Published in final edited form as:
J Struct Biol. 2018 October ; 204(1): 96–108. doi:10.1016/j.jsb.2018.07.006.

Evaluation System and Web Infrastructure for the Second Cryo-EM Model Challenge

Andriy Kryshtafovych[1], **Paul D. Adams**[2,3], **Catherine L. Lawson**[4], and **Wah Chiu**[5]

[1]Genome Center, University of California, Davis, 451 Health Sciences Drive, Davis, CA 95616, USA

[2]Molecular Biophysics & Integrated Bioimaging, LBNL, CA 94720, USA

[3]Department of Bioengineering, University of California Berkeley, CA 94720, USA

[4]Institute for Quantitative Biomedicine and Research Collaboratory for Structural Bioinformatics, Rutgers, The State University of New Jersey, 174 Frelinghuysen Road, Piscataway, NJ 08854, USA

[5]Department of Bioengineering, Microbiology and Immunology and Photon Science, Stanford University, James H. Clark Center, MC5447, 318 Campus Drive, Stanford, CA 94305-5447, USA

Abstract

An evaluation system and a web infrastructure were developed for the second cryo-EM model challenge. The evaluation system includes tools to validate stereo-chemical plausibility of submitted models, check their fit to the corresponding density maps, estimate their overall and per-residue accuracy, and assess their similarity to reference cryo-EM or X-ray structures as well as other models submitted in this challenge. The web infrastructure provides a convenient interface for analyzing models at different levels of detail. It includes interactively sortable tables of evaluation scores for different subsets of models and different sublevels of structure organization, and a suite of visualization tools facilitating model analysis. The results are publicly accessible at http://model-compare.emdatabank.org.

Keywords

cryo-EM; model challenge; protein structure modeling; protein structure verification

Introduction

The second cryo-Electron Microscopy Model Challenge (EMMC) was organized to bring together the cryo-EM structure determination community, learn about available approaches for generating atomic coordinates from three-dimensional electron microscopy (3DEM) density maps reported to be at 3.0–4.5 Å resolution, reveal abilities of the current modeling methods, and establish working protocols for validating the accuracy of models.

Corresponding author: Andriy Kryshtafovych, akryshtafovych@ucdavis.edu, Tel.: (+1) 530-754-8977.

The challenge organizing committee selected eight modeling targets from the cryo-EM structures published in the period 2014–16 (http://challenges.emdatabank.org/?q=model_challenge). Challengers were invited to submit 3D coordinates of models created to fit the provided density maps. Models could be generated either *ab initio* or by optimizing existing models. The submissions were collected, analyzed, preprocessed (if necessary), and evaluated with a suite of measures agreed upon by the challenge committee. The models and the results of the evaluations were anonymized and provided to volunteer assessors via the web interface. After the assessors prepared initial analyses of the modeling results, a face-to-face meeting with the challenge committee was organized to discuss preliminary outcomes. The challenge culminated in a joint participants, assessors and organizers meeting in October 2017, where the results were reviewed and discussed, and plans for future challenges developed. The detailed timeline of the various events can be found in the accompanying paper of this journal issue (Lawson and Chiu, to be submitted, editorial this issue).

Since no dedicated data-handling infrastructure was available, it was necessary to develop a system for evaluation of submitted models and presentation of the evaluation results. CASP experiments have been assessing accuracy of *in silico* structure models for over two decades (Moult et al., 2018), and we used the CASP evaluation system (Kryshtafovych et al., 2016a) as a prototype for the cryo-EM model evaluation system. Here we describe components of this system, enumerate the measures used in the evaluation, give an overview of the web infrastructure designed to facilitate analysis of the submitted models, and provide a brief statistical overview of the scores obtained for the submitted models.

1. Evaluation system: architecture and components

1.1. Targets

Eight modeling targets were included in the EMMC experiment (see http://challenges.emdatabank.org/?q=model-challenge-targets). The targets were numbered consecutively, from T0001 to T0008. For several targets (T0002, T0006, T0007 and T0008), two independently determined 3DEM maps were provided (designated as mapA and mapB). In these cases, models were evaluated only against the target/map combination specified by the challengers and appear only in results tables corresponding to the specified target/map combination. The list of twelve evaluated target/map pairs can be found at the main EMMC evaluation web page: http://model-compare.emdatabank.org.

1.2. Reference structures

For each target, the organizing committee picked one or more reference structure(s), against which the submitted models were evaluated. These reference structures were selected from experimental structures with the highest resolution. Experimental structures are defined here as protein structures determined with X-ray or cryo-EM techniques and publicly available from the Protein Data Bank (PDB) (Berman et al., 2000). For some targets, several reference structures were assigned. The selected reference structures for the EMMC targets are listed in the Model Challenge Targets table at http://challenges.emdatabank.org/?q=model_challenge.

It is important to keep in mind that the chosen reference structures are experimentally-determined models themselves and may have small conformational differences from the optimal model owing to differences in sample preparation, quality of experimental data restraints, and/or specifics of the model building procedure. Thus, they are not necessarily of a better quality than the submitted models. However, since the selected reference structures were the best publicly available models for each target, we considered them as the most reliable points of reference at the time of evaluation.

1.3. Participants /groups

Participation in the experiment was open to everyone. Throughout the paper we use a term ‘group’ to designate both individual researchers and multiple collaborating researchers participating in the challenge. Sixteen groups from six countries registered for the challenge and submitted 106 entries. Each group was assigned a unique three-digit identification number. The correspondence between the group IDs and group names (http://model-compare.emdatabank.org/doc/em_participants_id.htm) was concealed until after the assessors reported on the evaluation results, to ensure the unbiased evaluation.

1.4. Submissions /models

106 submissions were deposited to the EMMC challenge through the pdb_extract system (Yang et al., 2004), and were further processed using Maxit (https://sw-tools.rcsb.org/) to produce both PDBx/mmCIF and PDB format files. Each submission was originally assigned an acceptance code (e.g., emcm102_GSec) encoding the submission’s number and target name (‘GSec’ for Gamma-Secretase). The acceptance tags were used internally and do not appear in the evaluation resource. Each submission was supplemented with basic information on the modeling technique that was collected using a Drupal webform and made available via a downloadable spreadsheet.

Usually, one submission contained one model (i.e., one set of coordinates unambiguously describing the location of each atom in the protein structure). However, some submissions contained more than one model, as judged by the presence of multiple MODEL records within the uploaded file (e.g., as typically used for structures determined by NMR). The multi-model entries presented evaluation problems for some software packages (see below) and were preprocessed. If a submission contained multiple MODEL blocks with structurally equivalent coordinates (identical conformations), only the first model was evaluated. If a submission contained multiple MODEL blocks with structurally distinct conformations, it was split into separate models, and each model was evaluated separately. All in all, 142 models were evaluated.

For convenience of analysis, model identifiers (e.g., T0007EM164_2) carried encoded information about the participating group (‘164’ in the name above), the target on which the model was submitted (‘T0007’), and the consecutive number of the model submitted by this author on this target (‘2’). This naming scheme enables grouping by target (to compare models from different groups on the same target) or by participating group (to analyze models from the same group).

The model challenge guide to submitters stated that uploaded models were required to be positioned within the target map, have the same symmetry as the map, and use a pre-defined sequence/residue numbering. Even so, some submitted models did not adhere to these requirements and/or had other issues that caused problems for selected evaluation packages. For cases where the model was positioned outside of the target map, submitters complied with our requests to supply revised coordinates. Other issues encountered included incorrect polypeptide sequence and/or residue numbering, Cα-only or polyalanine-only *ab initio* models, 2-symbol chain IDs, multi-model submissions, duplicated atoms, HETATM records in place of ATOM records in the PDB format, incorrect symmetry parameters or incorrect element symbols. When a model had one or more of these issues, we still attempted evaluation with all available software packages and reported the results for the tools that were able to deal with the specific formatting issues. To ensure a more rigorous evaluation of models in future challenge rounds, we recommend implementation of an acceptance system to verify model format upon submission and report errors to the authors at the submission stage.

1.5. Schematics of the EMMC evaluation

Different levels of model structure organization (multimers, monomers, domains) require different evaluation approaches. For example, in the evaluation of model subunits (monomers or evolutionary domains), establishing the similarity of their structures to those of reference structures is of paramount interest, while in the evaluation of the whole multimeric assembly, similarity of the subunits is just one piece of the entire picture, and the principal evaluation interest may be in assessing the relative orientation of the subunits and similarity of their interfaces. Combining evaluation results at different levels of model structural granularity can help provide a well-rounded picture of model accuracy. For each level of the model structure organization, we split the analysis into parallel evaluation tracks to estimate the accuracy of models based solely on the structure of the model, or with respect to density maps, reference structures, or other submitted models (see Figure 1). Different evaluation measures are applied in different evaluation tracks, and these are discussed below.

2. Evaluation measures

2.1. General overview

Building a computational model of a protein involves assumptions, approximations, and simplifications. Thus, every model is inherently imperfect. Reliable estimation of both overall and local accuracy is critical for determining the usefulness of a model to address specific structural biology problems. How can the accuracy of a model be estimated?

2.1.1. Exclusively from coordinates—In any situation, accuracy can be estimated directly from the coordinates of the model. This approach is the only choice if no experimental structural data for the modeling target is available. Several conceptually different software tools can handle this task. One group of approaches (Chen et al., 2010; Hooft et al., 1996; Laskowski et al., 1996) focuses on validating basic stereo-chemical features of models by comparing them to geometric parameters observed in high-resolution experimental structures. We picked the most recent and comprehensive MolProbity package (Chen et al., 2010), version 4.4, to represent this group of methods in our analysis (see section 2.2 below for specifics). Methods from this group are very useful for identifying models with atypical geometric features, however they cannot distinguish models that are similar to the native structure from those that are not (in other words, ideal geometry and an almost perfect MolProbity score do not guarantee that a model corresponds to the native state).

The native (folded) state of a protein is characterized by the global minimum of free energy, and therefore close-to-native structures potentially can be differentiated from decoys by using energy terms. Since it is currently unfeasible to build a perfect energy function for complex systems such as a protein, the task of distinguishing good models (close-to-native structures) from poor models (incorrectly folded structures) has been approached by approximating the energy function with empirical molecular mechanics force fields (Lu et al., 2008; Wiederstein and Sippl, 2007; Zhou and Zhou, 2002; Zhou and Skolnick, 2011). This group of methods is represented in our evaluation by the DFIRE energy function (Zhou and Zhou, 2002). Even though multiple energy functions are practically useful in different contexts, they did not show any advantage over other scoring approaches in CASP experiments (Kryshtafovych et al., 2018b).

At the same time, machine-learning approaches combining various structural features of the models and/or energy potentials demonstrated promising results (Elofsson et al., 2018; Kryshtafovych et al., 2011; Kryshtafovych et al., 2018b). These include the recently developed ProQ series of methods (Ray et al., 2012; Uziela et al., 2016), SVMQA (Manavalan and Lee, 2017), QMEAN (Benkert et al., 2009) and ModFOLD (McGuffin et al., 2013). All of these methods are capable of generating accuracy estimates based solely on the model and that is why they are called “single-model” methods. For our evaluation, we picked two such methods: the QMEAN and ProQ3, the latest version of the original ProQ method.

2.1.2. Comparing to reference structures—If there exists a “gold standard” structure, a model can be compared to this structure and their structural agreement can be evaluated. Tools to quantify such an agreement have evolved over the years, being stimulated in particular by developments in CASP experiments (Moult et al., 2018). For the EMMC evaluation, we picked several conceptually different measures from the CASP tool chest (Kryshtafovych et al., 2016a) in order to provide different perspectives on the model accuracy (see section 2.3 below).

2.1.3. Checking model-to-map fit—A number of measures for validating fitness of EM models with respect to experimental 3DEM density maps have been developed (section 2.4 below). Newly developed measures are included along with the real space cross-correlation function (CCF) and/or the Fourier Shell Correlation (FSC) function, which are routinely used to estimate resolution of EM models (Henderson et al., 2012).

2.1.4. Agreement between models—Finally, the accuracy of a model can be estimated by comparing it to other models submitted on the target (section 2.5 below). The methods of this class are called clustering (or consensus-based) methods. According to CASP tests, Pcons (Larsson et al., 2009), ModFOLD (McGuffin et al., 2013) and Multicom (Cheng et al., 2009) methods are among the most reliable clustering methods (Kryshtafovych et al., 2018b). In our evaluation system we use a locally implemented Davis-QAconsensus method (Kryshtafovych et al., 2014), which was shown to perform on par with the best performing clustering methods. Clustering methods require model sets with many (preferably diverse) models and assume that agreement of models (or regions of models) may be an indication of correctness. In general, methods of this class perform well and are more accurate than single-model methods (Kryshtafovych et al., 2016b). However, if models are too diverse (i.e., clusters of similar models are too loose) or only a very few models are correct (i.e., the most popular clusters are dominated by poor models), these methods may generate wrong results, and the single-model methods may be a better choice. Since submitted models (except perhaps *ab initio* ones) are expected to be close to their reference structures, the clustering methods are expected to provide reliable results in the EMMC evaluation.

In the next sections we discuss collections of model accuracy measures and list software packages used in each of the four evaluation tracks mentioned above. All software packages were installed as stand-alone applications at the dedicated evaluation server, and were run locally. For convenience of referencing, software packages are highlighted in the text in bold, and evaluation measures - in bold and italic.

2.2. Evaluation based exclusively on model coordinates

Single-model accuracy evaluation n tools are software packages capable of estimating model accuracy using no other input than the model file itself. In the EMMC, we used five conceptually different packages - **MolProbity**, **PHENIX**, **DFIRE**, **ProQ3** and **QMEAN** for this task.

