Title

Assessment of protein model structure accuracy estimation in CASP14: Old and new challenges

Journal

Proteins Structure Function and Bioinformatics, 89(12)

ISSN

0887-3585

Authors

Kwon, Sohee

Won, Jonghun

Kryshtafovych, Andriy

et al.

Publication Date

2021-12-01

DOI

10.1002/prot.26192

Peer reviewed

Published in final edited form as:

Proteins. 2021 December ; 89(12): 1940–1948. doi:10.1002/prot.26192.

Assessment of Protein Model Structure Accuracy Estimation in CASP14: Old and New Challenges:

Assessment of Model Accuracy Estimation in CASP14

Sohee Kwon[1,†], **Jonghun Won**[1,†,‡], **Andriy Kryshtafovych**[2], **Chaok Seok**[1,3]

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

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

[3]Galux Inc., Seoul 08826, Republic of Korea

Abstract

In CASP, blind testing of model accuracy estimation methods has been conducted on models submitted by tertiary structure prediction servers. In CASP14, model accuracy estimation results were evaluated in terms of both global and local structure accuracy, as in the previous CASPs. Unlike the previous CASPs that did not show pronounced improvements in performance, the best single-model method (from the Baker group) showed an improved performance in CASP14, particularly in evaluating global structure accuracy when compared to both the best single-model methods in previous CASPs and the best multi-model methods in the current CASP. Although the CASP14 experiment on model accuracy estimation did not deal with the structures generated by AlphaFold2, new challenges that have arisen due to the success of AlphaFold2 are discussed.

Keywords

CASP14 assessment; estimation of protein model accuracy; protein model quality assessment; protein structure prediction

INTRODUCTION

Estimating the accuracy of a protein model structure, or model quality assessment, is a crucial part of protein structure prediction and a gateway to proper usage of models in biomedical applications. Estimation of model accuracy (EMA, a.k.a. QA) has been a prediction category in CASP (Critical Assessment of techniques for protein Structure Prediction) since 2006[1–7]. The CASP Prediction Center has been providing a platform for evaluating EMA methods based on the protein model structures submitted by tertiary structure (TS) prediction servers.

The success of AlphaFold2 (AF2) in predicting the three-dimensional structures of single protein chains in CASP14 [CASP14 Ref: TS assessment paper] raises questions about the

Correspondence to: Chaok Seok, Phone: +82-2-880-9197, chaok@snu.ac.kr.
‡Present address: Galux Inc., Seoul 08826, Republic of Korea
†SK and JW should be considered joint first authors.

future role of EMA. Unfortunately, EMA methods were not tested on AF2 models in the regular season of CASP14 because AlphaFold2 was not registered as an automated server, but rather as a TS human group, and only TS server models were released for accuracy estimation. In CASP, human groups are given a longer deadline of three weeks (rather than three days allotted for automatic servers) and allowed to incorporate human intuition in various modeling steps such as in initial domain splitting or final model selection. In the CASP-COVID session, models submitted by AlphaFold (not identical to AlphaFold2) were evaluated along with lower-quality models by EMA methods for one target in a blind fashion and for three targets in a post-experiment [CASP14 Ref: CASP-COVID paper].

In CASP14, the EMA methods were assessed in a manner similar to that used in the previous CASPs, using the same metrics. It is notable that progress from CASP13 was observed for single-model EMA methods in terms of selecting the top models. By definition, single-model methods evaluate models without taking advantage of information on other server model structures, unlike multi-model methods, which use consensus. The best EMA method in CASP14 was from the Baker group[8], and it also performed better than other methods in CASP-COVID [CASP14 Ref: CASP-COVID paper].

When models are accurate, the choice of evaluation measure, such as GDT-TS[9] or LDDT[10] is less important. For models of intermediate accuracy, however, model rankings can vary depending on the measure. In CASP14 EMA, there remains a tendency that single-model methods estimate LDDT better than GDT-TS compared to multi-model methods. This is because it is more difficult to train an EMA method to estimate the superposition-dependent quantity, GDT-TS. A future challenge for EMA is now to distinguish high-quality models, such as those generated by AlphaFold2 from low-quality models, and to evaluate which regions of high-quality models are relatively incorrect and in need of improvement.

As the single-chain protein TS prediction problem is largely solved by the AlphaFold2 group in CASP14, the protein structure prediction problem may move on to the prediction of quaternary assembly structures. EMA methods have also been used to score oligomer model structures; for example, in the scoring rounds of CAPRI [CASP14 Ref: CAPRI assessment paper], and the role of oligomer EMA would become more critical in the future. Accurate prediction of protein quaternary structures from sequences is still a challenge because of the lesser amount of available sequence and structure data compared to single-chain proteins.

ASSESSMENT METHODS

Overview of the EMA experiment performed in CASP14

The EMA experiment in CASP14 was performed in the same manner as in the previous experiments[4–6]. Model accuracy predictors were asked to predict the global accuracy of each TS model as a value between 0 (inaccurate) and 1 (accurate), and the local accuracy of each residue as a distance error in Angstroms. The TS server predictions for 72 protein sequences (excluding 11 sequences that were canceled by the Prediction Center) were released in two stages. In the first stage, 20 diverse server models selected using the DAVIS-EMAconsensus method[4] were released for each protein target. In the second stage, the top 150 server models were released for each target, except for the target T1080, for which only 144 server models

were submitted. The accuracy estimation methods were classified into “single-model” and “multi-model” methods by computing the differences in the accuracy estimation results for the common models in the two stages.

