*CASP9

Automatic evaluation of the QA category

Andriy Kryshtafovych
Anna Tramontano
Krzysztof Fidelis
John Moult

* QA1: global quality of models

  46 groups participated (45 CASP8)

* QA2: local (per-residue) reliability of models

  22 groups participated (17 CASP8)

* Predictions submitted: 5490 (5483 CASP8)

* Interest to the problem

Assessment measures

* Correlation of predicted and observed model quality scores (MQAS vs GDT) on per-target basis (QA1.1)
* Correlation of predicted (MQAS) and observed (GDT_TS) model quality scores for all models pooled together (QA1.2)
* Average (per target) loss from the best available model / ability to pick the best model
* Correlation of per-residue distances in model-target superposition (actual and estimated) (QA2)

* Targets and TS prediction difficulty

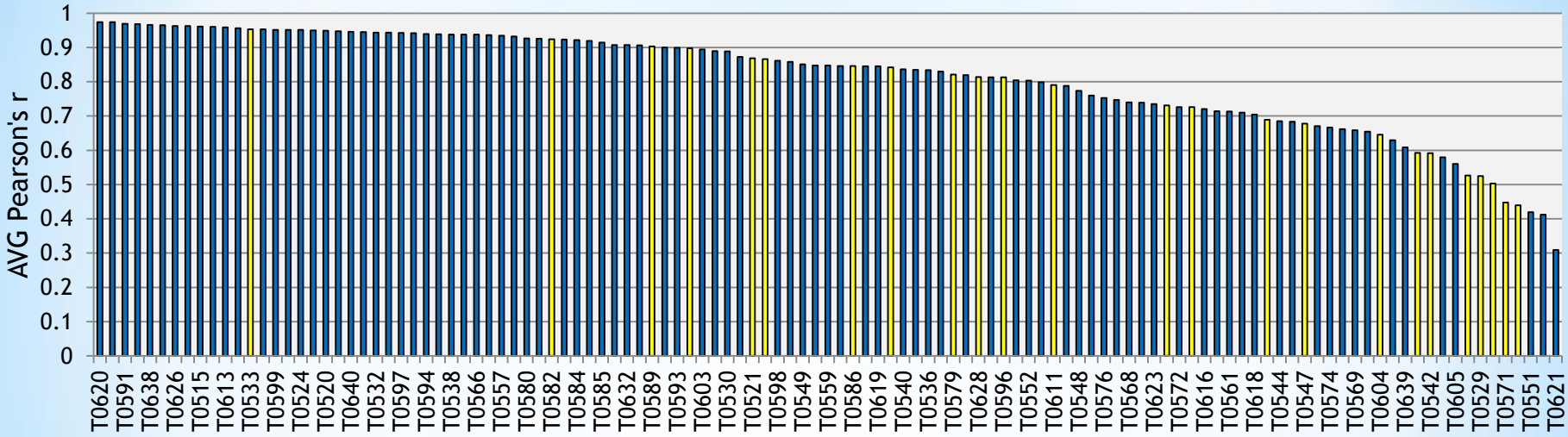
Is there any difference in QA methods performance on multi-domain vs single-domain targets?

Is there any difference in QA methods performance on hard vs easy for TS prediction targets?

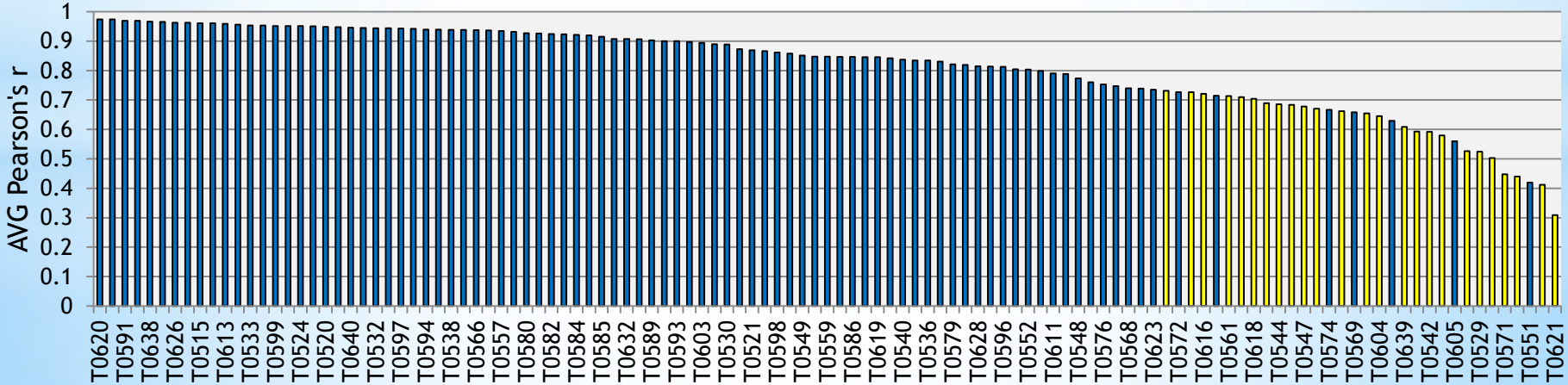
* Assessment measures

CASP9 QA target difficulty based on average Pearson's r

Multidomain targets highlighted (groups with r<0.5 not included)



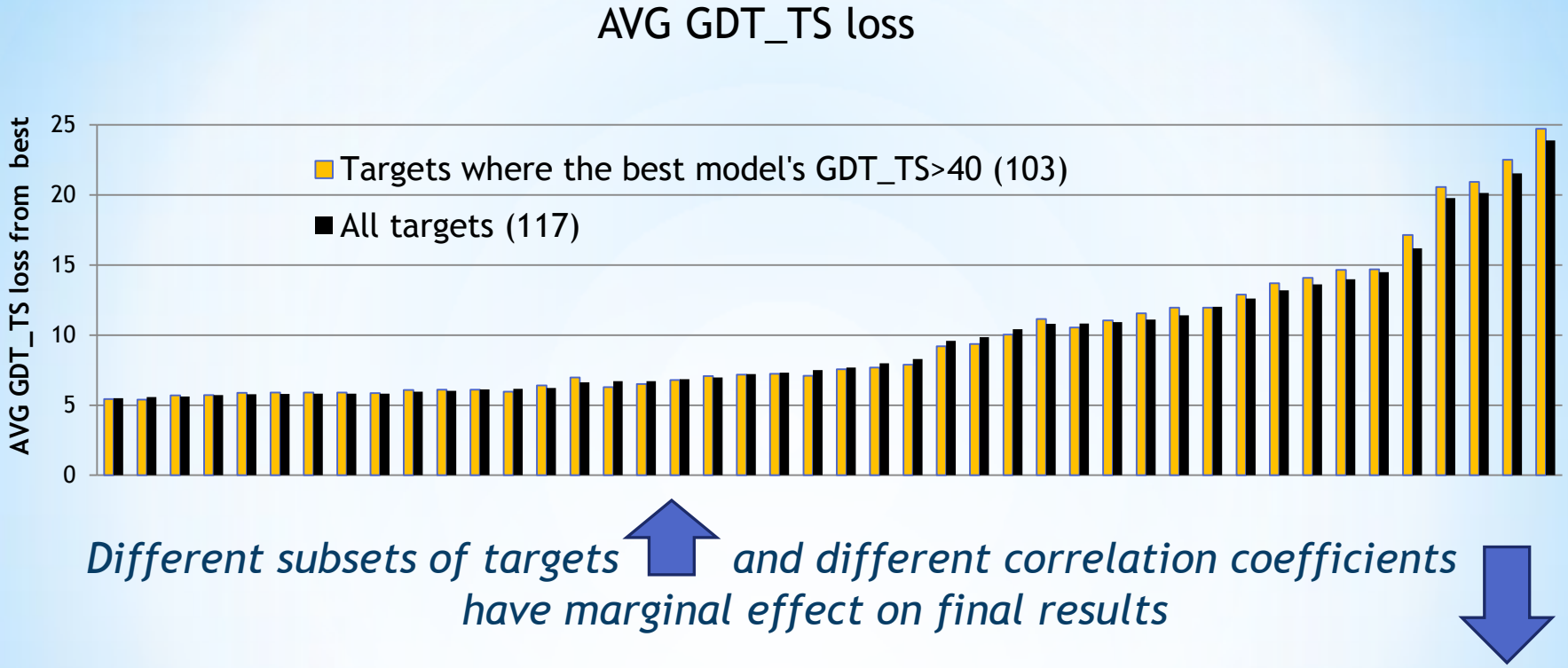
Targets containing FM domains or best model < 40 GDT_TS highlighted



