Title

Assessment of model accuracy estimations in CASP12

Journal

Proteins Structure Function and Bioinformatics, 86(S1)

ISSN

0887-3585

Authors

Kryshtafovych, Andriy

Monastyrskyy, Bohdan

Fidelis, Krzysztof

et al.

Publication Date

2018-03-01

DOI

10.1002/prot.25371

Peer reviewed

HHS Public Access

Author manuscript

Proteins. Author manuscript; available in PMC 2019 March 01.

Published in final edited form as:

Proteins. 2018 March ; 86(Suppl 1): 345–360. doi:10.1002/prot.25371.

Assessment of model accuracy estimations in CASP12

Andriy Kryshtafovych[1,*], **Bohdan Monastyrskyy**[1], **Krzysztof Fidelis**[1], **Torsten Schwede**[2,3], and **Anna Tramontano**[4]

[1]Genome Center, University of California, Davis, USA [2]Biozentrum, University of Basel, Switzerland [3]SIB Swiss Institute of Bioinformatics, Basel, Switzerland [4]Department of Physics, Sapienza University of Rome, Italy

Abstract

The record high 42 model accuracy estimation methods were tested in CASP12. The paper presents results of the assessment of these methods in the whole-model and per-residue accuracy modes. Scores from four different model evaluation packages were used as the 'ground truth' for assessing accuracy of methods' estimates. They include a rigid-body score - GDT_TS, and three local-structure based scores - LDDT, CAD and SphereGrinder. The ability of methods to identify best models from among several available, predict model's absolute accuracy score, distinguish between good and bad models, predict accuracy of the coordinate error self-estimates, and discriminate between reliable and unreliable regions in the models was assessed. Single-model methods advanced to the point where they are better than clustering methods in picking the best models from decoy sets. On the other hand, consensus methods, taking advantage of the availability of large number of models for the same target protein, are still better in distinguishing between good and bad models and predicting local accuracy of models. The best accuracy estimation methods were shown to perform better with respect to the frozen in time reference clustering method and the results of the best method in the corresponding class of methods from the previous CASP. Top performing single-model methods were shown to do better than all but three CASP12 tertiary structure predictors when evaluated as model selectors.

Keywords

CASP; EMA; QA; estimation of model accuracy; model quality assessment; protein structure modeling; protein structure prediction

INTRODUCTION

It has been ten years since CASP started evaluating model accuracy estimation[1–5]. The introduction of the EMA (a.k.a. QA) prediction category into CASP in 2006 instigated the development of the accuracy estimation methods: while there were no EMA-dedicated papers before then, around one hundred papers have been published since, including more

[*]To whom correspondence should be addressed: Andriy Kryshtafovych, Genome Center, University of California, Davis, 451 Health Sciences Dr., Davis, CA 95616, USA, akryshtafovych@ucdavis.edu, Tel/Fax: +1 5307548977.

than a dozen in the last year alone (2016). The ever growing number of the EMA participants in the CASP experiments reflected a high scientific interest in this problem: the latest, twelfth round of CASP tested 42 methods, five more than in the previous round. In addition, the CAMEO experiment evaluates performance of the automatic estimators of model accuracy on the continuous basis with currently 12 public and development servers participating (http://cameo3d.org/quality-estimation/).

We present an evaluation of model accuracy estimation in the twelfth round of CASP and assess progress in the field. A brief description of the best participating methods is provided in a separate paper in this issue. The overall conclusions of the paper are expected to be of interest to both, specialists developing the methods and researches using models of protein structure in their studies.

MATERIALS AND METHODS

Targets and predictions

Eighty two unique protein sequences were released as prediction targets in the latest CASP experiment. Eleven targets were canceled by the organizers and tertiary structure assessors due to the lack of reference structure at the time of the assessment (see CASP12 domain definition paper [Dal Peraro et al. - THIS ISSUE]); one additional target - T0865 - was canceled due to its inappropriateness for the assessment of monomeric predictions. The remaining 70 targets were evaluated.

7400 EMA predictions were submitted in CASP12, including 6095 predictions on targets that were evaluated. The latter included accuracy estimates for 11052 tertiary structure server models. All submitted QA predictions are accessible through http://predictioncenter.org/download_area/CASP12/predictions. Only groups submitting correctly formatted predictions on at least half of the evaluated targets were included in the analysis in this paper.

Testing procedure and prediction format

As in previous few CASPs, a two-stage target release procedure was applied. In the first stage, twenty tertiary structure models (out of more than 200 server models typically submitted on a target) were selected by the organizers and released to the EMA predictors. These models were selected to span the whole range of models accuracy for the target – from worst to best according to the estimates of the in-house Davis-EMAconsensus method [1]. Only after the QA predictions in the first stage had been collected, we released the top 150 server models in the second stage.

In both stages, predictors were asked to estimate the global and local accuracy of the provided models. In the global assessment mode (QAglob), each model has to be assigned a score between 0 and 1 reflecting accuracy of the model (the higher the score the better). In the local assessment mode (QAloc), each residue in the model has to be assigned an estimated distance error in Ångströms as would have been seen for that residue in the optimal model-target superposition. Details of the submission format and an example target

release timeframe can be found at the Prediction Center web page http://predictioncenter.org/casp12/index.cgi?page=format#QA.

Evaluation principles and measures

Accuracy estimation methods were evaluated separately in the global and local assessment modes. While the global model accuracy scores provide estimates of overall model quality, the local accuracy scores (per-residue scores) offer estimates of the correctness of model's local structure and geometry and can help recognizing well- and poorly- modeled regions.

42 groups were assessed in the QAglob mode and 24 groups in the QAloc mode. The main attention was paid to the analysis of the results submitted in the second stage of the prediction (best150 datasets). In the previous two CASPs we also assessed per-residue estimation of coordinate errors in the predictors' own structural models, but in this CASP this aspect of prediction was evaluated as an integral part of the tertiary structure assessment (see, e.g. the TBM assessment paper [THIS ISSUE]).

Global evaluation measures—Since global model accuracy estimates are submitted for whole models (and not domains), evaluation of the results is also carried out at the whole model level (differently from the tertiary structure prediction, which is evaluated at the level of domains.

More than a dozen measures are used in CASP to evaluate similarity of a model to the target, and each of these measures can be considered as a target function for model accuracy assessment. From CASP7 (2006) through CASP10 (2012), the measure of choice to assess EMA predictions was the GDT_TS measure[6]. Already in CASP11, we expanded the number of reference measures, adding three non-rigid-body based measures - LDDT[7], CAD[8] and SphereGrinder[9] to the evaluation tool chest. This way we were able to assess prediction results from different perspectives, recognizing the ability of EMA methods to not only properly estimate accuracy of the backbone, but also identify models with better local geometry or local structure context. In this CASP we again used all these measures in the evaluation pipeline. For the sake of uniformity, we multiplied LDDT and CAD scores by 100, this way having all the evaluation scores in the 0–100 range.

Local evaluation measures—Local model accuracy estimates (atom distance errors) are submitted for each residue separately and this allows us to carry out evaluation at both, the whole-target and domain levels. For single-domain targets, the results from both evaluation modes are identical. For multi-domain targets, the whole-target evaluation gives an extra credit to methods capable of correct identifying relative orientation of the constituent domains, while the domain-level evaluation gives advantage to methods being more accurate in prediction of the within-domain distance errors.

To evaluate the accuracy of predicted per-residue error estimates, we employed the ASE measure[10]. For each residue, the distance d is normalized to the [0;1] range using the S-function[11]

$$S(d)=\frac{1}{1+\left(\frac{d}{d_0}\right)^2}$$

and then averaged over the whole evaluation unit (target or domain) and rescaled to the [0;100] range using the formula

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

where e_i is the estimated distance as submitted by predictors, d_i is the actual distance from the LGA superposition, d_0 is a scaling factor set here to 5. The higher the score, the more accurate the prediction of the distance errors in a model. If error estimates for some residues are not included in the prediction, they are set to a high value so the contribution of that specific error to the total score is negligible.

Comparison to the baseline method—To provide a baseline for assessing performance of the participating methods, we used an in-house developed Davis-EMAconsensus method that has not changed since its first implementation in CASP9 [3] (2010). During the latest four CASP experiments this method was run as an ordinary EMA predictor, alongside with other participating methods. Ratio between the scores of the reference method in different CASPs may indicate the change in difficulty of targets. The change in the relative scores of the best methods with respect to the baseline method may reflect performance changes associated with the development of methods and not the change in the databases or target difficulty.

Comparison to TS methods—Global scores generated by EMA methods can be used to pick five highest scoring models out of the 150 server models released to predictors on every target. This way, every CASP EMA method can be considered as a tertiary structure meta-predictor (selector) and ranked alongside the TS prediction methods.

To insert EMA methods into ranking tables for tertiary structure methods, we calculated their pseudo z-scores using the mean and standard deviation computed from the distribution of tertiary structure prediction scores. This way z-scores of TS models are intact and have the same values both in TS-only and joined TS+EMA rankings. Note that tertiary structure prediction methods are ranked differently depending on the model comparison environment (i.e. group types (server or expert), target subsets (all or human; TBM or FM), model types (model_1 or best-of-five)), and so are the EMA methods.

RESULTS AND DISCUSSION

1. Classification of methods

Not all methods are created equal with regard to their input. Some methods require multiple models, while others can generate accuracy estimates for a single model alone. The methods

that require multiple models are usually called clustering methods. They are designed to take advantage of the consensus information extracted from a set of input models. The methods that take just one model as an input are called single-model methods. From the algorithmic perspective, this class of methods encompasses two different subclasses. If a method can generate an accuracy estimate going no further than coordinates of the model itself, it is called a pure single-model method (or, simply, a single-model method). Pure single-model methods score models based on the geometric and energetic analysis of their coordinates and therefore sometimes are also called ab initio or physics-based methods. If a method, besides the coordinates of the model, relies on additional information from the evolutionarily related proteins or specially generated ensembles of structural models, it is regarded as a quasi-single method.

Since pure single-model methods are based exclusively on the coordinates of the assessed models, it is expected that the accuracy estimates they produce would be the same every time we apply a method to the same model. On contrary, results of clustering methods are expected to differ as they would normally depend on the size and composition of the datasets the model belongs to. Quasi-single-model methods may expose both behaviors depending on the details of the method implementation and target specifics.