In CASP14, 72 and 38 groups submitted estimations of global and local accuracy, respectively. Among them, five groups submitted predictions for less than 60% of the targets and thus were not considered for a detailed assessment. Only the second-stage predictions were subjected to a detailed analysis and ranking. Group rankings were obtained in terms of the target-averaged Z-scores, as in CASP13. The Z-score of each group was computed based on the results of all groups for each target using the mean and standard deviation calculated by neglecting the samples with an initial Z-score < −2, and setting Z-scores < −2 to −2.

Methods for assessing global structure accuracy estimation

TS models are high-dimensional quantities, and it is not possible to describe the model accuracy using a single accuracy measure. The model accuracy measures used in previous CASP assessments include GDT-TS[9], LDDT[10], CAD[11,12], and SphereGrinder[13]. In this assessment, GDT-TS and LDDT were selected as evaluation metrics for measuring global fold accuracy and local environment accuracy, respectively, as in CASP13. Both the GDT-TS and LDDT scores were scaled to a range between 0 and 100. A total of 66 targets were considered for the assessment of global accuracy predictions, excluding T1048 and T1072s1 (single helices), T1062 (a broken helix), T1070 and T1080 (obligate oligomers showing no core structure within each monomer), and T1077 (experimental structure unavailable during the assessment period).

The results of the global accuracy estimation were analyzed in terms of the performance in ranking models and in absolute error estimation, as in CASP13. First, ranking performance was assessed by the accuracy of the top model selected by each EMA method, defined as the absolute difference of GDT-TS (and LDDT) of the top model selected by the predicted score and that by the best model with the knowledge of the experimental structure, and denoted by “top 1 GDT-TS loss” (and “top 1 LDDT loss”). For this analysis, only the targets with at least one higher-quality model with a score of > 40 were considered, resulting in 58 and 64 targets for GDT-TS and LDDT, respectively. Second, the absolute value of the predicted score was assessed by taking the difference between the predicted score and an accuracy measure (GDT-TS or LDDT) for all models. All 66 targets were included in this analysis.

As baseline methods for comparing progress over the CASPs, “GOAP” and “DAVIS-EMAconsensus” were employed for the single-model and multi-model methods, respectively. “GOAP” is a high-performance distance- and orientation-dependent statistical potential[14] that does not rely on protein-specific information. Although GOAP does not provide a score of 0–1, it can be used to rank different model structures. DAVIS-EMAconsensus estimates model accuracy purely based on consensus by scoring the i^{th} model by an average GDT-TS to all other models in the pool as $score_i = \frac{1}{N}\sum_{j \neq i}(GDT-TS)_{ij}$.

Methods for assessing local structure accuracy estimation

The local accuracy estimation was assessed at the level of evaluation units (EUs)[15]. Two EUs from obligate oligomers, T1070-D1 and T1080-D1, for which the hydrophobic core is absent within the monomer structure, were excluded. Only model structures with GDT-TS > 40, which corresponded to 10,308 models from 90 EUs, were considered.

Three types of analyses, namely, ASE, AUC, and ULR, were conducted, as in CASP13. ASE and AUC analyses were carried out in the same manner as in previous CASPs[6,7]. ASE measures the average residue-wise S-score error as $\mathrm{ASE} = \left(1 - \frac{1}{N}\sum_{i=1}^{N} |S(e_i) - S(d_i)|\right) \times 100$, where the S-score error $|S(e_i) - S(d_i)|$ for the i^{th} residue is calculated with the predicted (e_i) and the actual (d_i) distance errors of the i^{th} Cα atom of a given model after superposition of the model onto the experimental structure by LGA[9]. The S-function is $S(d) = 1/[1 + (d/d_0)^2]$ with $d_0 = 5$ Å, and N is the number of residues in the EU.

The AUC and unreliable local region (ULR) analyses evaluate how well the predicted local accuracy score distinguishes between accurately and inaccurately modeled residues in each model. A residue in a model is defined as accurately modeled if its Cα distance from the corresponding residue in the experimental structure is within 3.8 Å in the optimal model-target superposition. AUC measures the area under the ROC curve, which plots the true-positive rate against the false-positive rate in the prediction of accurate and/or inaccurate residues, varying the cutoff score for distinguishing between accurate and inaccurate residues.

The ULR analysis assesses the ability to detect stretches of inaccurately modeled residues. An ULR is a region consisting of three or more sequential model residues deviating by more than 3.8 Å from the corresponding target residues upon their optimal global superposition. Two ULRs separated by a single residue are merged into a single ULR. After assigning the ULRs, their accuracy and coverage are calculated. ULRs predicted within two residues from the boundaries of the actual ULRs are considered to be accurately predicted. For each accuracy prediction group, the $\mathrm{F1} = 2\,\frac{\mathrm{accuracy} \times \mathrm{coverage}}{\mathrm{accuracy} + \mathrm{coverage}}$ score is calculated. The cutoff score for inaccurately predicted residue and the sign of the score were adjusted for each group to maximize the target-averaged F1 score. This is because some groups did not submit local accuracy scores as predicted distance deviations.

