Supplementary Information

Assessment of Protein Model Structure Accuracy Estimation in CASP13: Challenges in the Era of Deep Learning

Jonghun Won[1†], Minkyung Baek[1†‡], Bohdan Monastyrskyy[2], Andriy Kryshtafovych[2], and Chaok Seok[1*]

Correspondence to: Chaok Seok, Phone: +82-2-880-9197, E-mail: chaok@snu.ac.kr

[1] Department of Chemistry, Seoul National University, Seoul 08826, Republic of Korea

[2] Genome Center, University of California, Davis, California 95616, USA

[†] JW and MB should be considered joint first author.

[‡] Present address: Department of Biochemistry, University of Washington, Seattle, Washington 98195, USA

.

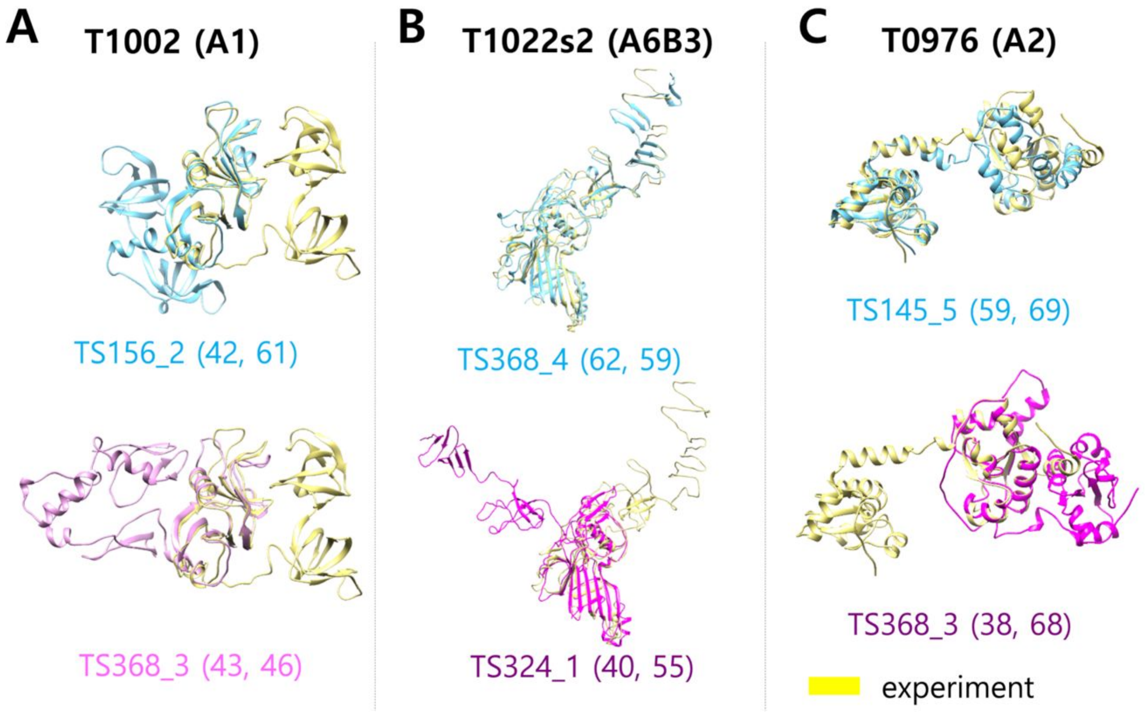
SUPPLEMENTARY TEXT

Examples of the cases in which models of similar GDT-TS but different LDDT and similar LDDT but different GDT-TS emphasize different aspects of model accuracy are described below. Model structures are compared in the next page for the targets whose experimental structures are public.