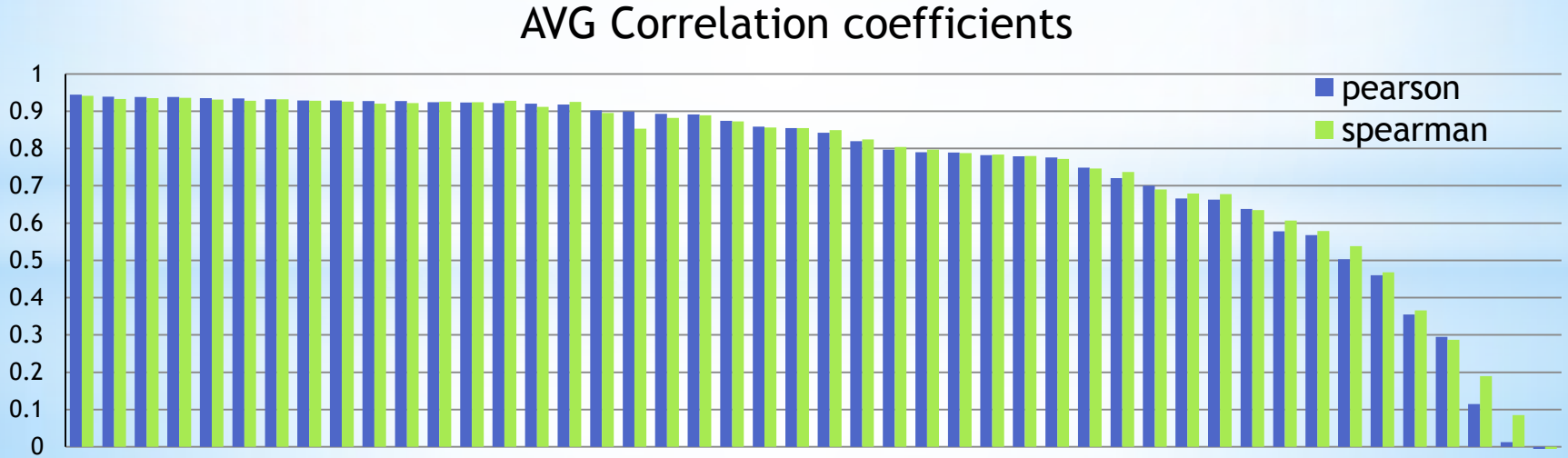
Assessment measures

* Models:
  - All targets (117)
  - Targets, where the best server's model GDT_TS>40 (103)
  - Targets, where the best server's model GDT_TS>50 (90)

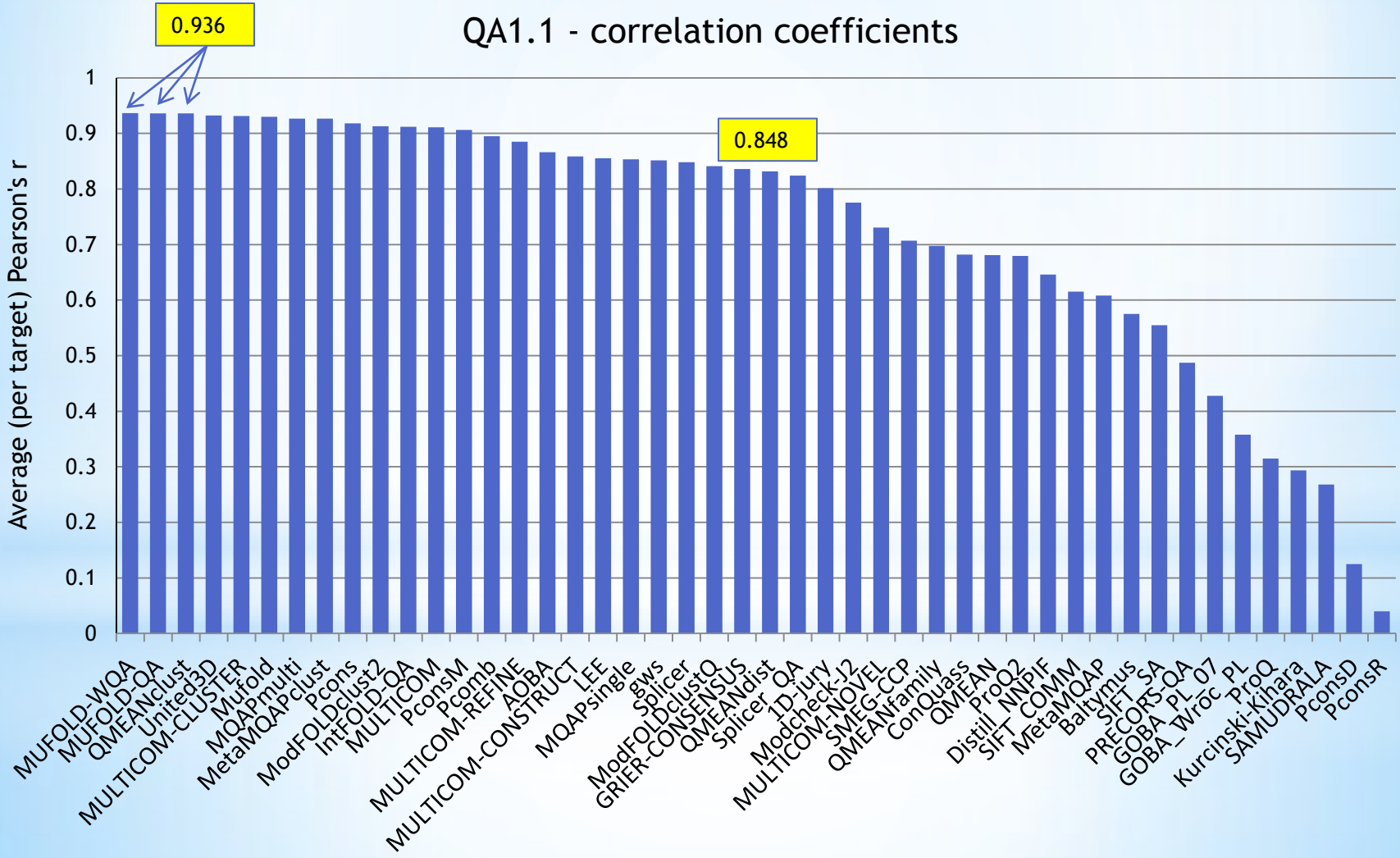
* Correlation measures:
  - Pearson's r
  - Spearman's ρ
  - Kendal's τ



Different subsets of targets and different correlation coefficients have marginal effect on final results



QA1.1 - assessment of quality estimates of whole models based on per-target correlation between MQAS and GDT_TS

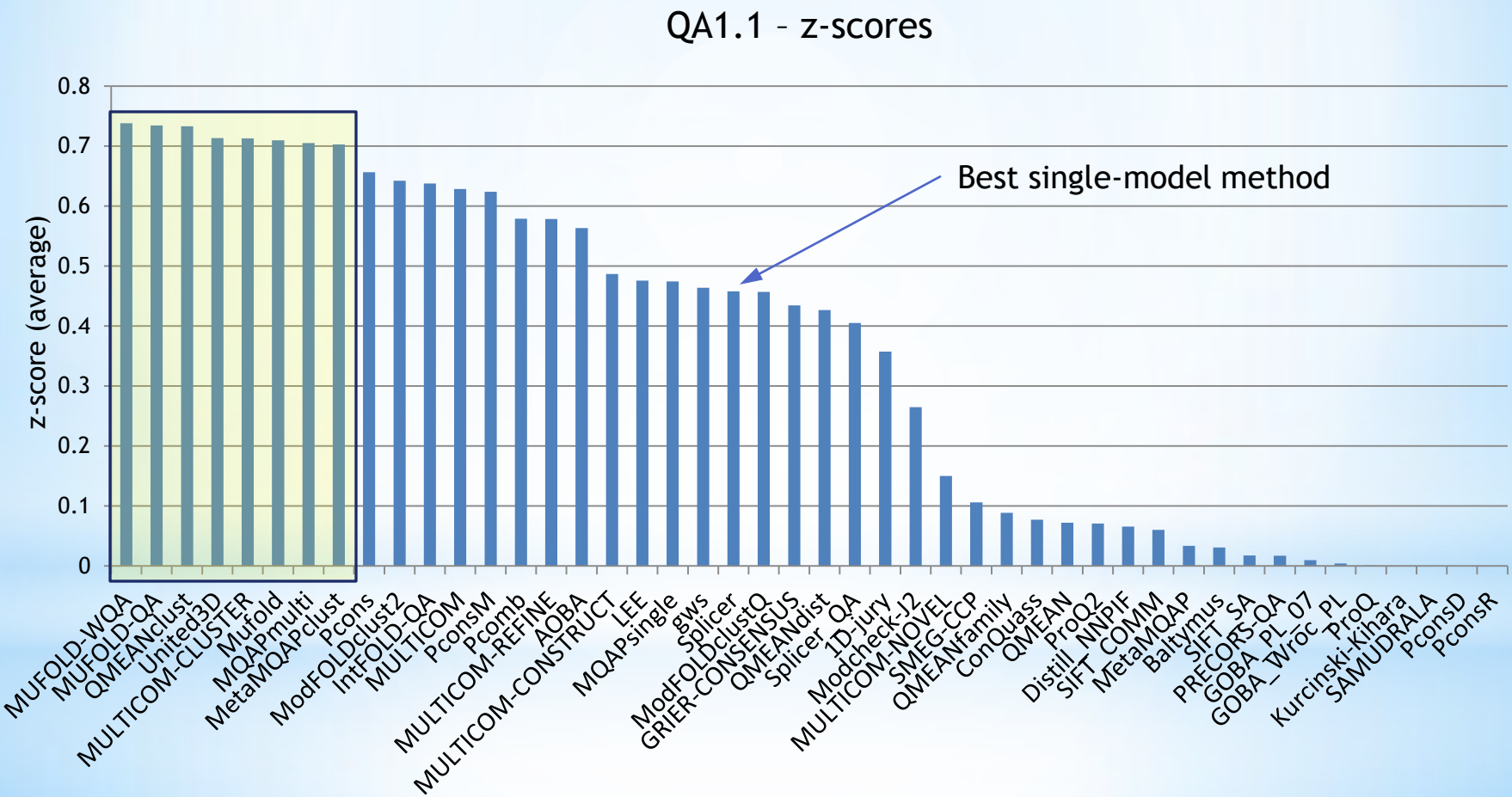


QA1: Paired t-test results: p-value \ number of common targets.

Shaded cells show statistically indistinguishable groups (p>0.01).
Darker shade show groups with very similar results.