MolProbity (Chen et al., 2010) is an all-atom structure validation package measuring agreement of a model with geometric parameters derived from high-resolution experimental structures (2 Å or better). In the EMMC evaluation, we ran MolProbity separately on the model of the entire complex and of its representative subunits. The results are reported correspondingly under the Multimers and Monomers tabs of the model comparison website, see sections 3.4 and 3.5 below. If a submitted model did not contain hydrogen atoms, a dedicated module within the software package added them automatically. Molprobity's ***clash score*** reports the number of serious steric clashes per 1000 atoms. A clash is considered "serious" if steric overlap between any two atoms is larger than 0.4 Å. A good quality structure usually has a clash-score below 20. ***Rot-out*** reports the percentage of sidechain conformations classified as poor rotamers, from those sidechains that can be evaluated. A sidechain conformation is considered to be poor if its set of torsion angles falls outside the bounds of the rotamer definition. ***Ram-out*** quantifies the percentage of backbone conformations classified as outliers (i.e., those for which the combination of φ and ψ torsion angles is unusual), while ***Ram-fav*** quantifies percentage of the conformations in favored Ramachandran plot regions, from those residues that can be evaluated. We also reported the cumulative MolProbity score ***(MPscore)***, which combines three of the four above-mentioned statistics giving one number that approximates the crystallographic resolution at which those values would be expected (Keedy et al., 2009):

$$MPscore = 0.426 * \ln(1 + clash\_score) + 0.33 * \ln(1 + \max(0, rot\_out - 1)) + 0.25 * \ln(1 + \max(0, (100 - Ram\_fav) - 2)) + 0.5$$

The coefficients were derived from a log-linear fit to crystallographic resolution on a filtered set of PDB structures. Lower *MPscores* correspond to better structures, with scores below 3 usually indicating models of acceptable polypeptide geometry.

The **PHENIX** (**P**ython-based **H**ierarchical **EN**vironment for **I**ntegrated **X**tallography) package (Adams et al., 2010) is built around the computational crystallography toolbox library (Adams et al., 2002), enabling its extensibility. Cryo-EM specific tools have recently been added, ranging from map analysis and improvement (phenix.mtriage (Afonine et al., 2018b); phenix.auto_sharpen (Terwilliger et al., 2018)) to model validation and real-space model optimization (phenix.real_space_refine (Afonine et al., 2013; Afonine et al., 2018a)). For single-model analysis, we used the PHENIX model validation tools to quantify deviations of five geometric parameters - ***bond distances, angles, chirality, planarity*** and ***dihedral angles*** - from ideal values (Vagin et al., 2004). For each parameter, three values are provided: the RMSD, the maximum deviation (in Ångstroms for distances or degrees for angles), and the number of bonds, angles, etc. measured.

DFIRE (**D**istance-scale **F**inite **I**deal-gas **Re**ference state) (Zhou and Zhou, 2002) is a potential of mean force that can be used for structure selection and stability prediction. The all-atom residue-averaged distance-dependent potential is derived from more than one thousand non-homologous protein structures with resolution better than 2 Å. The DFIRE-based evaluation was applied here to monomeric subunits of the submitted models. Since the native (folded) state of the protein corresponds to the lowest value of the free energy, lower potential values indicate better quality structures.

ProQ3 (**Pro**tein **Q**uality) (Uziela et al., 2016) and **QMEAN** (**Q**ualitative **M**odel **E**nergy **AN**alysis) (Benkert et al., 2009) are two of the most reliable single-model accuracy assessment methods according to the CASP large-scale evaluation of model quality estimators (Kryshtafovych et al., 2011; Kryshtafovych et al., 2018b). These methods differ from others reported earlier in this section in their usage of agreement terms between observed and predicted structural features, such as e.g. secondary structure elements. These methods were developed to estimate accuracy of monomeric models, and in our evaluation system we apply them correspondingly. Both methods return global (one per model) and local (one per residue) reliability scores scaled to [0–1] range (the higher the better). ***ProQ3*** is based on a machine learning algorithm that combines knowledge-based Rosetta energy terms (Alford et al., 2017) with comparison of predicted and observed structural features, including contacts between different atom types, secondary structure and surface accessibility, and features predicted from sequence profiles. Local, per-residue accuracy is described in terms of S-score (Gerstein and Levitt, 1998), and global accuracy is a normalized sum of the local values. ***QMEAN*** is a linear combination of four statistical potential terms and two agreement terms that evaluate the consistency of structural features with sequence-based predictions. The resulting global and local scores estimate the "degree of nativeness" of structural features observed in a model, reporting the likelihood that the model (or the residue environment) is of quality comparable to a reference structure.

2.3. Fit to density maps

The quality-of-fit of model atomic coordinates to the 3DEM density maps was evaluated using three packages recently developed (or extended) for cryo-EM: **TEMPy**, **EMRinger** and **PHENIX**.

TEMPy (**T**emplate and **E**lectron **M**icroscopy comparison using **Py**thon) (Farabella et al., 2015; Vasishtan and Topf, 2011) incorporates several scoring functions for assessing model-to-density fit. For the EMMC assessment we chose four global scores - cross-correlation coefficient (***CCC***); Laplacian-filtered cross-correlation (***LAP***), mutual information (***MI***), and envelope (***ENV***); and one local score - the segment-based Manders' overlap coefficient (***SMOC***). Depending on the map type, resolution, and the extent of overlap between volumes, one scoring function may be more useful than others. To calculate the *CCC, LAP, MI* and *SMOC* scores, the model's atomic structure is first blurred to the map resolution, see (Vasishtan and Topf, 2011) for details. ***CCC*** is calculated by the array multiplication of density values at the same points in the model and target maps. ***LAP*** is computed similarly, using density maps pre-processed with a Laplacian filter. ***MI*** is a statistical measure that quantifies the extent of register between two binned densities relative to their background distributions. The *MI* score can theoretically take any positive value, with larger values corresponding to better fits. A recent study (Joseph et al., 2017) showed that the *MI* score has better discriminatory power than the cross correlation coefficients especially at intermediate-low resolutions (>6 Å) when the maps overlap partially or have significant compositional differences. ***ENV*** estimates how much of the density map is filled with atoms, and penalizes protrusions from the map envelope. Larger *ENV* values denote better fits. ***SMOC*** is a per-residue model-to-map fit measure, which calculates the Mander's overlap coefficient (Joseph et al., 2017) for overlapping residue fragments and assigns the score to the central residue in the fragment. The score is in [0–1] range[1] with higher values indicating a better fit. The *SMOC* score can be generalized for the whole structure (by averaging the per-residue scores), and we report this averaged score in our evaluation system. A more recent version of TEMPy (Joseph et al., 2017) includes new measures (e.g., a combined local mutual information and overlap score) that were shown to perform consistently across different map categories, resolutions or the extents of overlap; we plan to use these in future EMMC experiments.

EMRinger (Barad et al., 2015) estimates global and local model-map fit based on the analysis of model side-chain placement within map density. Considering all potential positions of the side-chain Cγ atom around the $\chi 1$ dihedral angle, the most preferred position is determined based on the associated cryo-EM map density. If the most preferred position appears at a non-rotameric angle, it may indicate errors in the model backbone. The

[1] The score can also take negative values when the density values in one of the maps are negative. It is usually better to shift the densities to a positive scale before calculation.

EMRinger score is calculated by assessing the statistical enrichment of peaks in rotameric positions at the map density threshold with most significant enrichment. Map resolution and the ***global EMRinger score*** are strongly correlated. This is to be expected, given that EMRinger score reports on side chain density, which is only resolvable at about 4.5 Å or better. In general, for maps at around 3.5 Å resolution or better, the minimum expected score is around 1, and a very good score around 2. Most properly optimized models score above 1.5, with some scoring above 3. The ***local EMRinger score*** is in [0–1] range and reports the fraction of residues passing the rotameric threshold in the 21-residue sliding window around the central residue.

PHENIX tools have been described in section 2.2 with regards to the individual model analysis. Here we used them to evaluate the model-to-map fit by calculating the ***overall Fourier Shell Correlation (overall FSC)*** in reciprocal Fourier space and ***per-chain box cross-correlation (box_CC)*** in real space. The real-space cross-correlation coefficients produced by TEMPY and PHENIX are highly correlated, but not identical, owing to slightly different approaches in computing the scores. Both approaches use the entire map for the calculation, but TEMPY directly calculates the product of densities in the maps (see the description above), while PHENIX first offsets density values so that the mean of the density distribution is zero, and only then takes the product of the corresponding resulting values. Higher ***box_CC*** values usually signify a better fit to map. Low values do not necessarily mean that the model does not fit the map well, but may instead indicate that there are uninterpreted map regions or poorly connecting densities (Afonine et al., 2018b). ***Overall FSC*** is calculated between the complex-valued Fourier map and model coefficients binned in resolution shells. Model coefficients are obtained by sampling on the same grid as the experimental map, applying electron form-factors and atom model parameters including coordinates, occupancies, atomic displacement parameters, chemical atom types. The resulting curve is presented as a function of spatial frequency, and is used to define the FSC-based ***resolution*** of the model at 0.5 cutoff (Rosenthal and Henderson, 2003; van Heel and Schatz, 2005).

2.4. Similarity to reference structures

For each target, reference structures were selected by the challenge committee. Different measures were used for model evaluation at different levels of model structural organization. Monomeric subunits and constitutive structural domains were evaluated using four software packages that were extensively tested in CASP – **LGA** (Local-Global Alignment), **TM** (Template Modeling), **LDDT** (Local Distance Difference Test) and **CAD** (Contact Area Difference). Each reports on local and per-residue model accuracy from different perspectives: accuracy of the backbone - ***GDT_TS*** *(Global Distance Test – Total Score)*, ***GDT_HA*** *(Global Distance Test – High Accuracy)*, ***RMSD*** *(Root Mean Square Deviation)* and ***TM*** scores; quality of model-to-target alignment - ***LGA_S*** and ***TM-align***; similarity of inter-residue distance matrices (***LDDT***), or difference in inter-residue contact areas (***CAD***). For multimeric models, we analyzed similarity of model and target interfaces with the **QS-score** (Quaternary Structure score) and **IFaceCheck** programs; additionally, completeness of models and structural proximity of corresponding residues in models and reference structures was verified with the **phenix.chain_comparison** program.

The **LGA** package (Zemla, 2003) is a superposition-based rigid-body structure comparison tool, used here to evaluate accuracy of protein backbone modeling. We ran the program in sequence-dependent and sequence-independent modes. In the sequence-dependent mode, it superimposes a model onto the target using one-to-one correspondence between residues in the compared structures (i.e., residue 1 in the model corresponds to residue 1 in the target, and so on); while in the sequence-independent mode, the algorithm first finds an optimal alignment between two compared structures. ***GDT_TS*** is the LGA's sequence-dependent score reporting the average percentage of model Cα atoms that can be superimposed with the target structure under 1, 2, 4, and 8 Å distance cutoffs. Thus, only well-modeled regions contribute to the GDT_TS score, in contrast to RMSD, where all residues contribute, including superposition outliers. The GDT_TS score is in the range [0–100] with higher scores corresponding to better fit. GDT_TS scores over 50 indicate structures with significant similarity, while scores below 25 indicate unrelated structures (poor models). Extended GDT results can be plotted as a curve showing the percentage of fit residues for distance cutoffs in the range from 0 to 10 Å, with a larger area under the curve indicating a more accurate model[2]. ***GDT_HA*** is a modification of the GDT_TS score that uses tighter distance cut-offs (0.5, 1, 2 and 4Å) and thus is better suited for the evaluation of high accuracy models, as is the case with the majority of the EMMC models built using optimization procedures. GDT_HA scores are highly correlated with the GDT_TS scores and usually 10–20 points lower for the same models. In the sequence-dependent mode, LGA also outputs ***RMSD*** between the corresponding Cα residues in the model and the target. In the sequence-independent mode, the LGA algorithm finds an optimal alignment between a model and the target by combination of the GDT-based scores (see above) with scores promoting fewer gaps in constructed alignments – see (Zemla, 2003) for details. The reported ***LGA_S*** score reflects the percentage of residues that can be superimposed under 5 Å distance cutoff. The LGA_S score is in the range [0–100], with higher scores corresponding to structural matches with a higher percentage of fit residues and longer aligned fragments. LGA_S scores are similar to the GDT_TS scores for targets where alignment errors are insignificant. In the EMMC evaluation, LGA_S is particularly useful in cases where models are out of sequence register and therefore cannot be evaluated with sequence-dependent measures.

The **TM** package (Zhang and Skolnick, 2004; Zhang and Skolnick, 2005) is another rigid-body superposition-based tool complementing LGA. In its sequence-dependent mode, it reports the ***TM-score***, which evaluates distances between aligned residues, with length-dependent normalization to reduce dependence on protein size. TM-score is well correlated with the GDT_TS score, with better models exhibiting higher scores in [0–1] range. A TM-score below 0.2 indicates that the compared structures are unrelated whereas a score higher than 0.5 indicates that they have the same fold (Xu and Zhang, 2010; Zhang and Skolnick, 2005). In sequence-independent mode, the ***TM-align*** algorithm uses heuristics based on secondary structure assignments, TM-score guided threading, and dynamic programming to identify the best structural correspondence between the model and the target. The optimal

[2] Note that the GDT plot axes in the EMMC resource are swapped compared to the traditional CASP plots so that the graphs resemble typical ROC-curve shape.

alignment is then scored (we call the result 'TM-align' to differentiate it from the sequence-dependent TM-score).

LDDT (Mariani et al., 2013) is a superposition-independent measure based on the comparison of all-atom distance maps between model and target structures. The algorithm determines the percentage of preserved distances between all pairs of atoms in the target structure that are closer in space than a predefined cutoff. The final score is the average of the percentages of the preserved distances under four distance tolerance cutoffs (0.5, 1, 2 and 4Å). The LDDT score range is [0–1].