To check if classification of methods provided by their authors complies with this rule, we compared results of the methods in two prediction stages by calculating the average absolute difference between the stage-specific accuracies of the common models:

$$\mathrm{diff\,(METHOD)}=\frac{1}{\mathrm{N}}\sum\nolimits_{\mathrm{n}\in \mathrm{S}_1\cap \mathrm{S}_2}|\mathrm{QAscore}_1(\mathrm{n})-\mathrm{QAscore}_2(\mathrm{n})|,$$

where S*i* is the dataset of all models (on all targets) released in the *i*-th stage of the experiment (i.e., S_1=sel20 and S_2=best150); $QAscore_i(n)$ is METHOD's accuracy estimate for model *n* in the *i*-th stage; $n=1,\ldots,N$ – common models from S_1 and S_2 datasets. The results of the comparison are shown in Figure 1. Top twelve methods generate identical accuracy estimates in both stages (*diff* = 0); next 14 methods exhibit very low *diff* values ranging from 0.003 to 0.02; all the remaining methods have *diff* > 0.03. From the visual inspection of the graph one can identify a place where consecutive *diff* values undergo a significant jump (>150%, from 0.02 to >0.03) suggesting a natural separation point (shown as a horizontal red line) between the methods generating very similar accuracy scores for the same models in both stages of the experiment and those that do not. The *diff*-based separation of methods corresponds very well to the methodology-based classification of methods provided in the Abstracts (http://predictioncenter.org/casp12/doc/CASP12_Abstracts.pdf), as all single-model methods (here and henceforth colored blue) place above the separation line and all clustering methods (colored black) but one – below it. The clustering method that finds itself in the typical non-clustering zone is MESHI_con server, whose methodology is apparently less dependent on the dataset composition than that of other clustering methods. Quasi-single methods (colored green) can be found on both sides of the separation line showing their different levels of dependency on the number and distribution of the models in the datasets. Three of these methods (the ModFOLD6 series of

methods, L. McGuffin's group) seem unaffected by model environment in CASP datasets, while another three are quite sensitive to the environment (the main methodological reasons for the difference in their scores are highlighted in bold in the description of these methods provided inside Figure 1).

2. Estimation of global accuracy of models (QAglob)

An a-priori estimate of the global accuracy of a model can serve as the first filter in determining the usefulness of the model to address a specific biomedical problem. In this section, we assess the effectiveness of EMA methods to assign overall accuracy score to a model by evaluating their ability to (1) find the best model amongst many others, (2) reproduce model-target similarity scores, and (3) discriminate between good and bad models. All four evaluation scores described in the Methods section are used as the "ground truth" measures in these analyses.

2.1. Identifying the best models—To assess the ability of methods to identify the best models from among several available, for each target we calculated the difference between the scores of the model predicted to be the best (i.e. that with the highest predicted EMA score) and the model with the highest similarity to the native structure. Measuring the difference in accuracy between the predicted best model and the actual best model makes sense only if the actual best model is of good quality itself. As in previous CASPs, we performed this analysis only on targets for which at least one model was of 'good enough' quality, defined as 40% of the selected measures' top score.

Figure 2 illustrates the accuracy of CASP12 methods in selecting the best models according to the GDT_TS score (panel A), and shows cumulative ranking of methods according to four evaluation measures (panel A). In both panels of the Figure, two single-model EMA methods (SVMQA and ProQ3) hold top positions on the ranking ladder, while another single-model method – MESHI-server, sits only a couple of steps below. The best method according to the difference in the GDT_TS scores – SVMQA – is capable of identifying the best models in the datasets with an average error of 5.0 GDT_TS. The other two above mentioned single-model methods (ProQ3 and MESHI-server) demonstrate just slightly worse average errors of 5.6 and 5.8 GDT_TS units, correspondingly. This suggests the progress in this area of accuracy estimation, as in CASP11 the best single-model method (ProQ2) was only 4th in the ranking with the average accuracy loss of 6.4 GDT_TS units (22% worse than the best CASP12 result). It is also interesting to note that the ProQ2 method on CASP12 targets showed a larger accuracy loss of ΔGDT_TS=7.3, thus implicitly suggesting a more difficult nature of CASP12 targets for the accuracy assessment.

The data used to generate the bar plot in panel (B) show that the best two groups according to the GDT_TS score are also among the best according to other evaluation measures. It is also remarkable to see that nine out of twelve best performing groups are single-model methods and that all of these methods are ranked much higher than the reference Davis-EMAconsensus method.

To establish the statistical significance of the differences in performance we performed two-tailed paired t-tests on the common sets of predicted targets and models for each evaluation

measure separately. Table 1 summarizes the results of the t-tests for top10 groups from the cumulative ranking graph. The table shows that top 7 methods are statistically indistinguishable from each other, thus suggesting using any of them for practical applications.

In complement to the accuracy loss analysis (above), we carried out the recognition rate analysis, showing the success and failure rates of CASP12 EMA methods in identifying the best models. We assume that a method succeeds if the difference in scores between the best EMA model and the actual best model is small (within 2 score units) and fails if the difference is larger than 10. Figure 3A shows the percentage of targets for which the models identified as the best were 0–2, 2–10 and >10 GDT_TS units away from the actual best models. The top-performing EMA method (SVMQA) demonstrates 40% success rate with 10% failure rate. Comparing top 10 groups in CASP12 and CASP11, we can conclude that CASP12 methods have somewhat better average success rate (35% vs 30% in CASP11) at the expense of somewhat worse failure rate (22% vs 16% in CASP11). Results according to all evaluation measures are summarized in Figure 3B. Since high success rate and low failure rate are the desired features of an EMA method, we used the difference between these rates as the criterion to examine methods' efficiency. Panel B of Figure 3 looks largely like the panel B of Figure 2 confirming that single-model methods hold leading positions in the selection of the best models, with two methods of this type (SVMQA and ProQ3) found at the top of these classifications.

2.2. Reproducing model-target similarity scores—To assess overall correctness of global model accuracy estimates, we calculated the absolute difference between the actual evaluation scores and the predicted accuracies for every server model included in the best150 datasets. Smaller average difference over all targets signifies better performance of a predictor.

Figure 4A shows the average absolute difference between the submitted estimates of model accuracy (EMA) and one of the evaluation scores, GDT_TS, for each participating group. In CASP12, the best method could predict GDT_TS score of models with an average error of slightly higher than 5 GDT_TS units, which is 17% better than in CASP11, where the best ΔGDT_TS=6.0. Yet, the main progress between the two latest rounds of CASP in this aspect of evaluation lies not in the improvement of scores of the best methods (which both happen to be clustering methods), but in the growing performance accuracy of non-clustering methods. While in CASP11 there was not a single non-clustering method with the average difference of <10 GDT_TS, there were four of such methods in CASP12. Also, the gap in results between the best non-clustering method and the best clustering method shrank substantially to only 1.6 ΔGDT_TS compared to 4.8 ΔGDT_TS in CASP11.

Results according to other evaluation measures proved to be even more favorable to non-clustering methods. Figure 4B shows that top three places in the summary table are occupied by quasi-single methods from the ModFOLD6 family of methods, and places 5–7 – by pure single-model methods. While in CASP11 top seven positions in the cumulative ranking table were all occupied by clustering methods and these methods were well separated from the following non-clustering methods, in CASP12 only one clustering method (Pcomb-domain)

is among the top 7 groups (see Figure 4B) and practically all of the top performing methods are statistically indistinguishable from each other (see Table 2).

Summarizing these results, we can suggest that if a user wants to estimate overall correctness of a single model, quasi-single methods from the ModFOLD6 series might be the best choice. Since the best pure single-model methods (two variants of ProQ3 and Multicom-cluster) lag not that far behind, they could be the next best bet. If many models need to be estimated for their accuracy, best clustering methods can produce the results statistically similar to the best quasi-single methods.

2.3. Distinguishing between good and bad models—To assess the ability of methods to discriminate between good and bad models, we pulled together models for all targets and then carried out a Receiver Operating Characteristic (ROC) analysis using Measure=50 threshold to separate good and bad models. The area under the ROC curve (AUC) was used as a measure of the methods' accuracy.

Figure 5A shows ROC curves for the best 10 groups based on the GDT_TS scores. Since the ROC curves look very similar, we employed the DeLong tests[2] to establish statistical significance of the differences in the group performance. The results indicate that the top two methods - Wallner and QASproCL, are statistically similar to each other and better than the rest of the methods except for ModFOLDclust2.

We also built the ROC curves and performed the DeLong tests on other evaluation measures. Figure 5B provides a summary in terms of the cumulative ranking calculated on the areas under the ROC curves. In general, out of top 10 performing methods, eight are clustering and two – quasi-single methods. Therefore, when many models are available for a user and they need to be partitioned into two classes (good/bad), clustering methods (e.g., Wallner, Pcomb-domain, QASproCL) are particularly good and quasi-single methods (e.g., ModFOLD6_rank) are the next best choice. The best performing method is the Wallner method, which is proven to be statistically better than other ones but Pcomb-domain in separating good and bad models (see Table 3).

3. Estimation of local accuracy of models (QAloc)

The effectiveness of 24 CASP12 local model accuracy estimators is evaluated by verifying how well these methods (1) assign correct distance errors at the residue level, and (2) discriminate between reliable and unreliable regions in a model. Both analyses are carried out on the per-residue estimates submitted for all models and all targets.

3.1. Assigning residue error estimates—The accuracy of predicted per-residue error estimates was evaluated with the ASE measure (see Methods for description), which assesses how far away are the submitted error estimates from the actual errors defined as distances between the corresponding residues in the optimal LGA model-target superposition. Figure 6 shows average ASE scores for all participating methods in the whole-model (panel A) and domain (panel B) evaluation. Clustering methods are dominant in both modes of the analysis, with the best methods reaching an average ASE score of over 80%. Results in both evaluation modes are very similar, with single-model methods

deviating by 0.0% ASE between the whole-model and domain-based ASE scores (on average), quasi-single methods deviating by 1.5% and clustering methods by 2.9%. The statistical significance tests reveal that two groups - ModFOLDclust2 and Pcons are statistically indistinguishable between themselves and significantly better than all other participating groups (see Table 4 for whole-target tests and Table S1 in Supplementary for domain-based tests).

3.2. Discriminating between good and bad regions in the model—To evaluate how well CASP predictors can discriminate between accurately and poorly modeled regions in the model, we carried out the ROC analysis on the submitted distance errors setting the threshold for correct positioning of a residue at the 3.8Å level. The results of this analysis are very similar to the results reported in the previous section with three clustering methods – ModFOLDclust2, Wallner and Pcons – being on top of the ranking lists (Figure 7). Quasi-single methods follow the best clustering methods, while single model methods demonstrate a weaker differentiation power.

4. Progress in the accuracy estimation results

In this section we make an attempt to measure progress in the accuracy estimation field by comparing the results of the best methods and the frozen in time baseline Davis-EMAconsensus method.

To start, we want to note that the data from the latest two CASPs suggests that targets in CASP12 may have been more challenging for accuracy assessment than those in CASP11 as scores of the reference method dropped in all components of the analysis performed above.

