Table II. Statistical comparison of the best performing 12 groups and the NAÏVE_consensus method in QA1.2 mode (global quality estimates assessed on models from all targets pooled together)

		371	**0**	**2**	**397**	**426**	**78**	**386**	**407**	**359**	**369**	**119**	**490**	**319**
QMEANclust	**371**	35052	0.08	0.02	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
NAÏVE-consensus	**0**	2.53	35193	0.62	<0.01	<0.01	<0.01	<0.01	0.02	0.01	0.06	<0.01	<0.01	<0.01
Multicom-cluster	**2**	8.49	5.96	35158	0.05	0.01	<0.01	<0.01	0.07	0.04	0.15	<0.01	<0.01	<0.01
ModFOLDclust2	**397**	8.42	5.9	0.06	34964	0.18	0.02	0.6	0.9	0.94	0.61	<0.01	<0.01	0.34
MetaMQAPclust	**426**	8.37	5.85	0.12	0.06	35102	0.32	0.41	0.14	0.2	0.06	0.24	<0.01	0.68
IntFOLD-QA	**78**	13.54	11.03	5.08	5.13	5.2	34749	0.07	0.02	0.02	<0.01	0.86	<0.01	0.16
Mufold	**386**	13.23	10.75	4.86	4.91	4.97	0.17	33363	0.52	0.65	0.3	0.05	<0.01	0.68
United3D	**407**	16.37	13.88	7.99	8.03	8.1	2.94	3.08	33517	0.84	0.71	<0.01	<0.01	0.29
MUFOLD-QA	**359**	17.18	14.67	8.7	8.75	8.82	3.6	3.73	0.61	35152	0.56	0.01	<0.01	0.38
MQAPmulti	**369**	19.37	16.87	10.95	10.99	11.06	5.87	5.98	2.89	2.31	34141	<0.01	<0.01	0.14
Multicom-refine	**119**	18.64	16.13	10.17	10.21	10.28	5.05	5.17	2.06	1.46	0.86	35158	<0.01	0.11
MULTICOM	**490**	17.65	15.15	9.23	9.27	9.34	4.15	4.28	1.19	0.59	1.71	0.87	34196	<0.01
Pcons	**319**	23.48	20.99	15.05	15.09	15.16	9.95	10.02	6.93	6.39	4.05	4.94	5.77	34420

Table II.1. Results of the *Z*-tests (below diagonal) and DeLong tests (above diagonal). Each group submitted QA1 quality estimates for the number of models shown on the diagonal. The Pearson's *r* coefficients were computed on this set of models and compared based on the distributions of their corresponding Fisher's Z (formula 5). Values of the Z statistics are shown in the lower part of the table. Grey cells highlight pairs of statistically indistinguishable groups at the 10^{-2} significance level (Z<2.576). The upper part of the table displays *p*-values from the DeLong pairwise tests on the *AUC* scores, indicating probability that the difference in the performance of the best groups as binary classifiers (good/bad model) can be attributed to chance. Shaded cells highlight pairs of statistically indistinguishable groups at the 10^{-2} significance level.