Target	**Model A**			**Model B**		
(Oligomeric state)	Model	GDT-TS	LDDT	Model	GDT-TS	LDDT
Similar GDT-TS, different LDDT						
T1002 (A1)	TS156_2	42	61	TS386_3	43	46
T1004 (A3)	TS324_3	53	77	TS116_5	53	63
T0974s1 (A1B1)	TS368_1	98	91	TS312_3	95	78
Similar LDDT, different GDT-TS						
T0973 (A2)	TS156_2	84	70	TS386_5	56	66
T1022s2 (A6B3)	TS386_4	62	59	TS324_1	40	55
T0976 (A2)	TS145_5	59	69	TS368_3	38	68

In the two cases, T1002 (**Figure A**) and T1004, more accurate local structures of the region not superposed to the native structure are not reflected in GDT-TS but in LDDT. In the case of T0974s1, high accuracy of local side chain packing is reflected only in LDDT but not in GDT-TS when the global model accuracy is very high.

The case of T0973 is a pathological example which originates from the property of LDDT that does not penalize contacts that are not present in the reference structure. The relatively low GDT-TS compared to LDDT of TS386_5 is due to a large non-native contact between secondary structure elements, not penalized in LDDT. In the two cases T1022s2 (**Figure B**) and T0976 (**Figure C**), models of similar local structure accuracy show very different global structure accuracy due to different domain orientations.


A
T1002 (A1)
TS156_2 (42, 61)
TS368_3 (43, 46)
B
T1022s2 (A6B3)
TS368_4 (62, 59)
TS324_1 (40, 55)
C
T0976 (A2)
TS145_5 (59, 69)
TS368_3 (38, 68)
experiment