Figure 8 shows relative scores of the best performing methods (overall and single-model) normalized by the scores of the reference Davis-EMAconsensus method in CASP12 and CASP11. Cases where CASP12 bars (darker colors with outline) are higher than the corresponding CASP11 bars (lighter colors, no outline) may indicate methodological improvements. It is easy to notice that overall best predictors (red bars) show better results in CASP12 than in CASP11 in assigning absolute accuracy estimates to a model and distinguishing between good and bad models. Results in other categories are practically the same. At the same time, the graph suggests a substantial progress in performance of single-model methods (blue bars) in all aspects of the analysis. Visually, a considerable progress can be noticed in the estimation of global model accuracy (the first three scores), while in the local accuracy assessment (last two) the dark blue bars are only slightly higher than the light blue ones. Nevertheless, it should be mentioned that in the local analysis even smaller differences in the scores may result in statistically significant differences in the group performance, as there we process millions of observation points (residues) compared to a few orders of magnitude lower number of observation points (models) in the global analysis.

5. Ranking accuracy assessment methods alongside the tertiary structure prediction methods

Since every CASP EMA method can be considered as a tertiary structure meta-predictor (see Methods), we thought it interesting to compare the accuracy of the models selected by

these methods with the accuracy of models submitted by the TS prediction groups. We compared the TS models designated as first vs EMA models with the highest accuracy estimates scores, and also the highest scoring models out of the five submitted (TS) vs five selected (EMA). Figure 9 shows the joined GDT_TS-based rankings of the top12 methods alongside with the upper bound for the EMA-based methods in CASP12 (META-ideal).

For first models, in each target category (TBM; FM+FM/TBM; all) there are only three expert methods (from the list of four groups - Baker, Zhang, LEE and LEEab) that can outperform the best EMA meta-selectors (panel A), and no server predictor that can do that (panel B)! A single-model SVMQA method is the best method among server predictors in all target difficulty categories outscoring the best tertiary structure prediction server, Zhang-server, by 7% on easier targets and 13% on harder targets (panel B). Among all groups, the SVMQA ranks 4th on difficult targets (trailing the Baker group by 27%) and 7th on easier targets (trailing the Lee group by only 8%) (panel A).

For the best-out-of-five models (panels C and D), VoroMQAsr, ProQ3 and qSVMQA methods are consistently among the top12 groups for different target sets and difficulty categories. Interestingly, no TS prediction servers made it to the joined TS+EMA top12 list (panel D).

The results of the perfect model selector, META-ideal, show that such a method would have outscored all CASP12 methods, including the best human-expert ones, in all target difficulty categories. Thus, the optimal model selection in CASP can more than offset the advantage of modern human-expert methods over automatic server methods. The advantage of the perfect selector is most pronounced for difficult targets (the rightmost set of bar plots in every panel from A to D), where the set of overall best server models from all server groups outscores sets of the best-of-five models from individual groups and the EMA-based predictors by 25% in the server-only analysis (panel D) and 14% in the all-group analysis (panel C).

The results of the EMA selectors are equally impressive according to other evaluation measures. Figure 10 summarizes the comparison of EMA-based methods and TS methods according to all evaluation measures. Similarly to the GDT_TS-based analysis, the single-model EMA meta-predictors lead cumulative rankings among server methods (SVMQA on the harder targets and overall, ProQ3 on the easier targets - see panel B) and are high in the all group ranking (SVMQA is 3rd on the harder targets and overall, while ProQ3 is 4th on the easier targets – panel A). The ideal meta-predictor also holds the lead in the cumulative multi-measure ranking (data not shown).

6. Analysis of the results from the target perspective

While previous sections discuss the results from the group-centric perspective, here we analyze the data from the target perspective.

Starting with statistics on the per-target model scores and accuracy estimates, Figure 11 shows boxplots of five distributions: model-target GDT_TS scores (accuracy of models), model-model pairwise GDT_TS scores (similarity of models), and the estimates of model accuracy for different types of EMA methods. The distribution of model-target GDT_TS

scores (panel A) is the most similar to the distribution of EMA scores of clustering methods (panel C - Pearson's correlation coefficient of 0.97 for medians and 0.91 for interquartile ranges) and the least similar to that of single-model methods (panel E - correlation of 0.92 for medians and 0.47 for IQRs). Visual comparison shows that there are no targets that were easy to predict (panel A), but which received low EMA scores (panels C–E), and vice versa. The biggest difference is for the easiest target T0867, for which half models scored above GDT_TS=95.9 (panel A), while EMA median for the single-model methods was only at GDT_TS=67.5 (panel E). In general, single-model methods show the lowest variability in the median and spread of the EMA scores (panel E). Due to this, they are less competitive in predicting absolute accuracy of models than clustering methods (panel C), which follow the trend in panel (A) more closely (see also analysis of Figure 12 below). Usually the targets with a wider spread of model accuracy (e.g. T0867, T0860, T0895 or T0948) have also a wider spread of accuracy estimates. The most broad distribution is for the pairwise model-model GDT_TS scores (panel B), but one should remember that this distribution contains many more data points (>22,000) than the other ones (150). The multi-domain targets (designated with letter 'M' next to the target name) are harder to predict, and they concentrate at the harder end (right) of the target difficulty spectrum.

Figure 12 shows the difference between the predicted and actual accuracy scores, |EMA-GDT_TS|, as a function of (A) target difficulty (represented by the median GDT_TS score of the submitted models) and (B) similarity of models (represented by the interquartile width of inter-model GDT_TS scores). Both panels confirm that clustering methods (black) are better than single (blue) and quasi-single methods (green) in predicting absolute model accuracy scores. Also, both panels show that clustering methods are insensitive to target difficulty and spread of models submitted on the target, as their trend lines stay almost flat. On the contrary, single and quasi-single methods predict absolute accuracy of models better on targets with narrower spread of models (blue and green lines stay lower in the left hand side of panel B), and marginally better on easier for tertiary structure prediction targets (panel A). These conclusions are in accordance with the results of Figure 11 showing more similarity of data in panels 11A and 11B to those in panel 11C than to those in panels 11D and 11E. Targets T0862 and T0866 are examples of poor EMA performance by all types of methods. Both targets are difficult for TS prediction and have a relatively high diversity of models (see Figure 11A). The EMA methods did not recognize this diversity and assigned accuracy estimates in a much narrower interval. On the contrary, target T0867 is an example of an extraordinary good EMA prediction in the context of average deviations between the actual scores and accuracy estimates. High accuracy of models and quite wide spread of the submitted models (see Figure 11, first target from the left) were well reproduced in the EMA predictions.

Interestingly, while problematic for absolute accuracy estimation, target T0866 (see above) is an example of success for recognition of the best models. Figure 13 shows dependence of the loss in accuracy from the imperfect model selection on the separation between the best model and the distribution mean. Intuitively, targets where best models are further separated from the rest of the pack should be harder for identification of the best models, at least for clustering methods or by chance. This assumption is well supported in the data as all lines in Figure 13 demonstrate an upward trend. However, while single-model methods are just

marginally dependent on the best model's z-score, clustering and especially quasi-single methods show much stronger dependency (larger slope of the line). In general, single model methods are shown to perform better on this task than other types of methods as the blue line runs lower than the other two. This resonates well with the results of section 1 (Figures 2 and 3). All in all, for 41 out of 52 analyzed targets, a single-model method was either the absolute best or a tied best (blue color of the marker in the graph). Returning to the target T0866, we can assert that even though this target has one of the largest separations between the best model and the distribution mean (large z-score in the x-axis), the best EMA methods are able to perfectly identify the highest scoring model from among the 150 available. This is also the case for targets T0868, T0884 and T0885, all of which had extraordinarily good predictions for their level of difficulty. In opposition to these results, targets T0890, T0900 and T0942 appear to be hard for identification of the best model. Multi-domain nature of T0890 and T0942 likely played a role in their difficulty for EMA prediction and evaluation.

CONCLUSIONS

CASP12 witnessed yet another encouraging step forward in the development of accuracy assessment methods. There was a measurable progress in almost all areas of the assessment according to both, absolute evaluation scores and relative scores with respect to the baseline EMA method. Single-model methods excelled in picking the best models from decoy sets, with particularly impressive results demonstrated by the newly developed SVMQA and ProQ3 methods. Quasi-single approaches from the ModFOLD6 series of methods proved to be the best in estimating absolute scores of models. Consensus methods are still dominating in distinguishing between good and bad models (Wallner, Pcomb-domain), or reliably and unreliably predicted regions of models (ModFOLDclust2, Wallner, Pcons). These methods also hold the lead in predicting local accuracy of models; the quasi-single methods from the ModFOLD6 family are the next best choice. The top performing accuracy estimation methods were shown to outperform all but three CASP12 tertiary structure predictors (when evaluated as 'meta-predictors' selecting top five models per target from those submitted by the CASP servers), while a hypothetical perfect model selector would have outscored all participating methods. This shows that applying a good accuracy assessment method to a set of CASP server models, one can achieve performance better than that of any individual server predictor and rivaling state-of-the-art expert methods. Therefore, further development of the EMA methods can open additional predictive potential that has not been fully exploited yet.

It is worth mentioning that the top scoring EMA methods in many analyses performed here are the single-model methods. CASP has been continuously emphasizing the importance of the development of single-model methods, and it is encouraging to see a very strong response of the CASP community: the number of the developed methods more than quadrupled in four years (from only 5 in CASP10 to 22 in CASP12) and their accuracy matured to the point of practical applicability.

In a one-line summary, the paper shows abilities and limitations of modern accuracy estimation methods and convincingly asserts the advantages of their practical application to protein structure prediction.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This article is dedicated to the memory of Anna Tramontano, who unexpectedly passed away in March 2017. Anna was instrumental to establishing the evaluation of accuracy estimation predictions in CASP and was an assessor of five previous rounds of the EMA experiment.

This work was partially supported by the US National Institute of General Medical Sciences (NIGMS/NIH) – grant R01GM100482 to KF.