CAD-score (Olechnovic et al., 2013) is another superposition-free measure that estimates similarity of two structures based on the differences in their residue-residue contact areas. The inter-residue contact areas can be defined for any subset of residue atoms (e.g., backbone, side-chain only). In our system we report a variant of the CAD-score that is based on comparison of contact areas for all atoms in a residue. The contact areas are calculated using the Voronyi tessellation approach in the target and the model separately, and then their differences for the same pairs of residues are summed and normalized to the [0–1] interval. Based on CASP evaluation data, the CAD-score is bell-shape distributed with around 90% of scores falling in the range [0.3; 0.7]. It is worth noting that CAD score has a desired feature of favoring models with better stereo-chemical arrangements (Olechnovic et al., 2018, submitted to Bioinformatics).

Both *LDDT* and *CAD* scores are superposition-free measures of local structure and therefore can be directly applied to assessing quality of submitted models on multi-domain targets. While rigid body superposition-based scores (e.g., GDT_TS or TM-score) are very sensitive to relative domain orientation (as superposition of two multi-domain structures is usually dominated by one of the domains) and require split of multi-domain targets into separate domains for a fair assessment, the local measures are practically insensitive to spatial inter-domain arrangements and therefore are well suited for evaluation of model quality in such cases. (Olechnovic et al., 2018, submitted to Bioinformatics).

QS-score (Bertoni et al., 2017) was applied for reference-based evaluation of multimeric structures. The score quantifies the similarity between quaternary structures in terms of shared interfacial contacts of their subunits. The package first finds the best mapping between the target and model chains using the structure symmetry, and then calculates four scores: ***QS_best*** - the fraction of interchain contacts that are shared between two structures for the best fitting interface; ***QS_global*** - the fraction of interchain contacts that are shared between two structures for all interfaces; ***RMSD*** calculated on the whole aligned structure (Cα's of all common chains); and the ***LDDT*** score described above and adopted for multimeric structures in such a way that it does not penalize for over-prediction, e.g. a tetrameric model (containing a perfect dimeric model) vs the dimeric target is giving a perfect score. QS-scores are ranked in [0–1] interval. The scores above 0.7 indicate highly similar quaternary structures, while scores below 0.3 indicate structures of low assembly similarity.

Similarly to QS-score, the **IFaceCheck** compendium of statistical measures evaluates accuracy of multimeric models based on the similarity of their interfaces (Lafita et al., 2018). The tool first clusters inter-chain contacts in the target and a model based on the identity of the interface residues (in terms of the Jaccard distance), and then reports statistics on similarity of the interface clusters in different structures. For each model interface, we calculate its similarity to the corresponding target interface in terms of the ***precision (Prec)*** = TP/(TP+FP), ***recall*** = TP/(TP+FN), ***F1-score*** = 2*Prec*Recall/(Prec+Recall) and ***Jaccard distance (Jd)*** = (FP+FN)/(TP+FP+FN), where TP is the number of interface target contacts reproduced in the model, FP is the number of model contacts not present in the target and FN is the number of target contacts missing in the model. An interface contact is defined as the distance <5Å between any two non-hydrogen atoms from residues belonging to different chains. The scores are reported for the best scoring interface from each of the corresponding interface clusters and also as the average from all pairwise scores from all interfaces in the cluster. Also, the ***interface RMSD*** between the residues belonging to target interfaces and corresponding residues in the model (not necessarily belonging to an interface) is calculated together with the ***coverage*** of the target interface residues by the modeled residues (i.e., the percentage of target interface residues used in the interface RMSD calculation).

The **phenix.chain_comparison** program was suggested as an evaluation criterion by Tom Terwilliger in order to calculate the proximity of model and target structures, once coordinates of both are optimally fit to the density. This is important when analyzing *ab initio* models, which may be incomplete, have sequence errors, or have regions of unassigned sequence. Fit to maps was ensured using phenix.get_cc_mtz_pdb and phenix.superpose_pdbs modules. The method reports the number of Cα atoms in the model within 3Å of the target (***Nclose***); number of Cαs further than 3Å (***Nfar***); the number of Cαs within 3Å of the target divided by the rmsd (***CA score***); and the percentage of Cα atoms that have the correct residue name (***Seq. match %***).

2.5. Agreement among submitted models

Davis_QAconsensus method (Kryshtafovych et al., 2014) assigns accuracy score to a model based on the average pair-wise similarity of the model to all other models submitted on that target. The method superimposes all models submitted on the target by running LGA with default parameters in the sequence dependent mode. For each model, the quality score is calculated by ***averaging the GDT_TS scores*** from all pairwise comparisons. In the local mode, per-residue scores are obtained by averaging the ***S-function-transformed distances*** (Gerstein and Levitt, 1998) between the corresponding residues in pairwise LGA superpositions of the selected model with the other models submitted on the target.

3. Presentation of the evaluation results

We generated scores for each of the submitted models using measures described in section 2. Parameters for running software packages were either suggested by committee members, software developers, or as commonly used in CASP. With such a wide variety of parameters, it was important to organize the resulting data in a way that could be readily comprehended and evaluated. Though the assessors were free to develop their own metrics for evaluations, EMMC-determined scores provided initial reference points for in-depth assessments of model accuracy.

3.1. Technical implementation of the web infrastructure

Evaluation results calculated with the above-mentioned measures were parsed, uploaded to the relational Postgresql database, and served to web users using Python and Perl/CGI scripts, Javascript, HTML, and CSS. The results are presented in form of plain text files, sortable tables, scatter plots, interactive graphs, histograms, and 3D renderings of model-target superpositions. The graphs are plotted using the c3.js/d3.js libraries. Protein molecule visualization tools use the WebGL-technology implemented in the biopv.js library (http://pv.readthedocs.io/en/v1.8.1/).

3.2. Main page

The EMMC results website (http://model-compare.emdatabank.org) provides an access to raw data and processed evaluation results for each target.

The *data repositorium* link (http://model-compare.emdatabank.org/data) takes a user to the directory containing submitted models, maps and reference structures used in the evaluation. The data in the directory are explained in the README file.

The *model info* link shows information on the details of all submitted models as provided by the authors.

The *participants* link shows the correspondence between modelers' ids and names.

The *score distributions* link is a gateway to plots showing distribution of the evaluation scores separately for *ab initio* and optimization-built models.

Structure pictographs (with target and map IDs) are gateways for browsing evaluation results for each target.

3.3. Target-specific pages

Target-specific pages show evaluation results at three levels of target /model structural organization: quaternary structure (***Multimers*** tab), tertiary structure (***Monomers*** tab) and constituting domains (Cheng et al., 2014), if applicable (***Domains*** tab). For each of the structural organization levels, a user can browse the results by checking tabs corresponding to different assessment tracks. To switch between targets, a user does not need to go back to the starting page, but instead can use the *Target* drop-down menu. A set of three *Filter by method type* checkboxes allows separate analyses of different subsets of models (*ab initio*, optimized cryo-EM models, and/or optimized other known models). User-provided information about all models submitted on the target is accessible by clicking on the 'model info' link in the upper right portion of the page (shows a sub-table of the general 'model info' table from the main page). Also, the model's method type and the name of the group's leading author are displayed when hovering the mouse over the model id in the results tables.

The *help* link provides basic information on the organization of the web resource and the evaluation measures. In addition to this, a short description of each evaluation measure is provided as a tooltip with the mouse over the score column title in the results tables.

3.4. Multimers

At the multimeric level, models are evaluated in three tracks: using reference-free statistical potentials, versus experimental cryo-EM maps and versus reference structures. Results in each of the evaluation modes are presented under separate tabs.

3.4.1. Reference-free results—The results of stereo-chemical validation of multimeric models are presented under the ***Self (reference-free)*** tab in the form of tables and histograms.

The *Scores* tab reports PHENIX-generated deviations of model geometric parameters from the values observed in ideal models, and MolProbity-based scores for the whole multimeric model (see section 2.2 for details). Clicking on the title of a column (here and everywhere in the interactive tables) resorts the table according to the selected score. MolProbity scores for separate chains are provided under the Monomers tab.

The *Histograms* tab provides binned distributions of deviations from ideal bonds, deviations from ideal angles, deviation of non-bonded distances, and atom displacement parameters (ADPs). Values in the x-axis show numbers of examples in the bins specified in the y-axis.

3.4.2. Results of the model-to-map fit assessment—The ***vs EM maps*** tab shows results of the evaluation of global and local model-to-map fitness with the tools discussed in section 2.3.

The *Global Accuracy* tab presents evaluation *Scores* table, and *Plots* of the Fourier Shell Correlation (FSC) as a function of spatial frequency (see Figure 2 for an example). Dashes in the tables (here and elsewhere) indicate that particular models could not be evaluated with the corresponding software tool. The FSC curves are built from the PHENIX output data. The FSC plot page can also be reached by clicking on any FSC value directly in the table of results (under the *Scores* tab). Another type of plots, EMRinger scores for different Electron Potential Thresholds, can be brought up by clicking on the corresponding score in the results table.

The *Local Accuracy* tab contains a summary table and interactive line plots illustrating per-residue model-to-map fit as evaluated with three software packages. The *Summary* table shows TEMPy's cumulative SMOC score and PHENIX's cumulative box_CC score for each chain in the model separately. Clicking on the values in the table brings up per-residue line plots, which can also be reached from the dedicated *TEMPy* and *PHENIX* tabs (as discussed below). The *TEMPy*, *PHENIX* and *EMRinger* tabs provide access to scatter plots based on the per-residue SMOC, box_CC and EMRinger calculations, correspondingly (see section 2.3). The plots for all three measures are conceptually similar, and can be perused in a one-model mode (see Figure 3 for an example) or a multiple-model mode (Figure 4). The one-model mode allows user to check details of a specific model for every structural chain in this model. The multiple-model mode allows comparison of models submitted on the same target by different authors. In the latter mode each model is represented by a single subunit to avoid overcrowding of graphs. In case of SMOC and box_CC scores the representative chain is the highest scoring chain, in case of EMRinger - the first chain alphabetically. Some of the actions that a user can perform on the graphs are described in the captions to Figures 3 and 4.

3.4.3. Results of the evaluation versus reference structures—

Results of the evaluation of multimers vs reference structures are presented under the ***vs Structure*** tab. Three separate sub-tabs provide sortable tables of results generated with *QS score*, *IfaceCheck* and *phenix.chain_comparison* packages (see section 2.4). While the tables are intuitive to analyze, we want to emphasize one feature that is easy to overlook in the *IFaceCheck* tab. As it was described in section 2.4, IFaceCheck first clusters the interfaces based on their similarity to each other and then operates on the representatives from each cluster. Thus, the web table shows the results for the best individual interface pair among the clusters of corresponding interfaces. The full list of interfaces included in the model and/or target clusters can be displayed below the table by clicking on the interface pair in the *Corresponding Interfaces* column.

3.5. Monomers

For monomeric evaluation, the submitted multimeric models are first split into separate chain-based models, which are then checked for similarity. All significantly different structures are evaluated separately. Models are evaluated in three tracks: using reference-free measures, versus reference structures and using all-model consensus. Since density maps are not directly involved in the evaluation of model subunits (and therefore the Monomer track does not require a separate map-related tab), it is possible to present results from all three tracks in the same table. Thus, for monomers we skip the level of track-related tabs and present all the results under two tabs corresponding to *Global* and *Local Accuracy* analyses of subunits.

3.5.1. Global accuracy—

The table with the results of overall model accuracy evaluation contains three sections corresponding to three evaluation tracks in the monomeric assessment (*Scores* tab). The *Single-model validation* section includes two subsections with scores from two knowledge-based programs-MolProbity and DFIRE, and machine learning algorithms - ProQ and QMEAN (see chapter 2.2). The *Comparison to the reference structure* section also contains two sub-sections: one reporting results from superposition-free evaluation (LDDT and CAD) and the other reporting superposition-based scores. Note that the LGA_S and TM-align columns usually contain scores for more models than the other columns due to the sequence-independent nature of their underlying algorithms (i.e., they can evaluate models with wrong or unassigned sequence register – see 2.4). The *Comparison to other submitted models* section contains scores from the DAVIS_QAconsensus method (section 2.5), with higher scores indicating higher level of model similarity to other models.

The *Plots* tab show results of the extended GDT analysis (see 2.4). The GDT plots (Figure 5) show percentage of residues in the model that can be superimposed onto the target under the specified residue-residue distance cutoff. The higher the area under the curve – the better the model. An ideal model would be represented by a curve going straight up and then staying horizontally across the whole range of distance cutoffs. Graphs are interactive so that the lines can be switched on and off, and the underlying scores can be displayed using the techniques described in section 3.4.2.

3.5.2. Local accuracy—

There are five sub-tabs under the *Local Accuracy* tab, each corresponding to the selected evaluation package (*LGA*, *LDDT*, *ProQ*, *QMEAN* and *DAVIS_QA*). Clicking on each of the tabs shows color-coded bars illustrating per-residue accuracy of models according to the selected evaluation approach. For example, clicking on the LGA tab shows Cα-Cα distances between corresponding residues in models and the target after their optimal LGA superposition (Figure 6A), while clicking on the LDDT tab gives per-residue LDDT scores from the comparison of model and target distance patterns in the vicinity of the selected residue. Clicking on the color-coded bar shows structural LGA superposition of the model and the target colored the same way as the underlying bar (Figure 6B). For convenience, we always show the reference structure superimposed with the model using LGA next to 3D renderings of models, even for the approaches not using model-target superposition. The *DAVIS_QA* tab shows bar plots illustrating similarity of each of the models to all other models submitted on the target (Figure 7).

