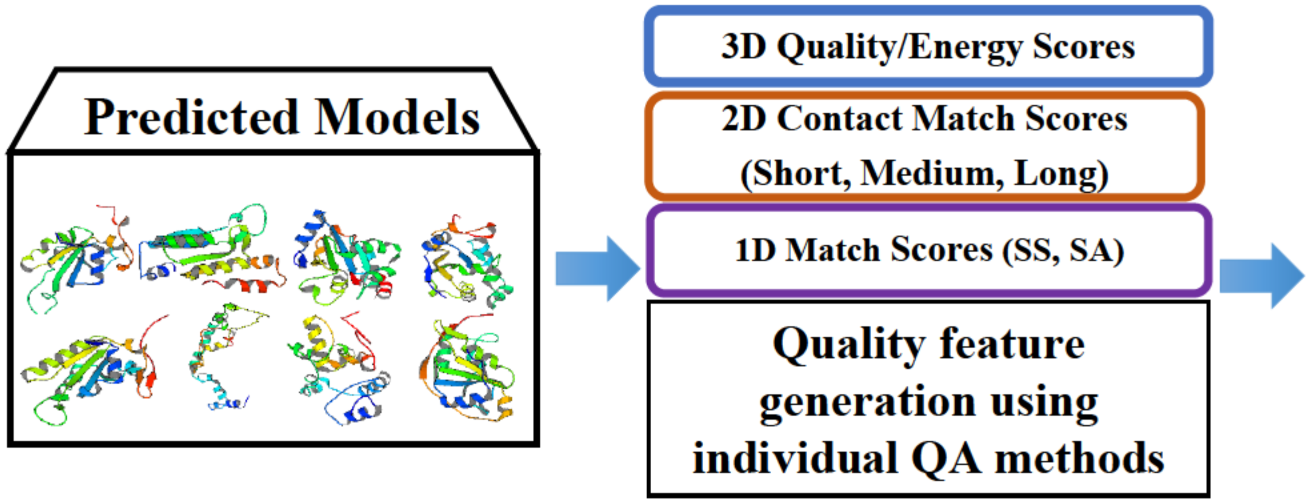
Protein models of higher global structure accuracy appear even for FM targets, and some of the models are not well locally optimized.

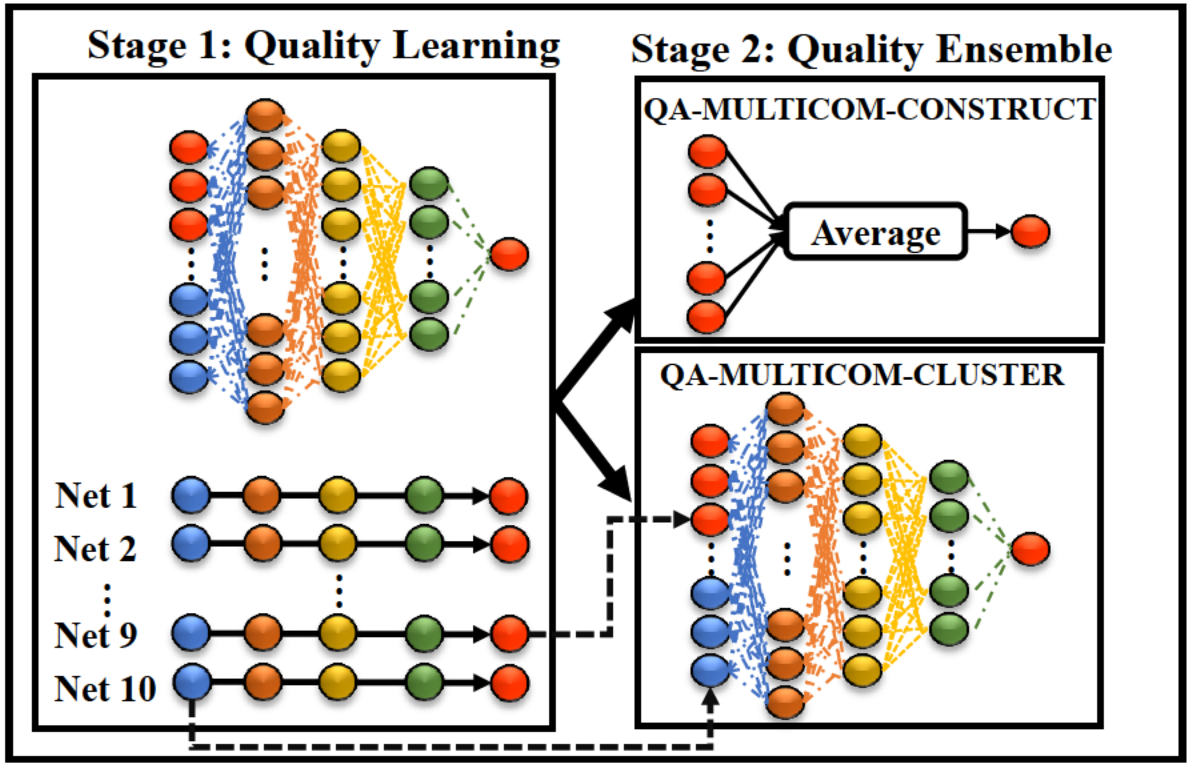
Round Table

Consensus method groups:
- **MULTICOM_CLUSTER** Jie Hou (a member of Jianlin Cheng group)
- **UOSHAN** Kun-Sop Han

Single-model method groups:
- **ModFOLD7_rank** Liam McGuffin
- **ProQ3D** Arne Elofsson
- **FaeNNz** Gabriel Studer
- **VoroMQA** Kliment Olechnovič (a member of Ceslovas Venclovas group)

Large-scale integration of protein model quality assessment using deep learning and contact predictions (MULTICOM-CLUSTER, MULCOM-CONSTRUCT)





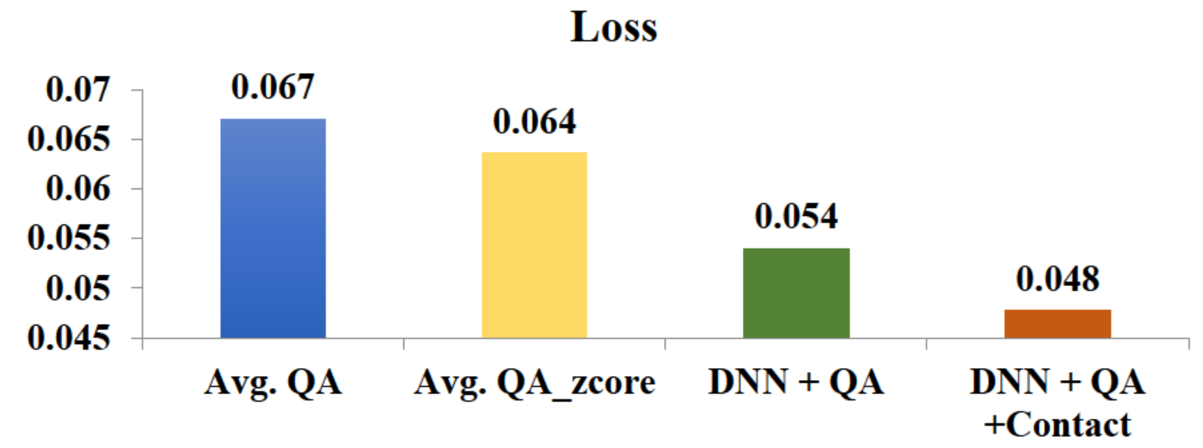
Quality Prediction via Deep Learning models

