Improved protein model quality assessments by changing the target function.

Karolis Uziela¹, David Menéndez Hurtado¹, Nanjiang Shu^{1,2},
Björn Wallner³ and Arne Elofsson^{1,†}

February 23, 2018

¹ Department of Biochemistry and Biophysics and Science for Life Laboratory, Stockholm University, 171 21 Solna, Sweden,

² Bioinformatics Short-term Support and Infrastructure (BILS), Science for Life Laboratory, 171 21 Solna, Sweden

³ Department of Physics, Chemistry and Biology (IFM) /Bioinformatics. Linköping University, 581 83 Linköping, Sweden.

[†]To whom correspondence should be addressed at arne@bioinfo.se

Short title: Target function for ProQ3D

Keywords:

- Protein Structure Prediction
- CASP
- Model Quality Assessments
- Estimation of Model Accuracy
- Deep Learning

Abstract

Protein modeling quality is an important part of protein structure prediction. We have for more than a decade developed a set of methods for this problem. We have used various types of description of the protein and different machine learning methodologies. However, common to all these methods has been the target function used for training. The target function in ProQ describes the local quality of a residue in a protein model. In all versions of ProQ the target function has been the S-score. However, other quality estimation functions also exist, which can be divided into superposition- and contact-based methods. The superposition-based methods, such as S-score, are based on a rigid body superposition of a protein model and the native structure, while the contact-based methods compare the local environment of each residue. Here, we examine the effects of retraining our latest predictor, ProQ3D, using identical inputs but different target functions. We find that the contact-based methods are easier to predict and that predictors trained on these measures provide some advantages when it comes to identifying the best model. One possible reason for this is that contact based methods are better at estimating the quality of multi-domain targets. However, training on the S-score gives the best correlation with the GDT_TS score, which is commonly used in CASP to score the global model quality. To take the advantage of both of these features we provide an updated version of ProQ3D that predicts local and global model quality estimates based on different quality estimates.

Introduction

Protein quality estimations has a long history. In the 1990s knowledge-based energy functions using the local environment or residue-residue contacts were developed^{1, 2}. These methods were developed to detect the native structure, or to detect the correct fold by threading a sequence through all known protein folds³. Later models based on atom-atom contacts were also developed⁴.

Although these methods were quite good at identifying errors in protein structures, they only had limited success when used for model quality estimation. One reason for this is that these energy functions are developed to distinguish between the native structure and an unfolded protein. However, incorrect models share many features of the native structure—making it difficult to distinguish them from the native structure using these energy functions.

To solve this problem model quality estimation methods were developed. These methods are often developed using a machine learning approach trained not only to detect the native structure but to actually predict the quality of a model. One of the first methods using this methodology was ProQ⁵.

For more than a decade we have developed protein model quality assessment methods using various input features^{6, 5, 7}. The ProQ methods have consistently been among the best individual protein model quality assessment methods in CASP evaluations^{8, 9, 10}. During the last few years we have been able to provide a small, but significant, improvement to the methods by (i) including additional descriptions of a protein model¹¹ and (ii) applying deep learning techniques¹². Although the overall prediction accuracy has improved substantially the ability to identify the best model has not improved significantly for some time.

When developing a model quality assessment method it is necessary to define the target function to be used. All ProQ as well as many other methods have been trained using the same target function, the S-score^{13, 14}. The S-score is calculated after a superposition of the model and the native structure and then describes the displacement of each residue. A global S-score can then be

calculated as the average S-score over a model. The global score is similar in nature to other widely used measures, including the main target function in CASP, GDT_TS^{15, 8}, MaxSub¹⁶ and TM-score¹⁷. The main difference between these methods is how the sum of the scores is normalised by model and protein target length.

During the last few years contact-based approaches to model quality assessments have been introduced; ModlDDT¹⁸, CAD-score¹⁹. These functions have also been used for consensus based MQAPs^{20, 21, 22}. All these methods assess the conservation of local interactions between a model and the native structure, i.e. they are not dependent on a global superposition, making the contact-based methods insensitive to variability in relative domain arrangements. This in turn might make them more suitable as target functions for single-model MQAPs that are using the local environment of individual residues as their main descriptor.