3.6. Domains

If the monomeric unit consists of several structural domains, the models are additionally evaluated at the level of domains. Targets are split into domains by consulting the DomainParser (Guo et al., 2003), DDomain2 (Zhou et al., 2007) programs, and the ECOD (Cheng et al., 2014) database of structural domains. Organization of the web resource for the *Domains* evaluation is similar to that for the *Monomers*.

4. Brief analysis of the results

In this section we present statistical analyses of scores obtained from evaluation of the submitted models and calculate correlations between them. We do not attempt to rank models or methods, leaving this task to the challenge assessors.

Distributions of selected scores for different types of models (*ab initio* and optimized) are presented as box plots in Figure 8 (the plots for all measures used in the EMMC evaluation are provided in Figures 1–3 of Ref. (Kryshtafovych et al., 2018a) and also accessible from the main Results web page through the *score distributions* link http://model-compare.emdatabank.org/em_score_boxplots.cgi). The box plots clearly show that score distributions on models built starting from reference models versus *ab initio* are very different. In vast majority of cases, the inter-quartile ranges (containing middle 50% of the data) even do not overlap. This highlights some of the challenges assessing *ab initio* models, which are often incomplete in structure and/or sequence. Figures 4 and 5 in (Kryshtafovych et al., 2018a) show distribution of evaluation scores when all models are grouped together (i.e., without splitting them into *ab initio* and optimization categories). Outliers in the graphs for the complete set of models are almost all ab-initio models.

To investigate the similarity of scores, we calculated the correlation between each pair of scores used in this evaluation. Since score distributions in general do not follow a Gaussian pattern, we used Spearman rank correlation analysis, which is suitable for comparing both normal and non-normal distributions (Altman and Krzywinski, 2015). Figures 9 and 10 show Spearman correlation coefficients between scores used to evaluate the accuracy of models in multimeric and monomeric regimes, respectively. As can be seen from the figures, some scores are closer to each other than others. For example, reference-based scores in both figures are much more similar to each other than to the reference-free scores. Similar methods tend to have high correlation (e.g., within superposition-based scores, local structure-based scores or interface similarity scores). MolProbity scores have very weak correlation to other types of scores, confirming that better model stereochemistry in general does not guarantee better similarity of the model to reference structures or a better fit to the corresponding map. The correlations above were calculated on scores for all submitted models. When we compare these correlation coefficients with the coefficients calculated for optimized models only (Figures S1 and S2 in Supplementary material), we see that the correlation tables are quite similar.

Conclusions

The paper provides a description of the evaluation system and the web resource for assessment of models submitted to the second cryo-EM model challenge. The resource may be useful to authors, assessors and research scientists for analyzing model details, estimating goodness of model-to-map fit and comparing models with each other and to reference structures.

For evaluation of the EMMC models we selected state-of-the art methods for assessing accuracy of models at different levels of granularity - whole multimeric structure, constitutive monomeric units, sub-domains and interfaces. Scores from all evaluation methods are posted on the web and statistically analyzed in this paper. In particular, we studied distributions of the scores and their similarity in ranking models. These data may be useful for specialists developing and benchmarking methods for building models from cryo-EM maps, and for developing new deposition guidelines for cryo-EM models and maps.

The web infrastructure of the second model challenge is publicly available at http://model-compare.emdatabank.org.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We are grateful to Pavel Afonine and Tom Terwilliger (PHENIX), Benjamin Barad (EMRinger), Maya Topf and Agnel P. Joseph (TEMPy) for providing their software for the cryo-EM model challenge evaluation, for the help in adjusting the software to the needs of the EMMC evaluation and for discussing details of their tools. Big thanks go to Bohdan Monastyrskyy (UC Davis) and Kenneth Dalenberg (Rutgers) for their help with setting up the web server. Authors acknowledge support of the National Institute of General Medical Sciences grant R01GM079429 to WC, and grant P01GM063210 to PDA.

References

Adams PD, Grosse-Kunstleve RW, Hung LW, Ioerger TR, McCoy AJ, Moriarty NW, Read RJ, Sacchettini JC, Sauter NK, Terwilliger TC, 2002 PHENIX: building new software for automated crystallographic structure determination. Acta Crystallogr D Biol Crystallogr 58, 1948–54. [PubMed: 12393927]

Adams PD, Afonine PV, Bunkoczi G, Chen VB, Davis IW, Echols N, Headd JJ, Hung LW, Kapral GJ, Grosse-Kunstleve RW, McCoy AJ, Moriarty NW, Oeffner R, Read RJ, Richardson DC, Richardson JS, Terwilliger TC, Zwart PH, 2010 PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr D Biol Crystallogr 66, 213–21. [PubMed: 20124702]

Afonine PV, Grosse-Kunstleve RW, Adams PD, Urzhumtsev A, 2013 Bulk-solvent and overall scaling revisited: faster calculations, improved results. Acta Crystallogr D Biol Crystallogr 69, 625–34. [PubMed: 23519671]

Afonine PV, Poon BK, Read RJ, Sobolev OV, Terwilliger TC, Urzhumtsev A, Adams PD, 2018a Real-space refinement in Phenix for cryo-EM and crystallography. bioRxiv

Afonine PV, Klaholz BP, Moriarty NW, Poon BK, Sobolev OV, Terwilliger TC, Adams PD, Urzhumtsev A, 2018b New tools for the analysis and validation of Cryo-EM maps and atomic models. bioRxiv

Alford RF, Leaver-Fay A, Jeliazkov JR, O'Meara MJ, DiMaio FP, Park H, Shapovalov MV, Renfrew PD, Mulligan VK, Kappel K, Labonte JW, Pacella MS, Bonneau R, Bradley P, Dunbrack RL, Jr., Das R, Baker D, Kuhlman B, Kortemme T, Gray JJ, 2017 The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. Journal of Chemical Theory and Computation 13, 3031–3048. [PubMed: 28430426]

Altman N, Krzywinski M, 2015 Association, correlation and causation. Nat Methods 12, 899–900. [PubMed: 26688882]

Barad BA, Echols N, Wang RY, Cheng Y, DiMaio F, Adams PD, Fraser JS, 2015 EMRinger: side chain-directed model and map validation for 3D cryo-electron microscopy. Nat Methods 12, 943–6. [PubMed: 26280328]

Benkert P, Kunzli M, Schwede T, 2009 QMEAN server for protein model quality estimation. Nucleic Acids Res 37, W510–4. [PubMed: 19429685]

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE, 2000 The Protein Data Bank. Nucleic Acids Res 28, 235–42. [PubMed: 10592235]

Bertoni M, Kiefer F, Biasini M, Bordoli L, Schwede T, 2017 Modeling protein quaternary structure of homo-and hetero-oligomers beyond binary interactions by homology. Sci Rep 7, 10480. [PubMed: 28874689]

Chen VB, Arendall WB, 3rd, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS, Richardson DC, 2010 MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr D Biol Crystallogr 66, 12–21. [PubMed: 20057044]

Cheng H, Schaeffer RD, Liao Y, Kinch LN, Pei J, Shi S, Kim BH, Grishin NV, 2014 ECOD: an evolutionary classification of protein domains. PLoS Comput Biol 10, e1003926. [PubMed: 25474468]

Cheng J, Wang Z, Tegge AN, Eickholt J, 2009 Prediction of global and local quality of CASP8 models by MULTICOM series. Proteins 77 Suppl 9, 181–4.

Elofsson A, Joo K, Keasar C, Lee J, Maghrabi AHA, Manavalan B, McGuffin LJ, Menendez Hurtado D, Mirabello C, Pilstal R, Sidi T, Uziela K, Wallner B, 2018 Methods for estimation of model accuracy in CASP12. Proteins 86 Suppl 1, 361–373. [PubMed: 28975666]

Farabella I, Vasishtan D, Joseph AP, Pandurangan AP, Sahota H, Topf M, 2015 : a Python library for assessment of three-dimensional electron microscopy density fits. J Appl Crystallogr 48, 1314–1323. [PubMed: 26306092]

Gerstein M, Levitt M, 1998 Comprehensive assessment of automatic structural alignment against a manual standard, the scop classification of proteins. Protein Sci 7, 445–56. [PubMed: 9521122]

Guo JT, Xu D, Kim D, Xu Y, 2003 Improving the performance of DomainParser for structural domain partition using neural network. Nucleic Acids Res 31, 944–52. [PubMed: 12560490]

Henderson R, Sali A, Baker ML, Carragher B, Devkota B, Downing KH, Egelman EH, Feng Z, Frank J, Grigorieff N, Jiang W, Ludtke SJ, Medalia O, Penczek PA, Rosenthal PB, Rossmann MG, Schmid MF, Schroder GF, Steven AC, Stokes DL, Westbrook JD, Wriggers W, Yang H, Young J, Berman HM, Chiu W, Kleywegt GJ, Lawson CL, 2012 Outcome of the first electron microscopy validation task force meeting. Structure 20, 205–14. [PubMed: 22325770]

Hooft RW, Vriend G, Sander C, Abola EE, 1996 Errors in protein structures. Nature 381, 272. [PubMed: 8692262]

Joseph AP, Lagerstedt I, Patwardhan A, Topf M, Winn M, 2017 Improved metrics for comparing structures of macromolecular assemblies determined by 3D electron-microscopy. J Struct Biol 199, 12–26. [PubMed: 28552721]

Keedy DA, Williams CJ, Headd JJ, Arendall WB, 3rd, Chen VB, Kapral GJ, Gillespie RA, Block JN, Zemla A, Richardson DC, Richardson JS, 2009 The other 90% of the protein: assessment beyond the Calphas for CASP8 template-based and high-accuracy models. Proteins 77 Suppl 9, 29–49. [PubMed: 19731372]

Kryshtafovych A, Fidelis K, Tramontano A, 2011 Evaluation of model quality predictions in CASP9. Proteins 79 Suppl 10, 91–106. [PubMed: 21997462]

Kryshtafovych A, Monastyrskyy B, Fidelis K, 2014 CASP prediction center infrastructure and evaluation measures in CASP10 and CASP ROLL. Proteins 82 Suppl 2, 7–13. [PubMed: 24038551]

Kryshtafovych A, Monastyrskyy B, Fidelis K, 2016a CASP11 statistics and the prediction center evaluation system. Proteins 84 Suppl 1, 15–9. [PubMed: 26857434]

Kryshtafovych A, Adams P, Lawson C, Chiu W, 2018a Distribution of evaluation scores for the models submitted to the Second Cryo-EM Model Challenge. Data in Brief Submitted

Kryshtafovych A, Monastyrskyy B, Fidelis K, Schwede T, Tramontano A, 2018b Assessment of model accuracy estimations in CASP12. Proteins 86 Suppl 1, 345–360. [PubMed: 28833563]

Kryshtafovych A, Barbato A, Monastyrskyy B, Fidelis K, Schwede T, Tramontano A, 2016b 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, 349–69.

Lafita A, Bliven S, Kryshtafovych A, Bertoni M, Monastyrskyy B, Duarte JM, Schwede T, Capitani G, 2018 Assessment of protein assembly prediction in CASP12. Proteins 86 Suppl 1, 247–256. [PubMed: 29071742]

Larsson P, Skwark MJ, Wallner B, Elofsson A, 2009 Assessment of global and local model quality in CASP8 using Pcons and ProQ. Proteins

Laskowski RA, Rullmannn JA, MacArthur MW, Kaptein R, Thornton JM, 1996 AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. J Biomol NMR 8, 477–86. [PubMed: 9008363]

Lu M, Dousis AD, Ma J, 2008 OPUS-PSP: an orientation-dependent statistical all-atom potential derived from side-chain packing. J Mol Biol 376, 288–301. [PubMed: 18177896]

Manavalan B, Lee J, 2017 SVMQA: support-vector-machine-based protein single-model quality assessment. Bioinformatics 33, 2496–2503. [PubMed: 28419290]

Mariani V, Biasini M, Barbato A, Schwede T, 2013 lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. Bioinformatics 29, 2722–8. [PubMed: 23986568]

McGuffin LJ, Buenavista MT, Roche DB, 2013 The ModFOLD4 server for the quality assessment of 3D protein models. Nucleic Acids Res 41, W368–72. [PubMed: 23620298]

Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A, 2018 Critical assessment of methods of protein structure prediction (CASP)-Round XII. Proteins 86 Suppl 1, 7–15. [PubMed: 29082672]

Olechnovic K, Kulberkyte E, Venclovas C, 2013 CAD-score: a new contact area difference-based function for evaluation of protein structural models. Proteins 81, 149–62. [PubMed: 22933340]

Ray A, Lindahl E, Wallner B, 2012 Improved model quality assessment using ProQ2. BMC Bioinformatics 13, 224. [PubMed: 22963006]

Rosenthal PB, Henderson R, 2003 Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. J Mol Biol 333, 721–45. [PubMed: 14568533]

Terwilliger TC, Sobolev O, Afonine PV, Adams PD, 2018 Automated map sharpening by maximization of detail and connectivity. bioRxiv 10.1101/247049.