SUPPLEMENTARY FIGURES

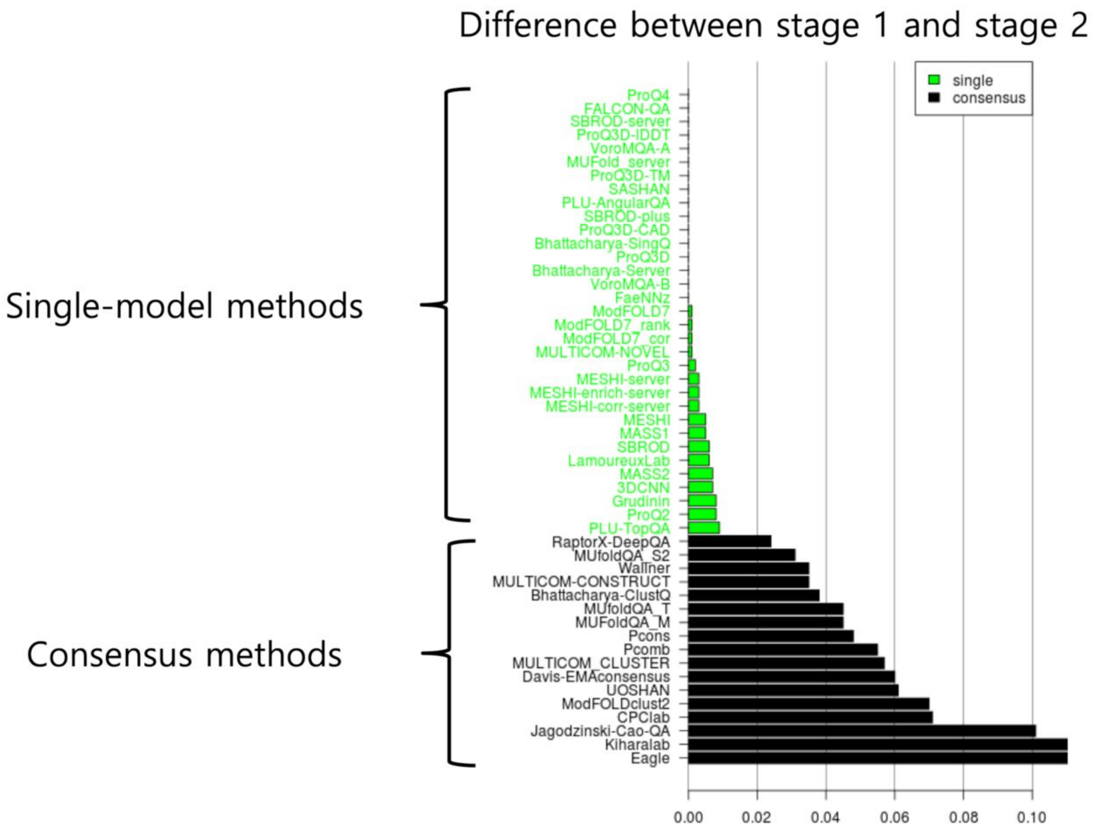


Figure S1. Classification of the EMA methods into single-model and consensus methods by the difference of scores submitted in the first stage and the second stage.

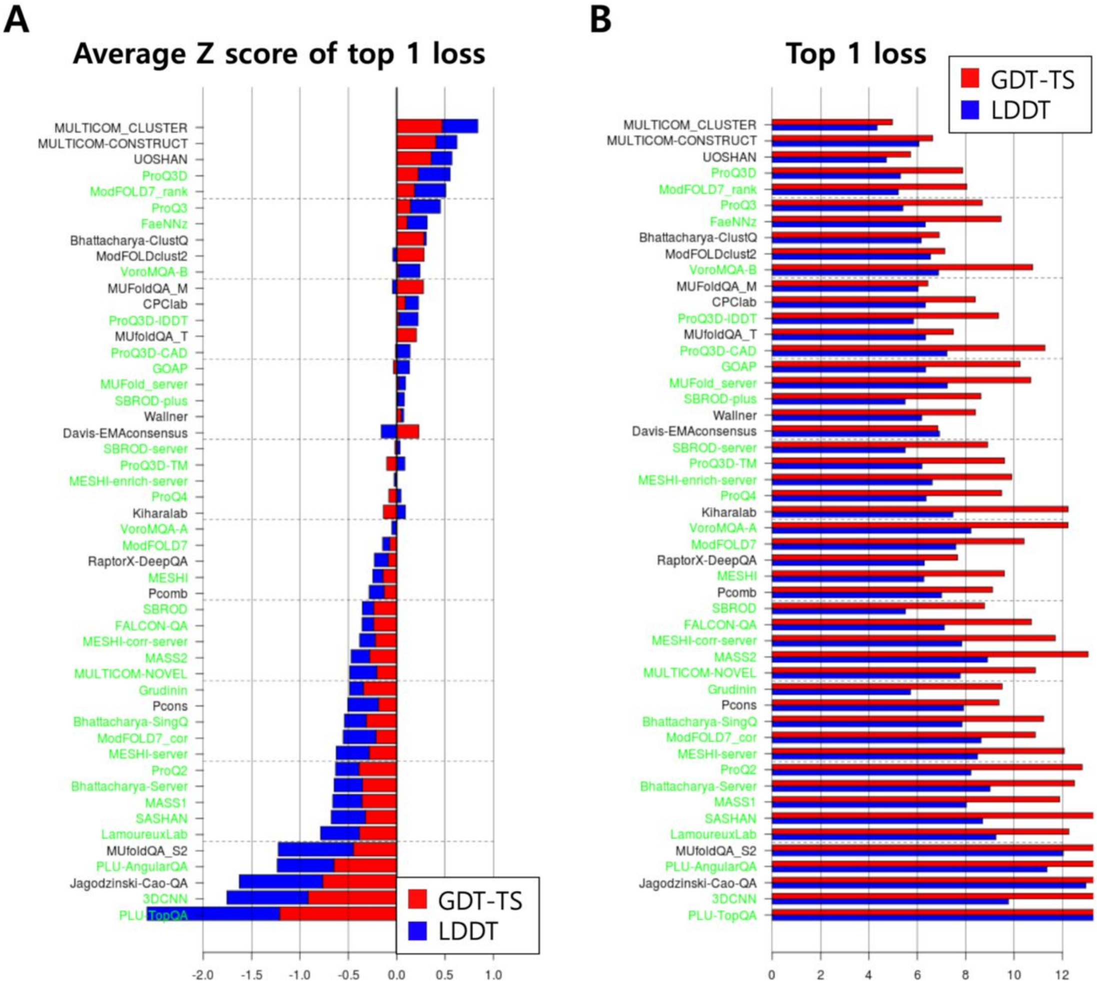


Figure S2. Ranking of the EMA methods in global accuracy estimation in terms of top 1 loss when only single-EU targets are considered.