Here, we first set out to examine the relationship between the different target functions for model quality assessment. Then, we retrained ProQ3D using the different measures as target functions and evaluated their performance. We show that, as expected, the version of ProQ3D that is trained on one target function also performs best when that target function is used for evaluation. However, when examining the ability to identify the best model out of a set of candidates, it is clear that training ProQ3D on contact-based functions is advantageous.

Methods

Below, we describe briefly the different model quality estimation methods used in this paper, for complete descriptions see the original papers.

Superposition-based quality estimators

Superposition-based estimators are using a structural alignment of a model to the native structure. In contrast to traditional measures of the Root Mean Square Deviations (RMSD) the goal of

these methods are to only superpose the most similar regions of the two structures. The different superposition-based methods differ on the exact algorithm to identify the most similar region and on the exact target function.

We have used three different superposition-based quality estimators for training and testing: S-score^{14, 13}, TM-score²³ and GDT_TS²⁴. The GDT_TS measure was only used for testing, because ProQ3D is trained on the local (residue) level while GDT_TS is defined only on the global (protein) level.

The quality estimator S-score used in ProQ^{7, 11, 12, 25} is defined as:

$$S_i = \frac{1}{1 + (d_i/d_0)^2} \quad (1)$$

where d_i is the distance between residue i in the model and in the native structure after superposition. Here, the distance threshold d_0 is set to 3Å. The global S-score for the whole model is the sum of all the local scores divided by the protein length.

TM-score is using the same equation (1) as in S-score but the d_0 parameter in TM-score depends on the protein length (L_N) and is defined as:

$$d_0 = 1.24 \sqrt[3]{L_N - 15} - 1.8. \quad (2)$$

By default, TM-score does not provide the local (per-residue) quality estimates. Therefore, we have run TM-score with the `-o` parameter that outputs the structural superposition between the model and the native structure and then calculated the local TM-scores using the formula (1) replacing d_0 with the expression in (2).

GDT_TS was calculated the same way as in the CASP experiments using the LGA package¹⁵. The program was run with the following parameters:

```
-3 -ie -o1 -sda -d:4 -gdc_sc -swap
```

Contact-based quality estimators

In contrast to superposition-based quality estimators the contact-based predictors do not use a superposition of a model and the native structure. Instead these methods use the local environment, e.g. described by contacts or residue surface areas, for each residue as a descriptor of the quality.

Two contact-based quality estimators were used: IDDT¹⁸ and CAD-score¹⁹. IDDT compares atom-atom distances between the model and the native structure, while CAD-score compares residue contact surface areas. Both IDDT and CAD-score were run with default parameters, but ignoring residue name consistency checks (`-x` parameter in IDDT and `-b` parameter in CAD-score). For CAD-score we have used the “all atom - sidechain” (A-S) contact score. The reason is that this score behaves better for models with missing residues.

Both CAD and IDDT utilizes all atoms in a protein model while the superposition-based ones only use the $C\alpha$ -coordinates. To examine if the improved performance in first rank loss is due to the fact that CAD and IDDT uses full atom representation we also trained a version of ProQ3D using the version of IDDT that only used the $C\alpha$ atoms (IDDT-CA¹⁸) as a target function.

Training deep learning method

We trained the deep learning methods in exactly the same way as in ProQ3D¹² using different local quality estimators, S-score, a local version of TM-score, IDDT or CAD-score as target functions. For details of the target function see the Methods section. Briefly, ProQ3D uses a multilayer perceptron with two dense hidden layers (600 and 200 neurons). The optimisation function used was Adadelta with a 10^{-11} penalty for the L^2 regularisation on the weights. Training was done using Keras²⁶ library with Theano²⁷ backend in Python.

Below, we will refer to the different predictors as: ProQ3D-S, ProQ3D-TM, ProQ3D-CAD and ProQ3D-IDDT, when trained on S-score, TM-score, CAD-score and IDDT, respectively, as target functions. ProQ3D-S is obviously identical to the original ProQ3D but we use the name ProQ3D-S

for clarity.

Datasets

We have used exactly the same training and test data sets as in ProQ3D¹². The training set consisted of CASP9 and CASP10 models and the two test sets consisted of CASP11 and CAMEO²⁸ (2014-06-06–2015-05-30) models. Cancelled targets and targets shorter than 50 residues were filtered out. For simplicity we choose to show CAMEO data in the main text and CASP11 data in the supplementary material, as the results are very similar.