Uziela K, Shu N, Wallner B, Elofsson A, 2016 ProQ3: Improved model quality assessments using Rosetta energy terms. Sci Rep 6, 33509. [PubMed: 27698390]

Vagin AA, Steiner RA, Lebedev AA, Potterton L, McNicholas S, Long F, Murshudov GN, 2004 REFMAC5 dictionary: organization of prior chemical knowledge and guidelines for its use. Acta Crystallogr D Biol Crystallogr 60, 2184–95. [PubMed: 15572771]

van Heel M, Schatz M, 2005 Fourier shell correlation threshold criteria. J Struct Biol 151, 250–62. [PubMed: 16125414]

Vasishtan D, Topf M, 2011 Scoring functions for cryoEM density fitting. J Struct Biol 174, 333–43. [PubMed: 21296161]

Wiederstein M, Sippl MJ, 2007 ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Res 35, W407–10. [PubMed: 17517781]

Xu J, Zhang Y, 2010 How significant is a protein structure similarity with TM-score = 0.5? Bioinformatics 26, 889–95. [PubMed: 20164152]

Yang H, Guranovic V, Dutta S, Feng Z, Berman HM, Westbrook JD, 2004 Automated and accurate deposition of structures solved by X-ray diffraction to the Protein Data Bank. Acta Crystallogr D Biol Crystallogr 60, 1833–9. [PubMed: 15388930]

Zemla A, 2003 LGA: A method for finding 3D similarities in protein structures. Nucleic Acids Res 31, 3370–4. [PubMed: 12824330]

Zhang Y, Skolnick J, 2004 Scoring function for automated assessment of protein structure template quality. Proteins 57, 702–10. [PubMed: 15476259]

Zhang Y, Skolnick J, 2005 TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 33, 2302–9. [PubMed: 15849316]

Zhou H, Zhou Y, 2002 Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci 11, 2714–26. [PubMed: 12381853]

Zhou H, Skolnick J, 2011 GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction. Biophys J 101, 2043–52. [PubMed: 22004759]

Zhou H, Xue B, Zhou Y, 2007 DDOMAIN: Dividing structures into domains using a normalized domain-domain interaction profile. Protein Sci 16, 947–55. [PubMed: 17456745]

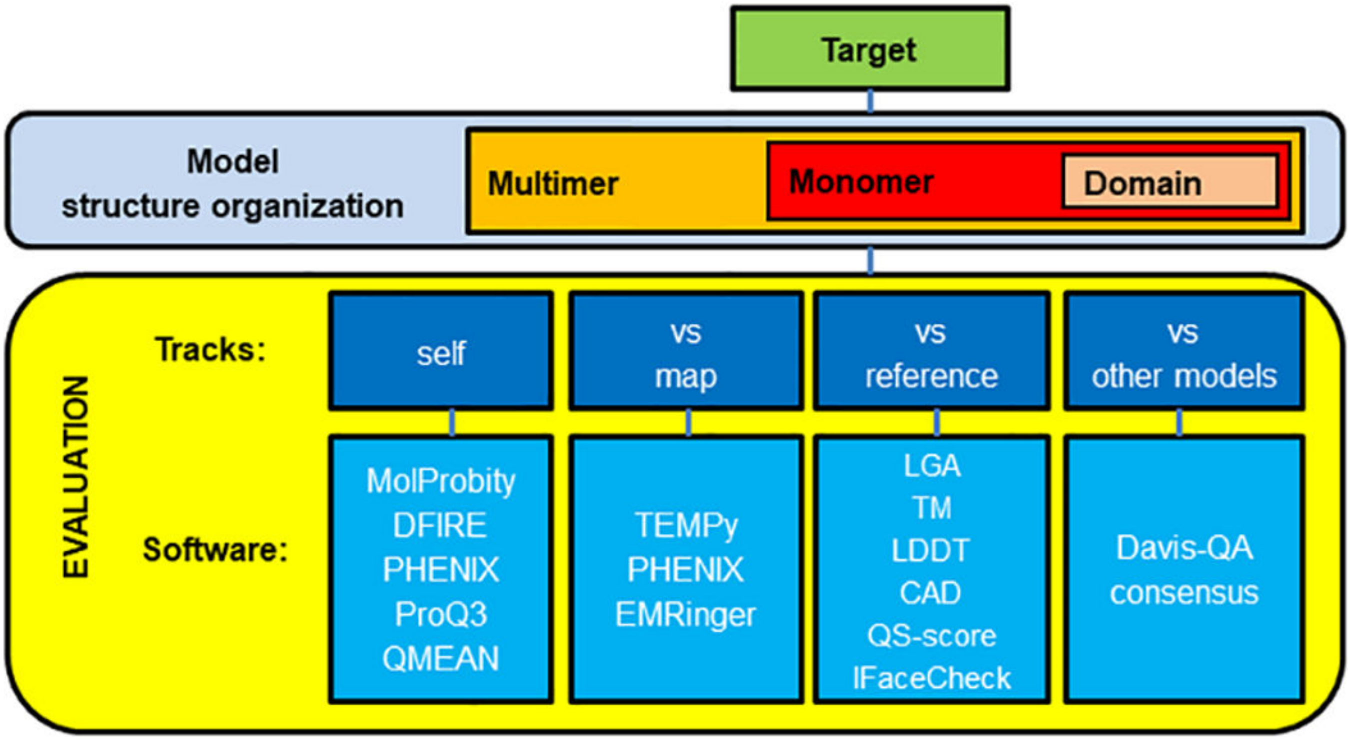


Figure 1.
General schema of the cryo-EM model challenge evaluation system.

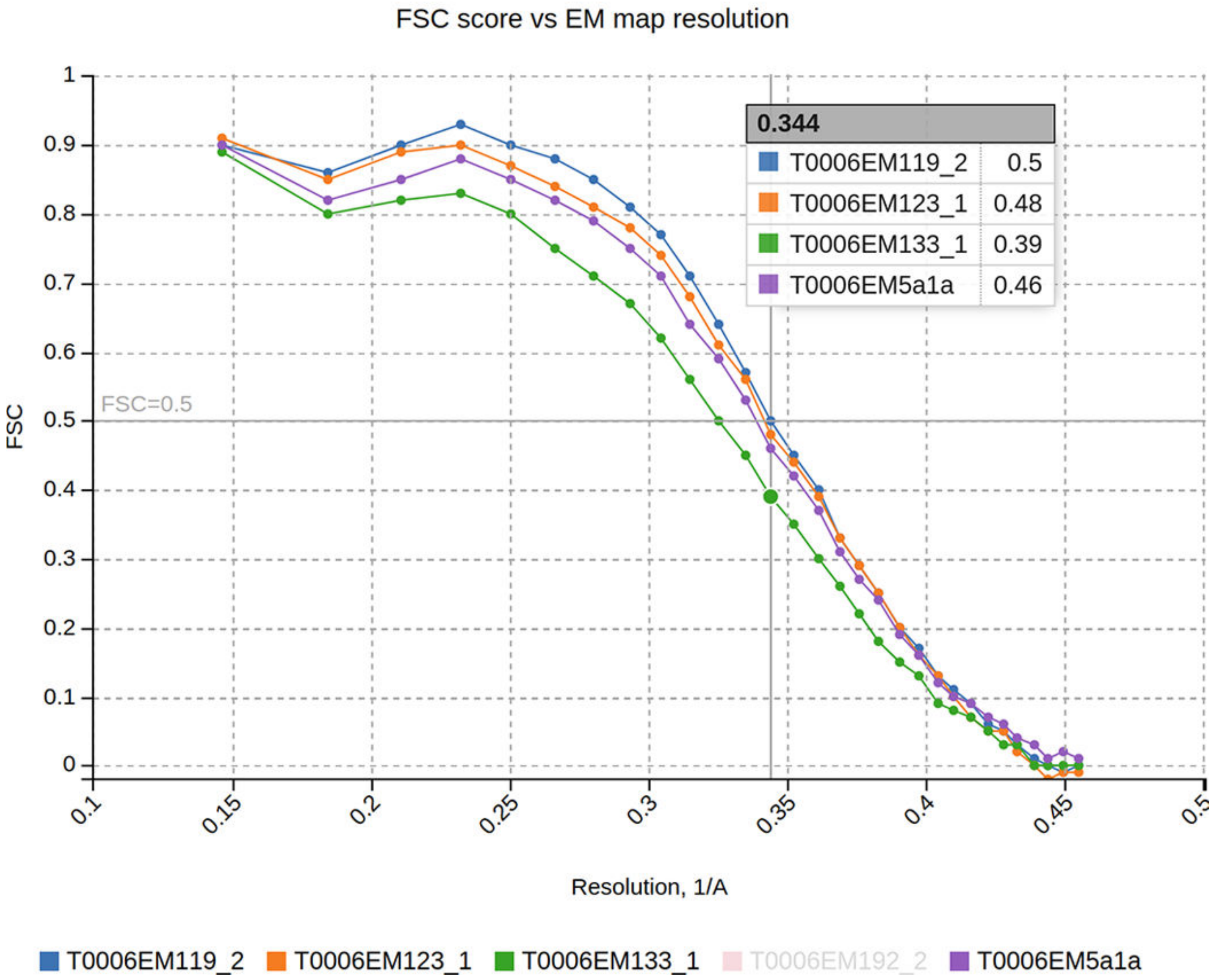
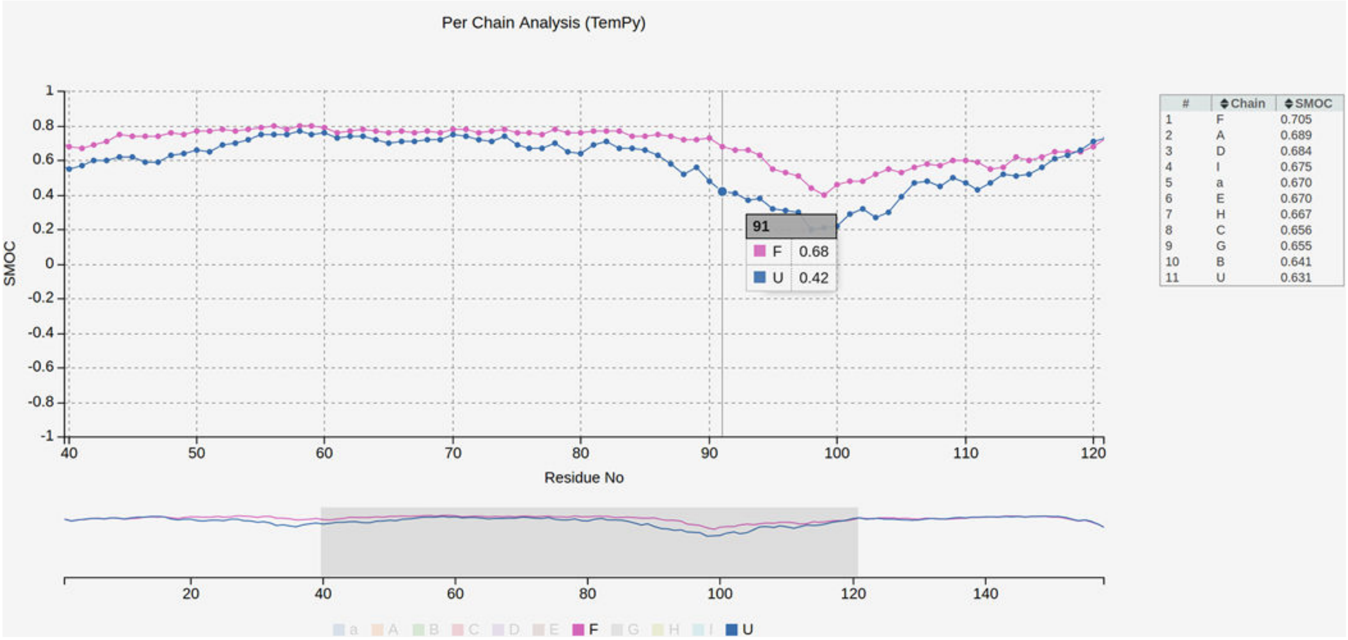


Figure 2.

FSC curves plotted for models of β-Galactosidase (T0006, map "B" emd_2984). Placing the cursor over the line shows y-axis values for each curve at the selected x-axis value (as shown in the example for x=0.344). Horizontal lines corresponding to FSC values of 0.5 and 0.143 are drawn for reference. X-values for all curves at y=0.5 can also be found in the results table (Multimers -> vs EM maps -> Global accuracy -> Scores -> Phenix -> Resol. (FSC=0.5)). Clicking on the model label (under graph) hides the corresponding curve and greys the group name (as shown here for model T0006EM192_2); clicking on the greyed model label makes the curve visible again. Placing the cursor over the model label (under graph) highlights the curve for the selected model and greys out other curves.



#	⬍Chain	⬍SMOC
1	F	0.705
2	A	0.689
3	D	0.684
4	I	0.675
5	a	0.670
6	E	0.670
7	H	0.667
8	C	0.656
9	G	0.655
10	B	0.641
11	U	0.631

Figure 3.

A plot showing local per-residue model-to-map fit based on the TEMPy’s per-residue SMOC scores for a selected model (T0001EM133_1) submitted on the TMV target. Cumulative scores for all chains of the selected model are shown in the table to the right of the plot. By default, the plot is displayed for the chain with the best cumulative SMOC score (all other chains are greyed out). To see per-residue data for other chains of the same model, click on the chain IDs at the bottom of the graph (the provided screenshot shows the result of clicking on chain ‘U’). Moving the mouse over the name of one of the selected chains dims other lines in the graph. Mouse over the line shows data values for each curve (SMOC values for residue #91 in the example). The graph can be explored in more details by selecting a specific region on the lower line-only graph. Clicking on the starting residue of the region and drugging the mouse to the last residue highlights the desired region in the lower graph and rescales the upper plot accordingly (region 40–120 in the example).