As a baseline method for local accuracy estimation, the “NAIVE_LOOP” measure was newly introduced in this CASP. It simply takes the distance along the sequence between the residue whose accuracy to be predicted and the closest residue belonging to any secondary structure element.

ASSESSMENT RESULTS AND DISCUSSION

Classification of EMA methods to single-model and multi-model methods

Accuracy estimation methods were classified into single-model and multi-model methods based on the difference between the first- and second-stage accuracy estimation results for the same model (see Methods). Considering stochastic numerical errors, a maximum

margin of 0.02 (in the 0–1 score scale), was allowed for the difference defining single-model methods. Among the 70 methods evaluated, 46 were designated as single-model methods and 24 as multi-model methods. Details of the method classification are presented in Supplementary Figure S1. Throughout the figures in the paper, single-model methods are colored green and multi-model methods black.

Assessment of global accuracy estimation

Ranking in the top 1 loss: The CASP14 ranking in global accuracy estimation in terms of the top 1 loss is presented in Figure 1. The ranking is based on the sum of the average Z-scores of the top 1 GDT-TS and LDDT losses. The Z-scores are averaged over the 58 and 64 targets for which at least one model had GDT-TS > 40 and LDDT > 40, respectively. The best methods according to this ranking are “BAKER-experimental” and “BAKER-ROSETTASERVER.” Overall, the Z-score of the top 1 GDT-TS loss and that of the top 1 LDDT loss are highly correlated (Pearson correlation coefficient of 0.914). A statistical analysis of the performance differences among the top groups is provided in Supplementary Table S1.

Notably, two single-model methods, “BAKER-experimental” and “BAKER-ROSETTASERVER,” performed better than the best multi-model method in this CASP, unlike CASP13 where the best multi-model method performed better than the best single-model method. Compared to CASP13, the best single-model method improved both in the top 1 GDT-TS and LDDT losses in CASP14. Among the single-model methods, “BAKER-experimental” performed the best in terms of top 1 GDT-TS loss (target-averaged loss = 8.4) and “BAKER-ROSETTASERVER” in terms of top 1 LDDT loss (target-averaged loss = 4.0).

The reference methods “DAVIS-EMAconsensus” and “GOAP” showed about average performance in terms of top 1 loss in this CASP.

Performance of EMA methods as meta-servers in the top 1 GDT-TS loss: When EMA methods are considered as meta-servers (i.e., model selectors represented by tertiary structure models they selected as the best), these methods perform better than the best TS servers, but not better than the best TS human method (Figure 2). The average top 1 GDT-TS loss of EMA meta-servers and TS servers on all targets in shown in Figure 2A; the comparison of EMA meta-servers with all TS methods on human targets is illustrated in Figure 2B.

Although there is no practical value of the meta-servers in real-life applications because not all CASP TS servers are available in non-CASP situations, such a comparison enables us to check the status of current EMA methods relative to the TS methods and to the EMA methods of previous CASPs.

The difference in the top 1 GDT-TS loss between the top TS human method “AlphaFold2” and the best EMA method is pronounced, unlike in previous CASPs, confirming that the current top tertiary prediction human group achieved results beyond possible from consensus.

Ranking in the absolute accuracy estimation: The CASP14 assessment of EMA methods in terms of the absolute score value was performed separately using the GDT-TS and LDDT measures, as presented in Figures 3A and 3B. The Z-scores were averaged over all 66 targets. The best absolute GDT-TS estimation was made by the naïve consensus method, "DAVIS-EMAconsensus," with an average GDT-TS difference of approximately 6.8; and the best LDDT estimation by the single-model method, "ModFOLD8_rank" with an average LDDT difference of 6.7. A statistical analysis of the performance differences among the top groups is provided in Supplementary Table S2.

Assessment of local accuracy estimation

Ranking in the local accuracy estimation: The CASP14 ranking of EMA methods in local accuracy estimation is presented in Figure 4. The ranking is based on the sum of the average Z-scores of the ASE, AUC, and ULR scores, as presented in Figure 4A. The Z-scores were averaged over 90 EUs. According to this ranking, the best multi-model method is "ModFOLDclust2," and the best single-model method is "BAKER-ROSETTASERVER." "BAKER-ROSETTASERVER" achieved a performance (ULR Z-score = 0.596) comparable to that of the best method ("VoroMQA-A") in CASP13 (ULR Z-score = 0.587). The Pearson correlation coefficients for the ULR-AUC, ULR-ASE, and AUC-ASE pairs of assessment measures are 0.70, 0.16, and −0.03, respectively, thus showing no correlation of the ASE measure to the other two. A statistical analysis of the performance differences among the top groups is provided in Supplementary Tables S3–S5 for each of the three local accuracy measures.