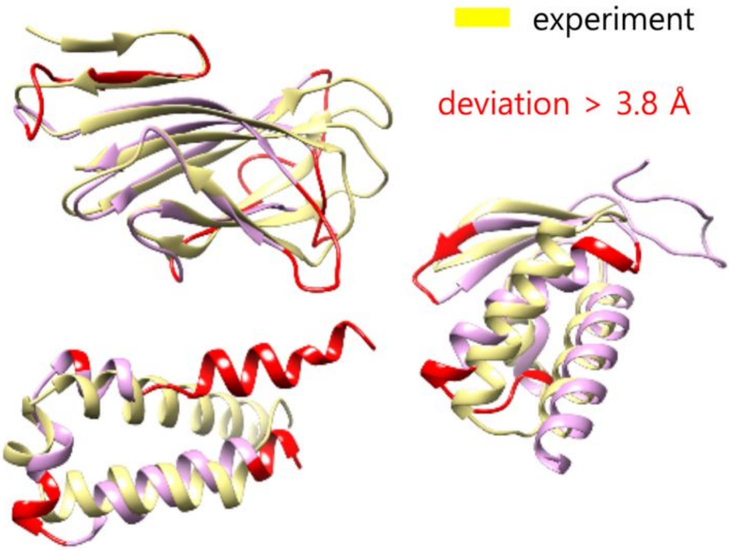


Figure S3. Examples of ULRs for which Cα deviations of the residues are greater than 3.8 Å after superposition to the experimental structure. Experimental structures are colored in yellow, model structures in pink and red, where red indicate ULRs.