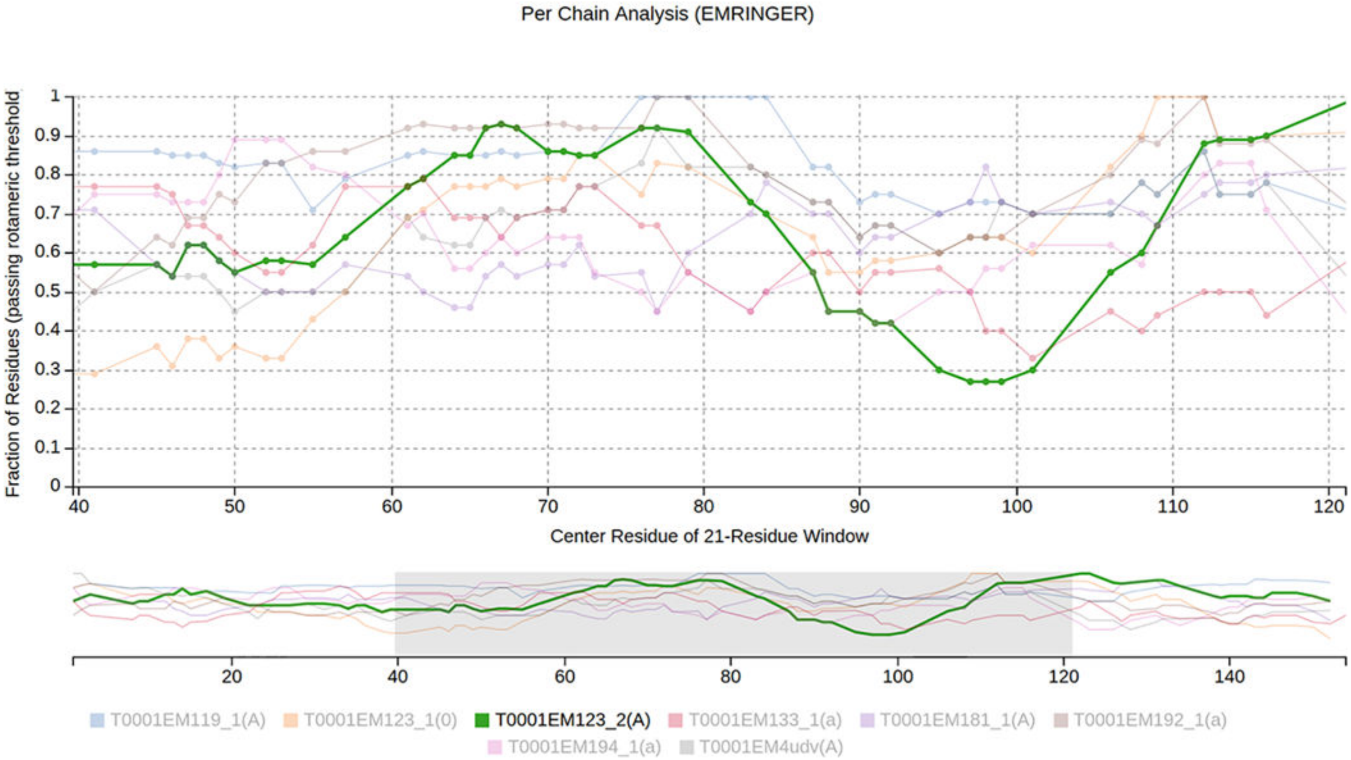


Figure 4.

An example of the local model-to-map fit graph based on the EMRinger per-residue scores (fraction of residues passing the rotameric threshold) for all models submitted on the Tobacco Mosaic Virus target (T0001). To see the plot with all models at once, a user has to select the topmost option '_all_' in the 'Model' dropdown menu. In order not to overcrowd the plot, only one chain from each model is shown (name of the selected chain is provided in parenthesis next to the model name below the graph). Mouse over the line shows data values for each curve at the selected x-value (as shown in Figures 2 and 3). Mouse over the model name (T0001EM123_2(A) in our example) highlights this model and dims all other lines in the graph. Clicking on the model name turns the line invisible and greys the model name; clicking on the greyed model name makes the line visible again. Specific region of the graph can be explored in more details using the procedure described in the Figure 3 caption (region 40–120 is selected in the current example).

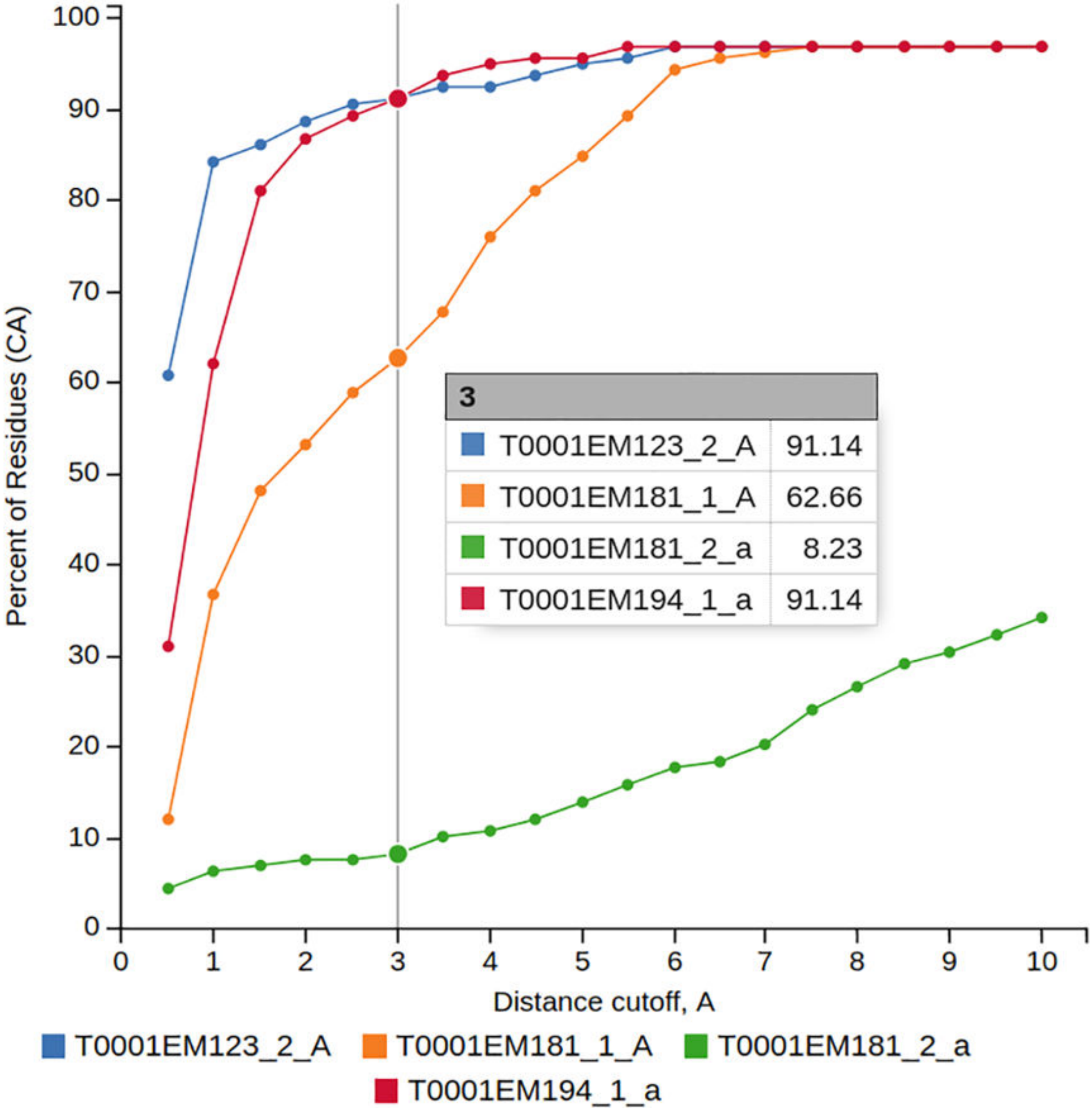
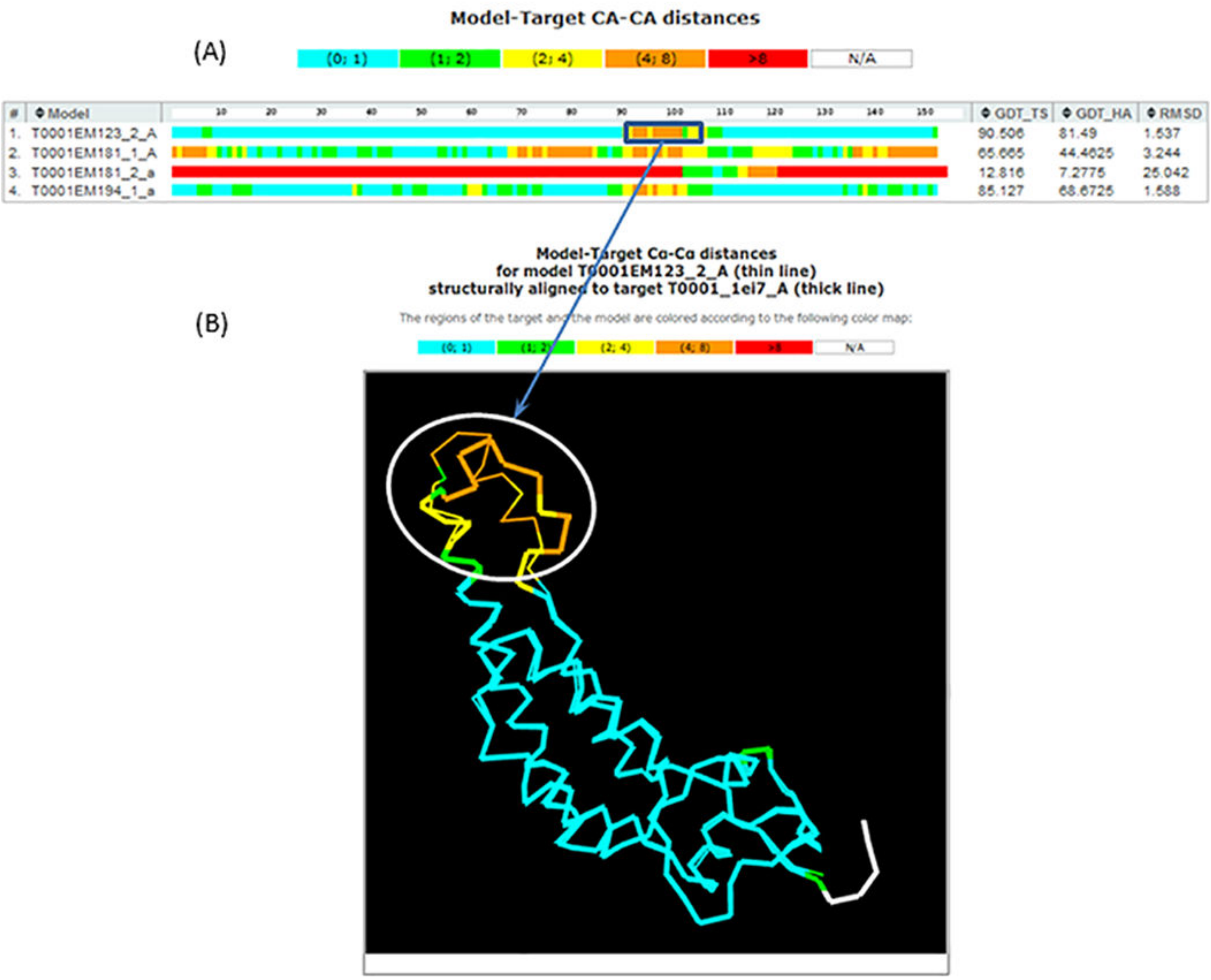


Figure 5.

Detailed GDT analysis plot of ab initio models submitted on the Tobacco Mosaic Virus target (T0001). Two top curves correspond to better models, for which 91% of Cα atoms are within 3 Å of the corresponding backbone atoms of the reference structure 1ei7; a model corresponding to the middle curve has 62% of Cα atoms within 3Å of the target.

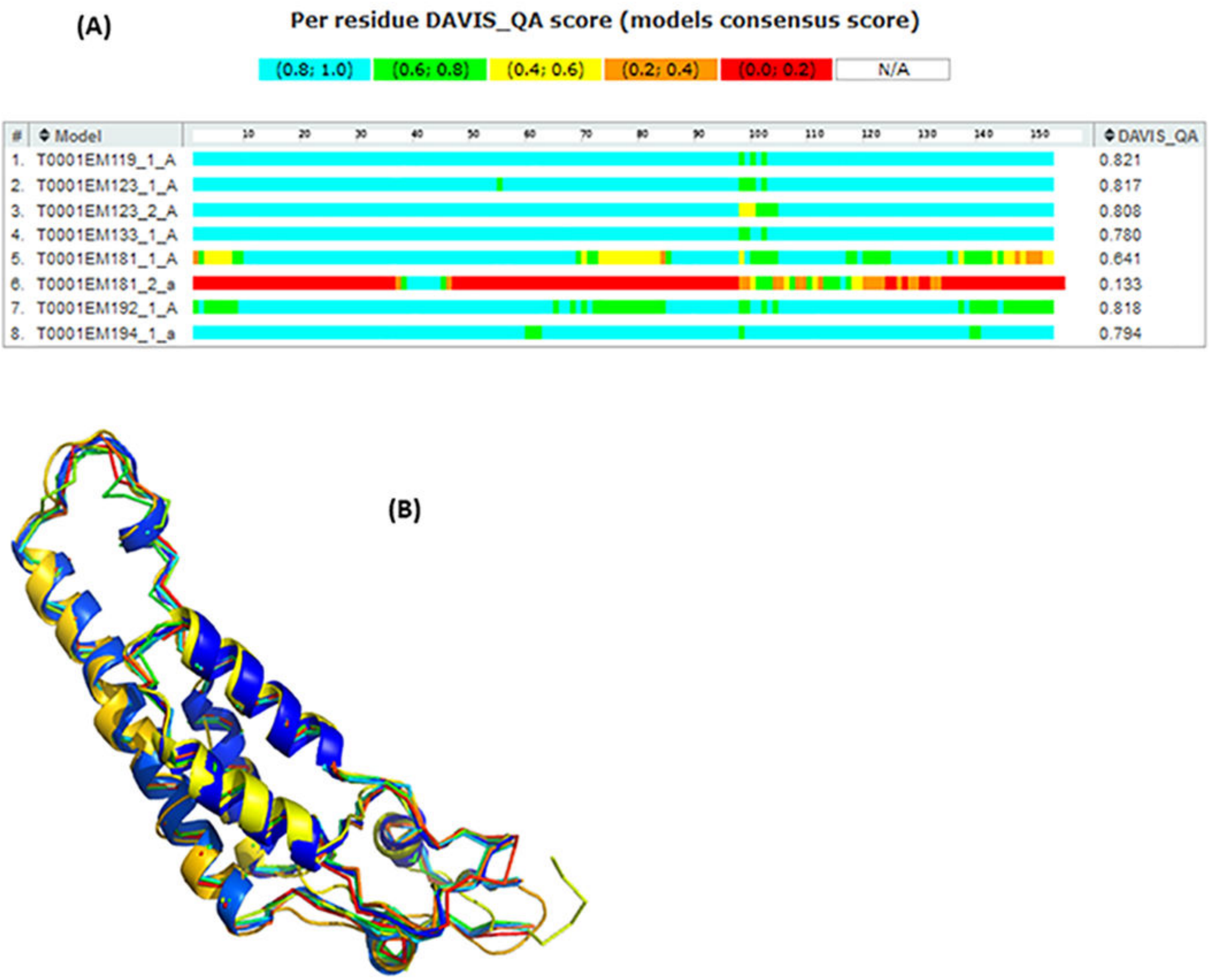

Model-Target CA-CA distances
(A)
(0; 1)
(1; 2)
(2; 4)
(4; 8)
>8
N/A

#	◆Model	10 20 30 40 50 60 70 80 90 100 110 120 130 140 150	◆GDT_TS	◆GDT_HA	◆RMSD
1.	T0001EM123_2_A		90.506	81.49	1.537
2.	T0001EM181_1_A		65.665	44.4625	3.244
3.	T0001EM181_2_a		12.816	7.2775	25.042
4.	T0001EM194_1_a		85.127	68.6725	1.588