Prediction of inaccurately modeled regions: The over-all-targets averages of ULR F1 scores for the best multi-model method, "Yang_TBM," the best single-model method "BAKER-ROSETTASERVER," and the baseline method "NAIVE_LOOP" are 0.24, 0.19, and 0.17, respectively. This implies that inaccurately modeled regions (ULRs) in a model structure may be detected with an accuracy and coverage of approximately 20% with current EMA methods. Two examples of ULR prediction results for which ULR-F1 scores are very good (= 1) and poor (0.2) are illustrated in Figures 5A and 5B, respectively.

ULRs occur due to the differences between the target and evolutionarily related proteins, the information from which is used for structure prediction. Some ULRs may be relevant to the functional specificity of the target protein, while others may only correspond to flexible regions. Some flexible regions may be involved in functional interactions with other biomolecules or drug-like molecules. Hence, it is difficult to determine their importance without further information on the interactions.

Global accuracy estimation using the local accuracy estimation: Local accuracy scores were used to calculate global accuracy score for each EU, and the top 1 loss analysis was performed, as presented in Supplementary Figure S2. For groups that submitted local accuracy estimates between 0 and 1, an average of the residue-based scores was used as global accuracy score, assuming residue LDDT was estimated. For groups that submitted scores > 1, GDT-TS was directly calculated assuming residue error in Angstrom was

estimated. The global accuracy score generated in this way performed worse than the best methods presented in Figure 1, both in terms of the top 1 GDT-TS and LDDT losses.

Progress over the previous CASPs

Progress in CASP EMA methods relative to reference methods: Performances in terms of the top 1 GDT-TS/LDDT loss for the best methods relative to the reference methods in CASP12, 13, and 14 are presented in Figures 6A and 6B for the multi-model and single-model methods, respectively. A statistical potential, "GOAP,"[14] was introduced as a reference method for single-model methods to exclude the effect of consensus. The best methods in Figure 6 in terms of the GDT-TS and LDDT can vary.

Figure 6 shows that the best multi-model methods performed better than the reference multi-model method, "DAVIS-EMAconsensus," with the ratio of top 1 loss > 1.0 in all CASPs for both GDT-TS and LDDT. However, the relative performance worsened from CASP13 to CASP14.

As shown in Figure 6B, the best single-model methods also performed better than the reference method "GOAP." The performance of the best single-model method improved both in terms of the top 1 GDT-TS and the top 1 LDDT. The progress in the top 1 LDDT in this CASP is notable, considering that there was no progress in the previous CASP.

CONCLUSION

An increased number of model accuracy estimation methods (73) participated in CASP14, compared to 51 in CASP13. Definite progress was observed in CASP14 in terms of the global accuracy estimation of the best single-model EMA method relative to the reference method, GOAP, the local accuracy estimation of the best single-model EMA method was comparable to that of the single-model EMA in CASP13.

A new challenge appeared in this CASP with the advent of AlphaFold2. TS models will become much more accurate in the coming years, and EMA should be able to deal with such changes to contribute to the field. In CASP14, TS models submitted by AlphaFold2 were not evaluated by EMA methods because AlphaFold2, registered as a human group, was not included in the current format of the EMA experiment. Post-experiments could be performed by individual EMA method developers with the rich data provided by the Prediction Center. With the significant advance in the structure prediction of single protein chains by AlphaFold2, an immediate next challenge would be the prediction of quaternary structures. Therefore, EMA methods that can provide an accurate evaluation of oligomer structures are required.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

The work was supported by the National Research Foundation of Korea (NRF) grants funded by the Korea government (Nos. 2020M3A9G7103933 and 2019M3E5D4066898) and by the US National Institute of General Medical Sciences (NIGMS/NIH), grant R01GM100482.

DATA AVAILABILITY:

The data that supports the findings of this study are available in the supplementary material of this article.