To account for sub-optimal side-chain orientations, we have repacked the side-chains of all models as described in the development of ProQ3¹¹. Models with repacked side-chains were used for all four predictors: ProQ3D-S, ProQ3D-TM, ProQ3D-CAD and ProQ3D-IDDT.

Evaluation

We used Pearson correlation for evaluating the agreement between the predicted values and the target values. The correlations are evaluated on local (residue) level and global (model) level. The global correlations are calculated in two ways: the correlation for the whole data set and the average correlation calculated for each target separately. We evaluated both direct correlations, i.e. when the method is trained and tested on the same quality measures, and indirect correlations when the method is trained on one quality measure but tested on another.

Model selection was assessed by the *first ranked loss*⁹, defined as the average difference between the model with the best quality and the selected model according to a given quality estimator.

Contact order is the average sequence separation between residues in contact in the structure. It was calculated by Perl script available at http://depts.washington.edu/bakerpg/contact_order/²⁹. The script was run with `-r` parameter that returns the relative contact order and `-c 6` parameter that sets 6 Å cutoff for contacting residues (default settings).

Results and Discussion

Here, we study the effect of the choice of a target function for protein model quality estimation. We use five different protein quality evaluation estimates, three based on superposition (GDT_TS, TM-score and S-score) and two superposition independent contact-based methods (CAD-score and IDDT). All methods except GDT_TS provide both local and global quality estimates.

Correlation between quality estimates.

First, we examine the correlation between the four measures that provide local quality estimate (Figure 1). All quality measures correlate with each other. However, it is also clear that the measures using similar principles, i.e. superposition (S-score and TM-score) or contacts (CAD-score and IDDT) are more similar to each other ($cc = 0.88$ and $cc = 0.79$). The correlations between the two groups are lower ($cc = 0.52$ to 0.76). In particular CAD-score only shows a weak correlation with the superposition-based methods.

It can also be noted that the quality estimated from both TM-score and S-score is almost bimodal with a clear majority of the residues evaluated as good (quality close to one) or bad (quality close to zero). This is because both these measures are using Eq. 1 to calculate the score, effectively assigning a score close to one to most superimposed residues and close to zero to non-superimposed residues.

Next we examined the global quality estimates (Figure 2). The correlations between the global measures are higher than for the local measures. The correlations between the three superposition-based measures are excellent ($cc > 0.95$). In fact, the correlation between S-score and GDT_TS is astonishingly high ($cc > 0.996$) possibly explaining the success of the ProQ methods in CASP, where GDT_TS is used for model quality evaluation. The correlation between the two contact-based measures, CAD-score and IDDT, is also excellent ($cc > 0.98$), while the correlations between the contact-based and the superposition-based measures are lower ($cc = 0.86$ to 0.92).

There are some differences in how the quality measures treat models that are not full length, i.e models that do not include all residues in the native structure. The models that are shorter than 50% of the full length are colored in red in Figure 2 and are often the outliers between different quality measures.

It can be noted that although the correlations are high between the different quality measures it is not higher than the correlation between predicted and estimated quality for the best quality predictors. In CASP12 the correlation of the consensus-based model quality predictors^{30, 31} is higher to GDT_TS than the correlation between GDT_TS and CAD-score¹⁰.

Training using different target functions

Next we set out to retrain ProQ3D using the four different quality measures as target functions. A deep learning network was trained on CASP9 and CASP10 models and evaluated on CAMEO (in main text) and CASP11 (see supplementary data). Below, the quality measures are referred to as ProQ3D-S, ProQ3D-TM, ProQ3D-CAD and ProQ3D-IDDT while the evaluation methods are referred to by their names (GDT_TS, S-score, TM-score, CAD-score and IDDT).

As expected, the best correlation is obtained when the same quality estimator is used both for training and testing (Figure 3). Clearly, on the local level it is easier to predict IDDT than any of the other measures. Independently, on the training method the correlation is always highest when using IDDT to evaluate the local quality.

After transferring the local predictions to global quality estimates it is still clear that the highest correlation is obtained when using the same training and evaluation function. However, the differences are smaller, with correlations in the range $cc = 0.78 - 0.88$, for all combinations of training and evaluation functions (Figure 4). Also, for global quality estimations both of the contact-based evaluations seems slightly easier to predict.