Model-Target Cα-Cα distances
for model T0001EM123_2_A (thin line)
structurally aligned to target T0001_1ei7_A (thick line)
(B)
The regions of the target and the model are colored according to the following color map:
(0; 1)
(1; 2)
(2; 4)
(4; 8)
>8
N/A

Figure 6.

Local accuracy of the ab initio models featured in Figure 5 in terms of Cα-Cα distances between the corresponding residues in models and the target, after their optimal LGA superposition (the LGA tab). (A) Bar plots showing proximity of models to the target 1ei7. The best model according to the global GDT_TS score, T0001EM123_2_A, has a large stretch of residues being closer than 1 Å to the corresponding target residues (teal); the biggest deviation of the model from the target structure is in the region 93–100 (boxed in the plot), where the deviation reaches values in the 4–8 Å range (orange). Clicking on the model-specific colored bar brings up an LGA-based superposition of the selected model and the reference structure. Superposition in panel (B) clearly shows that the sub-par modeled region in panel A corresponds to the helix-loop-helix region in the target. The coloring keys in both panels are the same and shown above the graphs.



Per residue DAVIS_QA score (models consensus score)

(0.8; 1.0) (0.6; 0.8) (0.4; 0.6) (0.2; 0.4) (0.0; 0.2) N/A

#	◆ Model	◆ DAVIS_QA
1.	T0001EM119_1_A	0.821
2.	T0001EM123_1_A	0.817
3.	T0001EM123_2_A	0.808
4.	T0001EM133_1_A	0.780
5.	T0001EM181_1_A	0.641
6.	T0001EM181_2_a	0.133
7.	T0001EM192_1_A	0.818
8.	T0001EM194_1_a	0.794

Figure 7.

Local and overall similarity of submitted models. Panel (A) shows per-residue scores (illustrated by different colors in bar plots) and overall consensus Davis_QA score (number next to the bar) for representative subunits of submitted models, reflecting their similarity. Panel (B) shows Pymol rendering of the superposition of seven out of eight models shown in panel (A) onto the eighth model (T0001EM119_1_A, blue) displaying the largest similarity to other models submitted on the target. Model EM181_2 is a polyalanine CA-only model and not displayed in the figure. The second least similar model EM181_1_A is shown in yellow (cartoon). All other models are displayed as backbone traces.

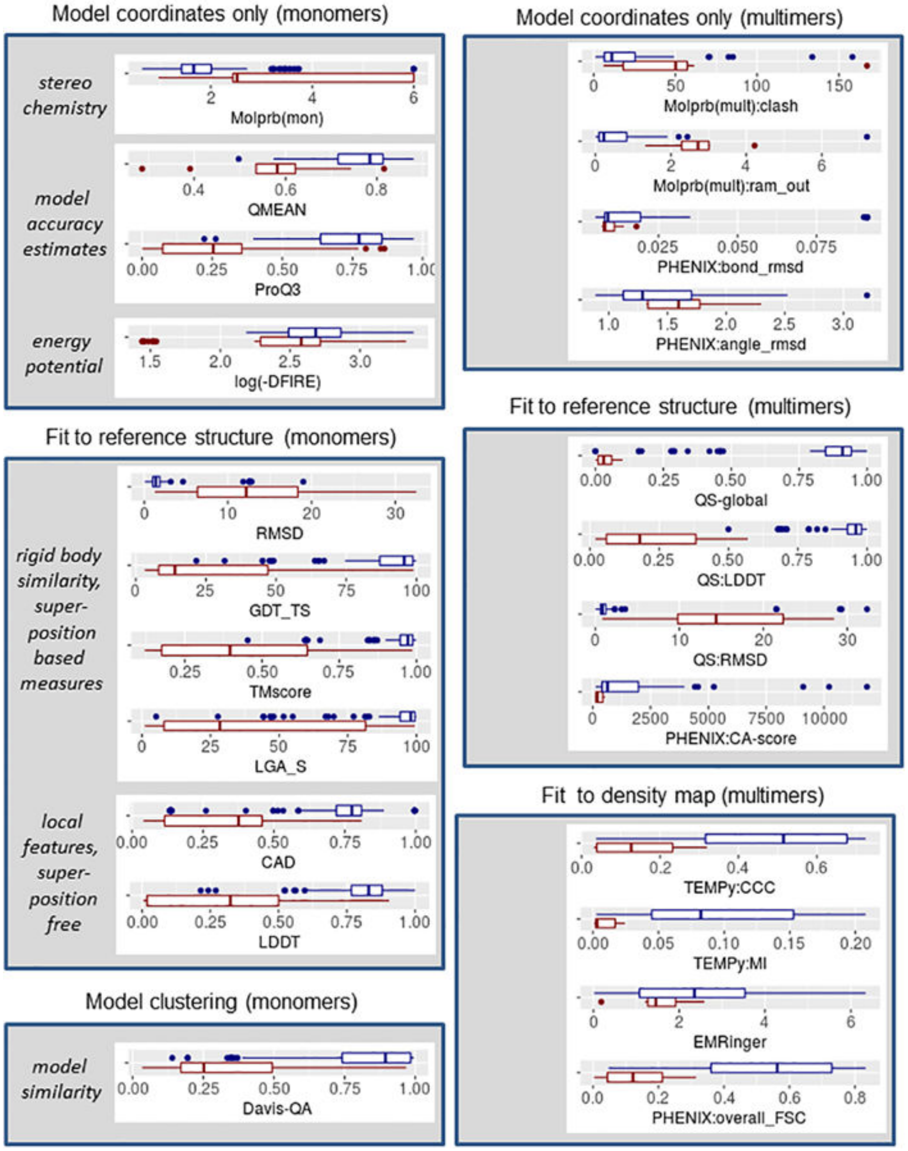
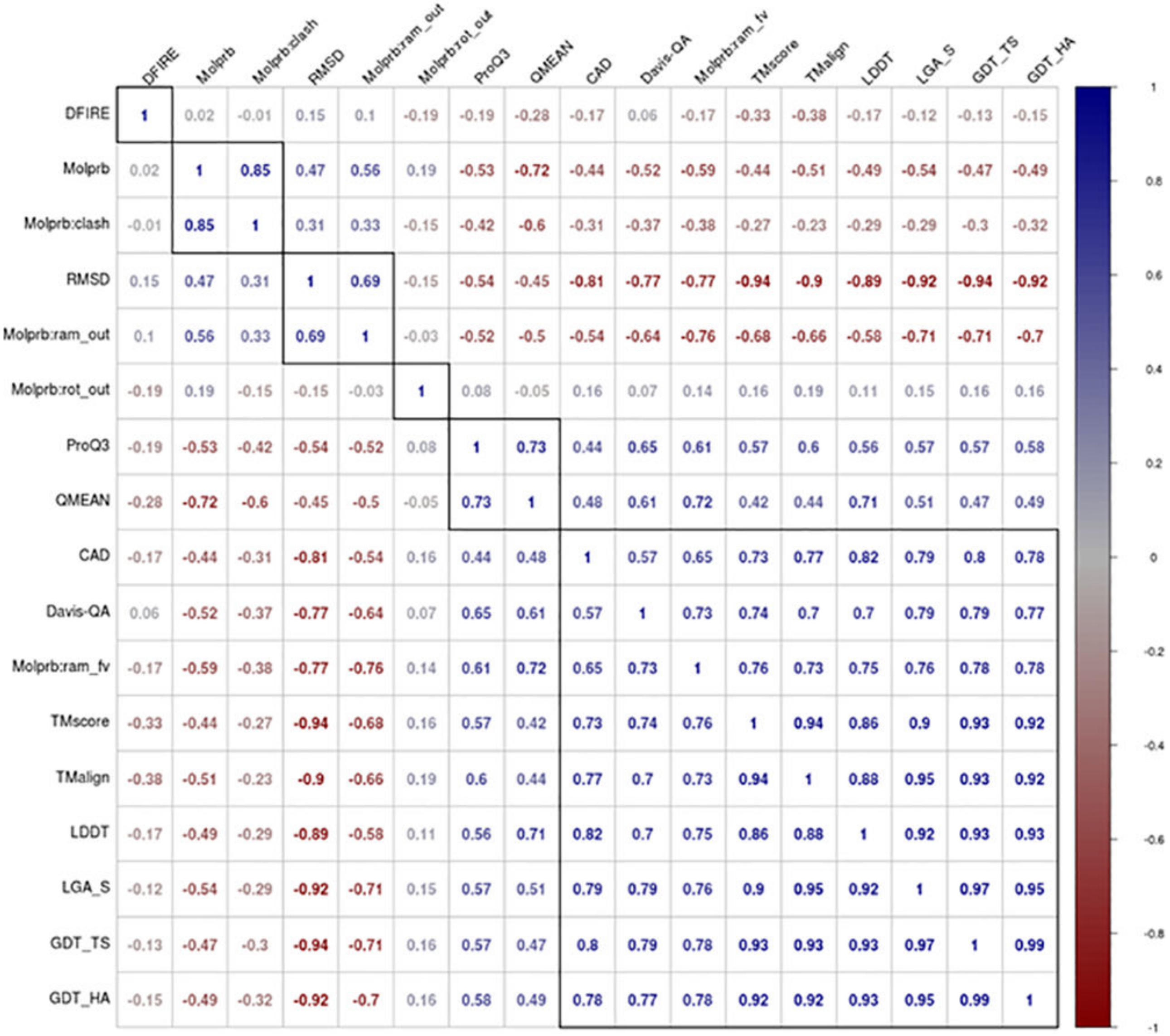


Figure 8.

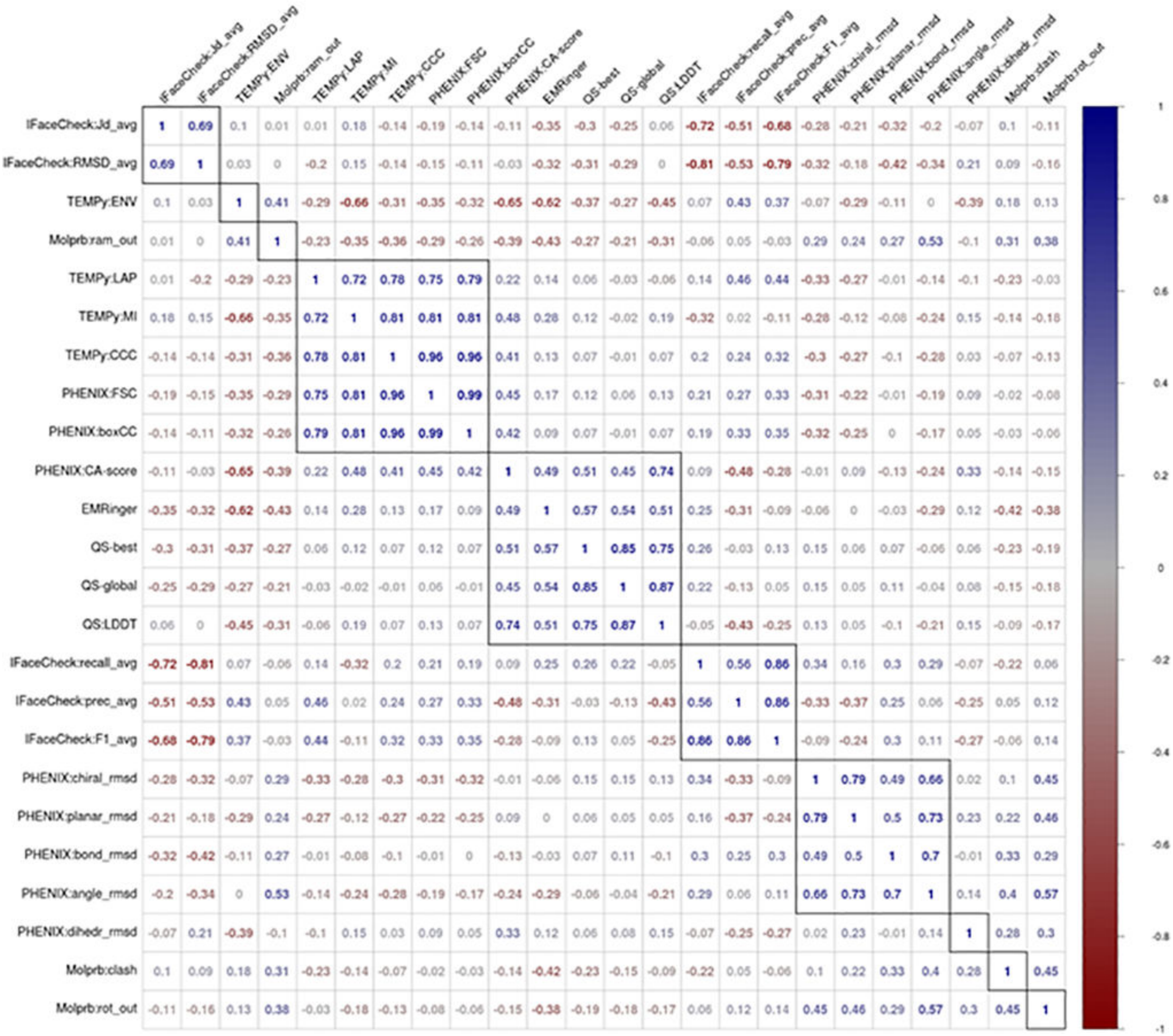
Distribution of selected evaluation scores for different types of models. Score plots for conceptually similar measures are clustered together (e.g., upper left box encompasses measures used in evaluation of monomers based on coordinates only). For each measure (specified in the x-axis title), a blue boxplot shows the score distribution for models built starting from reference structure, while a red boxplot -for models built ab initio. Left set of boxplots shows scores from monomeric evaluations, right set – from multimeric ones. Box boundaries correspond to the Q1=25th (bottom) and Q3=75th (top) percentiles in the data; the vertical line inside the box corresponds to the median (Q2). The width of the box defines the interquartile range (IQR=Q3-Q1). The length of the whiskers shows the range of the values outside the interquartile range, but within 1.5 IQR. The dots correspond to outliers, i.e. values outside the 1.5 IQR range.