- **11 Single-QA:**
  OPUS-PSP, RF_SRS, Rwplus, SBROD QMEAN, Voronota, ModelEvaluator, Dope, DeepQA,ProQ2, ProQ3
- **3 Consensus-QA:**
  Pcons, Apollo, ModFoldclust2
- **Contact Match Score**
  DNCON2

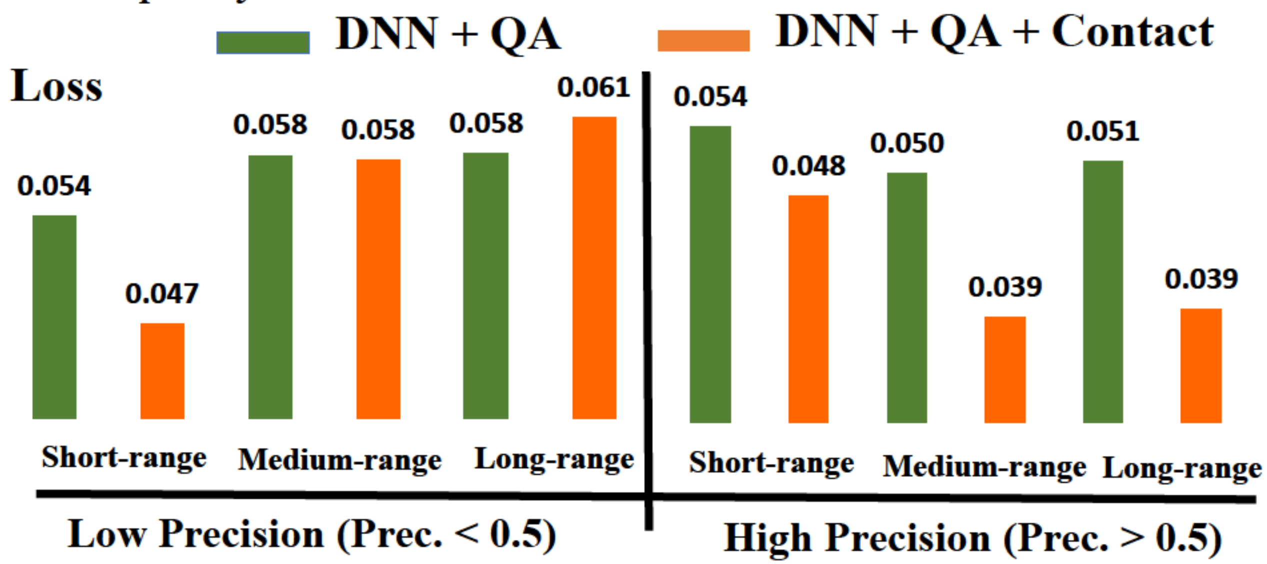
- ❑ Use deep learning to integrating the power of multiple complementary model features
- ❑ Train deep neural networks on CASP 8-11 datasets
- ❑ Benchmarked on the CASP12 and CASP13 dataset

Evaluation of Deep Learning Model Ranking on CASP12 and CASP 13

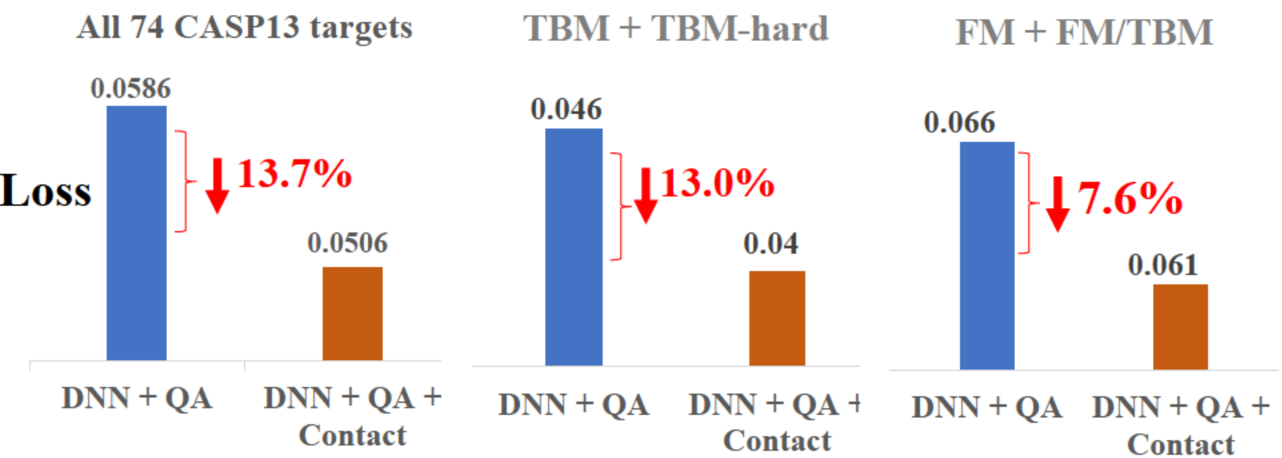
Result 1: Deep learning and contact prediction improve protein model quality assessment in CASP12 dataset.



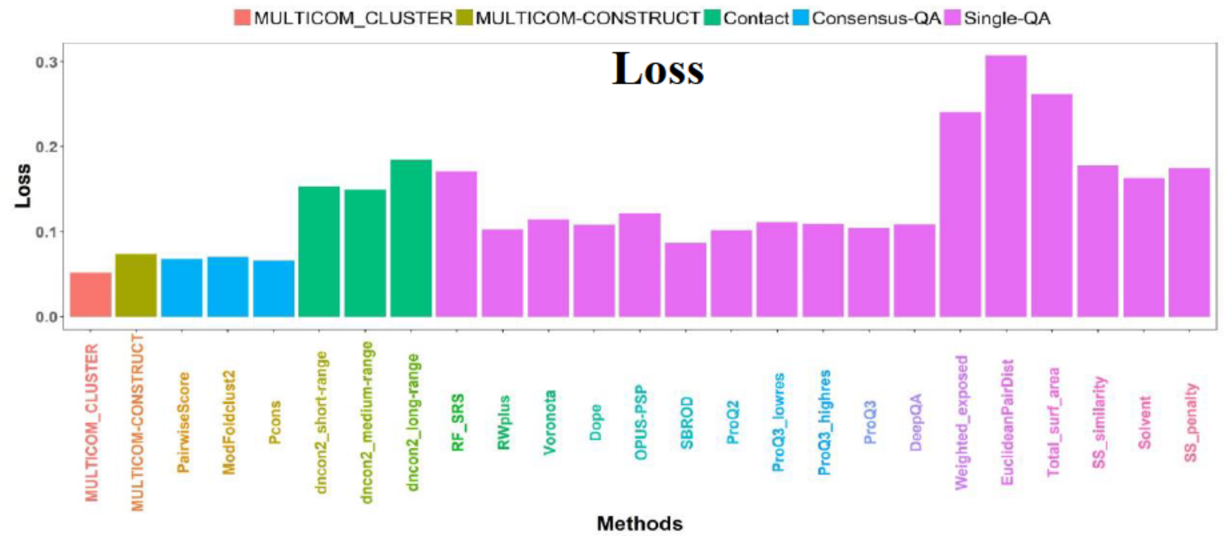
Result 2. Impact of contact prediction accuracy on protein model quality assessment in CASP12 dataset.