Low contact order models differentiate the quality measures

Superposition-based quality estimators have problems with models that contain multiple domains, where the individual domains are correctly modeled but their relative arrangement is wrong. As mentioned above a superposition-based quality estimation methods superimpose one of the domains and assign low quality to all residues in the other domain. In contrast, the contact-based measures assign more appropriate domain level scores to the residues, if each domain is considered in isolation. It should be noted that this problem does not only relate to classical multi-domain proteins, but also to proteins with a fibrillar (or extended) shape, or actually any protein with a low contact-order. Therefore, we decided to study the targets with respect to their contact order in the native structure.

An example of a low contact order protein where CAD-score gives a much higher evaluation than S-score is CASP11 target T0821 shown Figure 5. This target has the lowest contact order in CASP11 data set (0.026) (see Figure S6) and almost all of the submitted models for this target have higher CAD-score than S-score, see red dots in Figure 5a. T0821 is a helical solenoid and most models are locally correct but the exact orientation of the N- and C-terminal parts is wrong. Evaluation of the model BAKER-ROSETTASERVER_TS4 highlights this. It has both units modelled more or less correctly, but their relative orientation is wrong. Therefore, S-score is only good for the part that is superposed (Figure 5b) while CAD-score gives relatively good scores to the entire protein (Figure 5c). As a result, the global CAD-score of the model is higher than its S-score.

In Figure 4 and upper right side of Figure 2, the models are colored in blue to yellow according to their contact order, with low contact order targets being coloured in blue. When we compare the superposition-based scores (S-score and TM-score) with the two contact-based measures (CAD-score and IDDT) we can see that low contact order are more frequent to the upper left side of the diagonal. This means that the (blue) low contact order models achieve relatively higher qualities using the contact-based measures, as expected if only a part of the model has been superposed.

We can also notice that the S-score vs TM-score plot has more low contact order proteins on the

lower right side of the diagonal. However, these differences rise mostly because of the different way how S-score and TM-score normalize the scores by protein length. TM-score gives higher score for longer proteins because of the way $d0$ value is calculated (see formula in Methods). Long proteins in turn tend to have lower contact order, despite the fact that we calculate a relative contact order which is normalized by protein length (see Supplementary Figure S6).

Ranking models

It is not only important to have a good correlation for all protein models but also for each individual target. A good per target correlation enables the identification of the best model for that particular target. Here, the advantage of contact-based measures is clear, see Table I and Supporting Table S1. There are significantly higher self-correlations for IDDT and CAD-score ($cc > 0.8$ vs. $cc < 0.6$). Further, when training on a contact-based measures, the per target correlation is higher even when using the superposition-based evaluation methods on the CAMEO set, Table I, while this is not the case on the CASP set, Supporting Table S1.

Also the contact-based measures provide an improved detection of the best models as measured by the first ranked loss⁹, Table II. Training on CAD-score or IDDT selects higher quality models across all evaluation measures. Thus, using CAD-score or IDDT as target functions seem to provide a small but significant improvement in model selection. It can be noted that when training on a version of IDDT that only uses the $C\alpha$ -atoms, IDDT-CA¹⁸, the performance is very similar to the results obtained by IDDT. The results are slightly better at CASP11 and slightly worse using CAMEO, and for both datasets better than when using S-score or TM-score, see Supplementary Tables S1-4. Therefore, we do believe that the increased performance is mainly due to the difference between contact- and superposition-based quality evaluation functions and not due to the difference in what atoms are included in the analysis.

When studying GDT_TS for the models that are first ranked by different predictors it can be seen that for most targets models of very similar quality are selected (Figure 6). However, there

appears to be a set of targets where the contact-based predictors consistently choose models that are significantly different from the ones selected by the superposition-based predictors. Four of these targets are analysed in more detail, see Table III.

For the target 4YZZ_A, ProQ3D-S ranks the best model first, while the other predictors rank significantly worse models first. It can be seen in Figure 7 that the ProQ3D-S model is very good while the model selected by ProQ3D-CAD has a different fold. Also, for another target, 4RO3_B, both of the superposition-based predictors, ProQ3D-S and ProQ3D-TM, rank better models first than contact-based predictors. For the remaining two targets 4V1G_C and 4Z1H_A the contact-based predictors, ProQ3D-CAD and ProQ3D-IDDIT, rank better models first.