Group name	#	312	359	371	407	2	386	369	426	319	397	78	490
MUFOLD-WQA	312	X	117	117	117	117	114	117	117	117	117	117	114
MUFOLD-QA	359	0.98	X	117	117	117	114	117	117	117	117	117	114
QMEANclust	371	0.95	0.81	X	117	117	114	117	117	117	117	117	114
United3D	407	0.43	0.29	0.12	X	117	114	117	117	117	117	117	114
MULTICOM-CLUSTER	2	0.26	0.06	<0.01	0.67	X	114	117	117	117	117	117	114
Mufold	386	0.21	<0.01	<0.01	0.32	0.33	X	114	114	114	114	114	114
MQAPmulti	369	0.05	0.06	0.05	0.30	0.36	0.44	X	117	117	117	117	114
MetaMQAPclust	426	0.16	0.06	0.07	0.37	0.40	0.44	0.96	X	117	117	117	114
Pcons	319	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	0.14	0.20	X	117	117	114
ModFOLDclust2	397	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	0.28	X	117	114
IntFOLD-QA	78	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	0.21	<0.01	X	114
MULTICOM	490	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	0.14	0.36	0.45	X

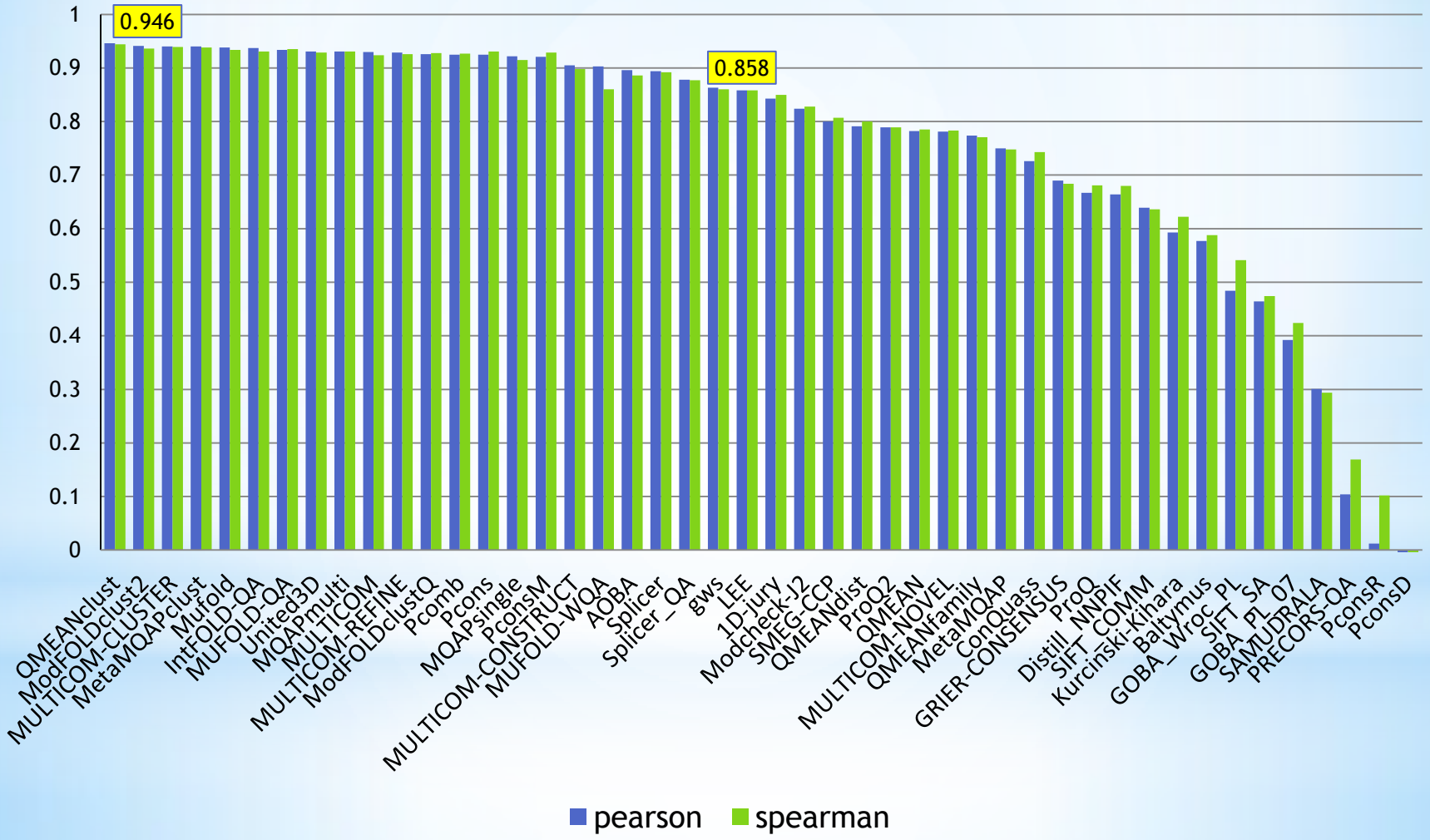
QA1.1 - assessment of quality estimates of whole models based on per-target correlation between MQAS and GDT_TS



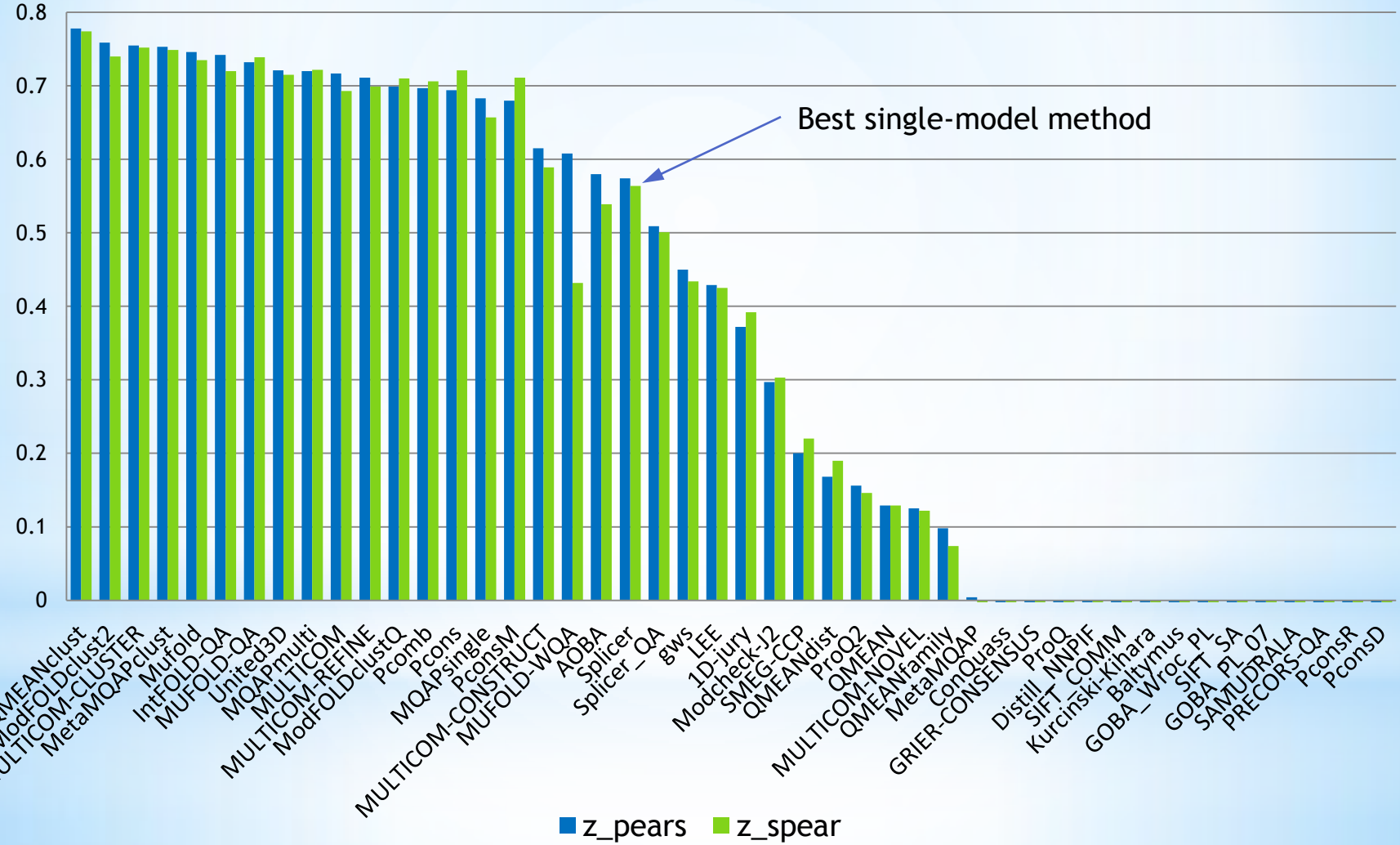
QA1.1 (per target analysis): best groups

	Group Name	Group Leader	Clst/Sng	Submt	Assess	r	Z
1	**MUFOLD-WQA**	Qingguo Wang, Univ. Missouri, USA	C	129	117	0.936	0.738
2	**MUFOLD-QA**	Yi Shang, Univ. Missouri, USA	C	129	117	0.936	0.734
3	**QMEANclust**	Pascal Benkert, Swiss Institute of Bioinformatics, Switzerland	C	129	117	0.936	0.733
4	**United3D**	M. Takeda-Shitaka, Kitasato Univ., Japan	C	128	117	0.932	0.713
5	**MULTICOM-CLUSTER**	Jianlin Cheng, Univ. Missouri, USA	C	129	117	0.931	0.713
6	**Mufold**	Dong Xu, Univ. Missouri, USA	C	125	114	0.930	0.709
7	**MQAPmulti**	Marcin Pawlowski, Genesilico – Warsaw, Poland	C	129	117	0.927	0.705
8	**MetaMQAPclust**	Marcin Pawlowski, Genesilico – Warsaw, Poland	C	129	117	0.926	0.703
9	**Pcons**	Arne Elofsson, Stockholm Univ.	C	129	117	0.918	0.656
10	**ModFOLDclust2**	Liam McGuffin, Univ. Reading, UK	C	129	117	0.913	0.642
…			C				
21	**Splicer**	Nakamura Yuuki, Kitasato Univ., Japan	S	128	117	0.855	0.458