Abbreviations

EMA	Estimation of Model Accuracy
QAglob	global quality assessment
QAloc	local quality assessment
TS	Tertiary Structure
ROC	Receiver Operating Characteristic
GDT_TS	Global Distant Test – Total Score
IQR	inter-quartile range
TBM	template-based modeling (easier target for tertiary structure prediction)
FM	free modeling (difficult target for tertiary structure prediction)

References

1. 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.
2. 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]
3. Kryshtafovych A, Fidelis K, Tramontano A. Evaluation of model quality predictions in CASP9. Proteins. 2011; 79(Suppl 10):91–106. [PubMed: 21997462]
4. Cozzetto D, Kryshtafovych A, Tramontano A. Evaluation of CASP8 model quality predictions. Proteins. 2009; 77(Suppl 9):157–166. [PubMed: 19714774]
5. Cozzetto D, Kryshtafovych A, Ceriani M, Tramontano A. Assessment of predictions in the model quality assessment category. Proteins. 2007; 69(Suppl 8):175–183. [PubMed: 17680695]
6. Zemla A. LGA: A method for finding 3D similarities in protein structures. Nucleic Acids Res. 2003; 31(13):3370–3374. [PubMed: 12824330]
7. 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. 2013; 29(21):2722–2728. [PubMed: 23986568]
8. 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]

9. 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]
10. Kryshtafovych A, Monastyrskyy B, Fidelis K. CASP11 statistics and the prediction center evaluation system. Proteins. 2016; 84(Suppl 1):15–19. [PubMed: 26857434]
11. Gerstein M, Levitt M. Using iterative dynamic programming to obtain accurate pairwise and multiple alignments of protein structures. Proc Int Conf Intell Syst Mol Biol. 1996; 4:59–67. [PubMed: 8877505]

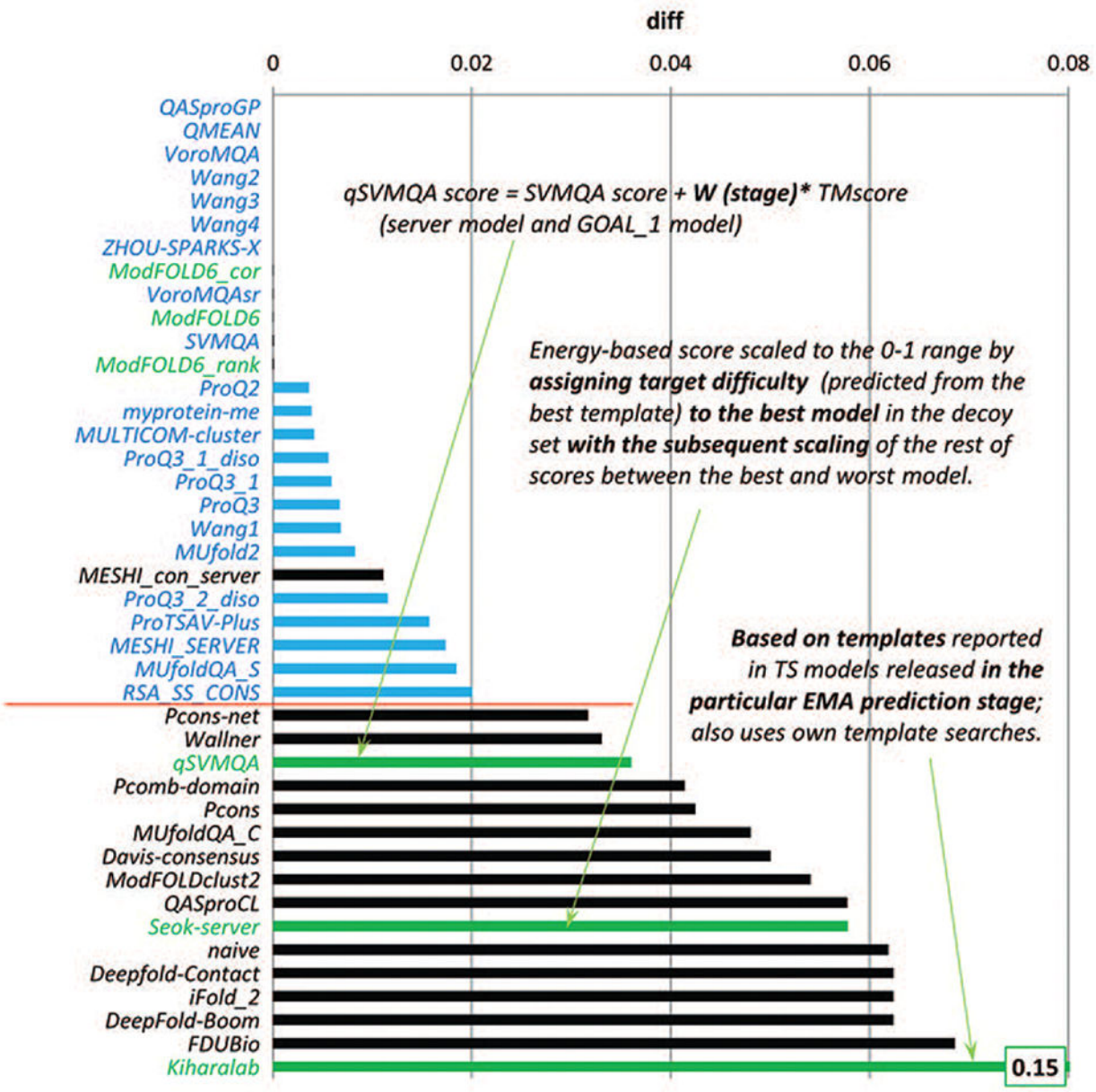

Figure 1.

Average difference in global accuracy estimates submitted by CASP12 predictors on the same models in two different stages of the EMA experiment. Groups are sorted by the increasing average absolute difference between the stage 1 and stage 2 scores. The red horizontal line (corresponding to a difference of 0.02) separates methods that generate approximately the same accuracy scores for the same models in both stages of the experiment (above) and those that do not (below). Single-model methods (blue) and clustering methods (black) are on different sides of the line. Quasi-single methods (green) can be found on both sides of the separation line.

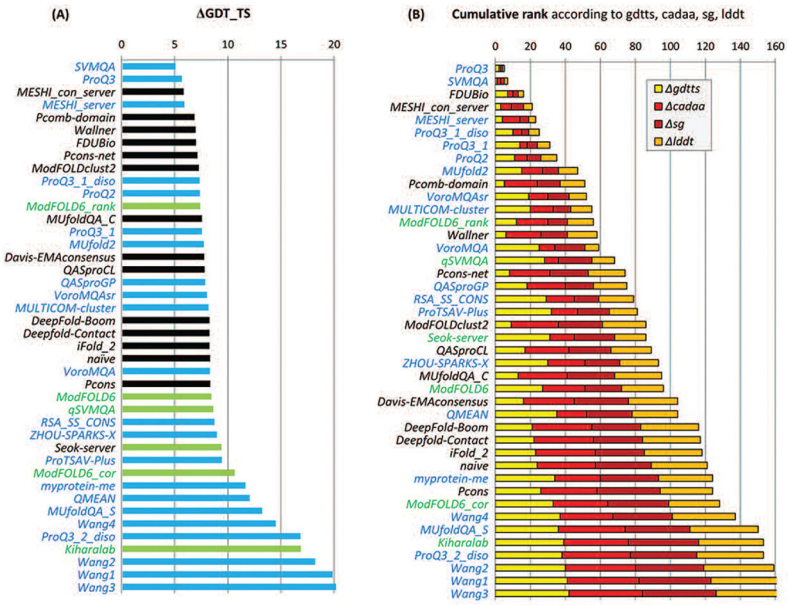

Figure 2.

Ability of CASP12 accuracy estimate methods to select the best model in decoy sets. (A) Average difference in accuracy between the models predicted to be the best and the actual best according to the GDT_TS score. For each group, the differences are averaged over all predicted targets for which at least one structural model had a GDT_TS score above 40. Clustering methods are in black, single-model methods in blue, and quasi-single model methods in green. Lower scores indicate better group performance. (B) A summary of the "best selector" results expressed as the cumulative ranking of the participating methods according to four evaluation scores –GDT_TS (yellow), CADaa (red), SphereGrinder (dark red) and LDDT (orange). Single-model methods are in leading roles with the ProQ3 and SVMQA ranked in the top two according to all evaluation measures.

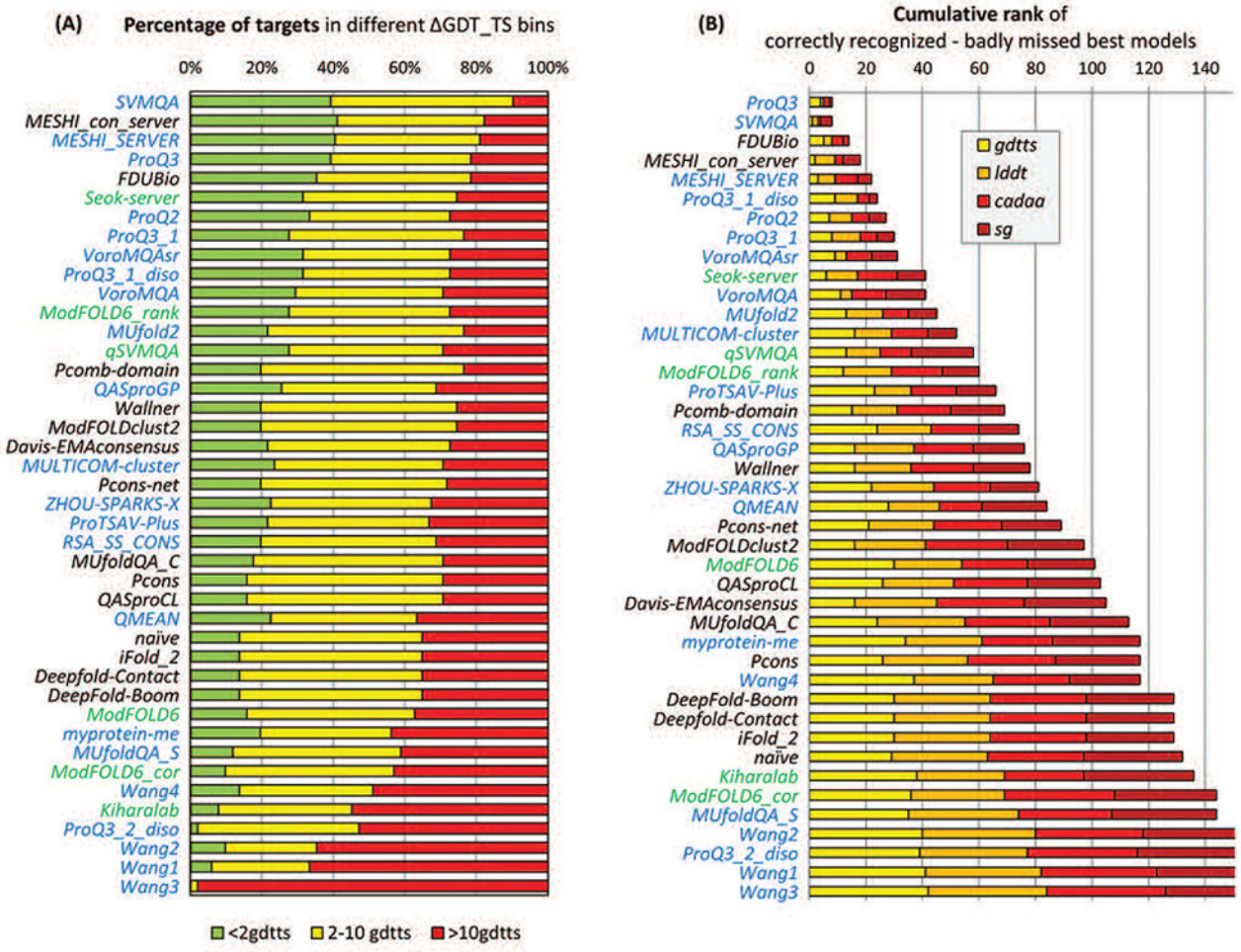

Figure 3.

