RESEARCH ARTICLE

PROTEINS: Structure, Function, and Bioinformatics
DOI10.1002/prot.25817

Cryo-EM targets in CASP13: overview and evaluation of results

Running title: Cryo-EM evaluation

Andriy Kryshtafovych [1*], Sony Malhotra [2*], Bohdan Monastyrskyy [1], Tristan Cragnolini [2], Agnel-Praveen Joseph [2], Wah Chiu [3], Maya Topf [2]

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

[2] Institute of Structural and Molecular Biology, Birkbeck, University College London, Malet Street, London WC1E 7HX, UK

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

***co-first authors**

Corresponding author

Andriy Kryshtafovych, akryshtafovych@ucdavis.edu

Acknowledgements

Received: Apr 17, 2019;Revised: Jul 29, 2019;Accepted: Sep 13, 2019

This work was partially supported by the US National Institute of General Medical Sciences (NIGMS/NIH) grant R01GM100482 to AK and BM, Wellcome Trust grants 208398/Z/17/Z and 209250/Z/17/Z to MT, and grant R01GM079429 to WC.

Abstract

Structures of seven CASP13 targets were determined using cryo-electron microscopy (cryo-EM) technique with resolution between 3.0 and 4.0 Å. We provide an overview of the experimentally derived structures and describe results of the numerical evaluation of the submitted models. The evaluation is carried out by comparing coordinates of models to those of reference structures (CASP-style evaluation), as well as checking goodness-of-fit of modeled structures to the cryo-EM density maps. The performance of contributing research groups in the CASP-style evaluation is measured in terms of backbone accuracy, all-atom local geometry and similarity of inter-subunit interfaces. The results on the cryo-EM targets are compared with those on the whole set of eighty CASP13 targets. *A-posteriori* refinement of the best models in their corresponding cryo-EM density maps resulted in structures that are very close to the reference structure, including some regions with better fit to the density.

Keywords

CASP, protein structure prediction, electron microscopy, cryo-EM, model evaluation

1. Introduction

Cryogenic electron microscopy (cryo-EM) is becoming increasingly instrumental in solving protein structures. By the end of 2018, the number of cryo-EM structure depositions to the Protein Data Bank (PDB) exceeded 2700, with almost 900 structures (or roughly 1/3 of the entries) submitted that year alone (http://www.rcsb.org/stats/growth/em). The cryo-EM determined structures made up around 8% of all protein structures deposited to the PDB in 2018. Incidentally, the share of cryo-EM structures in CASP13 was essentially the same with 7 out of 80 evaluated targets coming from the EM structural biology groups. Thus, CASP13 target dataset represents a proportional slice of the 2018 annual structure deposition to the PDB in sense of structure determination methods (Supplementary Figure S1).

Since cryo-EM targets are typically quite different from other CASP targets (in terms of their size, complexity of quaternary structure composition and resolution), CASP organizers thought that it would be useful to conduct a separate evaluation of the participated methods on such targets. In this article, we analyze performance of the CASP13 tertiary and quaternary structure prediction methods on the cryo-EM targets only, and compare the results with those on all CASP13 targets (discussed in detail elsewhere in this issue). Additionally, we carry out analyses specific to cryo-EM derived targets by checking the fit of the submitted models to the cryo-EM density maps and comparing the best-fitting models refined in the density with their corresponding reference structures (provided by the experimentalists).

2. Materials and Methods

2.1. Cryo-EM targets in CASP13

Structures of seven CASP13 targets were determined by cryo-electron microscopy (cryo-EM) and image processing with resolution between 3.0 and 4.0 Å. Six targets are multimeric (T0984, T0995, T0996, T1020, H1021, H1022) and one is monomeric (T0990). Names of homo-multimeric targets start with 'T', while names of hetero-multimers start with 'H'. Four of the six multimeric targets (T0984, T0995, T1020 and H1021) are also part of the CASP/CAPRI modeling experiment, where CASP participants are joined by members of CAPRI[1] community in modeling quaternary structure of proteins.

With regards to the target size, all cryo-EM targets are quite large. The monomeric target T0990 is 552-residue long. The multimeric targets vary in length from 1504 to 5088 residues for whole complexes, and from 149 to 848 residues for individual subunits. The average length of CASP13 cryo-EM targets is 2752 residues for assemblies and 462 residues for subunits. This is significantly different from CASP13 X-ray and NMR-derived targets, which are roughly five times shorter for whole targets (average length of 531 residues) and two times shorter for subunits (average length of 272 residues). Even though there is a substantial difference in the length at both whole-target and whole-subunit levels, the lengths of constitutive domains are comparable. The seven cryo-EM targets encompass 21 structural domains with an average length of 197 residues compared to 183 residues in 91 domains of the other 73 targets.

To ensure fair comparison of models, CASP13 targets and their domains are assigned to different prediction difficulty categories. Oligomeric targets are classified into three categories according to the principles outlined in the CASP13 assembly assessment paper [2]:

- Easy: templates can be identified by sequence homology for whole oligomeric assemblies;

- Medium: only partial templates can be found by sequence-based homology searches. Partial means that templates can be identified for subunits, but not the whole assembly (i.e., no hints on how to model the interface), or that information on only parts of the subunits or interfaces is known (e.g., a dimeric template is available for a tetrameric complex);
- Difficult: no templates are available for either the subunits or the assembly.

According to this classification, three out of six oligomeric cryo-EM targets are easy modeling, one – medium difficulty, and two more – hard modeling targets.

At the domain level, prediction targets are classified into difficulty categories following the same principles of the oligomeric categorization (i.e., based on the template availability) with an additional correction for the per-domain performance of the CASP participants (see paper [3] for details). All in all, 21 domains of CASP13 cryo-EM targets are split into two difficulty categories: 15 easier template-based modeling (TBM) targets, and 6 harder free modeling (FM) targets.