QA1.2 - all models pooled together (correlation coefficients)



QA1.2 - all models pooled together (Z-scores)

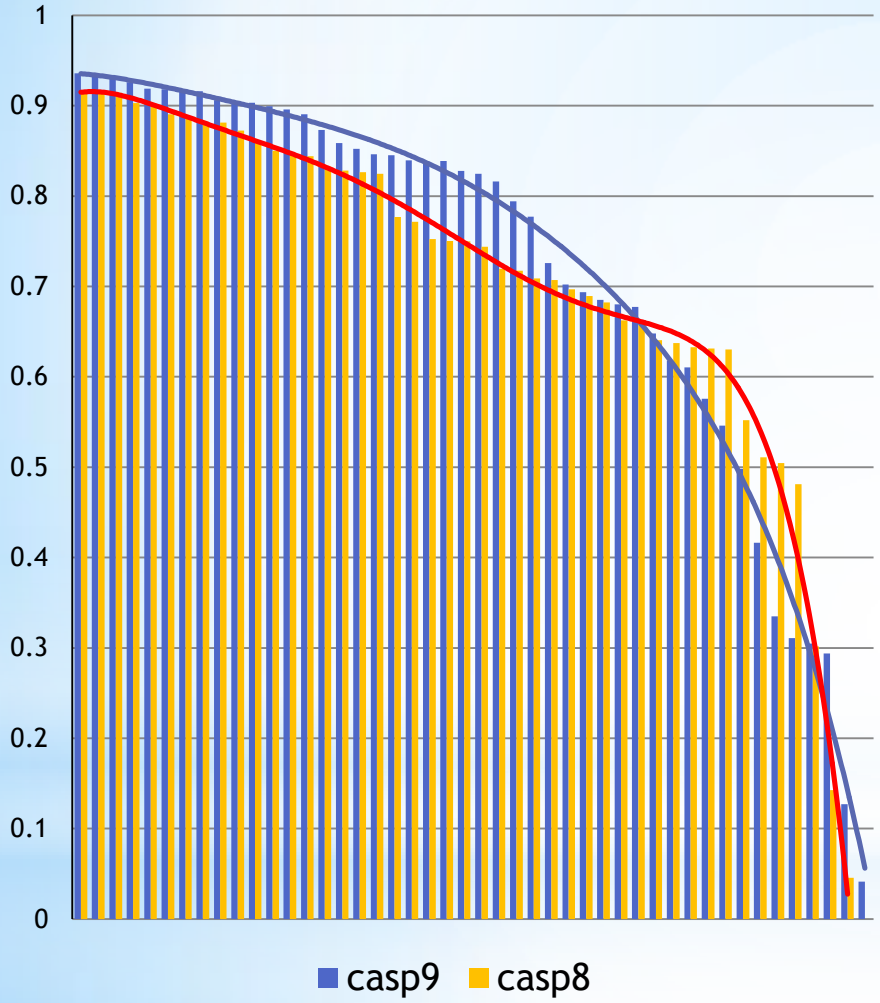


QA1.2 (all models together): best groups

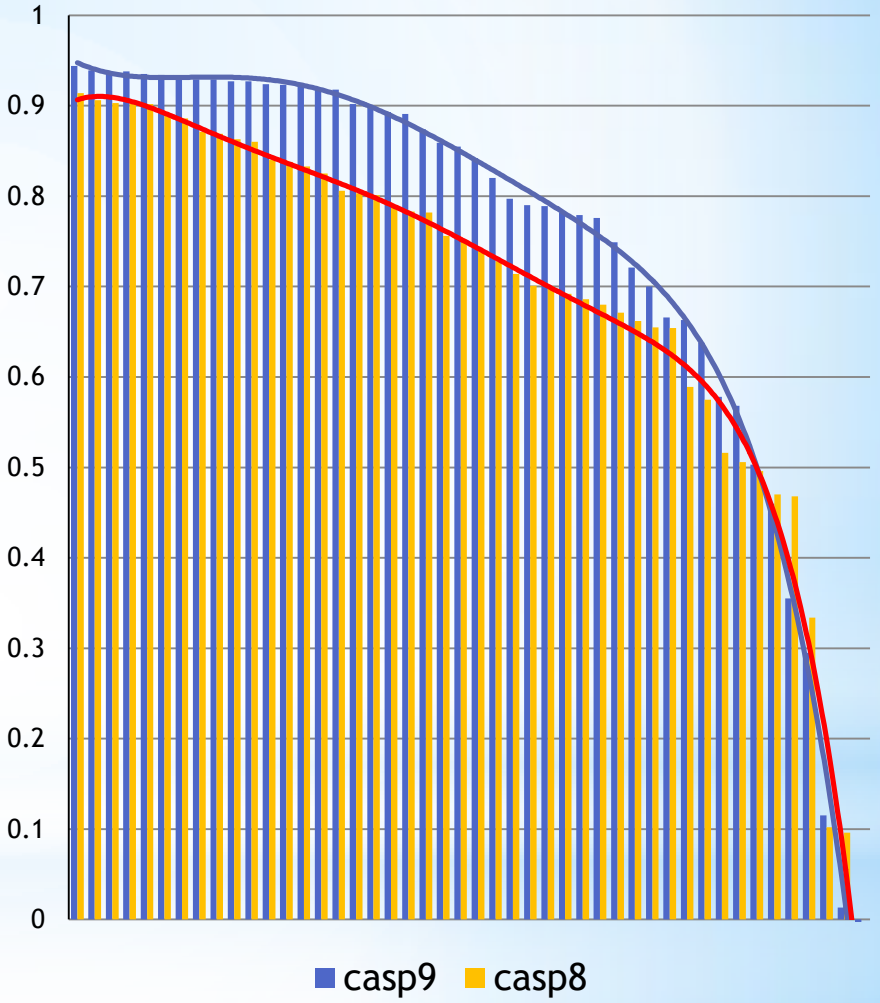
	Group Name	Group Leader	Clst/Sng	r
1	**QMEANclust**	Pascal Benkert, Swiss Institute of Bioinformatics, Switzerland	C	0.946
2	**ModFOLDclust2**	Liam McGuffin, Univ. Reading, UK	C	0.941
3	**MULTICOM-CLUSTER**	Jianlin Cheng, Univ. Missouri, USA	C	0.940
4	**MetaMQAPclust**	Marcin Pawlowski, Genesilico – Warsaw, Poland	C	0.940
5	**Mufold**	Dong Xu, Univ. Missouri, USA	C	0.938
6	**IntFOLD-QA**	Liam McGuffin, Univ. Reading, UK	C	0.937
7	**MUFOLD-QA**	Yi Shang, Univ. Missouri, USA	C	0.934
8	**United3D**	M. Takeda-Shitaka, Kitasato Univ.	C	0.931
9	**MQAPmulti**	Marcin Pawlowski, Genesilico – Warsaw, Poland	C	0.931
10	**MULTICOM**	Jianlin Cheng, Univ. Missouri, USA	C	0.930