Result 3. Impact of contact features on protein model quality assessment in CASP13 dataset.



Result 4. Comparison of DeepRank with individual features in CASP13 dataset.



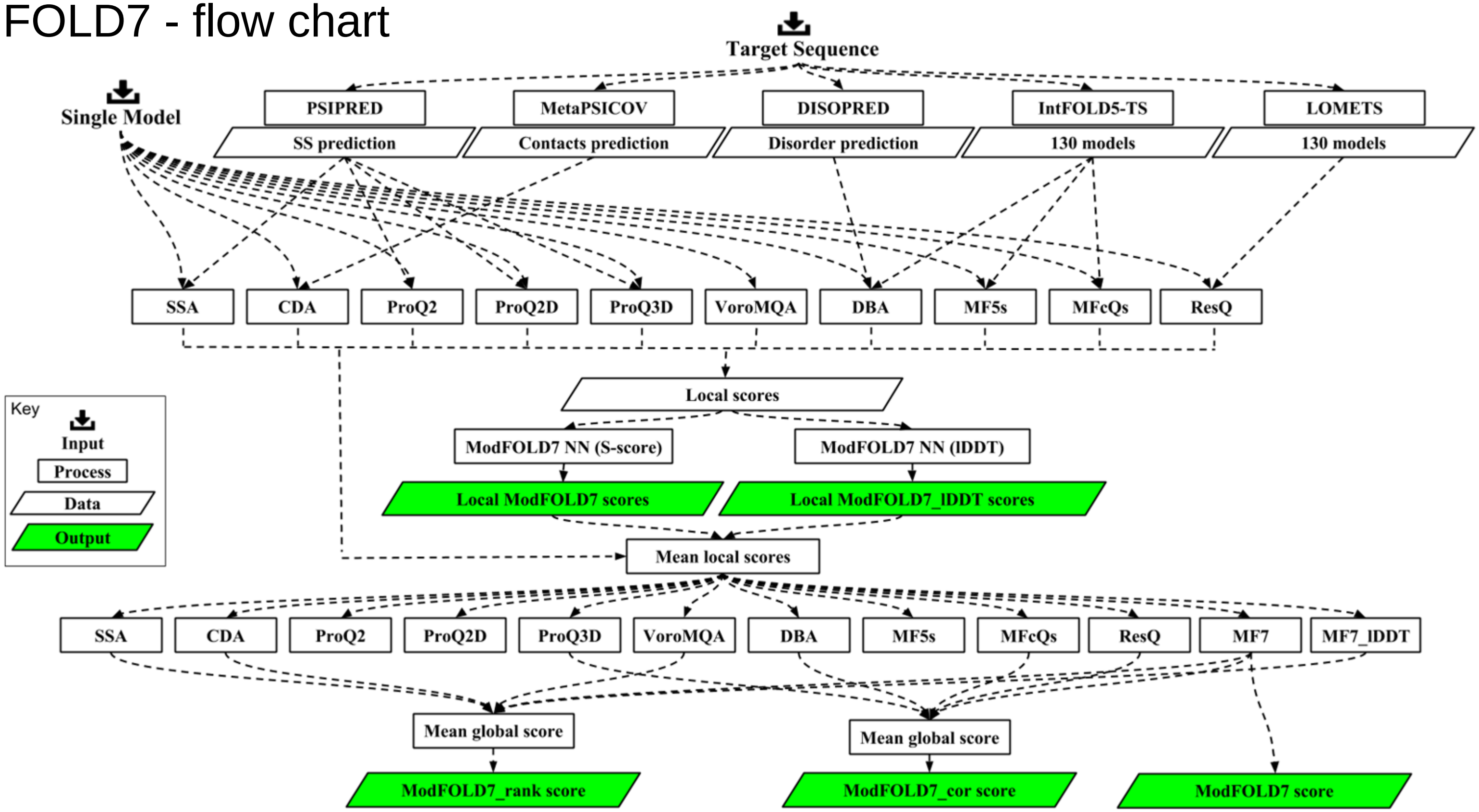
ModFOLD7

Liam McGuffin
University of Reading

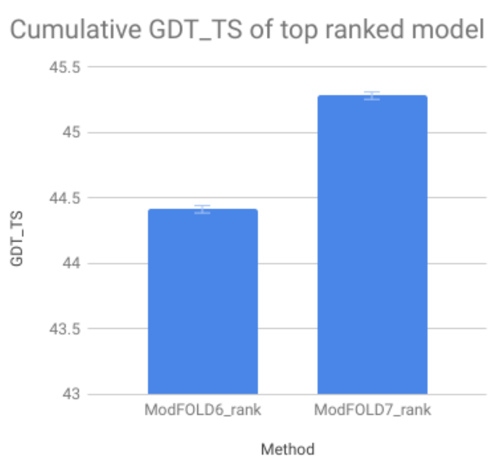
ModFOLD7 - Method Summary

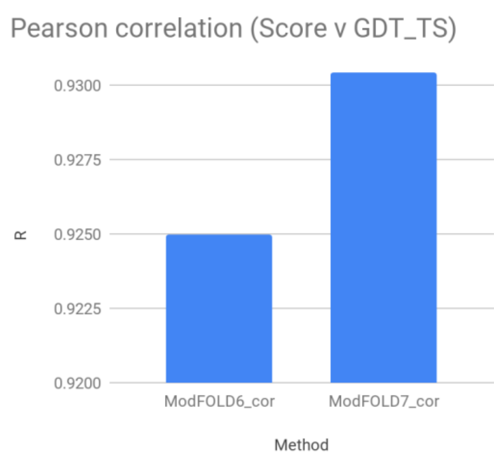
- **A single model approach combining inputs from 10 scoring methods**
- **6 pure-single model input methods:**
  - CDA = Contact Distance Agreement (MetaPSICOV versus contacts in model)
  - SSA = Secondary Structure Agreement (PSIPRED versus DSSP from model)
  - ProQ2, ProQ2D & ProQ3D
  - VoroMQA
- **4 quasi-single model input methods:**
  - MFcs = ModFOLDclust_single (input model versus <=130 IntFOLD5 models)
  - DBA = Disorder “B-factor” Agreement (DISOPRED versus MFcs score)
  - MFcQs = ModFOLDclustQ_single (input model versus <=130 IntFOLD5 models)
  - ResQ (input model versus LOMETS models)
- **Local score outputs - 2 variants** - 10 per-residue scores combined using a NN (MLP function in RSNNS) and trained using two target functions:
  - The S-score (included in **ModFOLD7** & **ModFOLD7_rank**)
  - The lDDT-score (included in **ModFOLD7_cor**)
- **Global score outputs** - **3 variants** - mean global scores that optimise for:
  - “Ranking” - selecting the best models (**ModFOLD7_rank**)
  - “Correlations” - estimating the absolute score (**ModFOLD7_cor**)
  - “Balanced” performance (**ModFOLD7**)