REFERENCES

1. Cozzetto D, Kryshtafovych A, Ceriani M, Tramontano A. Assessment of predictions in the model quality assessment category. Proteins: Structure, Function, and Bioinformatics. 2007;69(S8):175–183.
2. Cozzetto D, Kryshtafovych A, Tramontano A. Evaluation of CASP8 model quality predictions. Proteins: Structure, Function, and Bioinformatics. 2009;77(S9):157–166.
3. Kryshtafovych A, Fidelis K, Tramontano A. Evaluation of model quality predictions in CASP9. Proteins. 2011;79 Suppl 10:91–106. [PubMed: 21997462]
4. Kryshtafovych A, Barbato A, Fidelis K, Monastyrskyy B, Schwede T, Tramontano A. Assessment of the assessment: evaluation of the model quality estimates in CASP10. Proteins. 2014;82 Suppl 2:112–126. [PubMed: 23780644]
5. Kryshtafovych A, Barbato A, Monastyrskyy B, Fidelis K, Schwede T, Tramontano A. Methods of model accuracy estimation can help selecting the best models from decoy sets: Assessment of model accuracy estimations in CASP11. Proteins. 2016;84 Suppl 1:349–369.
6. Kryshtafovych A, Monastyrskyy B, Fidelis K, Schwede T, Tramontano A. Assessment of model accuracy estimations in CASP12. Proteins. 2018;86 Suppl 1:345–360. [PubMed: 28833563]
7. Won J, Baek M, Monastyrskyy B, Kryshtafovych A, Seok C. Assessment of protein model structure accuracy estimation in CASP13: Challenges in the era of deep learning. Proteins: Structure, Function, and Bioinformatics. 2019;87(12):1351–1360.
8. Hiranuma N, Park H, Baek M, Anishchenko I, Dauparas J, Baker D. Improved protein structure refinement guided by deep learning based accuracy estimation. Nature Communications. 2021;12(1):1340.
9. Zemla A LGA: A method for finding 3D similarities in protein structures. Nucleic acids research. 2003;31(13):3370–3374. [PubMed: 12824330]
10. Mariani V, Biasini M, Barbato A, Schwede T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. Bioinformatics (Oxford, England). 2013;29(21):2722–2728.
11. Olechnovic K, Kulberkyte E, Venclovas C. CAD-score: a new contact area difference-based function for evaluation of protein structural models. Proteins. 2013;81(1):149–162. [PubMed: 22933340]
12. Olechnovic K, Venclovas C. The CAD-score web server: contact area-based comparison of structures and interfaces of proteins, nucleic acids and their complexes. Nucleic acids research. 2014;42(Web Server issue):W259–263. [PubMed: 24838571]
13. Kryshtafovych A, Monastyrskyy B, Fidelis K. CASP prediction center infrastructure and evaluation measures in CASP10 and CASP ROLL. Proteins. 2014;82 Suppl 2:7–13. [PubMed: 24038551]
14. Zhou H, Skolnick J. GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction. Biophysical journal. 2011;101(8):2043–2052. [PubMed: 22004759]
15. Kinch LN, Kryshtafovych A, Monastyrskyy B, Grishin NV. CASP13 Target Classification into Tertiary Structure Prediction Categories. Proteins: Structure, Function, and Bioinformatics. 2019.

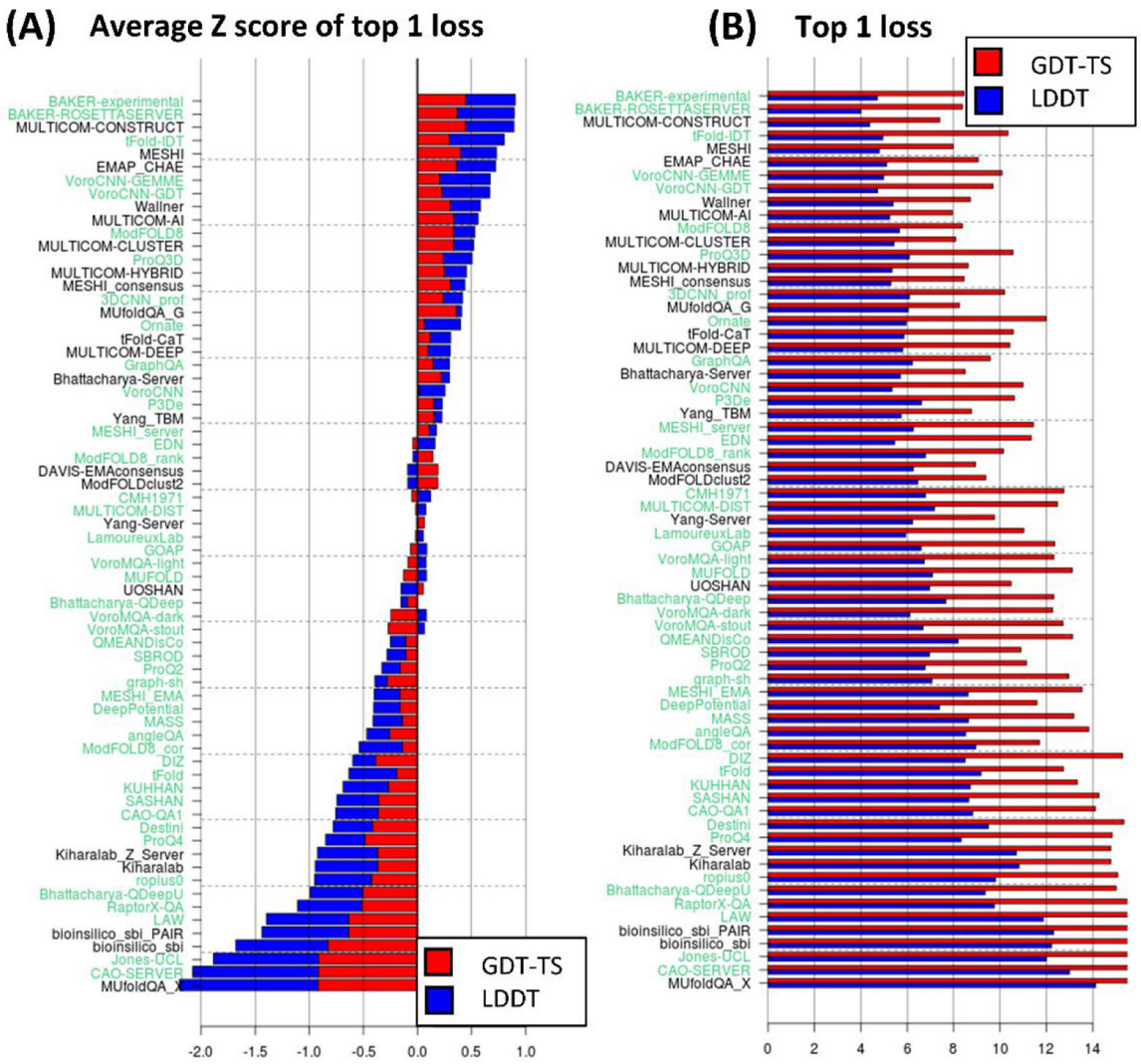

Figure 1.