In all of these four cases the difference in predicted scores between the good and the bad models are quite small. We did not find any obvious reason why one predictor ranked one model before the other one - indicating that when scores are similar and the models are different it is advisable to examine several models for a given target.

Conclusions

Here, we show that when ProQ3D is retrained using different target functions, i.e. different descriptions of the model quality of a protein, model selection can be improved. Target functions can be divided into two groups: superposition- and contact-based methods. We find that when using contact-based measures as target functions in training, we can improve performance in model selection. We also notice that it is easier to predict the quality when using a contact-based description, in particular IDDIT. However, the correlation with the “standard” CASP evaluation measure, GDT_TS, is best when the original superposition-based target function, S-score, is used for training ProQ3D.

Funding

This work was supported by grants from the Swedish Research Council (VR-NT 2016-03798 to AE and 2012-5270 to BW) and Swedish e-Science Research Center (BW). The Swedish National Infrastructure provided computational resources for Computing (SNIC) at NSC.

References

- [1] Luthy, R., Bowie, J., and Eisenberg, D. Assessment of protein models with three-dimensional profiles. *Nature* 356(6364):83–85, Mar, 1992.
- [2] Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Casari, G., and Sippl, M. Identification of native protein folds amongst a large number of incorrect models. The calculation of low energy conformations from potentials of mean force. *J Mol Biol* 216(1):167–180, Nov, 1990.
- [3] Jones, D., Taylor, W., and Thornton, J. A new approach to protein fold recognition. *Nature* 358(6381):86–89, Jul, 1992.
- [4] Colovos, C. and Yeates, T. Verification of protein structures: patterns of nonbonded atomic interactions. *Protein Sci* 2(9):1511–1519, Sep, 1993.
- [5] Wallner, B. and Elofsson, A. Identification of correct regions in protein models using structural, alignment, and consensus information. *Protein Sci* 15(4):900–913, Apr, 2006.
- [6] Wallner, B. and Elofsson, A. Can correct protein models be identified? *Protein Sci* 12(5):1073–1086, May, 2003.
- [7] Ray, A., Lindahl, E., and Wallner, B. Improved model quality assessment using ProQ2. *BMC Bioinformatics* 13:224, 2012.

- [8] Kryshchak, A., Barbato, A., Monastyrskyy, B., Fidelis, K., Schwede, T., and Tramontano, A. Methods of model accuracy estimation can help selecting the best models from decoy sets: Assessment of model accuracy estimations in CASP11. *Proteins* 84 Suppl 1:349–369, Sep, 2016.
- [9] Kryshchak, A., Monastyrskyy, B., Fidelis, K., Schwede, T., and Tramontano, A. Assessment of model accuracy estimations in CASP12. *Proteins* in press, Aug, 2017.
- [10] Elofsson, A., Joo, K., Keasar, C., Lee, J., Maghrabi, A. H. A., Manavalan, B., McGuffin, L. J., Hurtado, D. M., Mirabello, C., Pilstl, R., Sidi, T., Uziela, K., and Wallner, B. Methods for estimation of model accuracy in CASP12. *Proteins* in press, 2017.
- [11] Uziela, K., Shu, N., Wallner, B., and Elofsson, A. ProQ3: Improved model quality assessments using Rosetta energy terms. *Sci Rep* 6:33509, Oct, 2016.
- [12] Uziela, K., Menéndez Hurtado, D., Shu, N., Wallner, B., and Elofsson, A. ProQ3D: Improved model quality assessments using Deep Learning. *Bioinformatics* in press, Jan, 2017.
- [13] Levitt, M. and Gerstein, M. A unified statistical framework for sequence comparison and structure comparison. *Proc Natl Acad Sci U S A* 95(11):5913–5920, May, 1998.
- [14] Cristobal, S., Zemla, A., Fischer, D., Rychlewski, L., and Elofsson, A. A study of quality measures for protein threading models. *BMC Bioinformatics* 2:5, 2001.
- [15] Zemla, A. LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Res* 31(13):3370–3374, Jul, 2003.
- [16] Siew, N., Elofsson, A., Rychlewski, L., and Fischer, D. MaxSub: an automated measure for the assessment of protein structure prediction quality. *Bioinformatics* 16(9):776–785, Sep, 2000.