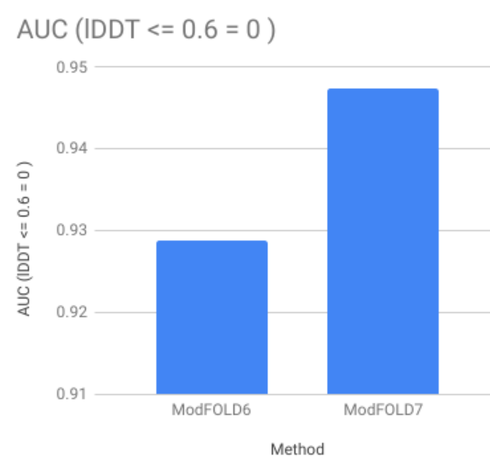
ModFOLD7 - flow chart



ModFOLD7 versus ModFOLD6







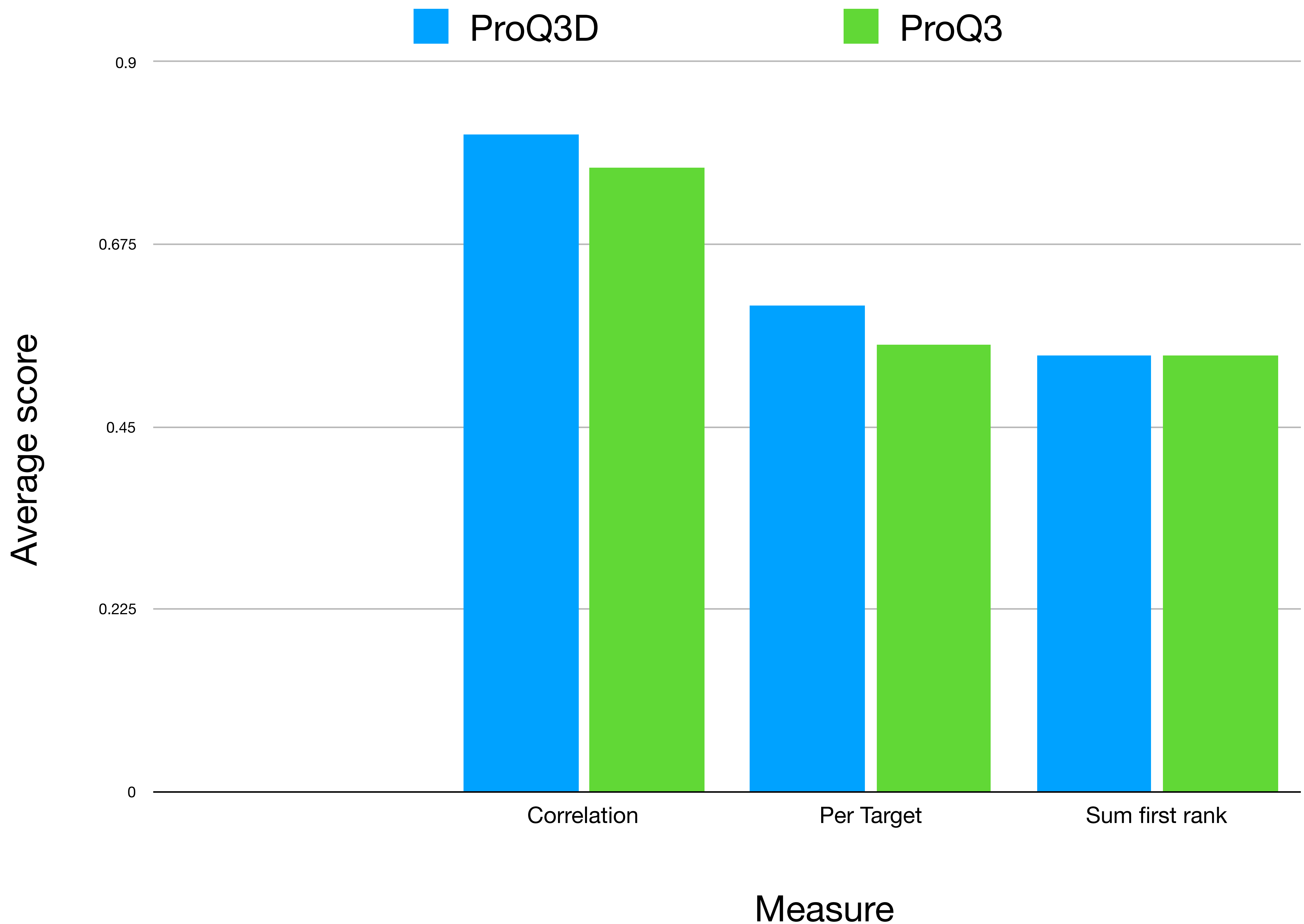
ProQ in CASP13

David Menendez-Hurtado, Karolis Uziela, Björn Wallner and Arne Elofsson

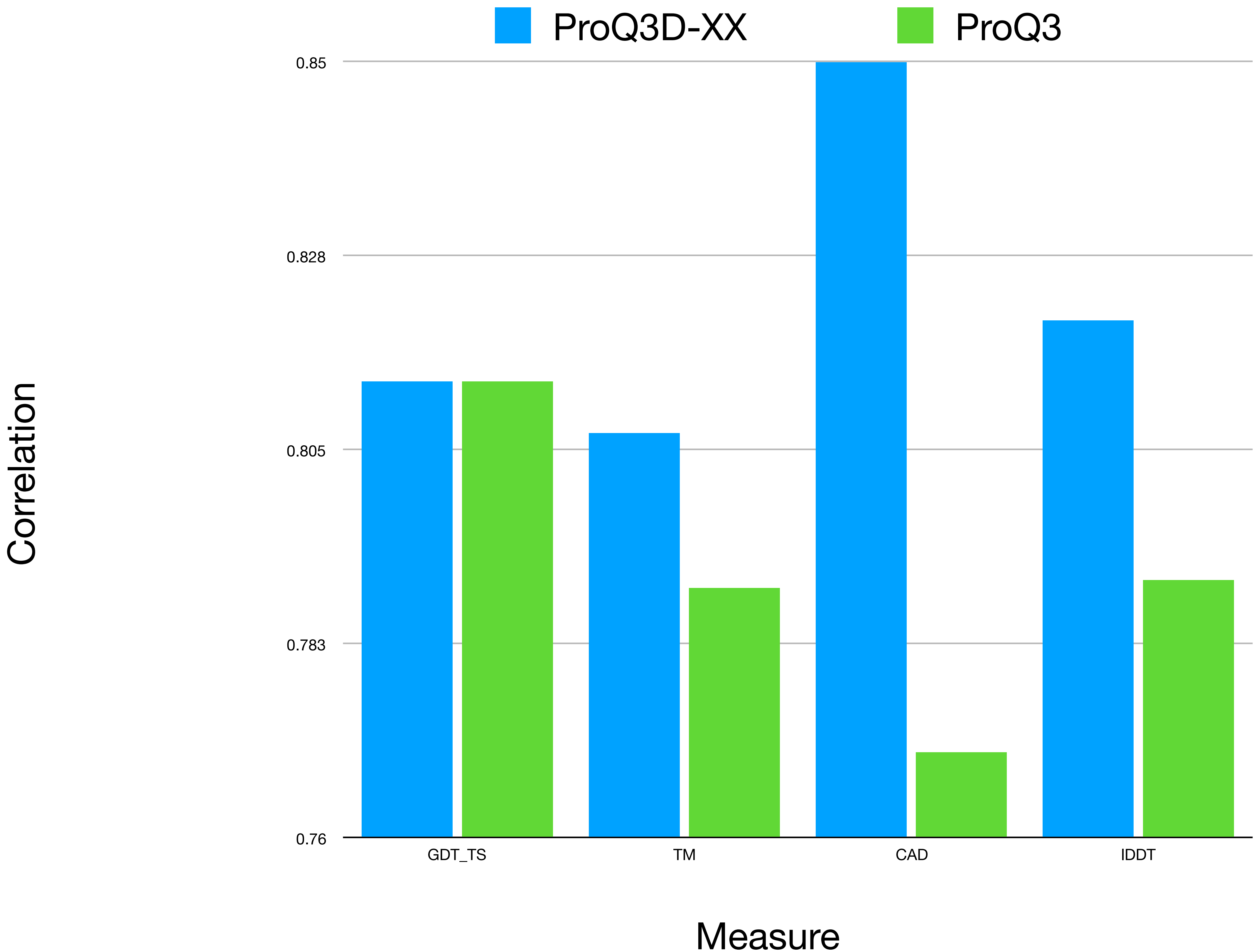
Overview

- ProQ3 = ProQ2 + Rosetta terms
- ProQ3D = ProQ3 using two-layer feed forward network.
    - ProQ3D: Trained on S-score (GDT_TS)