A summary on the CASP13 cryo-EM targets is provided in Table 1. All six oligomeric structures are symmetric; however, this information was not provided by the experimentalists in advance, and thus was not relayed to predictors or used in the analysis of the results.

2.2. Participants and predictions

In CASP13, 93 prediction groups submitted 4079 tertiary structure predictions of seven cryo-EM derived targets, and 20 groups submitted 343 quaternary structure predictions of six oligomeric cryo-EM targets. The models were generated without knowledge of the cryo-EM density maps, i.e., based solely on the sequence of the target.

2.3. Evaluation measures

The accuracy of models submitted for each cryo-EM target is evaluated with two broad classes of measures – those assessing accuracy of models with respect to their corresponding 'reference' structures, and those assessing quality of model fit in the experimental cryo-EM density map (model-to-map goodness-of-fit). Reference structures are the models generated by the experimentalists using the information in the cryo-EM map.

2.3.1. Accuracy of models with respect to reference structures

CASP has been using a wide suite of numerical measures to assess similarity of models to native structures[4-6]. Below we describe the measures that were chosen for the evaluation of cryo-EM targets.

2.3.1.1. Tertiary structure evaluation

To assess the accuracy of the tertiary structure of models, we employ five conceptually different measures - a rigid-body structure superposition measure *GDT_TS*[7,8], and four superposition-free measures – *LDDT*[9], *CADaa*[10], *SphereGrinder (SG)*[6] and *QCS*[11]. The chosen set of measures provides complementary information on the accuracy of a model: *GDT_TS* reports on conformation of model's backbone with respect to the target's backbone, *LDTT* on similarity of inter-residue distance patterns, *CADaa* on difference in all-atom contact areas, *SG* on similarity of corresponding local structural neighborhoods, and *QCS* on topological similarity and relative packing of secondary structure elements. Using all these scores helps provide a well-rounded opinion about the overall accuracy of the inspected models.

Importantly, for the ranking of the participating groups, absolute scores of models should not be considered in isolation from the target difficulty or the performance of other groups. For instance, a score of 0.6 can be considered 'outstanding' for a free modeling target

if all other groups score 0.4 or worse, but 'poor' for an easier template-based modeling target, where majority of modelers score 0.8 or better. Thus, using raw scores for group ranking can be misleading. A better practice is to work with the normalized scores quantifying relative performance of groups. To this end, we transform per-target raw scores into standard scores using the formula:

$$z\_score(model) = \frac{raw\_score(model) - Mean\_score}{StandardDeviation\_score}. \quad (1)$$

In CASP, each group can submit up to 5 models per target. Typically, groups are ranked either on the scores of their 'first models' (i.e., the models estimated to be the most accurate by the predictors), or their actual per-target best scores (based on the *a-posteriori* evaluation). In this study we use both ranking approaches, and formula (1) is applied to calculation of *z_scores* separately on each of the datasets. After the calculation of original *z_scores*, outliers that score two standard deviations or more below the mean (i.e. $z\_score \leq -2$) are excluded, and the standard scores are re-calculated based on the mean and standard deviation of the outlier-free model set (we call these new standard scores here Z_{scores} , starting with capital '*Z*'). Next, all models that score below the mean (i.e. those with negative Z_{scores}) and outliers from the first stage are assigned Z_{scores} of 0, in order not to over-penalize the groups attempting novel strategies[12]. If a group does not submit any predictions on a target, its per-target Z_{scores} are set to zero. Finally, the per-target Z_{scores} of different measures are combined and summed over the selected sets of targets.

Here, we adopted the cumulative ranking formulas from the latest CASP assessments [13-16]. For template-based modeling targets (15 in CASP13) the relative group performance is calculated as

$$TS_{ranking}(TBM) = \sum_{targ \in TBMdomains} \left[Z_{GDT\_TS} + 1/3 * (Z_{LDDT} + Z_{CADaa} + Z_{SG})\right]_{targ} , (2)$$

while for free modeling targets (6 in CASP13) it is calculated as

$$TS_{ranking}(FM) = \sum_{targ \in FMdomains} [Z_{GDT\_TS} + Z_{QCS}]_{targ} . \quad (3)$$

The TBM formula takes equal contributions from a global superposition measure (*GDT_TS*), and local-based measures (*LDDT*, *CADaa* and *SG*), while the FM formula weighs equally *GDT_TS* and *QCS*, a topology-based measure.

2.3.1.2. Quaternary structure evaluation

The accuracy of the quaternary structure of models is assessed relative to the subunit interfaces in the reference structures in terms of *F1* score (a.k.a. *Interface Contact Score*[17]), *JaccardCoefficient* (a.k.a. *Interface Patch score*[17]) and QS_{glob} score[18]; overall similarity of Cα traces in the model and the target (*GDT_TSo*, suffix 'o' stands for 'oligo'); and similarity of intra- and inter- chain distance patterns (*LDDTo*). Ranking of the participating groups is performed according to the procedure described above (i.e. removal of outliers and re-calculating of Z-scores) using the formula adopted from the CASP13 assembly assessment [2]:

$$QS_{ranking}(QS\_targ) = \sum_{targ \in QStargets} \left[Z_{F1} + Z_{Jacc} + Z_{GDT\_TSo} + Z_{LDDTo}\right]_{targ} . \quad (4)$$

2.3.2. Fit of model coordinates to cryo-EM density maps

To speed-up the calculation of the model-to-map goodness-of-fit, we first superimpose the models (using the Biopython's Bio.pdb module) onto the reference structures, which were produced by the target providers in the context of the cryo-EM density map. This procedure positions the models approximately in the correct region of their corresponding cryo-EM map.

Next, we fine-tune the position of the models in the density map using the fit-in-map tool from the UCSF Chimera package[19]. To this end, we use a Python script (accessible at https://gitlab.com/ccpem/ccpem/tree/master/src/ccpem_core/chimera_scripts) that utilizes Chimera's fitmap global search option, where 100 random initial positions in the map are searched and locally optimized. The solutions are then ranked based on the cross-correlation score. Note that for difficult targets, models are usually far away from the reference structure and thus grossly incompatible with the cryo-EM maps. In such cases, fine-tuning the fit and subsequent evaluation make little sense and, hence, is not attempted here.