- [17] Zhang, Y. and Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res* 33(7):2302–2309, 2005.
- [18] Mariani, V., Biasini, M., Barbato, A., and Schwede, T. IDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics* 29(21):2722–2728, Nov, 2013.
- [19] Olechnovic, K., Kulberkyte, E., and Venclovas, C. CAD-score: a new contact area difference-based function for evaluation of protein structural models. *Proteins* 81(1):149–162, Jan, 2013.
- [20] McGuffin, L. and Roche, D. Rapid model quality assessment for protein structure predictions using the comparison of multiple models without structural alignments. *Bioinformatics* 26(2):182–188, Jan, 2010.
- [21] Skwark, M. and Elofsson, A. PconsD: ultra rapid, accurate model quality assessment for protein structure prediction. *Bioinformatics* 29(14):1817–1818, Jul, 2013.
- [22] Roche, D. B., Buenavista, M. T., and McGuffin, L. J. Assessing the quality of modelled 3D protein structures using the ModFOLD server. *Methods in molecular biology (Clifton, N.J.)* 1137:83–103, 2014.
- [23] Zhang, Y. and Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins* 57(4):702–710, Dec, 2004.
- [24] Zemla, A., Venclovas, M., Mout, J., and Fidelis, K. Processing and evaluation of predictions in CASP4. *Proteins Suppl* 5:13–21, 2001.
- [25] Uziela, K. and Wallner, B. ProQ2: estimation of model accuracy implemented in Rosetta. *Bioinformatics* 32(9):1411–1413, May, 2016.
- [26] Chollet, F. Keras. <https://github.com/fchollet/keras>, , 2015.

- [27] Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. arXiv e-prints abs/1605.02688, May, 2016.
- [28] Haas, J., Barbato, A., Behringer, D., Studer, G., Roth, S., Bertoni, M., Mostaguir, K., Gumienny, R., and Schwede, T. Continuous Automated Model EvaluatiOn (CAMEO) complementing the critical assessment of structure prediction in CASP12. *Proteins*, Nov, 2017.
- [29] Plaxco, K., Simons, K., and Baker, D. Contact order, transition state placement and the refolding rates of single domain proteins. *J Mol Biol* 277(4):985–994, Apr, 1998.
- [30] Lundstrom, J., Rychlewski, L., Bujnicki, J., and Elofsson, A. Pcons: a neural-network-based consensus predictor that improves fold recognition. *Protein Sci* 10(11):2354–2362, Nov, 2001.
- [31] Ginalski, K., Elofsson, A., Fischer, D., and Rychlewski, L. 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics* 19(8):1015–1018, May, 2003.

Figure legends

Figure 1: Scatter plot comparing different quality estimators on the *local* level using 4,939,249 residues (points) from models submitted to CAMEO. The intensity is plotted using the Viridis color scale, i.e. the most dense regions are yellow. Contour lines based on the point density are marked. The correlation coefficient (cc) between the different quality estimators is given in the insets.

Figure 2: Scatter plot comparing different quality estimators on the *global* level using 19,985 models (points) submitted to CAMEO. The scatter plots on the upper right side are colored based on the contact order of the native structure, blue being low and yellow being high, see the color bar for actual values. The scatter plots on the lower left side are colored based on density of the points. The models that are shorter than 50% of the full length are colored in red. The correlation coefficient (cc) between the different quality estimators is given in the insets.

Figure 3: Scatter plot comparing real scores against the predicted scores on the *local* level using 4,939,249 residues (points) submitted to CAMEO. The intensity is plotted using the Viridis color scale, i.e. the most dense regions are yellow. Contour lines based on the point density are marked. The correlation coefficient (cc) between the different quality estimators is given in the insets.

Figure 4: Scatter plot comparing real scores against the predicted scores on the *global* level using 19,985 models (points) submitted to CAMEO. The scatter plots are coloured based on the contact order of the native structure, blue being low and yellow being high, see the color bar for actual values. The correlation coefficient (cc) between the different quality estimators is given in the insets.