Success rates of CASP12 methods in identifying best models. (A) The percentage of targets where the best EMA model is less than 2 (green bars), more than 2 and less than 10 (yellow), and more than 10 (red) GDT_TS units away from the actual best model. The percentages are calculated on targets for which at least one structural model had a GDT_TS score above 40. Groups are sorted by the difference between the rates of successful and failed predictions (green and red bars). Top performing groups can correctly identify the best models in approximately 40% of the test cases. (B) Cumulative ranking of the groups based on the differences between their success and failure rates calculated with GDT_TS, LDDT, CADaa, and SphereGrinder measures. Method coloring scheme is the same as in Figure 2.

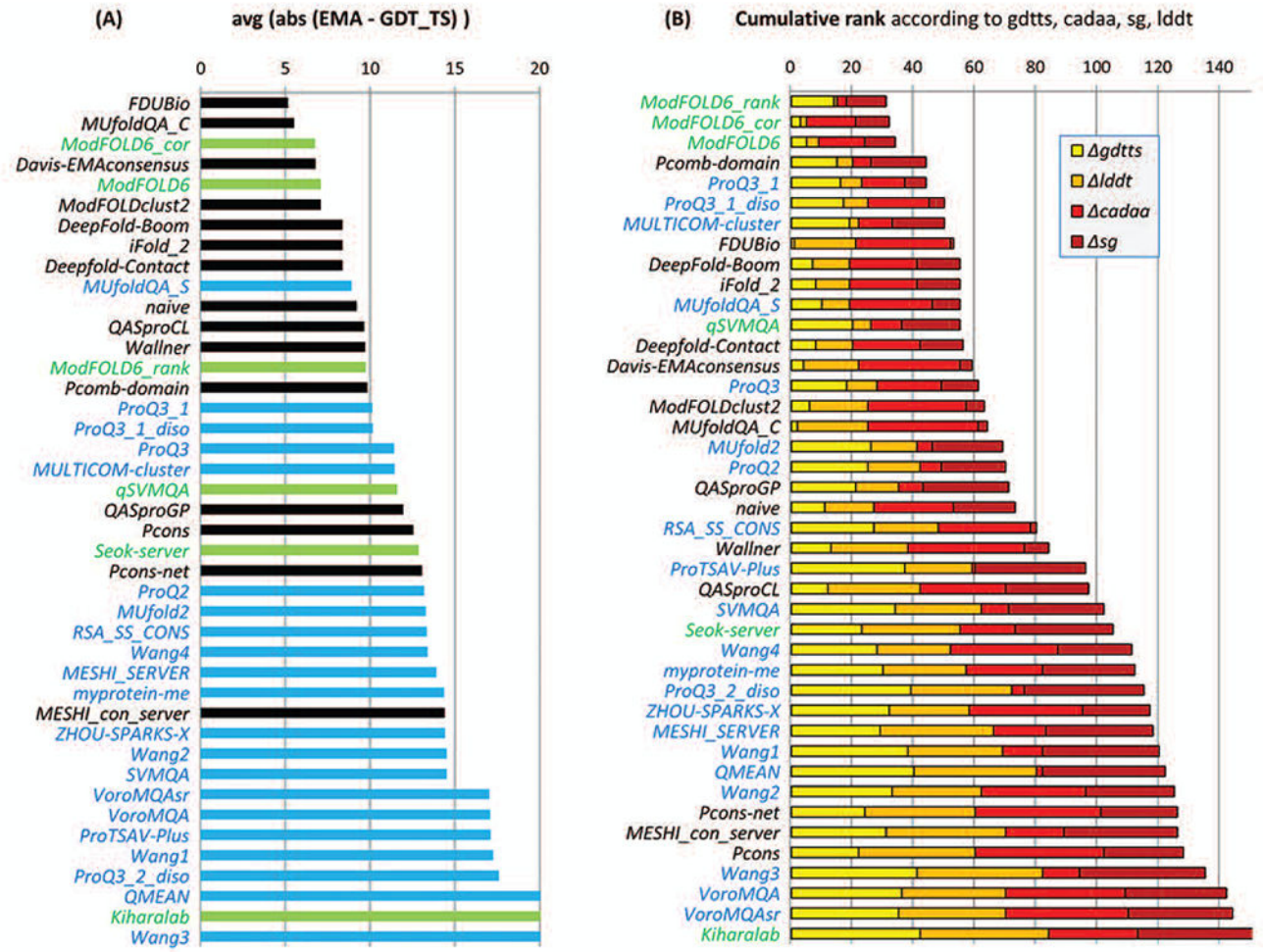

Figure 4.

(A) Accuracy estimates as compared to the GDT_TS scores of the assessed models. For each group, deviations are calculated for each model and then averaged over all predicted models. Group name colors in the plot distinguish different types of methods: clustering methods are in black, single-model in blue, and quasi-single in green. Lower scores indicate better group performance. The best performing methods are capable of predicting the absolute accuracy of models with an average per-target error of 5 GDT_TS. (B) Cumulative ranking of methods by the deviations of absolute accuracy estimates according to four evaluation measures - GDT_TS, LDDT, CADaa, and SphereGrinder. Method coloring scheme is the same as in Figure 2. Three quasi-single methods are leading the cumulative ranking.

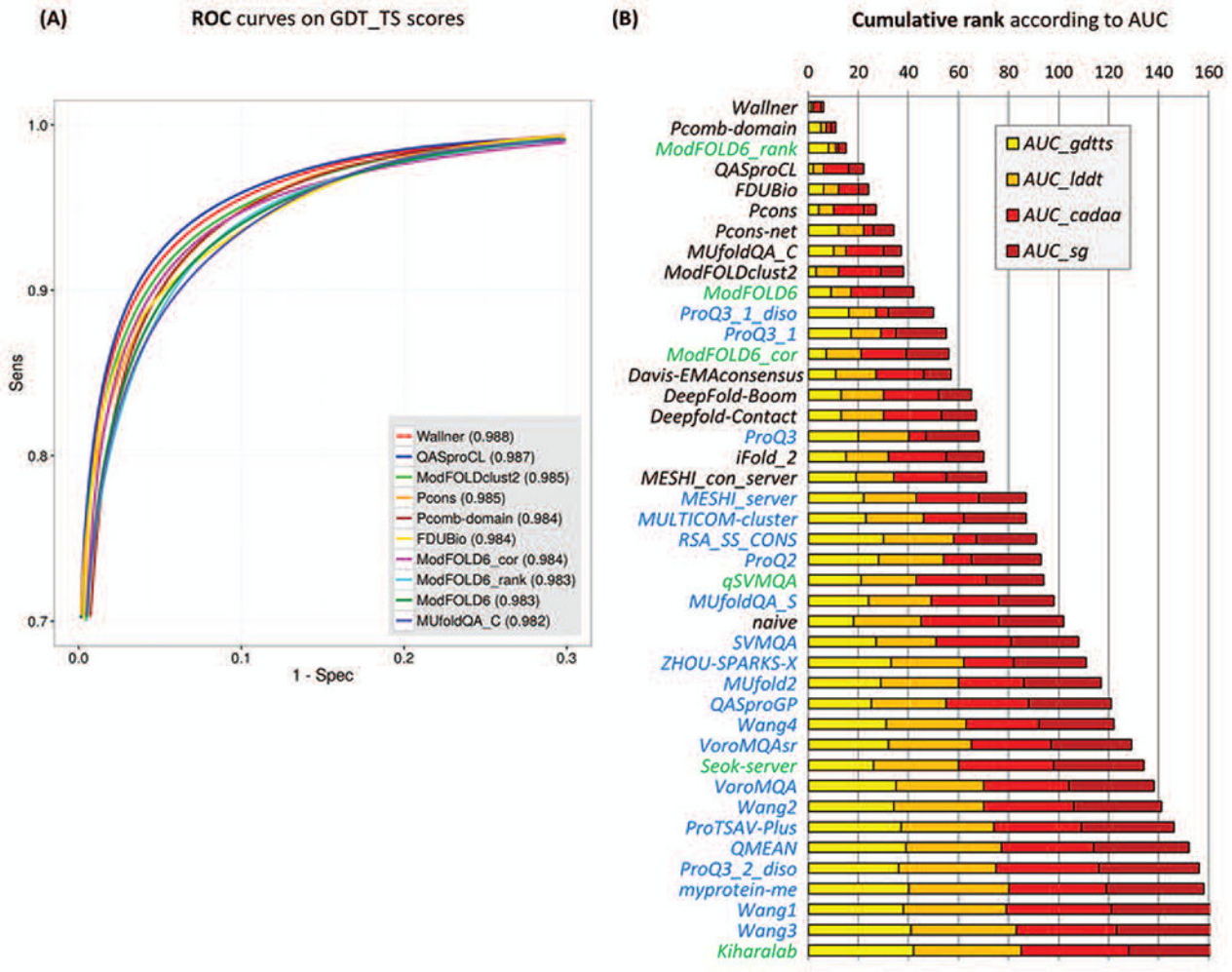

Figure 5.

Ability of methods to discriminate between good and bad models. (A) ROC curves for top 10 EMA groups on the GDT_TS data. The separation threshold between good and bad models is set to GDT_TS=50. Groups are ordered according to decreasing *AUC* score, which is provided in the legend after the group name. For clarity, only the left upper part of the ROC-curve graph is shown ($FPR \leq 0.3$, $TPR \geq 0.7$). (B) Cumulative ranking of groups based on the *AUC*s calculated on the GDT_TS, LDDT, CADaa and SphereGrinder data. Method coloring scheme is the same as in Figure 2. Clustering methods demonstrate dominance in this aspect of analysis.