Ranking of the methods in global accuracy estimation in terms of the top 1 loss. (A) Sum of average Z-scores of the top 1 GDT-TS and LDDT losses. Single-model methods are green and multi-model methods are black. (B) Average values of top 1 GDT-TS/LDDT loss are shown. Scores for GDT-TS are shown in red and those for LDDT in blue.

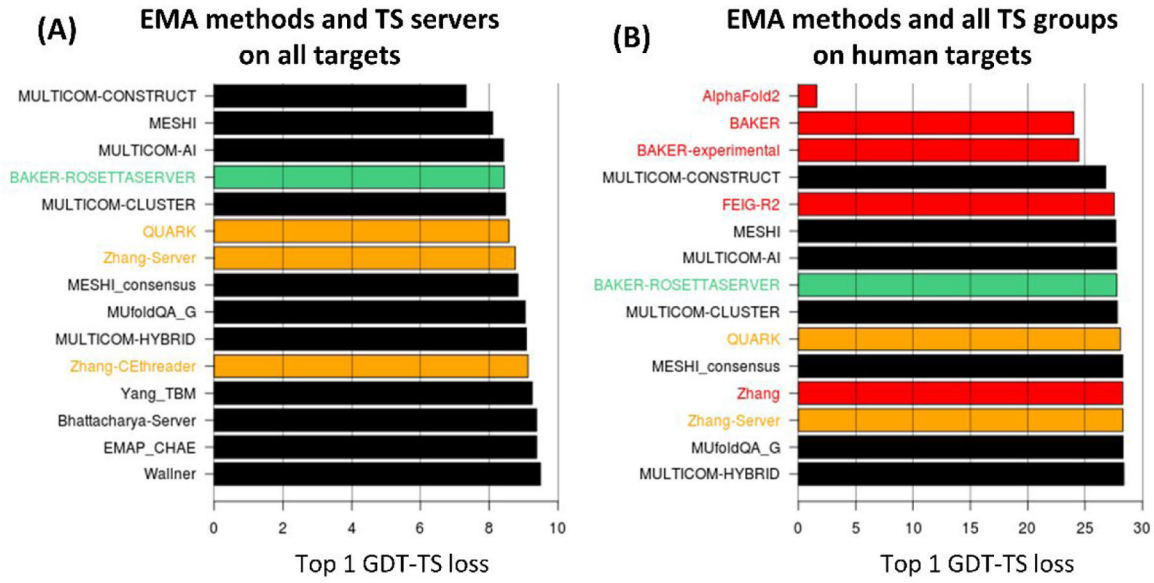

Figure 2.

Top 15 performers among EMA and TS methods based on the top1 GDT-TS loss. EMA methods are represented by models picked from the sets of 150 server TS models released in the second stage of the EMA experiment (see Methods). (A) Comparison with TS server groups on all targets, including “All group” targets (a.k.a. ‘human’) and “Server only” targets (see https://predictioncenter.org/casp14/targetlist.cgi). (B) Comparison with all TS groups (both server and human) on the subset of ‘human’ targets. Single-model EMA methods are colored in green, consensus EMA methods in black, TS servers in orange, and TS human groups in red.