    - ProQ3D-TM: Trained on TMscore
    - ProQ3D-CAD: Trained on CAD-score
    - ProQ3-lDDT: Trained on lDDT.
- ProQ4 = Using deep learning, few input features (only DSSP). Trained on pairs of models. Trained on lDDT.

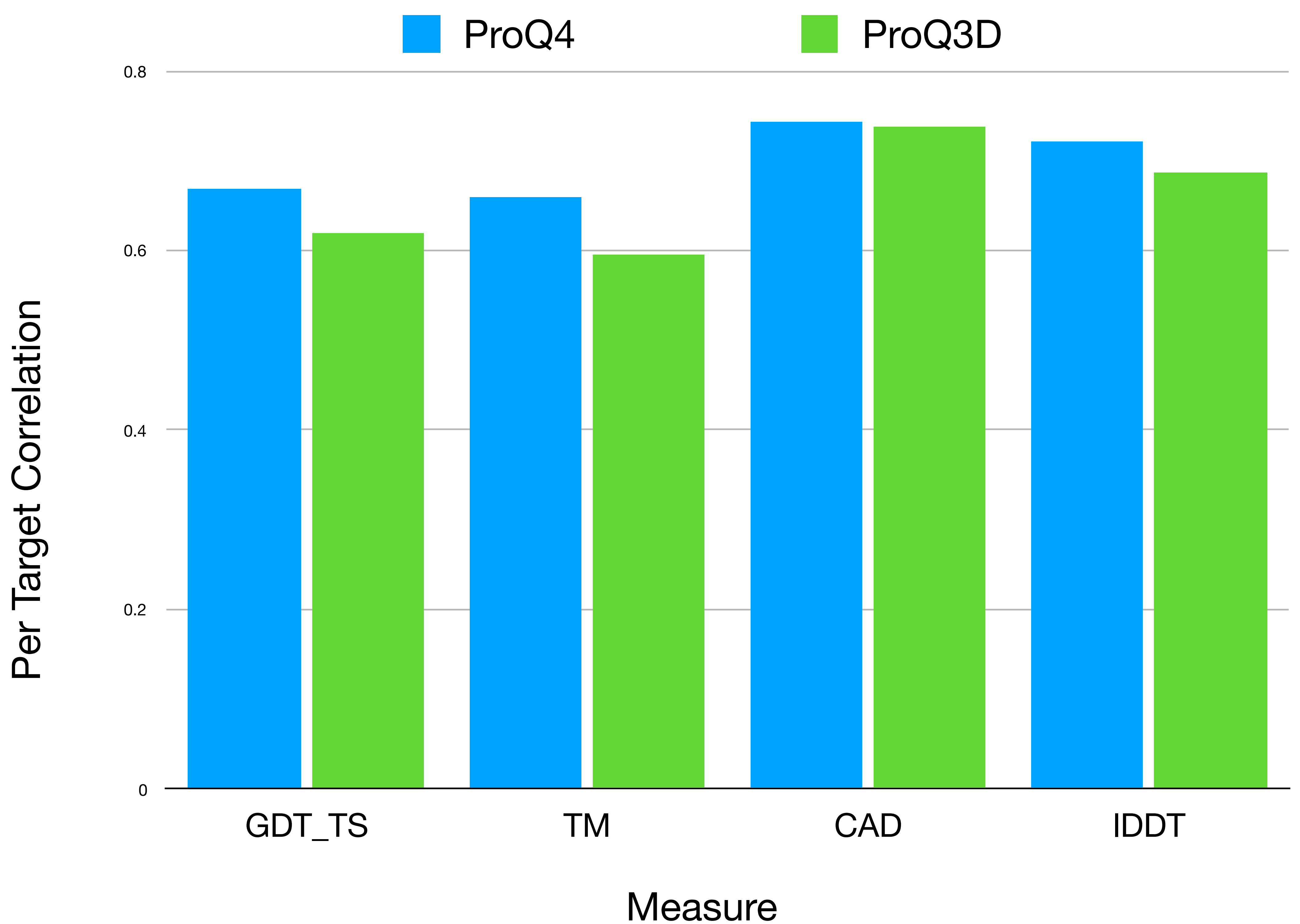
ProQ3D is better than ProQ3



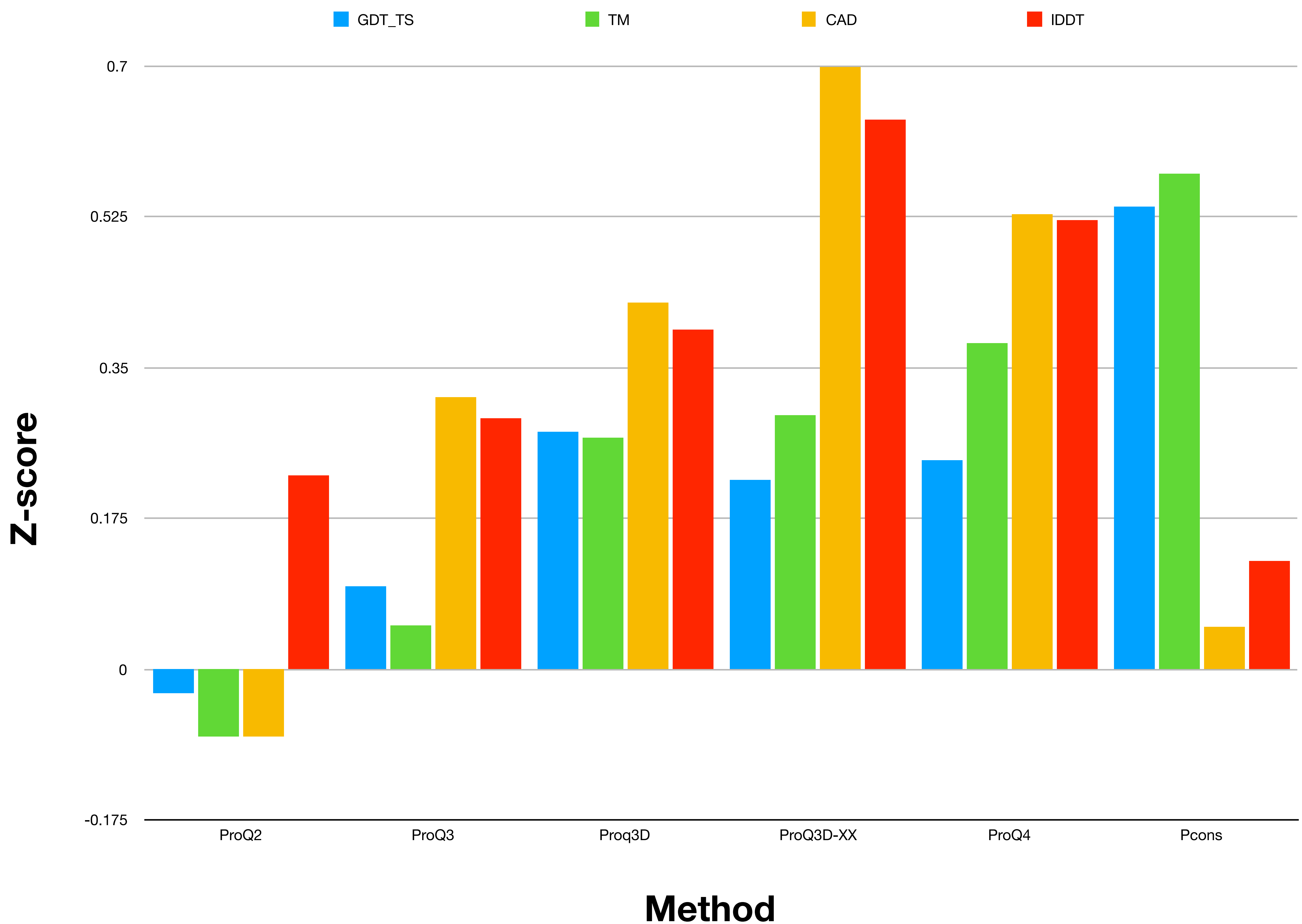
ProQ3D-XX i better than ProQ3 when evaluated on XX