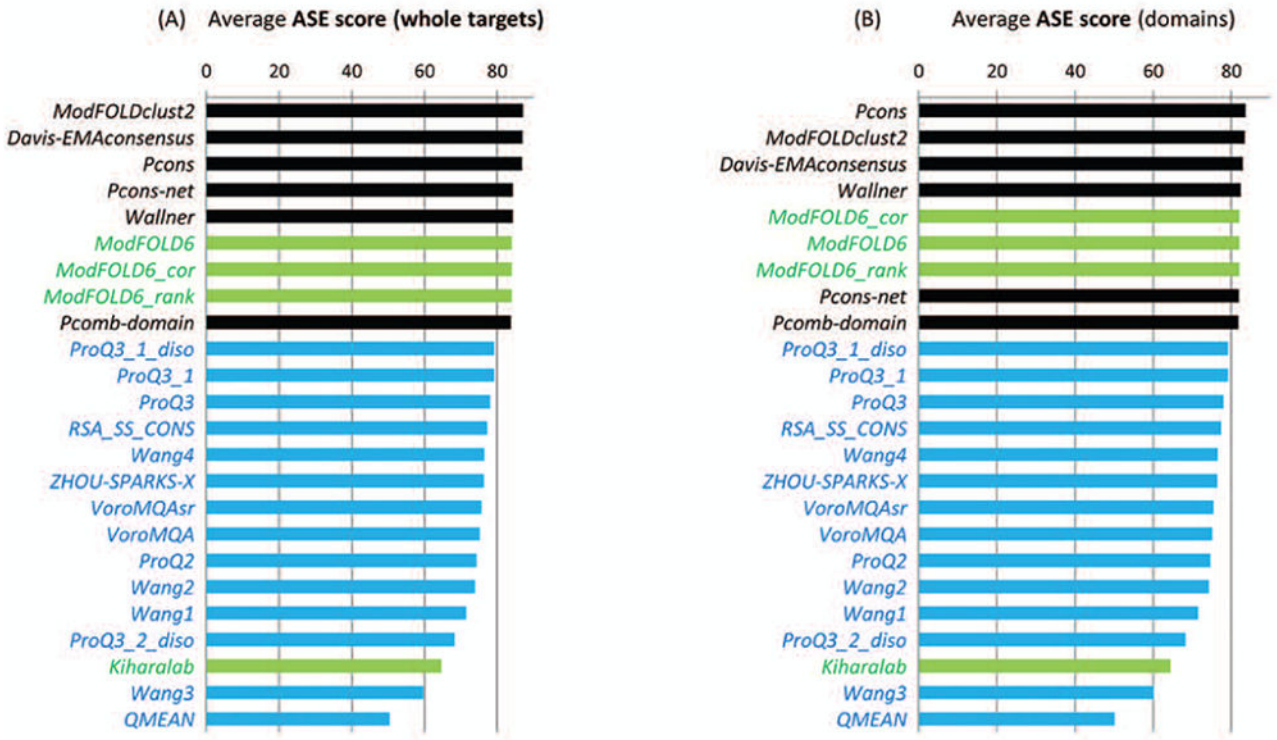

Figure 6.

Average ASE score calculated on (A) whole targets and (B) structural subdomains. Results in both evaluation modes are very similar, with the best methods exceeding ASE=80.

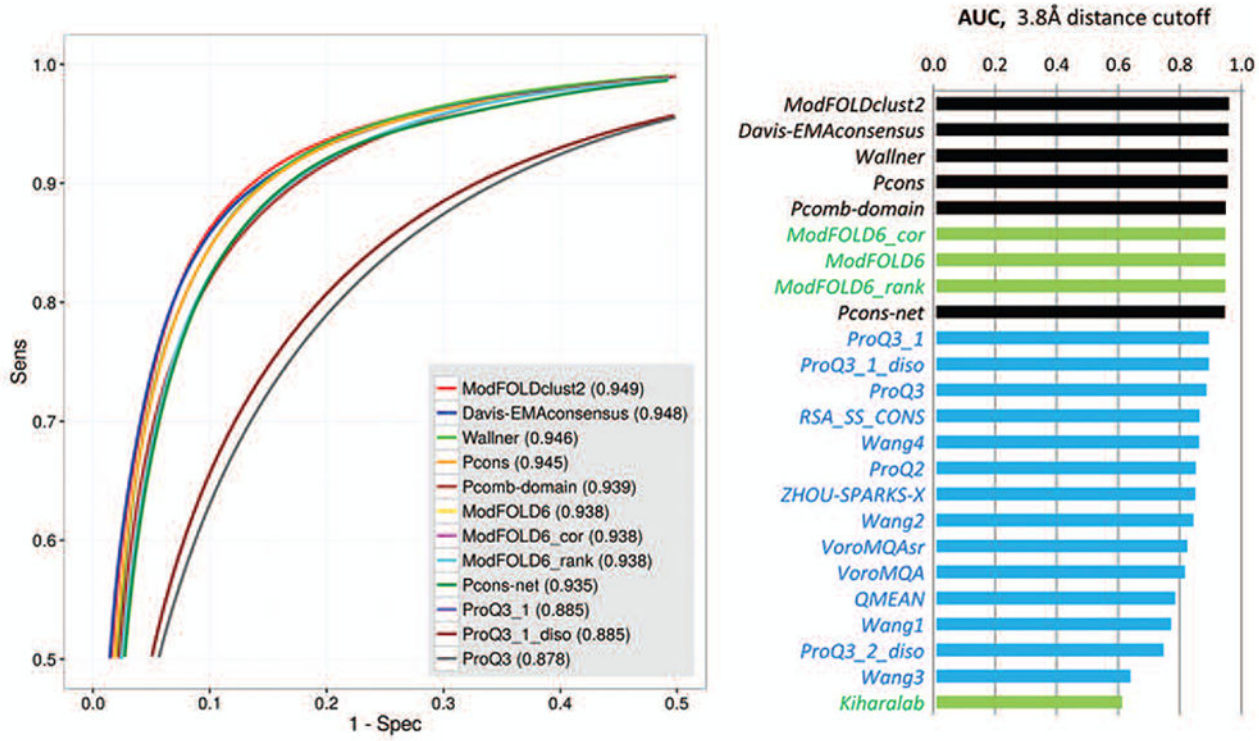

Figure 7.

Accuracy of the binary classifications of residues (reliable/unreliable) based on the results of the ROC analysis on whole targets. (A) ROC curves for top 12 EMA groups on the distance error data. A residue in a model is defined to be correct when its Cα is within 3.8Å from the corresponding residue in the target. Group names are ordered according to decreasing *AUC* scores, which are provided in the legend in parentheses. For clarity, only the left upper quadrant of a typical ROC-plot is shown (FPR≤0.5, TPR≥0.5). (B) AUC values for all participating groups. Clustering methods demonstrate better results, but cannot outperform the reference Davis-EMAconsensus method.

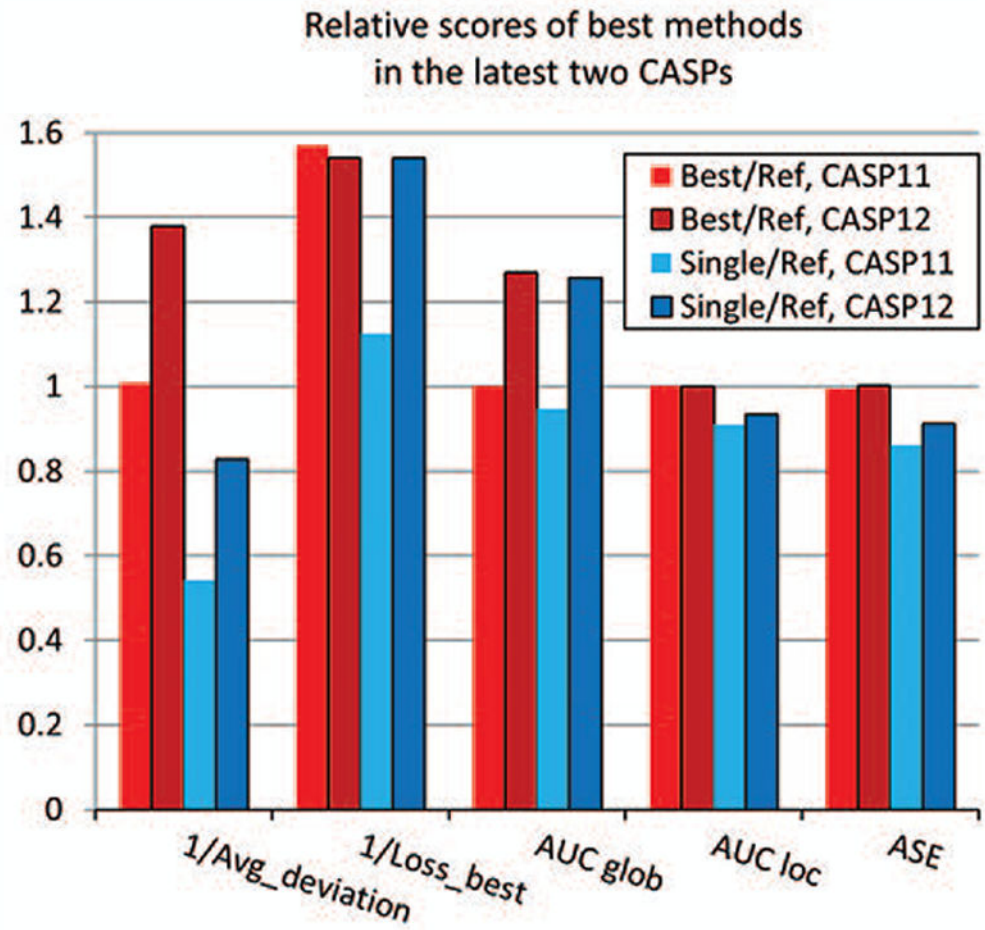

Figure 8.

Relative scores of the best overall methods (red) and single-model methods (blue) in CASP12 (dark color) and CASP11 (light color). The first three scores along the x-axis are based on comparison of the GDT_TS scores in the QAglob analysis (sections 2.1–2.3 in the text), the last two – on comparison of the distance errors in the QAloc analysis (sections 3.1.–3.2). For each of the five selected measures, the ratio between the score of the best participating method (overall or single-model) and the score of the Davis-EMAconsensus method is calculated. Two ratios - average deviation and loss from the best - are inverted so that higher bars in the graph always indicate a better result. Values above 1.0 mean that the best method outperforms the baseline method. Single-model methods in CASP12 demonstrate improved performance across the board.

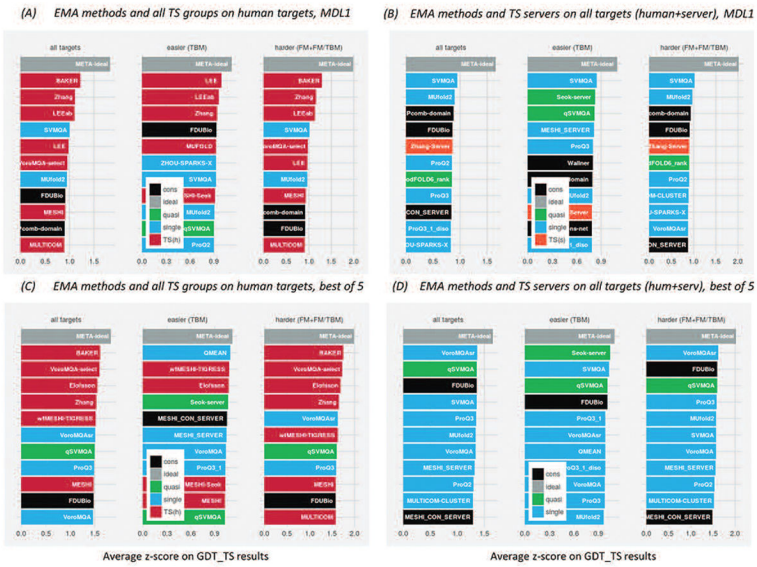

Figure 9.