Figure 5: CAD-score and S-score for all models in CASP11. All models from target T0821 are marked in red. (b) BAKER-ROSETTASERVER_TS4 model for T0821 target coloured by S-score and superposed to the native structure. Good regions are coloured in blue, bad regions are coloured in red. The native structure is coloured in grey. The isolated helix that is coloured in red in the model is not present in the native structure (c) BAKER-ROSETTASERVER_TS4 model for T0821 target coloured by CAD-score. Good regions are coloured in blue, bad regions are coloured in red. The isolated helix that is coloured in red in the model is not present in the native structure

Figure 6: Scatter plot comparing GDT_TS of selected models by ProQ3D trained on different quality measures. CAMEO data set. Targets where GDT_TS difference of selected models is higher than 0.2 are labelled.

Figure 7: Top ranked models of target four targets superposed to the native structure (green). Models selected by ProQ3D-S are blue and by ProQ3D-CAD magenta.

Tables

Table I: **Global predictions - average per target correlations on the CAMEO data set.** Best prediction for each quality measure is marked in bold. Predictions that do not differ significantly ($p\text{-value} < 0.001$) from the best one in a column are also marked in bold. Wilcoxon signed-rank test was used to evaluate the significance. Please, note that according to Wilcoxon signed rank test a higher difference in the average value between two samples does not always mean that the difference is more significant.

Predictor	CAD-score	IDDT	S-score	TM-score	GDT_TS
ProQ3D-CAD	0.82	0.80	0.59	0.61	0.61
ProQ3D-IDDT	0.80	0.80	0.60	0.62	0.61
ProQ3D-S	0.61	0.64	0.53	0.55	0.54
ProQ3D-TM	0.66	0.69	0.55	0.59	0.57

Table II: **First ranked loss on the CAMEO data set.** Best prediction for each quality measure is marked in bold. Predictions that do not differ significantly ($p\text{-value} < 0.001$) from the best one in a column are also marked in bold. Wilcoxon signed-rank test was used to evaluate the significance.

Predictor	CAD-score	IDDT	S-score	TM-score	GDT_TS
ProQ3D-CAD	0.024	0.022	0.039	0.034	0.036
ProQ3D-IDDT	0.025	0.022	0.040	0.035	0.037
ProQ3D-S	0.033	0.028	0.044	0.038	0.041
ProQ3D-TM	0.034	0.029	0.045	0.039	0.042

Table III: Four targets, where the top ranked models selected by the different ProQ3D methods have vastly different quality. The table shows the models that are top ranked by four different ProQ3D versions for four targets in CAMEO data set. GDT_TS score, as well as all four predicted scores are shown for each of these models. In parenthesis, the rank of the model is shown when sorted by the score of a particular predictor. In all and four examples ProQ3D-CAD and ProQ3D-IDDT selects the same top ranked model.

model	GDT_TS	ProQ3D-CAD	ProQ3D-IDDT	ProQ3D-S	ProQ3D-TM
4RO3_B					
server12-2	0.59 (#1)	0.36 (#3)	0.53 (#2)	0.44 (#1)	0.60 (#2)
server12-1	0.59 (#1)	0.36 (#4)	0.53 (#3)	0.44 (#2)	0.60 (#1)
server11-3	0.30 (#17)	0.40 (#1)	0.56 (#1)	0.41 (#6)	0.56 (#6)
4V1G_C					
server4-1	0.90 (#3)	0.38 (#1)	0.46 (#1)	0.26 (#3)	0.36 (#5)
server27-1	0.51 (#31)	0.37 (#4)	0.46 (#3)	0.27 (#1)	0.35 (#7)
server11-2	0.42 (#35)	0.35 (#12)	0.44 (#9)	0.26 (#5)	0.39 (#1)
4YZY_A					
server32-2	0.64 (#1)	0.34 (#7)	0.51 (#7)	0.53 (#1)	0.72 (#4)
server11-1	0.20 (#16)	0.42 (#1)	0.60 (#1)	0.49 (#3)	0.70 (#5)
server11-4	0.20 (#16)	0.43 (#2)	0.58 (#2)	0.46 (#6)	0.78 (#1)
4Z1H_A					
server11-5	0.64 (#6)	0.60 (#1)	0.78 (#1)	0.57 (#9)	0.77 (#12)
server4-1	0.42 (#19)	0.53 (#15)	0.71 (#15)	0.62 (#1)	0.83 (#1)