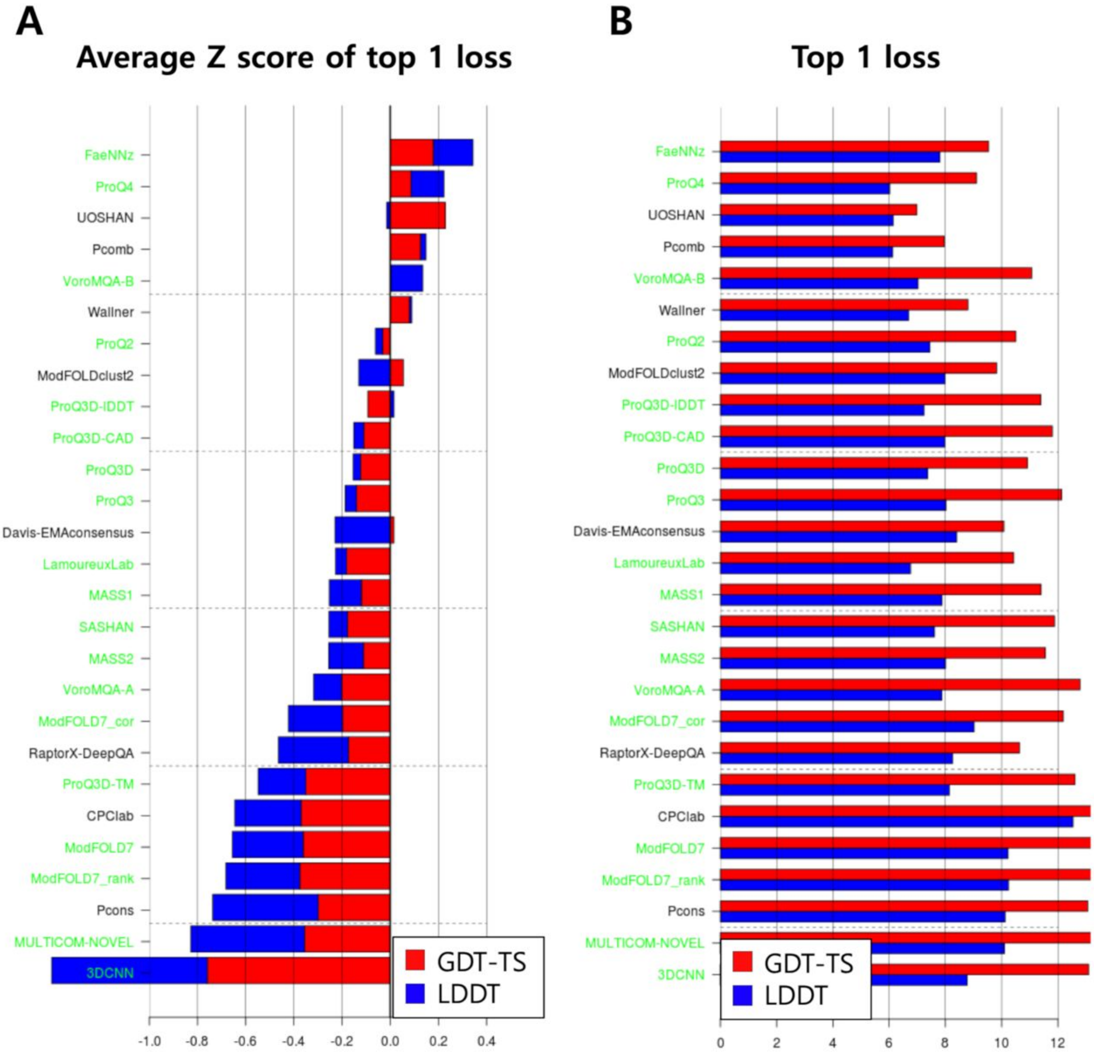


Figure S4. Performance of local EMA methods in global accuracy estimation in top 1 loss when local accuracy scores of the residues in each evaluation unit are converted to a global score and used to select top 1 model for the evaluation unit. A global score for each evaluation unit was calculated from the submitted local scores (which are supposed to be distance errors) as in GDT-TS calculation except for the cases in which all submitted local scores were less than one. In those cases, a global score was calculated as an average of the local scores as in LDDT calculation.