	DFIRE	Molprb	Molprb:clash	RMSD	Molprb:ram_out	Molprb:rot_out	ProQ3	QMEAN	CAD	Davis-QA	Molprb:ram_fv	TMscore	TMalign	LDDT	LGA_S	GDT_TS	GDT_HA
DFIRE	1	0.02	-0.01	0.15	0.1	-0.19	-0.19	-0.28	-0.17	0.06	-0.17	-0.33	-0.38	-0.17	-0.12	-0.13	-0.15
Molprb	0.02	1	0.85	0.47	0.56	0.19	-0.53	-0.72	-0.44	-0.52	-0.59	-0.44	-0.51	-0.49	-0.54	-0.47	-0.49
Molprb:clash	-0.01	0.85	1	0.31	0.33	-0.15	-0.42	-0.6	-0.31	-0.37	-0.38	-0.27	-0.23	-0.29	-0.29	-0.3	-0.32
RMSD	0.15	0.47	0.31	1	0.69	-0.15	-0.54	-0.45	-0.81	-0.77	-0.77	-0.94	-0.9	-0.89	-0.92	-0.94	-0.92
Molprb:ram_out	0.1	0.56	0.33	0.69	1	-0.03	-0.52	-0.5	-0.54	-0.64	-0.76	-0.68	-0.66	-0.58	-0.71	-0.71	-0.7
Molprb:rot_out	-0.19	0.19	-0.15	-0.15	-0.03	1	0.08	-0.05	0.16	0.07	0.14	0.16	0.19	0.11	0.15	0.16	0.16
ProQ3	-0.19	-0.53	-0.42	-0.54	-0.52	0.08	1	0.73	0.44	0.65	0.61	0.57	0.6	0.56	0.57	0.57	0.58
QMEAN	-0.28	-0.72	-0.6	-0.45	-0.5	-0.05	0.73	1	0.48	0.61	0.72	0.42	0.44	0.71	0.51	0.47	0.49
CAD	-0.17	-0.44	-0.31	-0.81	-0.54	0.16	0.44	0.48	1	0.57	0.65	0.73	0.77	0.82	0.79	0.8	0.78
Davis-QA	0.06	-0.52	-0.37	-0.77	-0.64	0.07	0.65	0.61	0.57	1	0.73	0.74	0.7	0.7	0.79	0.79	0.77
Molprb:ram_fv	-0.17	-0.59	-0.38	-0.77	-0.76	0.14	0.61	0.72	0.65	0.73	1	0.76	0.73	0.75	0.76	0.78	0.78
TMscore	-0.33	-0.44	-0.27	-0.94	-0.68	0.16	0.57	0.42	0.73	0.74	0.76	1	0.94	0.86	0.9	0.93	0.92
TMalign	-0.38	-0.51	-0.23	-0.9	-0.66	0.19	0.6	0.44	0.77	0.7	0.73	0.94	1	0.88	0.95	0.93	0.92
LDDT	-0.17	-0.49	-0.29	-0.89	-0.58	0.11	0.56	0.71	0.82	0.7	0.75	0.86	0.88	1	0.92	0.93	0.93
LGA_S	-0.12	-0.54	-0.29	-0.92	-0.71	0.15	0.57	0.51	0.79	0.79	0.76	0.9	0.95	0.92	1	0.97	0.95
GDT_TS	-0.13	-0.47	-0.3	-0.94	-0.71	0.16	0.57	0.47	0.8	0.79	0.78	0.93	0.93	0.93	0.97	1	0.99
GDT_HA	-0.15	-0.49	-0.32	-0.92	-0.7	0.16	0.58	0.49	0.78	0.77	0.78	0.92	0.92	0.93	0.95	0.99	1

Figure 9.

Spearman correlation coefficients between the scores used to evaluate accuracy of models in the multimeric regime. The calculations were performed on models from all targets clustered together. Rows and columns in the table are clustered according to the similarity between the scores. Deeper blue /red color illustrates stronger correlation /anti-correlation between the measures, boxes designate clusters.



	IFaceCheck:Jd_avg	IFaceCheck:RMSD_avg	TEMPy:ENV	Molprb:ram_out	TEMPy:LAP	TEMPy:MI	TEMPy:CCC	PHENIX:FSC	PHENIX:boxCC	PHENIX:CA-score	EMRinger	QS-best	QS-global	QS:LDDT	IFaceCheck:recall_avg	IFaceCheck:prec_avg	IFaceCheck:F1_avg	PHENIX:chiral_rmsd	PHENIX:planar_rmsd	PHENIX:bond_rmsd	PHENIX:angle_rmsd	PHENIX:dihedr_rmsd	Molprb:clash	Molprb:rot_out
IFaceCheck:Jd_avg	1	0.69	0.1	0.01	0.01	0.18	-0.14	-0.19	-0.14	-0.11	-0.35	-0.3	-0.25	0.06	-0.72	-0.51	-0.68	-0.28	-0.21	-0.32	-0.2	-0.07	0.1	-0.11
IFaceCheck:RMSD_avg	0.69	1	0.03	0	-0.2	0.15	-0.14	-0.15	-0.11	-0.03	-0.32	-0.31	-0.29	0	-0.81	-0.53	-0.79	-0.32	-0.18	-0.42	-0.34	0.21	0.09	-0.16
TEMPy:ENV	0.1	0.03	1	0.41	-0.29	-0.66	-0.31	-0.35	-0.32	-0.65	-0.62	-0.37	-0.27	-0.45	0.07	0.43	0.37	-0.07	-0.29	-0.11	0	-0.39	0.18	0.13
Molprb:ram_out	0.01	0	0.41	1	-0.23	-0.35	-0.36	-0.29	-0.26	-0.39	-0.43	-0.27	-0.21	-0.31	-0.06	0.05	-0.03	0.29	0.24	0.27	0.53	-0.1	0.31	0.38
TEMPy:LAP	0.01	-0.2	-0.29	-0.23	1	0.72	0.78	0.75	0.79	0.22	0.14	0.06	-0.03	-0.06	0.14	0.46	0.44	-0.33	-0.27	-0.01	-0.14	-0.1	-0.23	-0.03
TEMPy:MI	0.18	0.15	-0.66	-0.35	0.72	1	0.81	0.81	0.81	0.48	0.28	0.12	-0.02	0.19	-0.32	0.02	-0.11	-0.28	-0.12	-0.08	-0.24	0.15	-0.14	-0.18
TEMPy:CCC	-0.14	-0.14	-0.31	-0.36	0.78	0.81	1	0.96	0.96	0.41	0.13	0.07	-0.01	0.07	0.2	0.24	0.32	-0.3	-0.27	-0.1	-0.28	0.03	-0.07	-0.13
PHENIX:FSC	-0.19	-0.15	-0.35	-0.29	0.75	0.81	0.96	1	0.99	0.45	0.17	0.12	0.06	0.13	0.21	0.27	0.33	-0.31	-0.22	-0.01	-0.19	0.09	-0.02	-0.08
PHENIX:boxCC	-0.14	-0.11	-0.32	-0.26	0.79	0.81	0.96	0.99	1	0.42	0.09	0.07	-0.01	0.07	0.19	0.33	0.35	-0.32	-0.25	0	-0.17	0.05	-0.03	-0.06
PHENIX:CA-score	-0.11	-0.03	-0.65	-0.39	0.22	0.48	0.41	0.45	0.42	1	0.49	0.51	0.45	0.74	0.09	-0.48	-0.28	-0.01	0.09	-0.13	-0.24	0.33	-0.14	-0.15
EMRinger	-0.35	-0.32	-0.62	-0.43	0.14	0.28	0.13	0.17	0.09	0.49	1	0.57	0.54	0.51	0.25	-0.31	-0.09	-0.06	0	-0.03	-0.29	0.12	-0.42	-0.38
QS-best	-0.3	-0.31	-0.37	-0.27	0.06	0.12	0.07	0.12	0.07	0.51	0.57	1	0.85	0.75	0.26	-0.03	0.13	0.15	0.06	0.07	-0.06	0.06	-0.23	-0.19
QS-global	-0.25	-0.29	-0.27	-0.21	-0.03	-0.02	-0.01	0.06	-0.01	0.45	0.54	0.85	1	0.87	0.22	-0.13	0.05	0.15	0.05	0.11	-0.04	0.08	-0.15	-0.18
QS:LDDT	0.06	0	-0.45	-0.31	-0.06	0.19	0.07	0.13	0.07	0.74	0.51	0.75	0.87	1	-0.05	-0.43	-0.25	0.13	0.05	-0.1	-0.21	0.15	-0.09	-0.17
IFaceCheck:recall_avg	-0.72	-0.81	0.07	-0.06	0.14	-0.32	0.2	0.21	0.19	0.09	0.25	0.26	0.22	-0.05	1	0.56	0.86	0.34	0.16	0.3	0.29	-0.07	-0.22	0.06
IFaceCheck:prec_avg	-0.51	-0.53	0.43	0.05	0.46	0.02	0.24	0.27	0.33	-0.48	-0.31	-0.03	-0.13	-0.43	0.56	1	0.86	-0.33	-0.37	0.25	0.06	-0.25	0.05	0.12
IFaceCheck:F1_avg	-0.68	-0.79	0.37	-0.03	0.44	-0.11	0.32	0.33	0.35	-0.28	-0.09	0.13	0.05	-0.25	0.86	0.86	1	-0.09	-0.24	0.3	0.11	-0.27	-0.06	0.14
PHENIX:chiral_rmsd	-0.28	-0.32	-0.07	0.29	-0.33	-0.28	-0.3	-0.31	-0.32	-0.01	-0.06	0.15	0.15	0.13	0.34	-0.33	-0.09	1	0.79	0.49	0.66	0.02	0.1	0.45
PHENIX:planar_rmsd	-0.21	-0.18	-0.29	0.24	-0.27	-0.12	-0.27	-0.22	-0.25	0.09	0	0.06	0.05	0.05	0.16	-0.37	-0.24	0.79	1	0.5	0.73	0.23	0.22	0.46
PHENIX:bond_rmsd	-0.32	-0.42	-0.11	0.27	-0.01	-0.08	-0.1	-0.01	0	-0.13	-0.03	0.07	0.11	-0.1	0.3	0.25	0.3	0.49	0.5	1	0.7	-0.01	0.33	0.29
PHENIX:angle_rmsd	-0.2	-0.34	0	0.53	-0.14	-0.24	-0.28	-0.19	-0.17	-0.24	-0.29	-0.06	-0.04	-0.21	0.29	0.06	0.11	0.66	0.73	0.7	1	0.14	0.4	0.57
PHENIX:dihedr_rmsd	-0.07	0.21	-0.39	-0.1	-0.1	0.15	0.03	0.09	0.05	0.33	0.12	0.06	0.08	0.15	-0.07	-0.25	-0.27	0.02	0.23	-0.01	0.14	1	0.28	0.3
Molprb:clash	0.1	0.09	0.18	0.31	-0.23	-0.14	-0.07	-0.02	-0.03	-0.14	-0.42	-0.23	-0.15	-0.09	-0.22	0.05	-0.06	0.1	0.22	0.33	0.4	0.28	1	0.45
Molprb:rot_out	-0.11	-0.16	0.13	0.38	-0.03	-0.18	-0.13	-0.08	-0.06	-0.15	-0.38	-0.19	-0.18	-0.17	0.06	0.12	0.14	0.45	0.46	0.29	0.57	0.3	0.45	1

Figure 10.

Spearman correlation coefficients between the scores used to evaluate accuracy of models in the monomeric regime. The calculations were performed on models from all targets clustered together. Rows and columns in the table are clustered according to the similarity between the scores. Deeper blue /red color illustrates stronger correlation /anti-correlation between the measures, boxes designate clusters.