ProQ4 is better at ranking than ProQ3D.



ProQ performs relatively better on CAD and lDDT



Ref. No. SU FV-1696-18
Closing date: 15/01/2019

Assistant Professor in Computational biology

at the Department of Mathematics. Closing date: 15 January 2019.

Stockholm University is a leading European university and one of the world's top 100 institutes of higher education and research. Stockholm University has more than 60,000 students and 5,000 staff.

The Science for Life Laboratory (SciLifeLab) is a national center for large-scale biosciences with a focus on health and environmental research and is a collaboration between Stockholm University, Karolinska Institutet, the Royal Institute of Technology, and Uppsala University. SciLifeLab-Stockholm is located in a new building on the Karolinska Institutet campus.

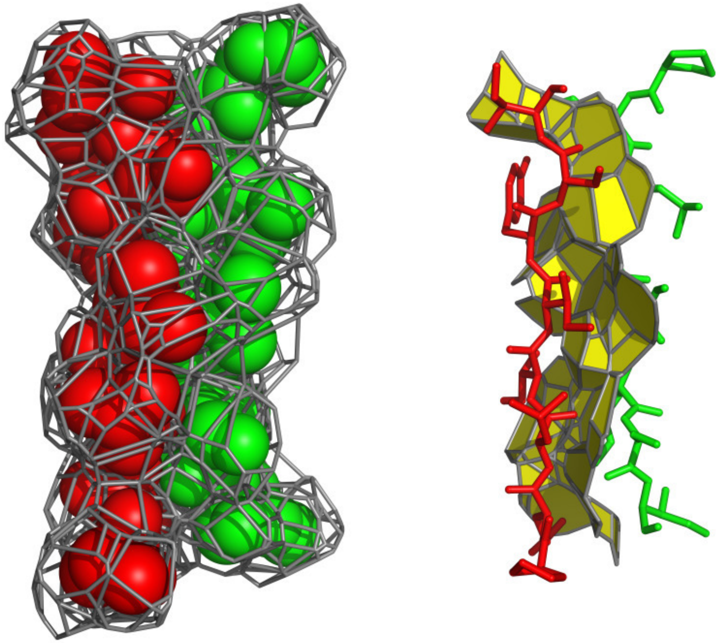
The last century of research has led the Department of Mathematics at Stockholm University to acquire a prominent place in Scandinavian mathematics. The department consists of three divisions: Mathematics, Mathematical statistics, and the recently formed Computational mathematics. The research in the division of mathematics include algebra, geometry and combinatorics, analysis and logic. The research in mathematical statistics include probability theory and statistical inference theory, with applications in biostatistics, climatology, econometrics, finance and insurance. Computational mathematics is a new direction for the department, with activities in computational biology, stochastic modelling, scientific computing for climatology, and logic of programs. During the first six years, the main workplace for this position will be at the Science for Life Laboratory. The formal employment will be at the Department of Mathematics.

Subject
Computational biology

VoroMQA - Voronoi tessellation-based Model Quality Assessment

Kliment Olechnovič and Česlovas Venclovas, Vilnius University Institute of Biotechnology

Method definition:



Pseudo-energy for contact type:

$$\mathrm{E}(a_i, a_j, c_k) = \log \frac{\mathrm{P}_{\mathrm{exp}}(a_i, a_j, c_k)}{\mathrm{P}_{\mathrm{obs}}(a_i, a_j, c_k)} = \log \frac{\mathrm{F}_{\mathrm{exp}}(\mathrm{area}(a_i), \mathrm{area}(a_j), \mathrm{area}(c_k))}{\mathrm{F}_{\mathrm{obs}}(\mathrm{area}(a_i, a_j, c_k))}$$

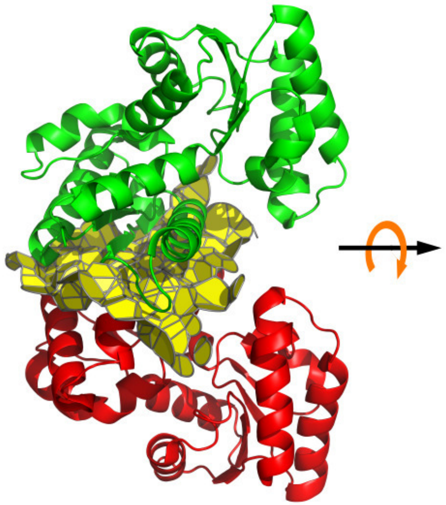
Normalized energy for atom:

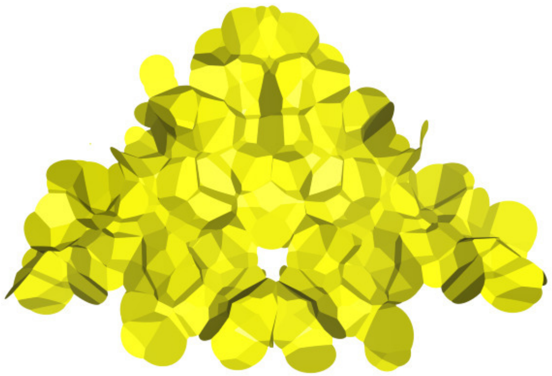
$$\mathrm{E_n}(\Omega_\phi) = \frac{\sum_{\omega \in \Omega_\phi} \mathrm{E}(\mathrm{type}_\omega) \cdot \mathrm{area}_\omega}{\sum_{\omega \in \Omega_\phi} \mathrm{area}_\omega}$$

Quality score for atom:

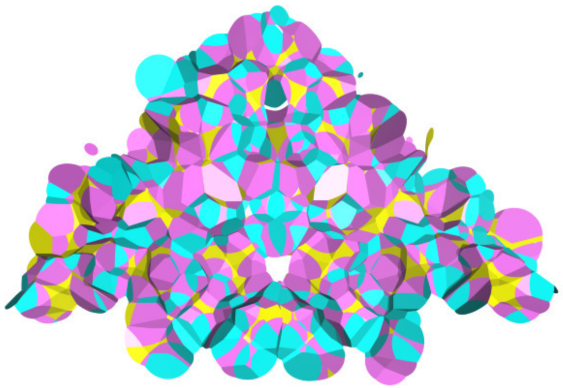
$$\mathrm{Q_a}(\Omega_\phi) = \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{\mathrm{E_n}(\Omega_\phi) - \mu_{\mathrm{type}_\phi}}{\sigma_{\mathrm{type}_\phi}\sqrt{2}}\right)\right)$$

Enhancement for CASP13:





CASP12
(contacts without hydrogens)



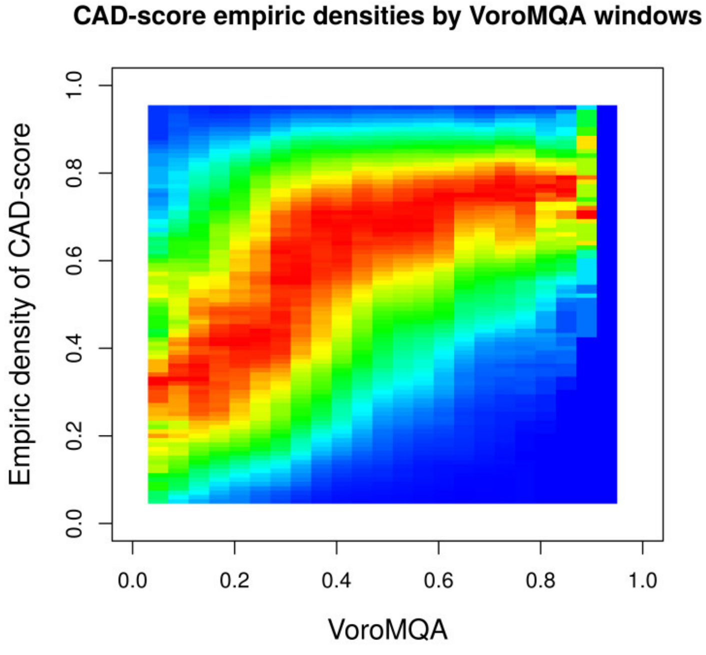
CASP13
(contacts with hydrogens)

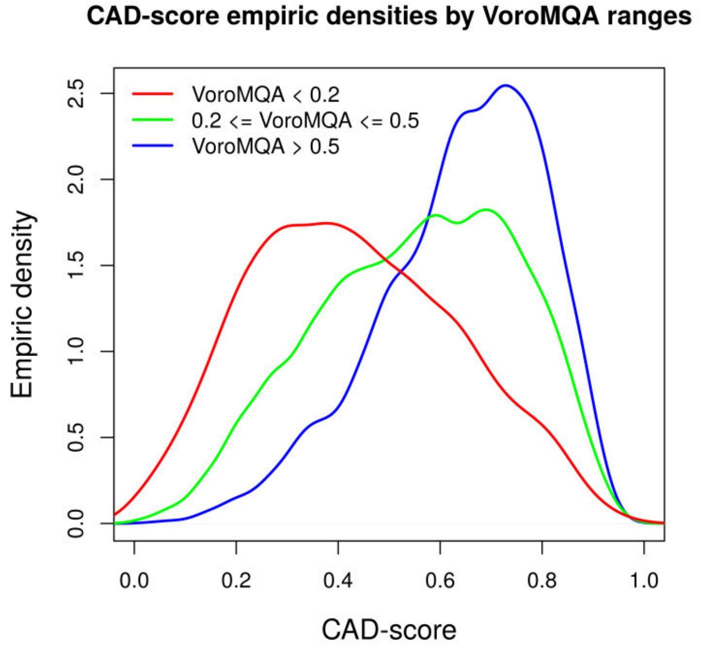
heavy - heavy

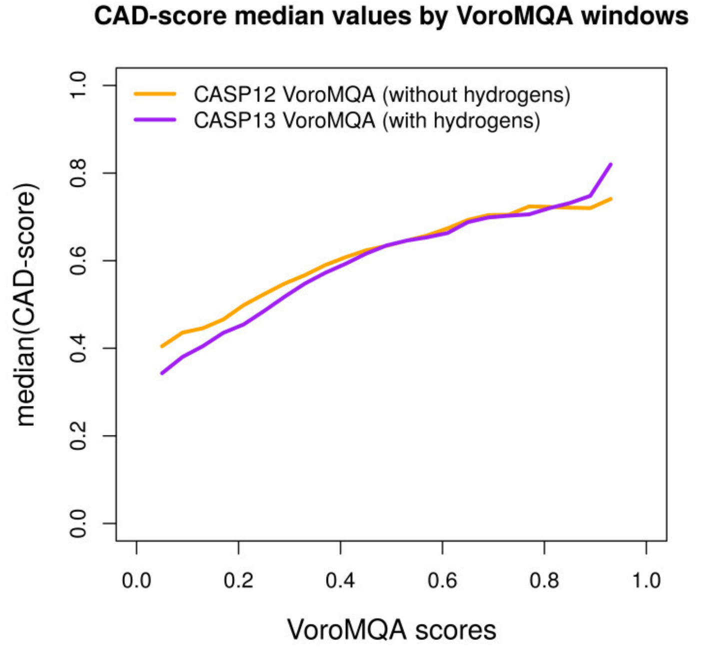
hydrogen - hydrogen

heavy - hydrogen

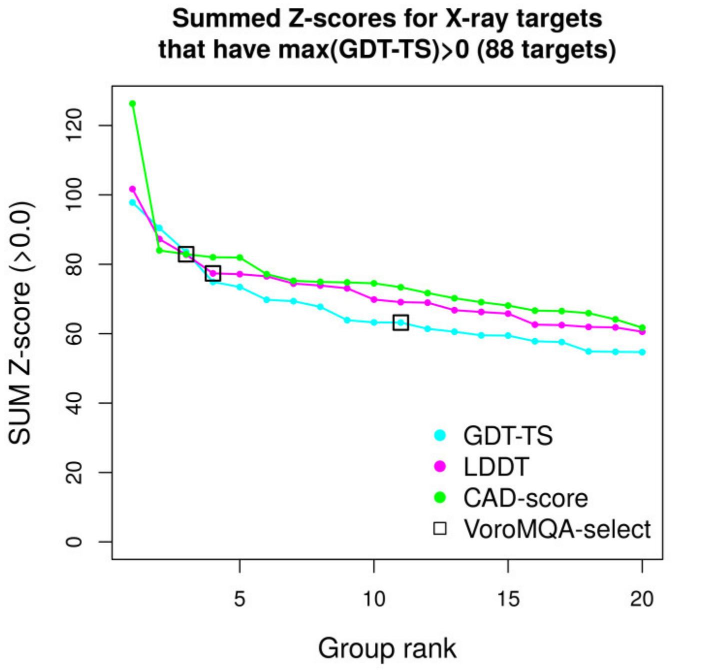
Local scoring:

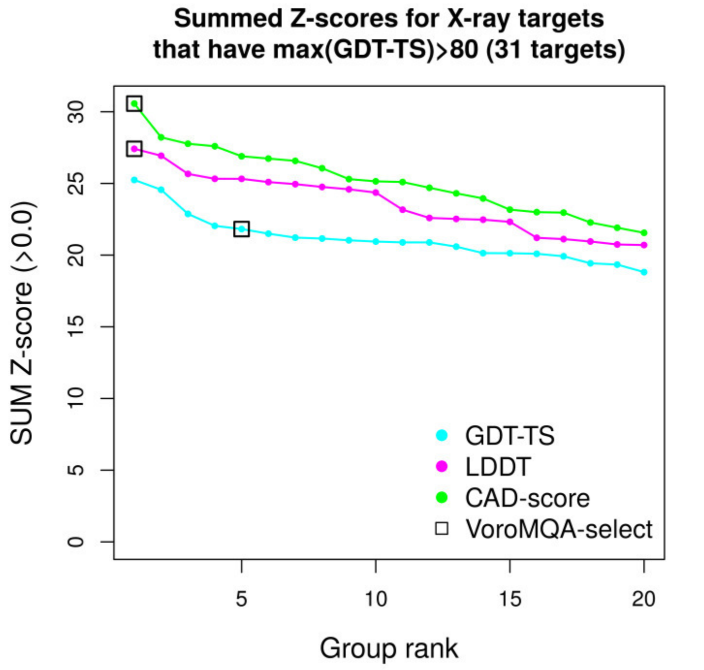






Global scoring:





Conclusions:

- VoroMQA local scores can be used to classify the structure into the accurate regions and those with the uncertain accuracy.
- VoroMQA global scores are more useful when selecting from models of higher quality.
- VoroMQA performs relatively well because it uses tessellation-derived contact areas.

FaeNNz

Combining Statistical Potentials with Consensus-Based Prediction of Local Quality

UNI BASEL
BIOZENTRUM
SIB Swiss Institute of Bioinformatics

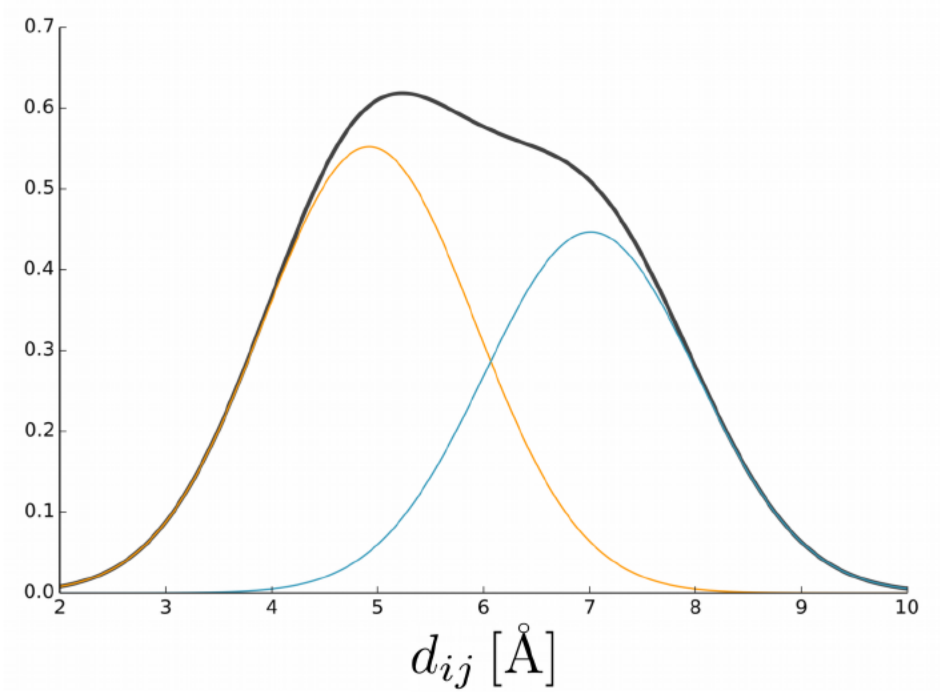
FaeNNz

Fast single model prediction of local model quality
Main target: scoring models for SWISS-MODEL

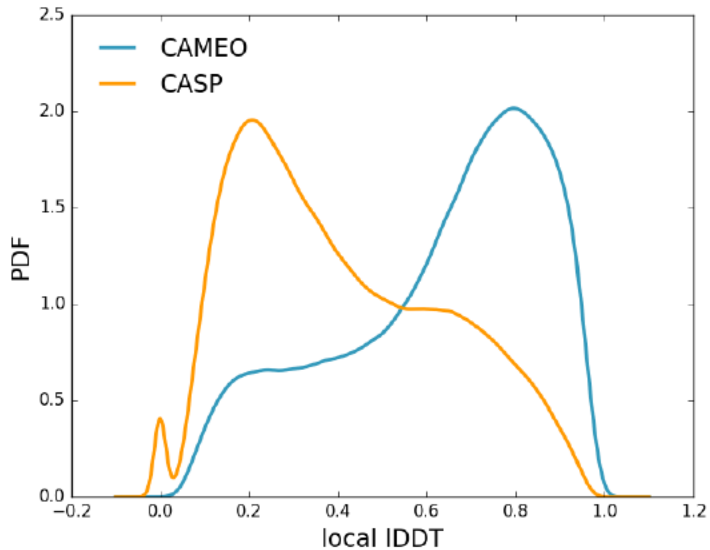
QMEAN
Statistical potentials

DisCo
Distance constraints

FaeNNz
Low resolution features
Mix all in NN



FaeNNz



- Constraints from found templates improve local quality estimates
- NN help to identify complex interdependencies in training data
- Low resolution features help to identify local regions with poor packing
- CASP and CAMEO targets are not the same thing

	CAMEO CrossVal		CASP CrossVal	
Predictor	Pearson R	ROC AUC	Pearson R	ROC AUC
QMEANDisCo	0.855	0.931	0.681	0.886
FaeNNz (CASP)	0.841	0.916	0.836	0.937
FaeNNz (CAMEO)	0.887	0.940	0.812	0.934
FaeNNz (Mixed)	**0.889**	**0.940**	**0.856**	**0.946**

Discussion Topics

(1) **Deep learning** has a clear impact in QA. How can this be pushed further?

(2) Is the current **number of models**, 150 per target in stage 2, enough? Would a larger number of models facilitate advance?

(3) Model qualities for **oligomer targets** have been evaluated using only monomer models. How should this be treated?

(4) What is the value of applying **consensus methods** to CASP server models that are available only in CASP season? How should it be treated in the future?

(5) In CASP13 we seem to have **little progress** over CASP12. Why? How should we proceed?

(6) Other topics

1. Consensus & Deep Learning

Consensus methods exploiting **pure consensus of CASP-specific** server models are not desirable for advance of the field.

One suggestion is to provide models that are more uniformly spaced in the conformational space. This needs **more models** from TS servers.

More structural decoy data may promote method developments in both QA and TS by providing more training data for **deep learning**.

Is the current **number of models**, 150 per target in stage 2, enough? Would more models facilitate advance?

2. Oligomer Targets

Qualities of only monomer models, not of full quaternary models, were evaluated for **oligomer targets**.

It makes sense to evaluate monomer models only for some oligomer targets for which monomer units are stable by themselves. In more general cases, oligomer models have to be evaluated as a whole.

CAPRI runs a **scoring round** in which ~1000 oligomer models are available for evaluation for each target. Would there be any problems if CASP QA predictors participate in the CAPRI scoring rounds?

3. Progress

Single-model methods performed relatively poorly in particular on FM targets.

This seems to be because globally more accurate, but locally less optimized models were generated by TS servers for FM targets.

How can this problem be treated?