CASP9-CASP8: comparison of correlation coefficients

QA1.1 - per-target analysis

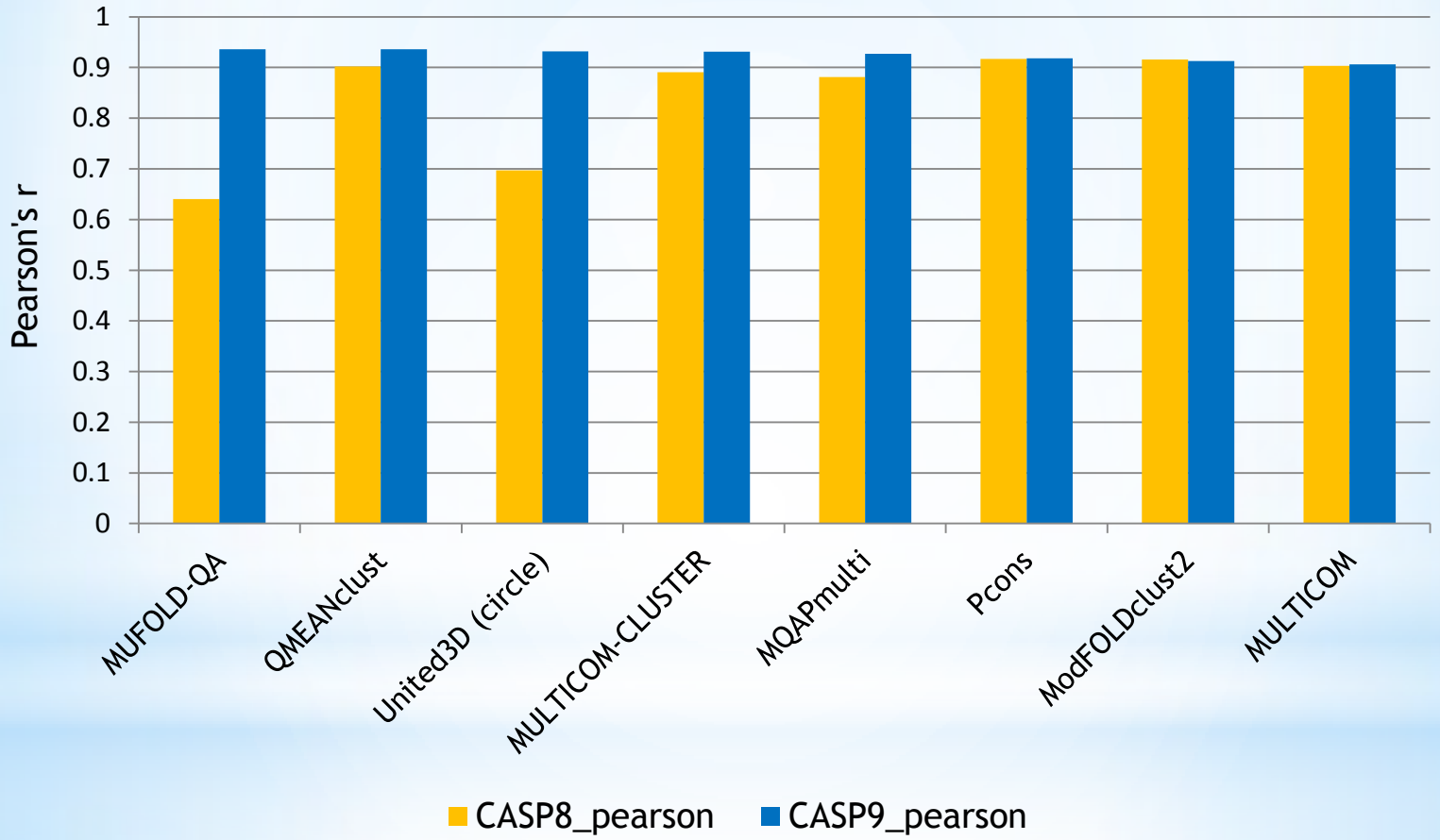


QA1.2 - all models together

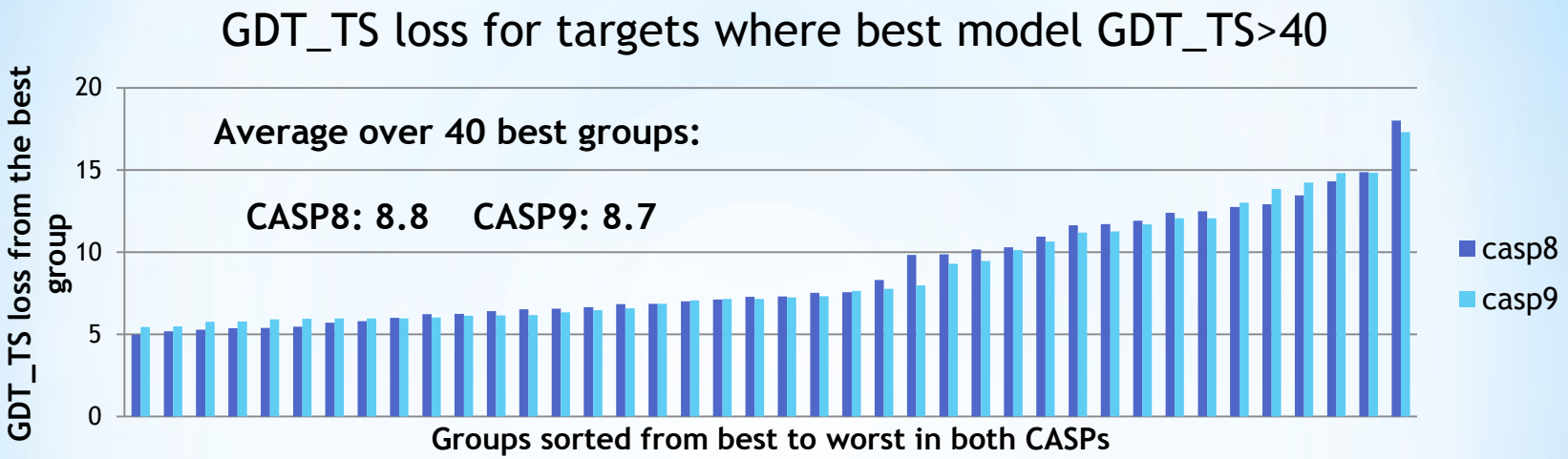


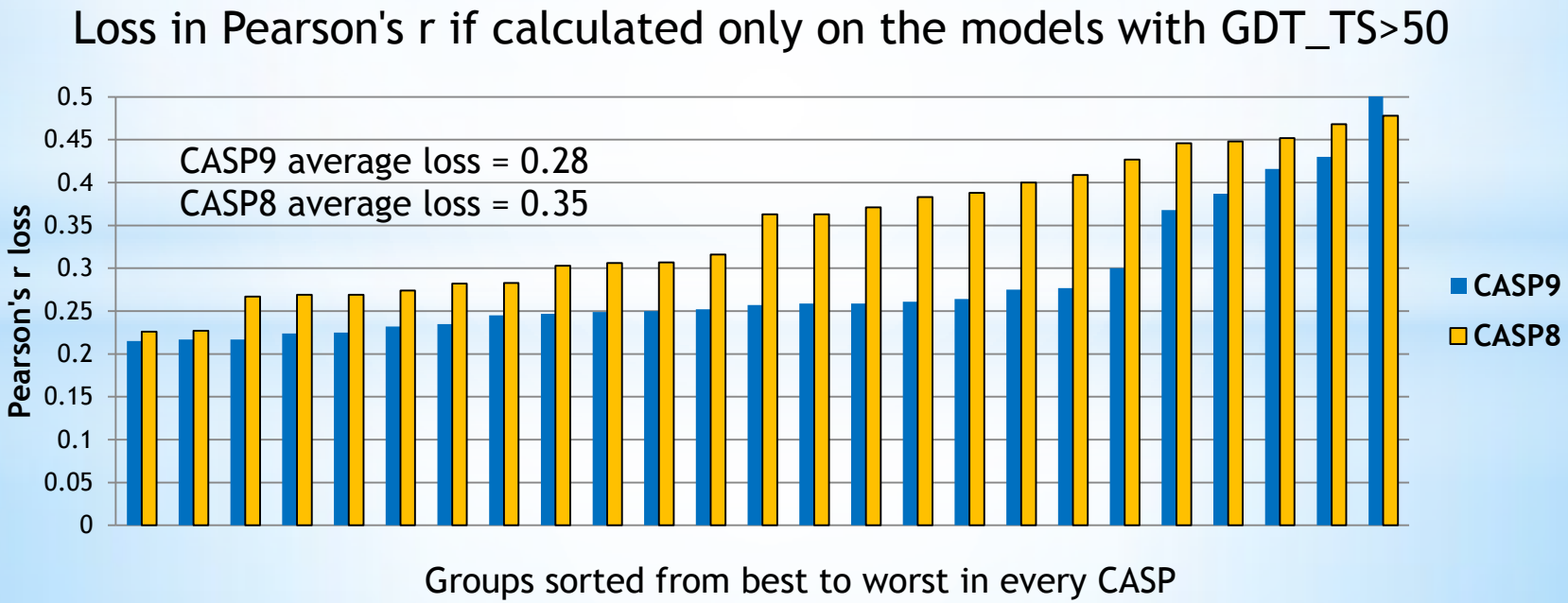
Groups sorted from best to worst in every CASP

CASP9-CASP8: progress of the best groups (QA1.1)



Loss on good models: CASP9 - CASP8 comparison



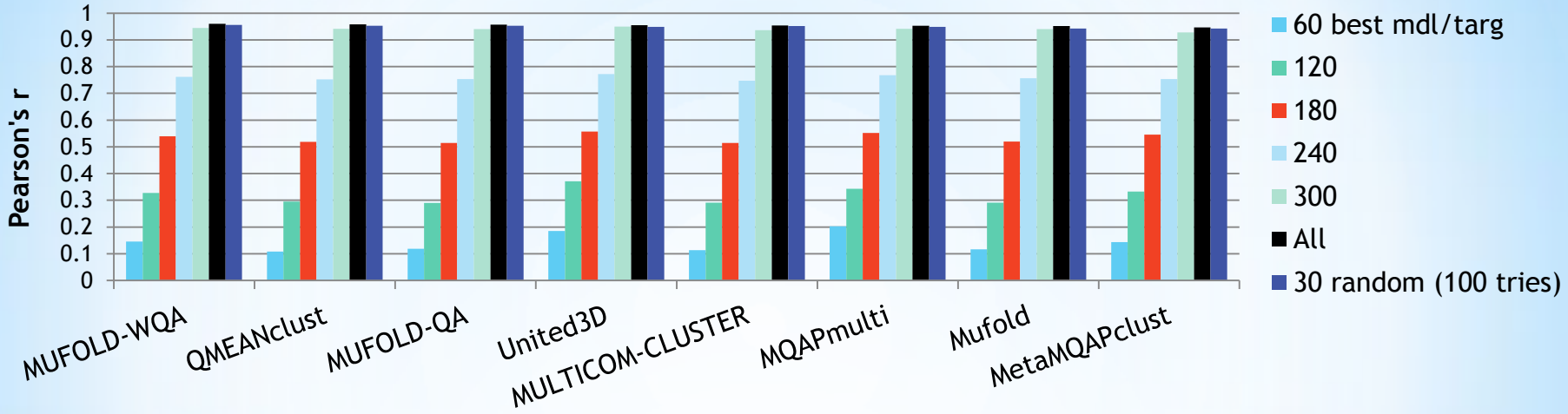


QA1 global quality of models
Average Pearson's *r*

		Target-based		All models	
		CASP8	CASP9	CASP8	CASP9
Best group		0.917	**0.936**	0.914	**0.946**
Median group		0.750	**0.848**	0.752	**0.858**
Best CASP8	*Pcons (TB)*	0.917	0.918	0.902	0.925
	ModFOLDclust (All)	0.915	0.913	0.914	0.941
Best CASP9	*Mufold-QA (TB)*	0.640	0.936	0.575	0.934
	QMEANclust (All)	0.902	0.936	0.903	0.946

So, QA1 is a solved problem *?