For assessing the goodness-of-fit, we used three software packages: PHENIX[20], TEMPy[21,22], and EMRinger[23]. The overall model-to-map goodness-of-fit is quantified using PHENIX's real space correlation coefficients – CC_{volume}, CC_{mask} and CC_{peaks}, – each probing different aspects of model-to-map fit[24]; TEMPY's cross-correlation coefficients – *CCC*, CCC_{ov}, the Laplacian-filtered correlation coefficient – LAP and the average per-chain Segment-based Mander's Overlap Coefficients – $SMOC_f$ and $SMOC_d$;[21,25]; and EMRinger's global score enumerating accuracy of side-chain placement within map density[24].

The per-residue (local) model-to-map goodness-of-fit is evaluated with PHENIX's local CC_{box} measure[24]; local *EMRinger* score[23]; and TEMPy's local $SMOC_f$ and $SMOC_d$ scores[25]. $SMOC_f$ is calculated on overlapping residue windows (sequence fragments), whereas $SMOC_d$ on the voxels occupied by the atoms of a specific residue.

CASP infrastructure for running the evaluation, reporting scores and visualizing evaluation results for cryo-EM targets (http://predictioncenter.org/casp13/cryoem_results.cgi) is based on the prototype of the evaluation infrastructure[26,27] developed for the cryo-EM model

challenge[28].

3. Results

3.1. Evaluation of tertiary structure

3.1.1. Comparison of results on EM and non-EM targets

By evaluating models versus reference structures, we want to address the question of whether the results on the cryo-EM targets are substantially different from those on the other CASP13 targets. To answer this question, we compare the raw scores and rankings of groups on these two subsets of targets.

First, we calculate the averages of per-target maximum scores (MAX) and mean scores (MEAN) for TBM and FM domains of EM and non-EM targets. Results of the calculations are provided in Supplementary Table S1 (panels A and B). Since the tendencies in the data are similar for the different scores, we discuss here only the *GDT_TS*-based results.

Comparing averages of the MEAN scores shows that TBM domains from EM targets are overall harder to predict than non-EM targets (GDT_TS=47.8 on TBM/EM targets vs 55.8 on TBM/non-EM), while FM domains are equally difficult, regardless of the experimental technique used for structure determination (GDT_TS=27.7 on FM/EM vs 28.3 on FM/non-EM). For the MAX scores, this tendency holds only for the TBM domains (76.5 on TBM/EM vs 81.8 on TBM/non-EM), while for the FM domains the scores on EM targets are higher than those on non-EM (65.7 on FM/EM vs 59.8 on FM/non-EM). Thus, from the analysis of the highest-scoring models, FM domains from cryo-EM structures might seem easier for modeling. However, the data show that the difference in the 'average of MEAN' versus 'average of MAX'

tendencies on the FM targets can be explained by the outstandingly good results for the cryo-EM targets (in particular T0990) by one group (A7D), which pulls the corresponding set of maximum scores up. To probe whether the difference in the predictive difficulty of EM and non-EM targets is statistically significant, we performed unpaired t-tests on the per-target MAX and MEAN scores. The results of the tests show that for the harder (FM) domains any difference in the predictive difficulty can be attributed to pure chance, while for the easier (TBM) domains the GDT_TS and LDDT measures are discriminative at the p=0.05 significance level, thus confirming the conclusion that TBM domains from CASP13 cryo-EM targets are in general harder to predict. The complete results of the statistical tests are provided in Supplementary Tables S1A and S1B.

3.1.2. Overall group performance

To compare group performance on cryo-EM targets, we apply the ranking procedure described in Materials and Methods, section 2.3.1.1. Figure 1 provides a summary of the relative performance of groups on the TBM and FM domains (panels A and B, respectively). On the TBM domains, several top groups demonstrated comparable results having cumulative Z_{scores} within 2 units from each other, both on first models (M1) and best results. On the FM domains, the top group (A7D) is an undisputable leader. Paired t-tests on the cumulative Z_{scores} and on the individual evaluation scores (Tables S2 and S3, Supplementary Material), show that on the TBM domains the performance of McGuffin, Zhang, Seok-refine, QUARK, Zhang-Server, A7D and MULTICOM groups is statistically indistinguishable. On the FM domains, the A7D group outscored all the other groups by a statistically significant margin. These results are in agreement with the results on all CASP13 targets reported by CASP13 TBM and FM

assessors[15,16].

3.2. Evaluation of quaternary structure

3.2.1. Comparison of results on EM and non-EM targets

Similarly to the evaluation of tertiary structure, we start our analysis on quaternary structure by calculating averages of the per-target maximum (MAX) and average (MEAN) scores, for the multimeric EM and non-EM targets. Results of the calculations are provided in Table S1 (panel C). Representing each CASP13 multimeric target by the average *GDT_TS* score from all groups (MEAN), and comparing the averages of the representative scores on different target sets shows that the interfaces in EM targets were in general easier to predict than those in the non-EM ones, as the interface-based scores are higher on the former (*F1*=20.4, *Jaccard*=30.2) than the latter (14.2 and 26.2, respectively). This result can be explained by lower modeling difficulty of CASP13 multimeric EM targets, where the fraction of easy targets (TBM) is 50% (3 out of 6) compared to only 33% (12 out of 36) for the non-EM targets. The overall shape of multimeric targets is predicted rather poorly on average, for both types of targets (EM and non-EM) as quantified by low MEAN *GDT_TSo* and *LDDTo* scores. If we analyze differences among the best models (MAX scores), we will see that the contact-based interface score (*F1*) is higher for the EM targets (44.9 vs 36.2 on non-EM), while all other scores are very similar for both types of targets (difference within 1.0 score unit). All MAX scores are significantly higher (around 20 units) than the corresponding MEAN scores, thus signifying that the best assembly predictors performed much better than the rest of the participants.