Comparison of the EMA methods with the tertiary structure prediction methods according to GDT_TS. Panels A and B show the data for first models, while panels C and D for best-out-of-five models. (A, C) Joined ranking of the EMA methods and all TS groups on human targets; (B, D) Joined ranking of the EMA methods and server TS groups on all targets. Rankings are provided separately for all targets, easier targets (TBM) and harder targets (FM and FM/TBM targets). Model accuracy estimation methods are colored as in the rest of the paper: single-model methods in blue, quasi-single in green, and clustering in black; tertiary structure prediction methods are colored as follows: human-expert groups in red, servers in orange. All graphs include the data for the perfect meta-predictor, which always picks the best server model (META-ideal, grey). EMA methods rival performance of the best TS methods in all target difficulty categories, with the perfect meta-predictor being consistently on top of rankings.

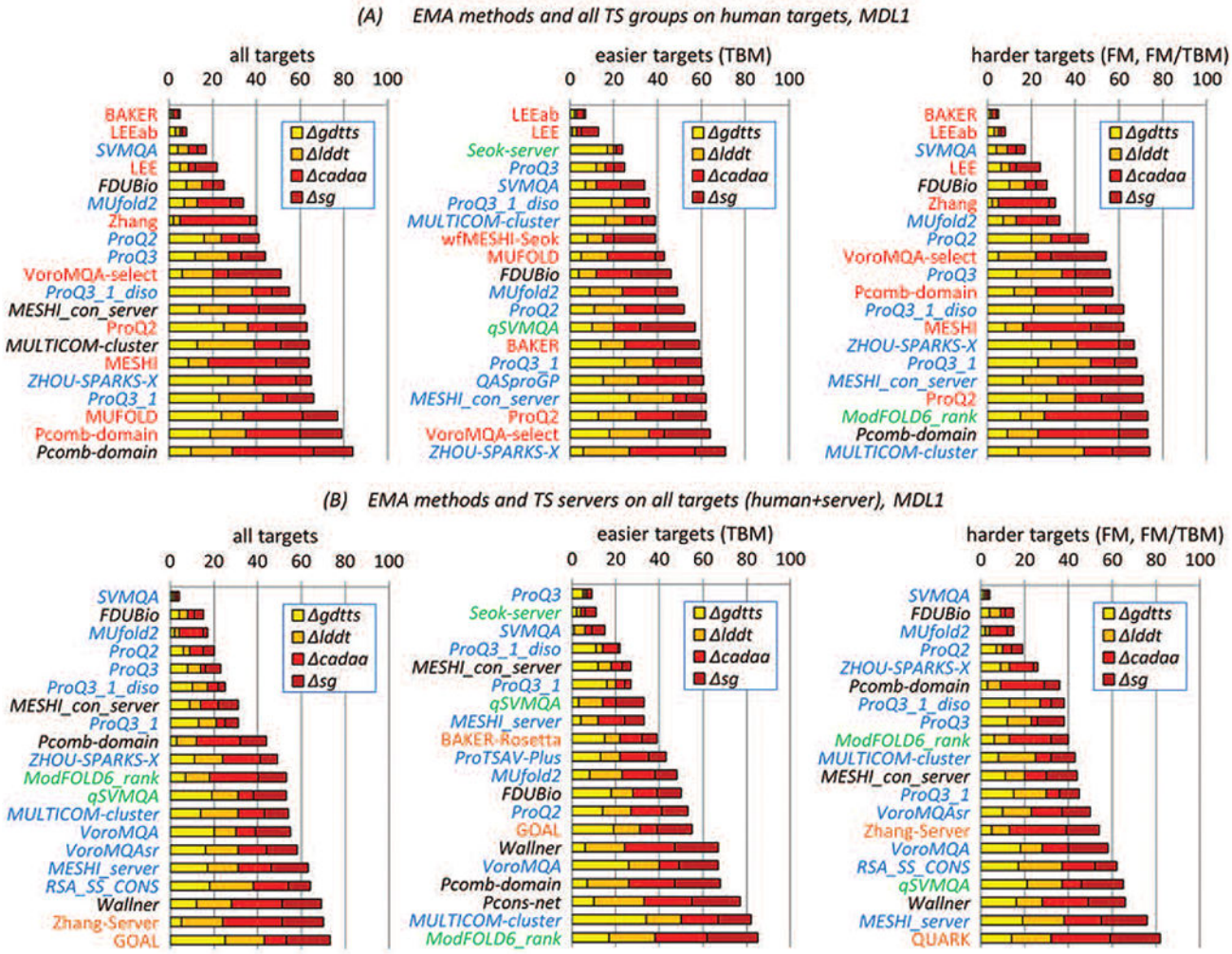

Figure 10.

Cumulative ranking of the EMA methods and the tertiary structure prediction methods on the first models according to four evaluation scores – GDT_TS (yellow), LDDT (orange), CADaa (red), and SphereGrinder (dark red). The best 20 methods in joined ranking are shown. Method coloring scheme is the same as in Figure 9. Being assessed as tertiary structure meta-predictors, accuracy assessment methods rival best expert groups and outperform CASP servers.

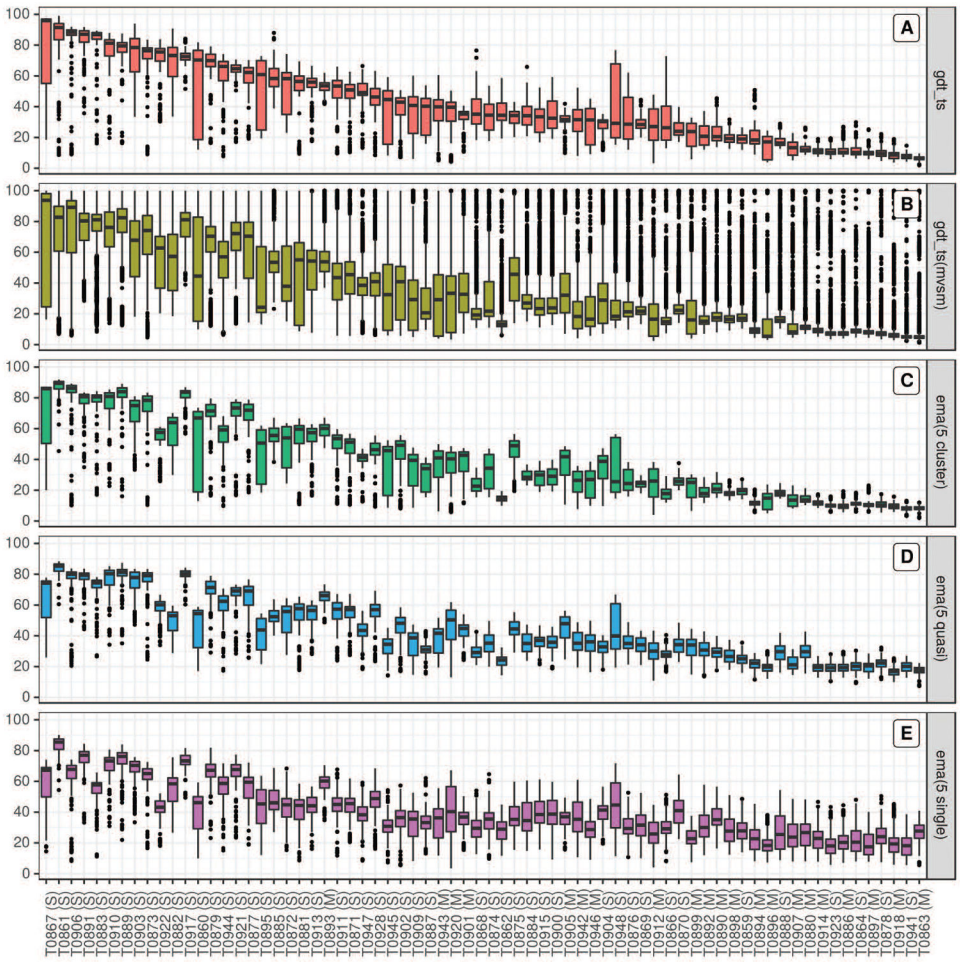

Figure 11.

Boxplots showing per-target distribution of the actual accuracy of server models in the best150 dataset (150 model-target GDT_TS scores, panel A), similarity of models in the best 150 dataset (150*149=22350 model-model pairwise GDT_TS scores, panel B), and the accuracy estimates from the top 5 clustering methods (panel C), quasi-single methods (panel D) and single-model methods (panel E). Each of the panels (C–E) contains 150 data points representing average EMA scores from the selected five methods on a particular target. Box boundaries correspond to the25th (bottom) and 75th (top) percentiles in the data; the horizontal line inside the box corresponds to the median. The height of the box defines the interquartile range (IQR). The height of the whiskers shows the range of values outside the interquartile range, but within 1.5 IQR. The black dots correspond to the outliers outside the 1.5 IQR range. Targets are sorted by the descending median GDT_TS score of the model set (panel A). Single-domain targets are marked with the letter (S) next to the target number, multi-domain with letter (M).

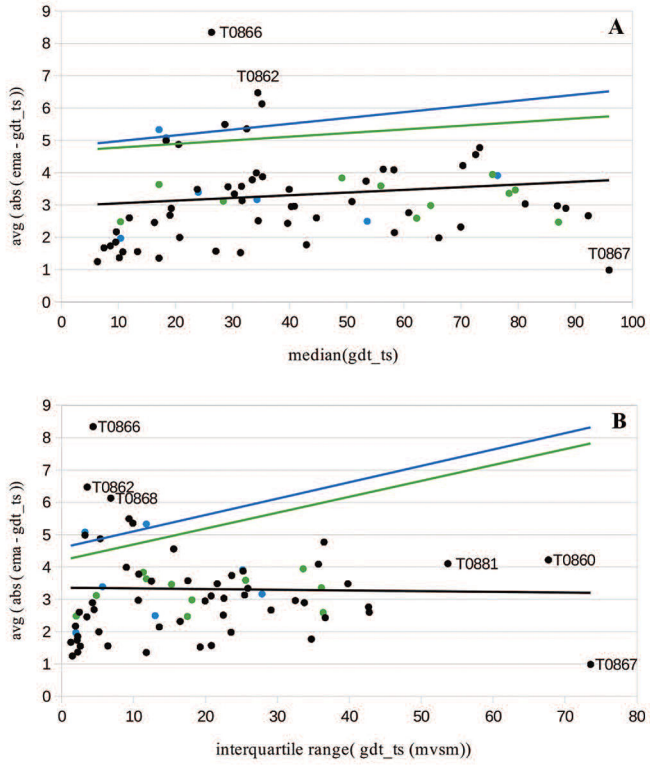

Figure 12.