QA1.1: Incremental r averaged over the 90 targets with at least one model >50GDT_TS



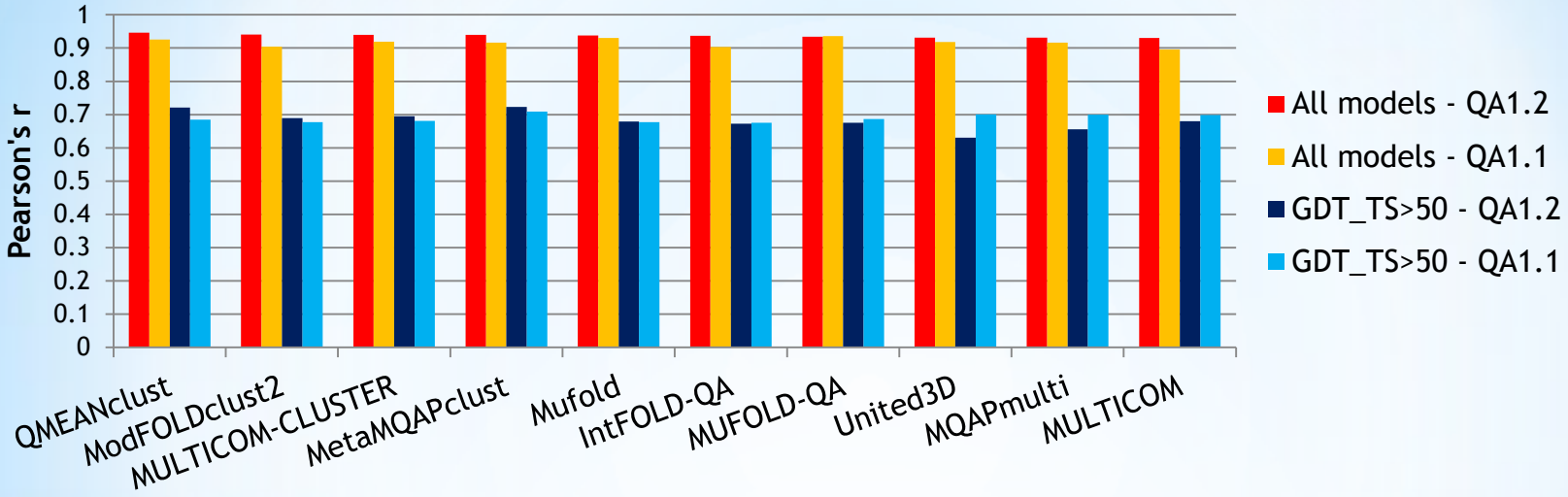
Objection #1 to the "solved" status of the problem:

To perform reasonably well, clustering QA methods need many models, which is unlikely in the real world

In CASP, MQAP results on only 30 randomly picked models show on the average the similar correlation coefficient as results on the whole set of ~300 models per target. Looks like there is no scaling effect here. But in fact, it is …

CASP naturally has hundreds of models of different quality from dozens of different servers. In real life you normally have several "best you can get" models and here clustering methods fail.

Pearson's r calculated on all models vs the models with GDT_TS>50



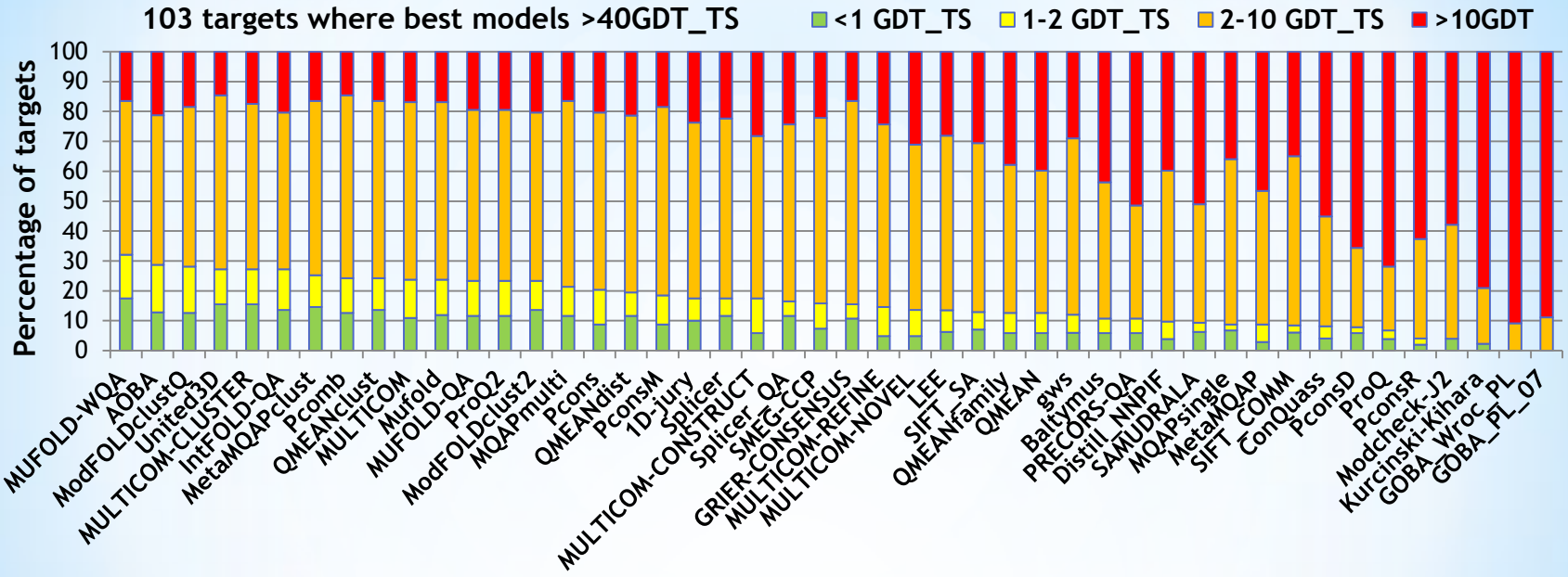
Objection #2 to the "solved" status of the problem:
To perform reasonably well, clustering QA methods need bad models

The tendency stays after confronting the "post-dictionally" curated data with the recalculated data for the same set of models

Objections 1-2 summary

* Do clustering methods make sense outside CASP-like exercises?
* Do we need clustering methods in CASP any more?
* If so, how can we test them under the conditions close to real life scenarios?