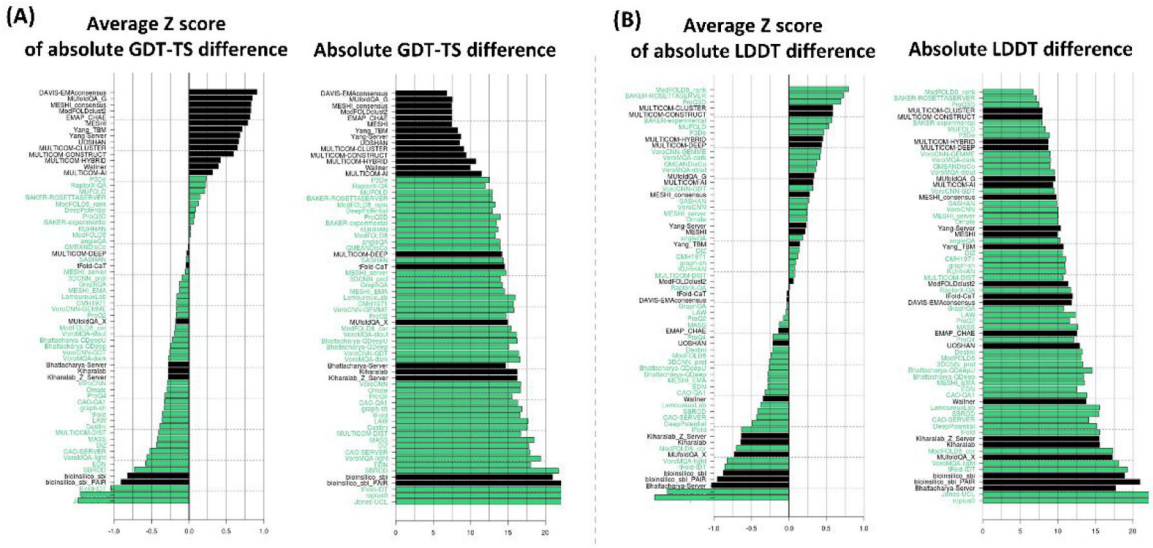

Figure 3.

Ranking of the methods in absolute accuracy estimation. The average Z-score of GDT-TS error is shown for each group along with average values of the absolute GDT-TS difference in (A), and the corresponding data for LDDT are presented in (B). Single-model methods are shown in green and multi-model methods in black.

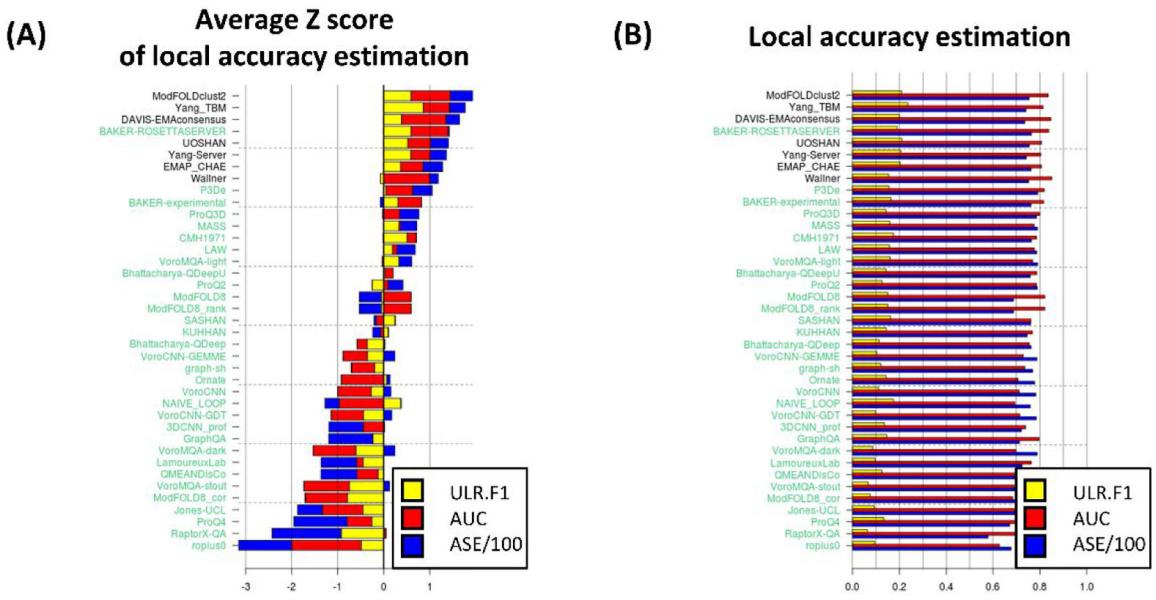

Figure 4.

Ranking of the methods for local accuracy estimation. (A) Sum of average Z-scores for ULR (yellow), AUC (red), and ASE (blue) used to rank the methods is shown for each group, with single-model methods listed in green and multi-model methods in black. (B) Average values of the individual measures.