3.2.2. Overall group performance

The relative performance of groups in predicting the quaternary structure of six CASP13 cryo-EM oligomeric targets is summarized in Figure 2. The cumulative ranking score was calculated using Eq. (4) from section 2.3.1.2. Figure 2A shows the ranking of CASP participants for these six targets, while Figure 2B shows the ranking of CASP and CAPRI groups for four out of the six targets, which were selected for the joined CASP/CAPRI experiment[1]. The Venclovas group leads the rankings among CASP-only participants and is also a member of a tight cluster of the top-performing groups on the CASP/CAPRI targets. Similarly to the outcome of the tertiary structure analysis, the results on the quaternary structure for cryo-EM targets are similar to the results for the complete set of all CASP13 targets[2].

3.3. Evaluation of model-to-map fit

Evaluating the goodness-of-fit of CASP models to the experimental cryo-EM density maps makes sense only for targets with good homology, where high-accuracy models are expected. Three out of seven CASP13 cryo-EM targets – T0984o, T0995o and T1020o - were classified as easy for modeling (see Materials and Methods). Below we concentrate our attention on these three targets, all of which are oligomeric. Density maps for these targets are in the 3.2-3.4 Å resolution range (Table 1). Typically, maps in this resolution range contain enough information to reliably trace the backbone and some of the side chains. However, in practice even models built on such well-resolved maps are not void of structural inconsistencies or errors[24,29].

In this section we analyze whether CASP models, which are built without the knowledge of the cryo-EM density, agree with the density, and compare their goodness-of-fit with that of the experimentally-derived structures. We also check if consensus between the

models can be an indicator of the reliability of the local goodness-of-fit of the reference structure. Finally, we probe the utility of the best CASP models as starting points for further real-space refinement in the cryo-EM density map.

3.3.1. Correlation between evaluation scores and selection of the assessment measures

Since the goodness-of-fit analysis is done for the first time in CASP, we want to check which scores provide complementary information for the assessment of CASP models. To this end, we calculate the pair-wise correlation between all goodness-of-fit scores (section 2.3.2) and the average correlation of each score with all other scores. Figure 3 shows that some pairs of scores (e.g., CC_{mask}/CC_{vol}, $SMOC_d$/$SMOC_f$, CCC/CC_{peaks}) are highly correlated on CASP models. Therefore, we leave only one score from each pair (CC_{mask} , $SMOC_f$ and CC_{peaks}) for the assessment. We also exclude the LAP and CCC_{ov} scores on the grounds of their high similarity to the rest of the measures (see corresponding diagonal values in Figure 3). On the other side of the correlation spectrum is the *EMRinger* score. The score has the lowest average correlation to all other model-to-map goodness-of-fit scores (0.45, Figure 3) as well as to the '*vs* the reference structure' scores (0.39, Supplementary Figure S2). The low correlation of this measure likely stems from the fact that EMRinger's effective usage requires an approximately correct placement of the backbone and side chains within the density, which often lacks in CASP models. Thus, the EMRinger score can be misleading in the CASP context and is not used here. Following this analysis, the fit scores used for the assessment of CASP models are $SMOC_f$ (from TEMPy), and CC_{mask} and CC_{peaks} (from PHENIX). The $SMOC_f$ score accounts for per-residue correlation of density values in sequential fragments surrounding each residue.

PHENIX's CC_{peaks} compares map regions with highest density values, while CC_{mask} uses values inside the mask calculated around the molecule of interest.

3.3.2. Overall (global) goodness-of-fit

The model-to-map global fit score is calculated as

$$fit\_score = 1/3\ [CC_{mask} + CC_{peaks} + SMOC_f\ ], \quad (5)$$

and ranges between -1 and 1.

Since the cryo-EM experimental models (reference structures) are built to fit the EM density, it is not surprising that their *fit_score* is substantially higher than that of models built without the knowledge of experimental data. The corresponding scores of the reference structure and the highest-scoring CASP model are: 0.69 and 0.30 for target T0984o; 0.52 and 0.26 for target T0995o; and 0.66 and 0.37 for target T1020o. The three main reasons behind such large differences are distortions and shifts of secondary structure elements, inaccurate modeling of interfaces (docking of subunits), and mistakes in modeling of loops and packing of side chains.

Next, we investigate how well the *fit_score* (Eq. 5) correlates with the overall quaternary structure accuracy score (*assembly_score*) calculated as an average of individual scores used in Eq. 4:

$$assembly\_score = 1/4\ [F1 + Jacc + GDT\_TSo + LDDTo]. \quad (6)$$

The assembly score ranges between 0 and 1.

We find that the answer strongly depends on the target (Figure 4). For targets with easier assembly organization, e.g. T0984o (dimer) or T1020o (trimer), the correlation is high (0.85 and 0.82, respectively) thus confirming intuitive assumption that models with better

overall fold should have better fit to map. However, for the target with a more complex organization – T0995o (octamer), the correlation between the two scores is weak (0.25). The latter target contains 16 outlier models with noticeably higher *fit_score* (>0.1) than the other 57 models, all of which score below 0.1 (see Figure 4). Groups that contributed the better-scoring models are Baker, Baker-ROSETTAserver, Yasara and Kiharalab_CAPRI.

3.3.3. Per-residue (local) goodness-of-fit

Local model-to-map fit scores can help identifying regions of poor fit, distortions and shifts. To evaluate the local fit, we use the per-residue $SMOC_f$ score.