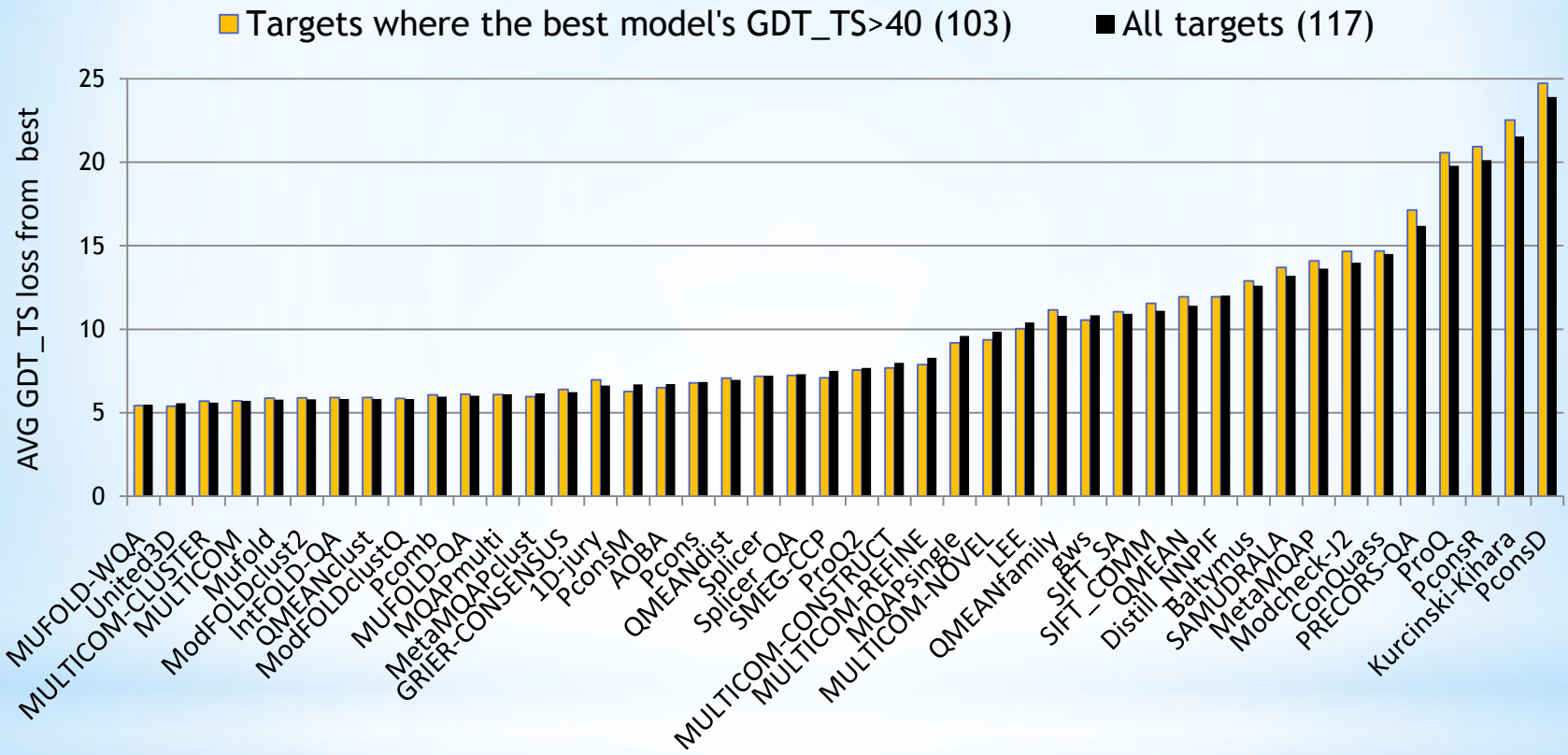
Ability to select the best models



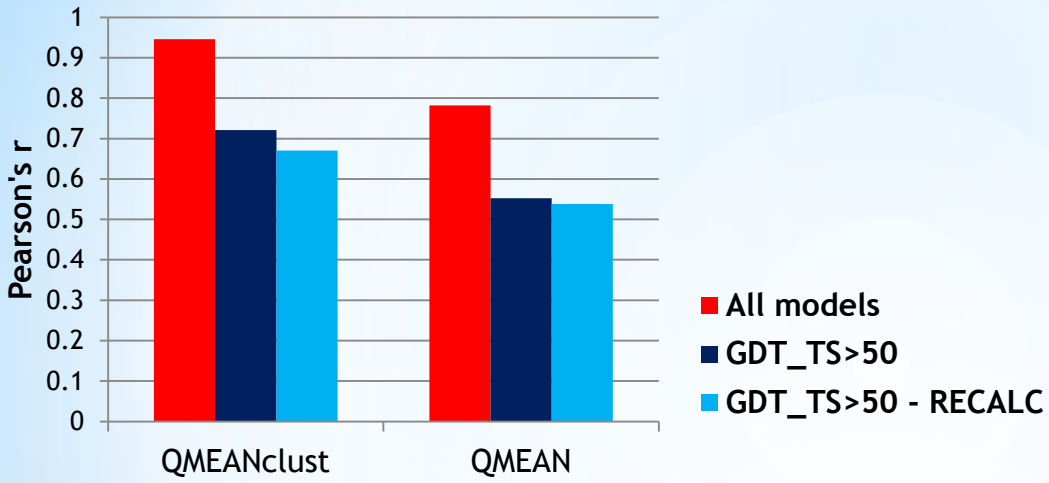
Objection #3 to the "solved" status of the problem:

QA methods are still far away from being able to select best models in the decoy set

Ability to select the best models



Pearson's r - confronting the “post-dictionally” curated data with the recalculated data



Note. Calculations were performed on 85/90 targets where there were at least 30 models over 50 GDT_TS

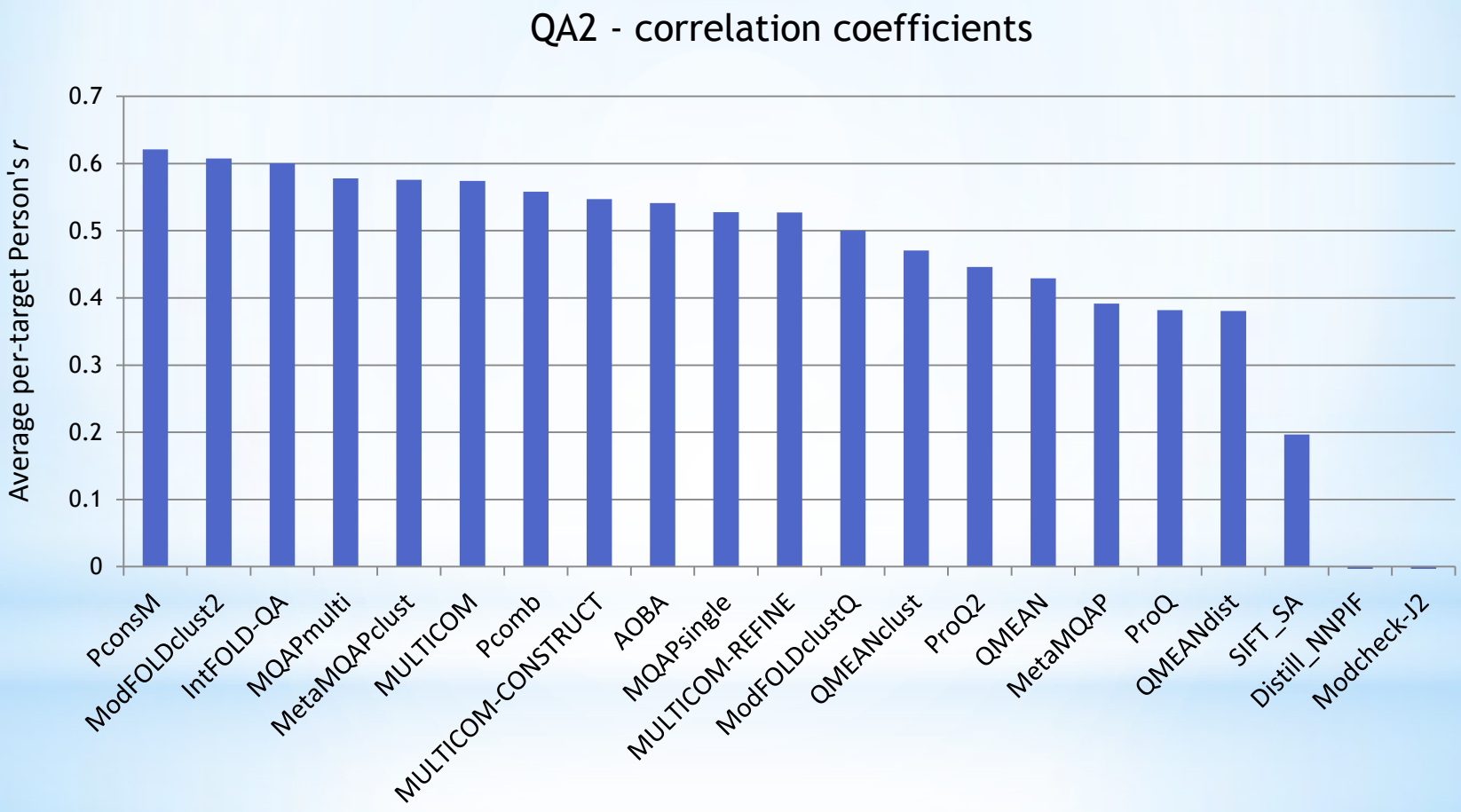
Question

Why does correlation for single-model methods drop while tested on good models only?

QA1 - problems

* There was an obvious, even though modest, progress in QA1 since CASP8. But still, do we still need QA1 in CASP?
* How can we switch to the domain-based evaluation for QA1 using QA2 data?
* Alternative QA measures (e.g., reliability of alignment in addition to structural fitness)
* Alternative evaluation measures (e.g., correlation with full-atom measures)

QA2 - assessment of local model quality estimates based on average per-target Person's *r* correlation between predicted and actual residue distances in model-target superposition