Deviation between the predicted (EMA) and actual (GDT_TS) scores of server models in the best150 dataset, as a function of (A) target difficulty represented by the median GDT_TS score and (B) similarity of models represented by the per-target interquartile width of the pairwise model-model GDT_TS scores. Each point corresponds to one target. For each target and each EMA group, absolute deviations are calculated for every TS model and then averaged over all predicted models. The minimum average deviation among all EMA groups submitting on the target is plotted with the color corresponding to the type of the best performing method (blue for single, green for quasi-single and black for clustering). Ties are resolved in the order: single, quasi-single, clustering. Lower scores indicate better predicted targets. Black lines run visibly lower than blue and green ones, indicating advantage of clustering methods over single and quasi-single methods in this aspect of analysis. Targets T0862 and T0866 are among the most challenging for predicting absolute accuracy scores, while T0867 is an example of target with very good EMA predictions.

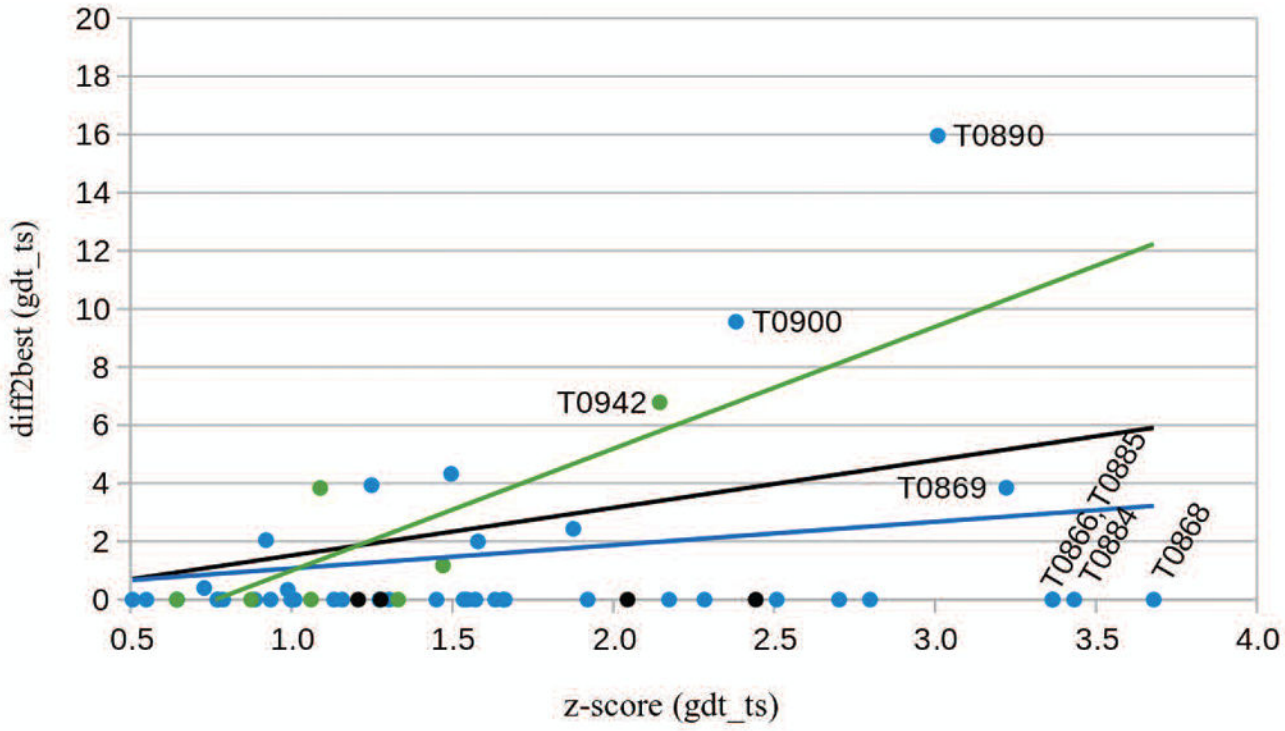

Figure 13.

Difference in accuracy of models predicted to be the best and the actual best according to the GDT_TS score as a function of the separation between the best model and the distribution mean (GDT_TS-based z-score). Each point corresponds to one target. The data are shown for targets with at least one structural model scoring GDT_TS>40. For each target, the minimum deviation among all EMA groups is plotted, with the color corresponding to the type of the best performing method (blue for single, green for quasi-single and black for clustering). Ties are resolved in the order: single, quasi-single, clustering. Lower scores indicate better predicted targets. Larger slope of the line indicates larger dependency of the methods on the separation between the best model and the mean model in terms of the GDT_TS score. Targets T0890, T0900 and T0942 are examples of the largest failures. Targets T0868, T0884, T0866 and T0885 are all examples of the successful identification of best models on the targets where only few models were much better than the others.

Table 1

Statistical comparison of the ability of best performing groups to identify the best models: summary of the two-tailed paired t-tests on per-target differences between the models predicted as the best and the actual best models. The summary is shown for top 10 groups according to the cumulative ranking. Groups are sorted according to the increasing cumulative rank. Single-model methods are in bold, quasi-single – in italic (not present in this table), clustering – in plain text. Each non-diagonal cell contains a 4-character string encoding relationship between the two groups. Each symbol of the string relates to a specific evaluation measure in order – GDT_TS, LDDT, CADaa and SG, and shows if the group in the row was statistically better (+), worse (−) or similar (=) to the group in the column. The data are provided for the statistical significance threshold of p=0.05. We define group A as statistically better/worse than group B if it was proven statistically better/worse according to at least two measures and was not proven worse/better according to any remaining measures. All other pairs of groups are statistically indistinguishable or such that unambiguous statistical difference in performance could not be established (shaded cells).

		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
ProQ3	1	X	====	====	====	==+=	=+==	=+==	=++=	=++=	=+++
SVMQA	2	====	X	====	====	====	====	+===	====	+=+=	=+++
FDUBio	3	====	====	X	====	====	====	====	====	====	=++=
MESHI_con_serv	4	====	====	====	X	====	====	====	====	====	=++=
MESHI_server	5	==−=	====	====	====	X	====	====	====	====	=++=
ProQ3_1_diso	6	=−==	====	====	====	====	X	====	====	====	==+=
ProQ3_1	7	=−==	−===	====	====	====	====	X	====	====	==+=
ProQ2	8	=−−=	====	====	====	====	====	====	X	====	==+=
MUfold2	9	=−−=	−=−=	====	====	====	====	====	====	X	==+=
Pcomb-domain	10	=−−−	=−−−	=−−=	=−−=	=−−=	==−=	==−=	==−=	==−=	X

Table 2

Statistical comparison of the ability of best performing groups to assign ‘correct’ accuracy estimates to models: summary of the two-tailed paired t-tests on differences between the predicted and observed model accuracy. Please see caption to Table 1 for explanation of the content.

		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
ModFOLD6_rank	1	X	−++−	−++−	=+++	+++−	+++−	++++	−++−	−++=	−++=
ModFOLD6_cor	2	+−−+	X	++==	++−+	++==	+++=	++−+	−++−	++++	++++
ModFOLD6	3	+−−+	−−==	X	++−+	++==	+++=	+=−+	−++−	++++	++++
Pcomb-domain	4	=−−−	−−+−	−−+−	X	+=+−	+=+−	+−+=	−++−	−++−	−++−
ProQ3_1	5	−−−+	−−==	−−==	−=−+	X	==+=	+−−+	−++−	−+++	−+++
ProQ3_1_diso	6	−−−+	−−−=	−−−=	−=−+	==−=	X	+−−+	−++−	−+++	−+++
MULTICOM-cluster	7	−−−−	−−+−	−=+−	−+−=	−++−	−++−	X	−++−	−++−	−++−
FDUBio	8	+−−+	+−−+	+−−+	+−−+	+−−+	+−−+	+−−+	X	+−−+	+−−+
Deepfold-Boom	9	+−−=	−−−−	−−−−	+−−+	+−−−	+−−−	+−−+	−++−	X	====
iFold_2	10	+−−=	−−−−	−−−−	+−−+	+−−−	+−−−	+−−+	−++−	====	X

Proteins. Author manuscript; available in PMC 2019 March 01.

Table 3

Statistical comparison of the ability of best performing groups to separate good and bad models: summary of the two-tailed DeLong tests on the ROC-curve data. Please see caption to Table 1 for explanation of the content.

		1	2	3	4	5	6	7	8	9	10
Wallner	1	X	+===	+==+	=+++	++++	++++	++++	++++	=+++	++++
Pcomb-domain	2	−===	X	====	−=+=	=++=	=++=	+++=	==+=	=+++	=+++
ModFOLD6_rank	3	−==−	====	X	−=+=	==+=	==+=	+=+=	==+=	=++=	=+++
QASproCL	4	=−−−	+=−=	+=−=	X	+===	+===	+===	+=+=	==+=	+=++
FDUBio	5	−−−−	=−−=	==−=	−===	X	====	+===	==+=	==+=	==++
Pcons	6	−−−−	=−−=	==−=	−===	====	X	+===	==+=	==+=	==++
Pcons-net	7	−−−−	−−−=	−=−=	−===	−===	−===	X	−=+=	−=+=	−=++
MUfoldQA_C	8	−−−−	==−=	==−=	−=−=	==−=	==−=	+=−=	X	−=+=	==−=
ModFOLDclust2	9	=−−−	=−−−	=−−=	==−=	==−=	==−=	+=−=	+=−=	X	==−=
ModFOLD6	10	−−−−	=−−−	=−−−	−=−−	==−−	==−−	+=−−	==+=	==+=	X

Table 4

Results of the two-tailed paired t-tests on ASE scores for per-residue error estimates evaluated in the whole-target mode. The upper right part of the table contains the numbers of common targets predicted. The lower part displays the probabilities that the differences between the two ASE scores are due to chance. Shaded cells highlight pairs of statistically indistinguishable groups at the 0.05 significance level. Single-model methods are in bold, quasi-single – in italic.

(A)											
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
ModFOLDclust2	**1**	-	70	70	63	70	70	70	70	70	70
Davis-EMAconsensus	**2**	0.19	-	70	63	70	70	70	70	70	70
Pcons	**3**	0.05	0.55	-	63	70	70	70	70	70	70
Pcons-net	**4**	<0.01	<0.01	<0.01	-	63	63	63	63	63	63
Wallner	**5**	<0.01	<0.01	<0.01	0.76	-	70	70	70	70	70
ModFOLD6	**6**	<0.01	<0.01	<0.01	0.23	0.19	-	70	70	70	70
ModFOLD6_cor	**7**	<0.01	<0.01	<0.01	0.23	0.19	0.89	-	70	70	70
ModFOLD6_rank	**8**	<0.01	<0.01	<0.01	0.23	0.19	0.18	0.27	-	70	70
Pcomb-domain	**9**	<0.01	<0.01	<0.01	0.25	0.04	0.48	0.48	0.48	-	70
ProQ3_1_diso	**10**	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	-

Proteins. Author manuscript; available in PMC 2019 March 01.