Figure 5 demonstrates the consensus among CASP models in predicting the local structure of the targets (the *IQR* score), and shows local fit scores ($SMOC_f$) for the highest scoring chain in the experimentally-derived reference structure and the highest-scoring CASP models. $SMOC_f$ scores for all individual chains of the three analyzed cryo-EM targets (T0984o, T0995o and T1020o) are provided in Supplementary Figure S3. It is evident from the figures that the best-scoring CASP models have worse local fit to the density than the reference structures (dashed blue line is consistently below the solid line). An interesting question to consider is whether regions of higher structural consensus correspond to regions of better local fit in the reference structure and the CASP models. If such a correspondence was present, then the blue and red lines in Figure 5 should be in 'anti-phase' (i.e. peaks of red lines should correspond to dips of the blue ones). With respect to the reference structures, we do not observe such a tendency and therefore cannot state that regions of higher agreement between models are more likely to correspond to regions where the target fits the map better. However, when we compare the best-fitting model lines (dashed blue) with the consensus lines (solid red), we

notice that the lines are in anti-phase in most regions, thus indicating dependency between the inter-model consensus and models' goodness-of-fit to the density, especially for targets T0995 and T1020.

For example, in the best oligomeric model for target T0995 (T0995TS368_5o, from the Baker-ROSETTAserver group) the four worst-fitting regions (dips where $SMOC_f$ <0.4) are regions of bad inter-model consensus (peaks where *IQR*>1.5 Å), while the five best-fitting regions (peaks where $SMOC_f$ >0.7) are regions of good consensus (valleys where *IQR*<0.5 Å).

Similar situation can be observed for the highest-scoring model of target T1020 (T1020TS004_2o from the Yasara group, shown in Figure 6A). Although this model is well fitted in the density overall, including the subunit interface (Figure 6B), it has several significant $SMOC_f$ dips, most of which correspond to distinct *IQR* peaks in Figure 5. The deepest and the widest dip of the $SMOC_f$ line is in the C-terminus region starting at residue 475. Detailed analysis of the structure reveals that poor fit to density in this region is due to inaccurate modeling of loop 475-478 (Figure 6C, left) and associated shift in helix 479-511 with respect to the reference structure (Figure 6D). Another example of the region that does not fit well to the T1020 density is loop 220-230 (Figure 6C, right). Figure 5 shows that $SMOC_f$ line has a local minimum in this region dipping to (low) $SMOC_f$ values around 0.4, while the *IQR* line attains one of its highest peaks, thus signaling substantial disagreement between the CASP models. Modeling problems in this region can be attributed to incorrect secondary structure prediction, where loop 220-230 was attempted to be modeled as a helix. It is worth mentioning that this loop is spatially close to the above discussed C-terminus region 475-511, potentially indicating a more global problem in modeling this area.

3.3.4. Refinement of models in the cryo-EM map

With hundreds of models submitted for cryo-EM targets in CASP, an interesting question to study is whether these models can be effectively used as starting points for the refinement into the EM map. To examine this, we apply automated refinement protocols to best-fitting models, and compare the resulting refined structures with the reference ones. For targets T1020o and T0995o, we use PHENIX real-space refinement with default parameters (five macro-cycles of global real-space refinement with rotamer, Ramachandran plot, secondary structure and C_β deviation restraints enabled), while for target T0984o we use Flex-EM[25] refinement followed by the PHENIX refinement. Flex-EM refinement was ran with rigid-bodies set as secondary structure elements. In all cases, the models were refined in their entirety, as oligomers.

Target T1020o: Upon refinement, the highest-scoring model for target T1020o (T1020TS004_2o, already analyzed in the previous subsection) shows considerable improvement in both the global and local fit (Figure 7). Figure 7A shows the original unrefined model (left), the refined model (middle) and the reference structure (right), all fitted to the map. It can be easily seen that quality-of-fit improves from left to right, and this is confirmed by the goodness-of-fit scores. The global fit score (CC_{mask}) increases from 0.38 (for unrefined model) to 0.69 (for refined), stopping less than 0.1 short of the reference structure's score of 0.78 (Figure 7B). The local fit score also increases substantially (Figure 7C), with the average per-residue $SMOC_f$ scores rising from 0.47 (unrefined) to 0.64 (refined) and approaching the reference structure's score of 0.70. Inspection of the per-residue $SMOC_f$ plot (Figure 7D) shows that the refinement improves the local fit of the original model in several regions. Interestingly,

in two of the regions (residues 273-285 and 410-425, marked as regions 2 and 5 in panel D) the local fit of the refined CASP model is better than that of the reference structure. Figure 7E demonstrates more accurate placement of the residues D422 and Y418, where the side chains move more into the density after refinement as compared to their positions in the reference structure. Furthermore, the refinement results in forming an intra-chain hydrogen bond between D422-OD2 and W414-NE1, which is not present in the reference structure. Figure 7D also identifies four other regions (residues 222-234, 287-303, 320-340 and 473-512 marked as 1, 3, 4 and 6, respectively), where the $SMOC_f$ scores improves by much, although the fit is still worse compared to that of the reference structure (see Figure 7F as an example). Additional cycles of refinement in PHENIX cannot improve the fits in these regions. Although not tested here, further improvement could potentially be achieved using other tools, such as Coot[30]. Similar results are observed in refining other chains (Supplementary Figure S4).

Target T0995o: Upon refinement, the highest scoring model for target T0995o, TS368_5o (dimeric), improves considerably in both the global and local goodness-of-fit (Figure S5). The global fit score (CC_{mask}) increases from 0.48 (for unrefined model) to 0.76 (for refined), reaching very close to the reference structure's score of 0.81 (Figure S5A). The local fit score also increases substantially (Figure S5B), with the average per-residue $SMOC_f$ scores rising from 0.54 (unrefined, average over both chains) to 0.71 (refined, average over both chains) and approaching the reference structure's score of 0.72 (average over all chains).