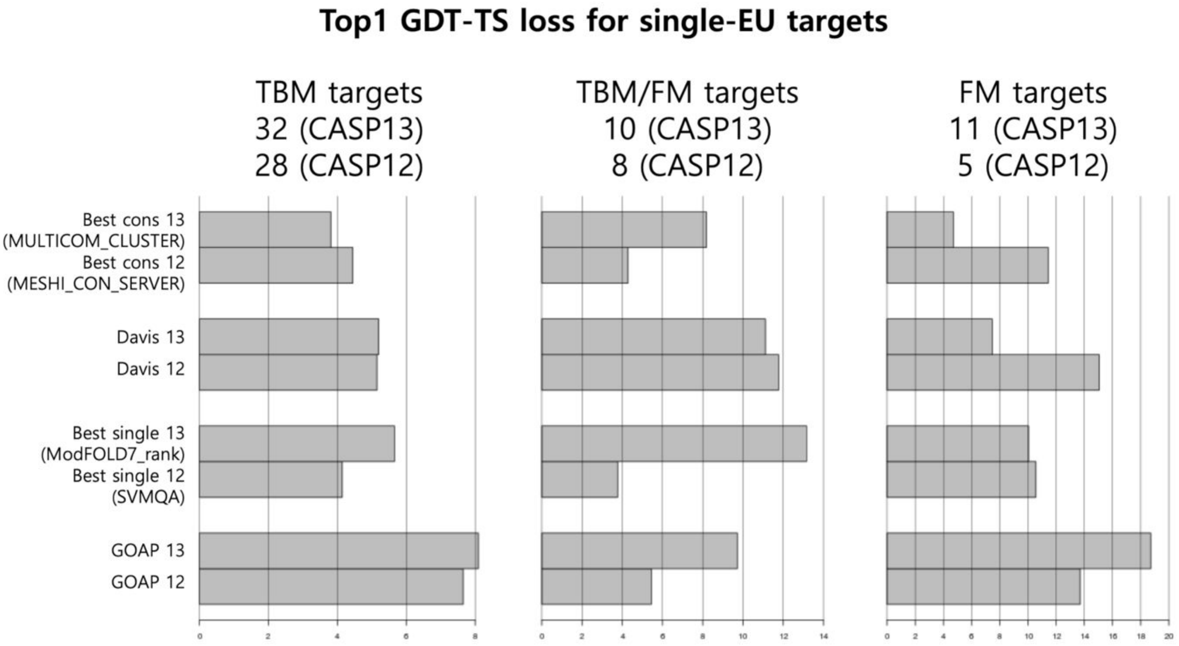


Figure S5. Performance comparison of the best EMA methods and reference methods in CASP12 and CASP13 for targets in different TS categories.

SUPPLEMENTARY TABLE

Table S1. Statistical comparison of top 1 GDT-TS/LDDT loss.

	MULTICOM_CLUSTER	ModFOLD7_rank	UOSHAN	MULTICOM-CONSTRUCT	Bhattacharya-ClustQ	ModFOLDclust2	MUfoldQA_T	MUFoldQA_M	ProQ3D	FaeNNz
MULTICOM_CLUSTER		+=	==	++	++	++	++	=+	++	++
ModFOLD7_rank			==	==	==	==	==	==	==	=+
UOSHAN				==	==	=+	==	=+	+=	++
MULTICOM-CONSTRUCT					==	==	==	==	==	==
Bhattacharya-ClustQ						==	==	==	==	==
ModFOLDclust2							==	==	==	==
MUfoldQA_T								==	==	==
MUFoldQA_M									==	+=
ProQ3D										==
FaeNNz										

The above table shows summary of the two-tailed paired t-tests on per-target differences between the models predicted as the best and the actual best models. Top 10 groups according to the cumulative ranking are mutually compared. Single-model methods are in green and consensus methods are in black. Each cell contains two characters representing the comparison between the two groups. "+" represents that the performance of the row group is statistically better than that of the column group. "-" represents the opposite case. "=" represents no significance. P-value threshold of 0.05 is used for all tests. The first character relates to GDT-TS and the second character relates to LDDT.

Table S2. Statistical comparison of absolute GDT-TS/LDDT accuracy estimation.

	MULTICOM-CONSTRUCT	MULTICOM_CLUSTER	ModFOLD7_cor	FaeNNz	ModFOLD7	MUfoldQA_T	MUFoldQA_M	UOSHAN	ModFOLD7_rank	ProQ3D-lDDT
MULTICOM-CONSTRUCT		=+	-+	+=	-+	-+	-+	-+	++	+=
MULTICOM_CLUSTER			=+	+-	=+	-+	-+	-+	+=	+=
ModFOLD7_cor				+-	==	-+	-+	-+	+-	+-
FaeNNz					-+	-+	-+	-+	=+	++
ModFOLD7						-+	-+	-+	+-	+-
MUfoldQA_T							=+	==	+-	+-
MUFoldQA_M								=-	+-	+-
UOSHAN									+-	+-
ModFOLD7_rank										=-
ProQ3D-lDDT										

The above table shows summary of the two-tailed paired t-tests on per-target differences between the predicted and observed model accuracy. Top 10 groups according to the cumulative ranking are mutually compared. Single-model methods are in green and consensus methods are in black. Each cell contains two characters representing the comparison between the two groups. "+" represents that the performance of the row group is statistically better than that of the column group. "-" represents the opposite case. "=" represents no significance. P-value threshold of 0.05 is used for all tests. The first character relates to GDT-TS and the second character relates to LDDT.