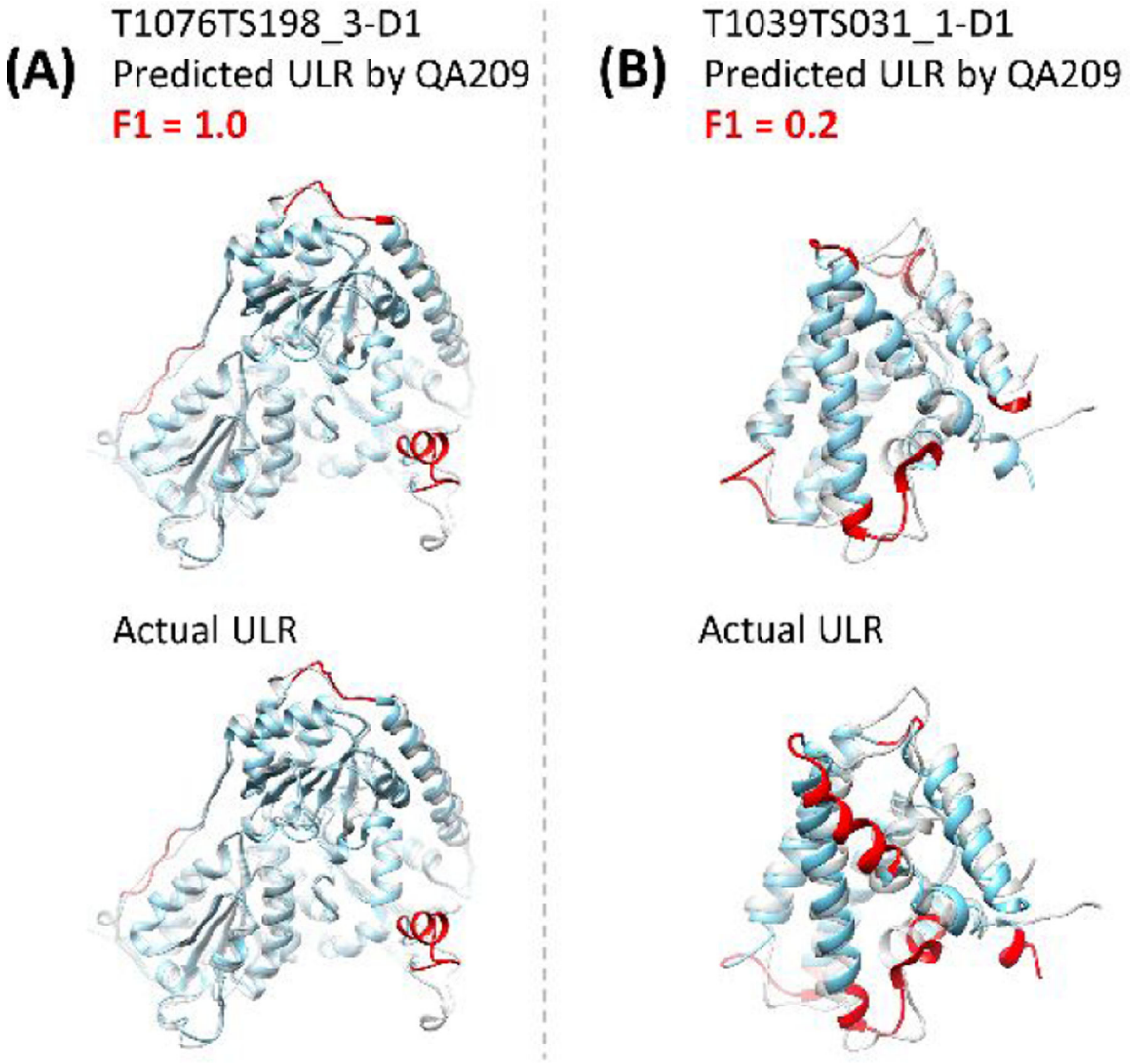

Figure 5.

Two examples of ULR prediction results and the corresponding actual ULR. Crystal structures are colored in gray, model structures in sky blue, and predicted and actual ULR regions in red.

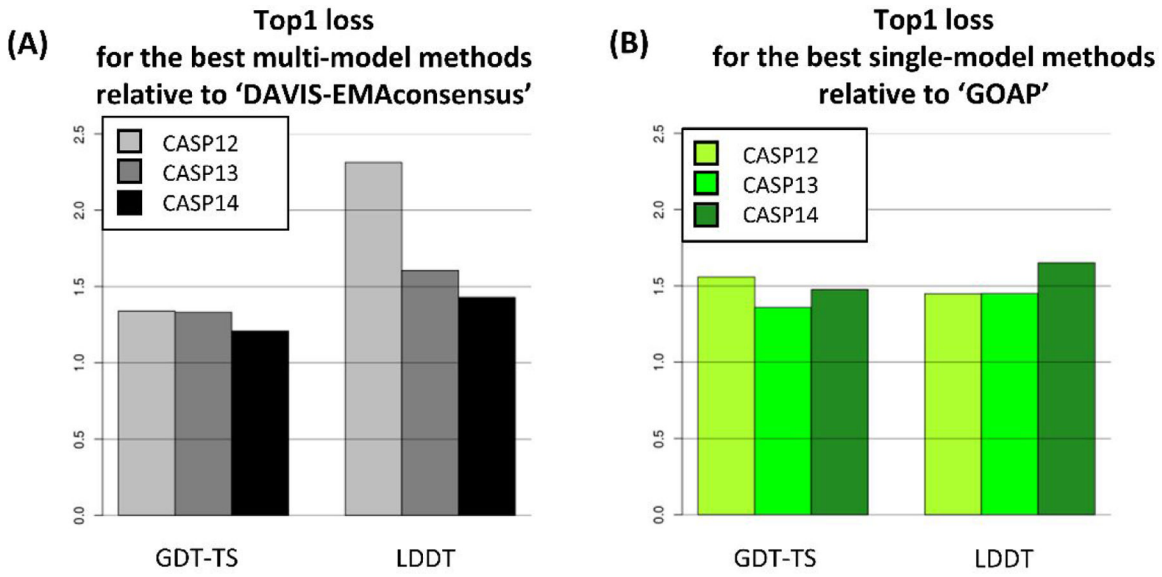

Figure 6.

Performance comparison of EMA methods relative to the reference methods over the last three CASPs. (A) Ratio of top 1 loss of "DAVIS-EMAconsensus" to that of the best consensus EMA method. (B) Ratio of top 1 loss of "GOAP" to that of the best single-model method.