Target T0984o: For this target, using real-space refinement in PHENIX also significantly improves the global fit of the best-fitting model (TS329_1). In particular, the CC_{mask} score of the model with the highest *fit_score* (TS329_1) increases from 0.30 to 0.58

(Figure 8A). However, this is still significantly lower than the CC_{mask} for the reference structure (0.79). Additionally, the $SMOC_f$ curve indicates that large regions of the structure are not improved upon refinement (Figure 8B), with an average $SMOC_f$ score of 0.51 and 0.58 before and after PHENIX-only refinement, respectively. Visual inspection reveals large rigid-body shifts relative to the reference structure (e.g. residues 419-559) (Figure 8C). Trying to remedy this, we perform real-space flexible fitting with Flex-EM[25,31], starting from the best-fitting model (TS329_1). The refined model obtained using Flex-EM is then subjected to further refinement with PHENIX as before. Although this protocol results in a similar global fit in the density as compared to the PHENIX-only refinement, the local fit improves significantly (see for example the region corresponding to residues 419-559, Figure 8B), with less shifts and distortions of rigid bodies (Figure 8D). The latter is also reflected in the average $SMOC_f$, which improves from 0.51 to 0.62 (for both chains).

Overall, we show here that the refinement of CASP models in their density maps clearly helps bringing the models closer to the experimentally-derived reference structures, as judged by the overall (multimeric) and per-chain accuracy of the backbone, packing of the side chains and similarity of inter-residue distances (Figure 9). Compared to the original CASP13 models, the backbone conformation of the refined models improves significantly as the overall (multimeric) *GDT_TSo* score increases by 13-25% (depending on target), and the per-chain (monomeric) *GDT_TS* score increases by 8-22%. The side-chain packing is also improved noticeably as quantified by the side-chain only version of the CAD-score (*CADss*)[10]. Interestingly, side-chain packing is enhanced most (25%) for the target with the smallest correction of the backbone (8%, T0995), and least (3%) for the target with the largest correction

(22%, T0984). Similarity of models' distance patterns with respect to the targets' ones is improved for all targets as the after-the-refinement *LDDT* scores are higher than the before-the-refinement *LDDT* scores by 4-12%. Finally, we want to mention that the improvement in 'versus the reference structure' scores does not necessarily translate into improvement of *MolProbity* scores[32], which are better for one target (T0984), but worse for the other two. The MolProbity scores for the target structures and the refined best CASP models are provided in Table S4; the corresponding Ramachandran maps are provided in Figure S6.

4. Discussion

For the first time in CASP, a sizeable portion of targets (8%) was determined with cryo-EM. Since cryo-EM structures are typically different from the structures derived by X-ray or NMR (for example, in their average size or complexity of quaternary structure), it is interesting to evaluate if the target determination technique affects accuracy of models, or performance of the predictors. Also interesting is to look at the quaternary structure prediction (although not unique to cryo-EM targets), and in particular at the accuracy of interfaces, as at present most cryo-EM structures represent large protein assemblies (here 6 out of 7 targets). This paper studies these issues and also explores additional assessment approaches specific to cryo-EM models /targets. These approaches examine fit of CASP models to the experimental density and check utility of the models for the refinement in cryo-EM maps.

In terms of tertiary structure, the paper demonstrates that in comparison with the structures derived by X-ray or NMR, the CASP13 cryo-EM structures are in general harder to model on template-based domains and of approximately the same difficulty on free modeling

domains. For the free modeling targets, results differ significantly, depending on whether the analysis is based on the average scores or maximum scores. Based on the maximum scores the results seem to be favorable to cryo-EM targets (i.e. models are of better accuracy), but this conclusion is heavily influenced by the outstanding models of one group on one difficult 3-domain target. It is worth noting that all results reported here should be taken with caution as there were significantly fewer cryo-EM targets in all categories of the analysis.

The rankings of the participating groups on cryo-EM targets are consistent with those on all CASP13 targets. The same groups that are leading cryo-EM rankings are the top performers on all-target datasets. On the template-based modeling targets, seven groups topped the ranking - McGuffin, Zhang, Seok-refine, QUARK, Zhang-Server, A7D and MULTICOM. These groups showed statistically similar results. On the free modeling targets, the A7D group is an apparent leader, outperforming other groups in a statistically significant manner. Among the assembly predictors, the Venclovas group is the best.

The comparison of different types of scores for top-ranked models shows that models that demonstrate the best fit to the cryo-EM density quite often have subpar 'versus the reference' scores, and vice versa. However, in general the accuracy of the best models and their fit to density are correlated well based on both global and local measures. Comparing local consensus of different models with fit to the density maps also reveals that structurally conserved regions are overall better fitted to experimental data. It should be noted though that all fit-to-map analyses are performed here on targets where models are of relatively good quality, and therefore the conclusions can be related only to this type of targets.

Refinement of the submitted CASP models in the experimental density shows that the

models could be improved to the point of approaching the quality of the reference structures (and beyond in some local structural regions), thus indicating that high-quality models from CASP predictors can be a good starting point for structure refinement. This could potentially save computer time and reduce the overall effort in reaching a good model.