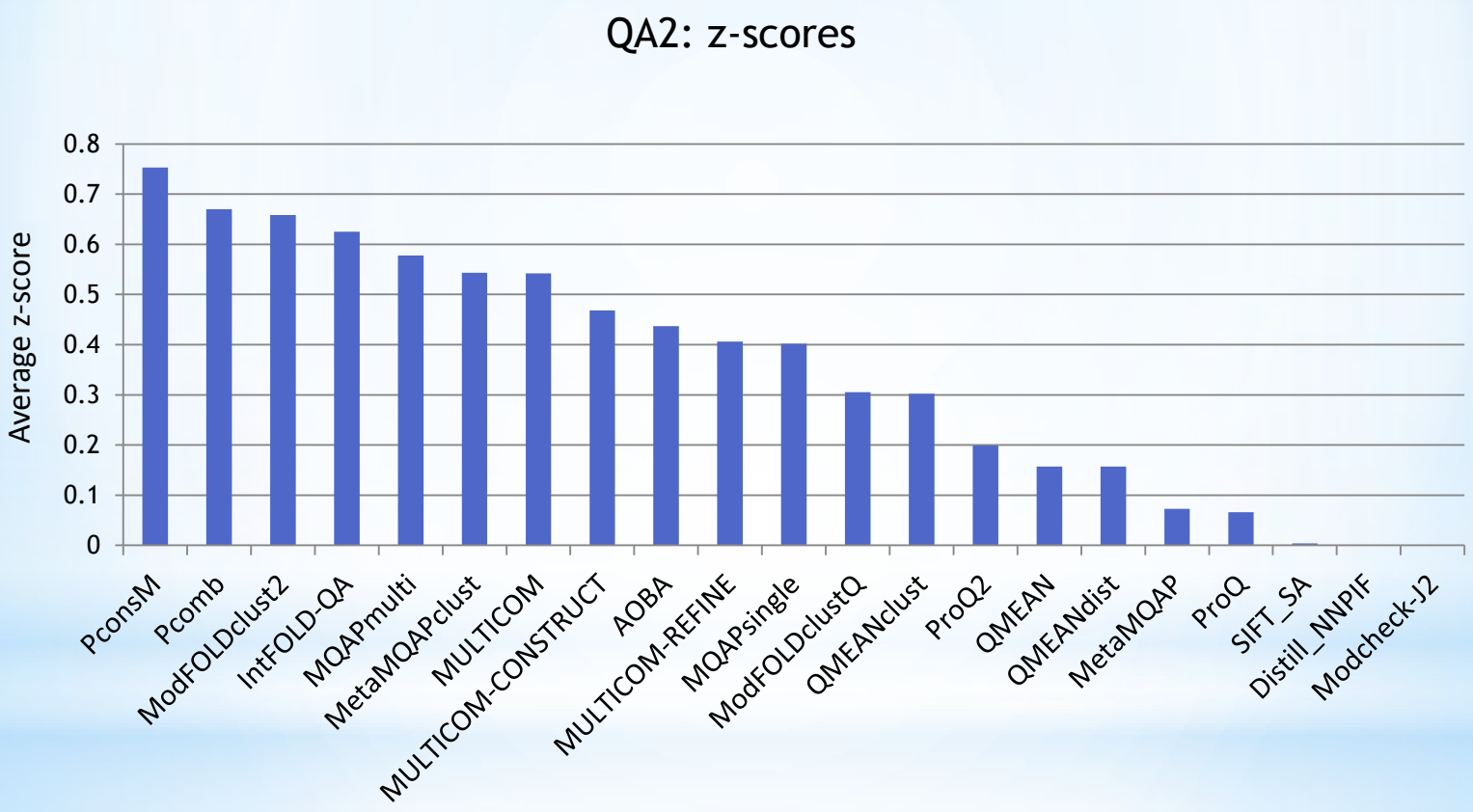


QA2: Paired t-test results: p-value \ number of common targets.

Shaded cells show statistically indistinguishable groups (p>0.01).

		56	397	78	369	426	490	273	80	324	308
PconsM	56	X	117	117	115	114	114	117	117	107	115
ModFOLDclust2	397	0.01	X	117	115	114	114	117	117	107	115
IntFOLD-QA	78	<0.01	<0.01	X	115	114	114	117	117	107	115
MQAPmulti	369	<0.01	<0.01	<0.01	X	114	112	115	115	107	115
MetaMQAPclust	426	<0.01	<0.01	<0.01	0.19	X	111	114	114	106	114
MULTICOM	490	<0.01	<0.01	<0.01	0.33	0.61	X	114	114	107	112
Pcomb	273	<0.01	<0.01	0.03	0.40	0.59	0.37	X	117	107	115
MULTICOM-CONSTRUCT	80	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	0.57	X	107	115
AOBA	324	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	0.36	0.69	X	107
MQAPsingle	308	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	0.08	<0.01	0.35	X

QA2 - assessment of local model quality estimates based on average per-target Person's *r* correlation between predicted and actual residue distances in model-target superposition

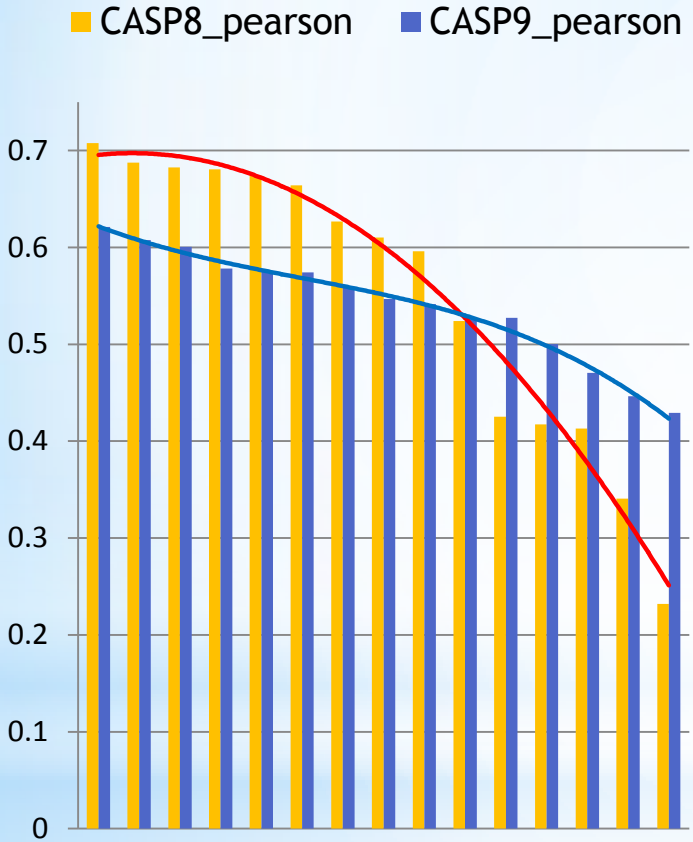


QA2 (per target analysis): best groups

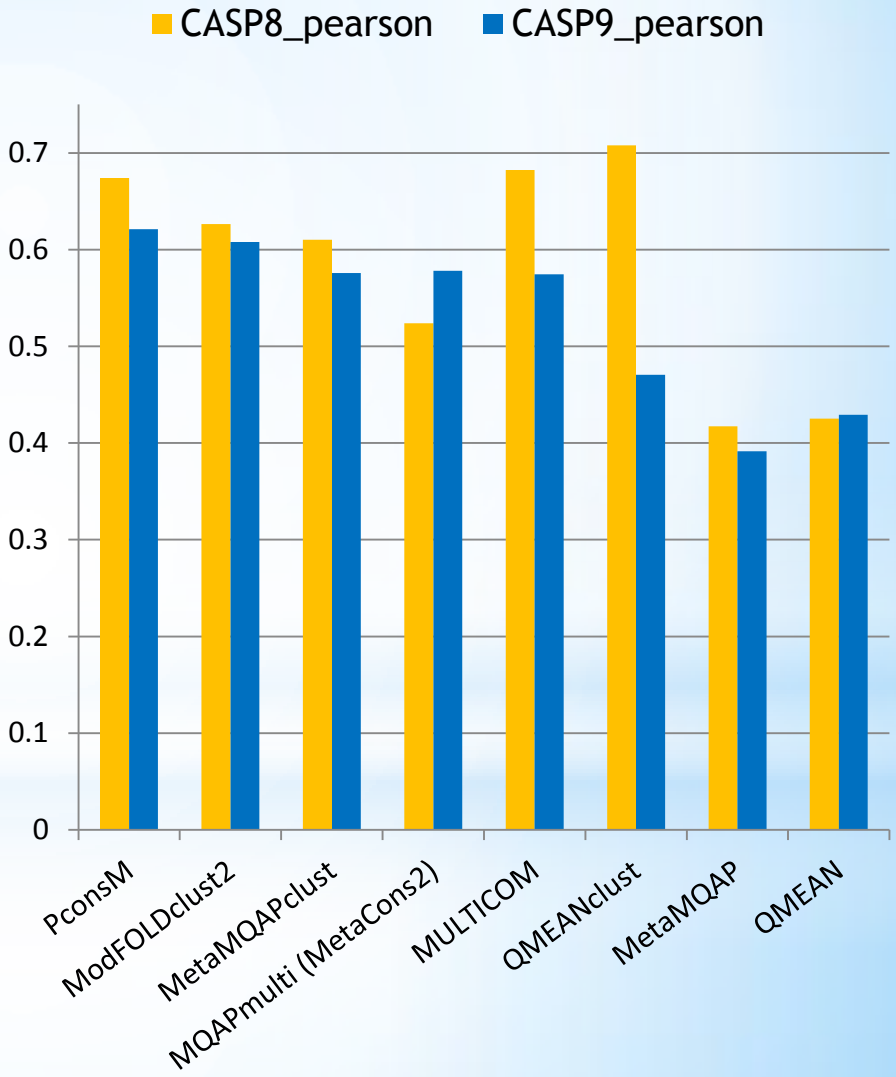
	Group Name	Group Leader	Clst/Sng	r
1	PconsM	Arne Elofsson, Stockholm Univ.	C	0.621
2	Pcomb	Arne Elofsson, Stockholm Univ.	C	0.558
3	ModFOLDclust2	Liam McGuffin, Univ. Reading, UK	C	0.608
4	IntFOLD-QA	Liam McGuffin, Univ. Reading, UK	C	0.601
5	MQAPmulti	Marcin Pawlowski, Genesilico – Warsaw, Poland	C	0.578
6	MetaMQAPclust	Marcin Pawlowski, Genesilico – Warsaw, Poland	C	0.576
7	MULTICOM	Jianlin Cheng, Univ. Missouri, USA	C	0.574
8	MULTICOM-CONSTRUCT	Jianlin Cheng, Univ. Missouri, USA	C	0.547
9	AOBA	Matsuyuki Shirota, Tohoku Univ.	C	0.541
10	MULTICOM-REFINE	Jianlin Cheng, Univ. Missouri, USA	C	0.527

CASP9-CASP8: comparison of correlation coefficients (QA2)

Groups participated in both CASPs and sorted from best to worst in every CASP



Groups participated in both CASPs and sorted from best to worst in CASP9



QA2
local (per-residue) quality of models

Average Pearson's *r*

	Target-based	
	CASP8	CASP9
Best group	**0.71**	0.62
Median group	**0.61**	0.53
Best CASP8 (QMEANclust)	*0.71*	*0.47*
Best CASP9 (PconsM)	*0.70*	*0.62*

QA2 - problems

* Even clustering methods performed not that great in QA2. Why they lag behind QA1 methods as logically global QA scores are obtained from per-residue model analysis?

* There is no progress in performance of QA2 methods since CASP8. What is holding the progress in this area? Any principal obstacles?

* Alternative QA measures for local model quality fitness and alternative results evaluation measures?

* Single-model methods look poor in QA1, but in QA2 they seem to be useless at the moment (r ~ 0.25). If we were to give up clustering methods, is there any light in the end of QA2 tunnel?