Table S3. Statistical comparison of ULR F1 values.

	VoroMQA-A	UOSHAN	VoroMQA-B	ModFOLDclust2	ProQ4	Davis-EMAconsensus	ModFOLD7	ModFOLD7_rank	ProQ3D-CAD	FaeNNz
VoroMQA-A		-	+	=	+	=	=	=	+	+
UOSHAN			+	+	+	+	+	+	+	+
VoroMQA-B				=	+	=	=	=	+	+
ModFOLDclust2					+	+	+	+	+	+
ProQ4						=	=	=	=	=
Davis-EMAconsensus							=	=	+	+
ModFOLD7								=	=	=
ModFOLD7_rank									=	=
ProQ3D-CAD										=
FaeNNz										

The above table shows summary of the two-tailed paired Wilcoxon-tests on per-target ULR F1 values. Top 10 groups according to the ULR F1 ranking are mutually compared. Single-model methods are in green and consensus methods are in black. Each cell contains two characters representing the comparison between the two groups. “+” represents that the performance of the row group is statistically better than that of the column group. “-” represents the opposite case. “=” represents no significance. P-value threshold of 0.05 is used for all tests.

Table S4. Statistical comparison of AUC of local error estimation.

	Davis-EMAconsensus	Pcomb	ModFOLDclust2	Wallner	ModFOLD7	ModFOLD7_rank	UOSHAN	ModFOLD7_cor	ProQ3	RaptorX-DeepQA
Davis-EMAconsensus		+	=	+	+	+	+	+	+	+
Pcomb			=	+	=	=	=	+	+	+
ModFOLDclust2				-	-	-	+	+	+	+
Wallner					=	=	=	+	+	+
ModFOLD7						=	=	+	+	+
ModFOLD7_rank							=	+	+	+
UOSHAN								=	=	+
ModFOLD7_cor									=	=
ProQ3										=
RaptorX-DeepQA										

The above table shows summary of the two-tailed paired Wilcoxon-tests on per-target AUC differences. Top 10 groups according to the AUC ranking are mutually compared. Single-model methods are in green and consensus methods are in black. Each cell contains two characters representing the comparison between the two groups. "+" represents that the performance of the row group is statistically better than that of the column group. "-" represents the opposite case. "=" represents no significance. P-value threshold of 0.05 is used for all tests.

Table S5. Statistical comparison of ASE.

	VoroMQA-A	VoroMQA-B	UOSHAN	ProQ4	ModFOLD7	ModFOLD7_rank	ProQ2	MASS2	MASS1	ProQ3
VoroMQA-A		+	+	+	+	+	+	+	+	+
VoroMQA-B			+	+	+	+	+	+	+	+
UOSHAN				=	=	=	=	=	=	=
ProQ4					+	+	+	+	+	=
ModFOLD7						=	-	-	-	-
ModFOLD7_rank							-	-	-	-
ProQ2								=	=	-
MASS2									=	=
MASS1										=
ProQ3										

The above table shows summary of the two-tailed paired t-tests on per-target ASE differences. Top 10 groups according to the ASE ranking are mutually compared. Single-model methods are in green and consensus methods are in black. Each cell contains two characters representing the comparison between the two groups. “+” represents that the performance of the row group is statistically better than that of the column group. “-” represents the opposite case. “=” represents no significance. P-value threshold of 0.05 is used for all tests.

Table S6. GDT-TS loss and MolProbity score of the top 1 models selected by three EMA methods 'Davis-EMAconsensus', 'GOAP', and 'ProQ3'

FM target	**Davis-EMAconsensus**			**GOAP**			**ProQ3**		
	Model	ΔGDT	MolP	Model	ΔGDT	MolP	Model	ΔGDT	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**	**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

† A model of high MolProbity score was selected