References

1. Lensink MF, Wodak SJ. CAPRI Evaluation in CASP13. Proteins 2019;This issue.
2. Guzenko D, Monastyrskyy B, Kryshtafovych A, Duarte J. CASP13 assembly evaluation paper. Proteins 2019;This issue.
3. Kinch L, Monastyrskyy B, Kryshtafovych A. CASP13 domain definition and classification. Proteins 2019;This issue.
4. Kryshtafovych A, Monastyrskyy B, Schwede T, Topf M, Moult J, Fidelis K. CASP evaluation measures. Proteins 2019;This issue.
5. Olechnovic K, Monastyrskyy B, Kryshtafovych A, Venclovas C. Comparative analysis of methods for evaluation of protein models against native structures. Bioinformatics 2018.
6. Kryshtafovych A, Monastyrskyy B, Fidelis K. CASP prediction center infrastructure and evaluation measures in CASP10 and CASP ROLL. Proteins 2014;82 Suppl 2:7-13.
7. Zemla A. LGA: A method for finding 3D similarities in protein structures. Nucleic Acids Res 2003;31(13):3370-3374.
8. Zemla A, Venclovas, Moult J, Fidelis K. Processing and evaluation of predictions in CASP4. Proteins 2001;Suppl 5:13-21.
9. Mariani V, Biasini M, Barbato A, Schwede T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. Bioinformatics 2013;29(21):2722-2728.
10. Olechnovic K, Kulberkyte E, Venclovas C. CAD-score: a new contact area difference-based function for evaluation of protein structural models. Proteins 2013;81(1):149-162.
11. Cong Q, Kinch LN, Pei J, Shi S, Grishin VN, Li W, Grishin NV. An automatic method for CASP9 free modeling structure prediction assessment. Bioinformatics 2011;27(24):3371-3378.
12. Tramontano A, Morea V. Assessment of homology-based predictions in CASP5. Proteins 2003;53 Suppl 6:352-368.
13. Kryshtafovych A, Monastyrskyy B, Fidelis K, Moult J, Schwede T, Tramontano A. Evaluation of the template-based modeling in CASP12. Proteins 2018;86 Suppl 1:321-334.
14. Abriata LA, Tamo GE, Monastyrskyy B, Kryshtafovych A, Dal Peraro M. Assessment of hard target modeling in CASP12 reveals an emerging role of alignment-based contact prediction methods. Proteins 2018;86 Suppl 1:97-112.
15. Abriata LA, dal Peraro M. CASP13 evaluation of FM targets. Proteins 2019;This issue.
16. Croll T, Read RJ. Evaluation of template-based models in CASP13. Proteins 2019;This issue.
17. Lafita A, Bliven S, Kryshtafovych A, Bertoni M, Monastyrskyy B, Duarte JM, Schwede T, Capitani G. Assessment of protein assembly prediction in CASP12. Proteins 2018;86 Suppl 1:247-256.

18. Bertoni M, Kiefer F, Biasini M, Bordoli L, Schwede T. Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by homology. Sci Rep 2017;7(1):10480.
19. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. UCSF Chimera--a visualization system for exploratory research and analysis. Journal of computational chemistry 2004;25(13):1605-1612.
20. 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. PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr D Biol Crystallogr 2010;66(Pt 2):213-221.
21. Farabella I, Vasishtan D, Joseph AP, Pandurangan AP, Sahota H, Topf M. : a Python library for assessment of three-dimensional electron microscopy density fits. J Appl Crystallogr 2015;48(Pt 4):1314-1323.
22. Vasishtan D, Topf M. Scoring functions for cryoEM density fitting. J Struct Biol 2011;174(2):333-343.
23. Barad BA, Echols N, Wang RY, Cheng Y, DiMaio F, Adams PD, Fraser JS. EMRinger: side chain-directed model and map validation for 3D cryo-electron microscopy. Nat Methods 2015;12(10):943-946.
24. Afonine PV, Klaholz BP, Moriarty NW, Poon BK, Sobolev OV, Terwilliger TC, Adams PD, Urzhumtsev A. New tools for the analysis and validation of cryo-EM maps and atomic models. Acta crystallographica Section D, Structural biology 2018;74(Pt 9):814-840.
25. Joseph AP, Malhotra S, Burnley T, Wood C, Clare DK, Winn M, Topf M. Refinement of atomic models in high resolution EM reconstructions using Flex-EM and local assessment. Methods 2016;100:42-49.
26. Kryshtafovych A, Adams PD, Lawson CL, Chiu W. Evaluation system and web infrastructure for the second cryo-EM model challenge. Journal of structural biology 2018;204(1):96-108.
27. Kryshtafovych A, Monastyrskyy B, Adams PD, Lawson CL, Chiu W. Distribution of evaluation scores for the models submitted to the second cryo-EM model challenge. Data in Brief 2018;20:1629-1638.
28. Lawson CL, Chiu W. Comparing cryo-EM structures. Journal of structural biology 2018;204(3):523-526.
29. Chen M, Baldwin PR, Ludtke SJ, Baker ML. De Novo modeling in cryo-EM density maps with Pathwalking. Journal of structural biology 2016;196(3):289-298.
30. Emsley P, Lohkamp B, Scott WG, Cowtan K. Features and development of Coot. Acta crystallographica Section D, Biological crystallography 2010;66(Pt 4):486-501.
31. Topf M, Lasker K, Webb B, Wolfson H, Chiu W, Sali A. Protein structure fitting and refinement guided by cryo-EM density. Structure 2008;16(2):295-307.
32. Chen VB, Arendall WB, 3rd, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS, Richardson DC. MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr D Biol Crystallogr 2010;66(Pt 1):12-21.

FIGURE LEGENDS

Figure 1. Relative performance of CASP13 groups in predicting tertiary structure of cryo-EM targets. Data are shown for top 12 groups on (A) 15 TBM domains and (B) 6 FM EUs. Cumulative Z_{scores} are calculated according to formulas (2) and (3) (see Materials and Methods) for TBM and FM targets, correspondingly. Blue and orange bars show the ranking scores calculated on first models (M1) and best results (best), correspondingly. Groups are sorted according to the first model scores (blue bars).

Figure 2. Relative performance of CASP13 and CAPRI groups in predicting quaternary structure of cryo-EM targets. Data are shown for (A) top 12 CASP13 groups for all six oligomeric cryo-EM targets and (B) CASP13 and CAPRI groups on four CASP/CAPRI oligomeric targets. Cumulative Z_{scores} are calculated according to formula (4) (see Materials and Methods). Blue and orange bars show the ranking scores calculated on first models (M1) and best results (best), respectively. Groups are sorted according to the best scores (orange bars). CAPRI groups in panel B are marked with an asterisk.

Figure 3. Correlation between model-to-map goodness-of-fit scores based on models submitted for all seven CASP13 cryo-EM targets. The under-the-diagonal part of the table shows Spearman correlation coefficients between each pair of scores. The correlation scores are visualized in the upper portion of the table with color and shape (deeper colors and thinner ovals relate to higher correlations). Diagonal cells (shaded) show average correlation versus all other scores.

Figure 4. The global model-to-map *fit_score* (Eq. 5) versus *assembly_score* (Eq. 6) for CASP13 models submitted for three cryo-EM targets. Each point corresponds to a model. Linear trend lines are threaded through the data. The value of the coefficient of determination R^2 is provided on the graphs.

Figure 5. Local model-to-map goodness-of-fit scores of the highest-scoring chain in the reference structure (solid blue line) and the highest-scoring CASP model (dashed blue line) versus the local consensus score of all CASP models (red line) for three of the cryo-EM targets. Score values for red lines are provided on the left of the plot, and for blue lines – on the right. The goodness-of-fit score is represented by the local $SMOC_f$ score (the higher the better). The inter-model consensus score (the lower the better) is represented by the interquartile range of Cα-Cα distances (in Ångstroms) between corresponding residues in top 100 models according to the *GDT_TS* score, after their optimal superposition. The best-fitting models for the shown targets are: T0984TS329_1o, T0995TS368_5o and T1020TS004_2o (see http://predictioncenter.org/casp13/cryoem_results.cgi).

Figure 6. The best CASP13 model (TS004_2o) for target T1020o fitted in the corresponding density map. (A) The best model colored according to the local $SMOC_f$ score (scale bar at the left). The region marked by a circle is zoomed-in in panel (B); the region marked by a rectangle is enlarged in panels (C) and (D), from slightly different spatial perspectives. (B) The hydrophobic residues at the trimer interface within the cryo-EM map. (C) Loops 475-478 and

220-230 within the density. (C) Helix 479-511 within the density. In panels (B), (C) and (D), the best model is colored according to the $SMOC_f$ score (red representing bad fit and blue representing good fit), and the reference structure is shown in green.

Figure 7. Assessment of the best-fitting model (TS004_2o) for target T1020o before and after PHENIX refinement in the cryo-EM map. Blue color in all panels correspond to the unrefined model, orange to the refined one, and green to the reference structure. (A) The original model (left), the refined model (middle) and the reference structure (right) fitted into the cryo-EM map. Regions that are encircled and numbered in the refined model (middle) correspond to the numbered regions in panel D. (B) Global CC_{mask} score for the unrefined model, refined model, and experimentally-derived structure. (C) Boxplots of per-residue $SMOC_f$ scores for chain A in the unrefined model, refined model, and target. (D) Per-residue $SMOC_f$ scores for chain A in the unrefined model, refined model, and target. Shaded strips show most notable areas of fit improvement. The pink-shaded strips (#2 and 5) mark areas that improved beyond the target structure fit, while the grey-shaded strips (#1, 3, 4 and 6) mark those that improved significantly, but remain still worse than the corresponding areas in the target structure. Plots for other chains are very similar and shown in Figure S4. (E) A region of the refined model that has improved over the reference structure. An intra-chain hydrogen bond between the side-chains of D422 and W414 in the refined best-fitting model is indicated for chain C. (F) Regions in the refined model that are poorly fit to density even after the real-space refinement.

Figure 8. Assessment of the best-fitting model (TS329_1o) for target T0984o before and after

refinement in the cryo-EM map using Flex-EM and PHENIX. (A) Global CC_{mask} score for the unrefined model (blue), refined model with PHENIX only (pink), refined model with Flex-EM and PHENIX (orange), and the experimentally-derived reference structure (green). (B) Average $SMOC_f$ scores for chain A of the best-fitting model before refinement (blue), after PHENIX-only refinement (pink), after Flex-EM and PHENIX refinement (orange), and for the reference structure (green). The region corresponding to residues 419-559 is shown in gray shade. (C) The fit of the model after PHENIX-only refinement (orange) and the reference structure (green) in the cryo-EM map. The zoomed panel shows residues 419-559, which are outside of the density after refinement. (D) The fit of the model after Flex-EM refinement followed by PHENIX refinement (orange) and the reference structure (green) in the cryo-EM map. Region 419-559 is better fit to the density if Flex-EM refinement is applied first.

Figure 9. Improvement in model accuracy as quantified by the multimeric *GDT_TSo* score and monomeric *GDT_TS*, *CADss*, *LDDT* scores for the best-fitting CASP13 models before (grey) and after (black) refinement in the cryo-EM map. For uniformity of the graph scale, the *GDT_TS* scores are presented as fractions rather than percentages (i.e., are divided by 100).

TABLES

Table 1. Overview of CASP13 cryo-EM targets

CASP ID	Protein description	Map resol, Å	Stoi-chiom	Length mono (cmplx)	Pred diffclty	# Dom	Author
T0984	A two-pore calcium channel protein playing an important role in regulating lysosomal membrane potential	3.4	A2	752 res (1504)	Easy	2	Xiaochen Bai, U. Texas Southwestern Medical Center, Dallas, TX, USA
T0990	A virulence factor modulating the innate immune response and influenza A virus pathogenicity	4.0	A1	552	Hard	3	Hong Zhou, U. California, Los Angeles, CA, USA
T0995	A cyanide dehydratase providing insight into substrate specificity and thermostability	3.15	A8	330 (2640)	Easy	1	Bryan T. Sewell, U. Cape Town, South Africa
T0996	A protein likely playing role in bacterial outer membrane lipid transport	3.0-3.5	A6	848 (5088)	Medium	7	Damian Ekiert, Skirball Institute, NY, USA
T1020	An anion channel with an important role in plant physiology	3.3	A3	577 (1731)	Easy	1	Oliver Clarke, Columbia U., NY, USA
H1021	A part of the anti-feeding prophage (AFP) complex, which is a contractile ejection system	varying	A6 B6 C6	A:149 B:354 C:295 (4788)	Hard	4	Ambroise Desfosses, Institut de Biologie Structurale, Grenoble, France
H1022	A part of the anti-feeding prophage (AFP) complex, which is a contractile ejection system	3.3-3.5	A6 B3	A:229 B:529 (2961)	Hard	3	Ambroise Desfosses, Institut de Biologie Structurale, Grenoble, France

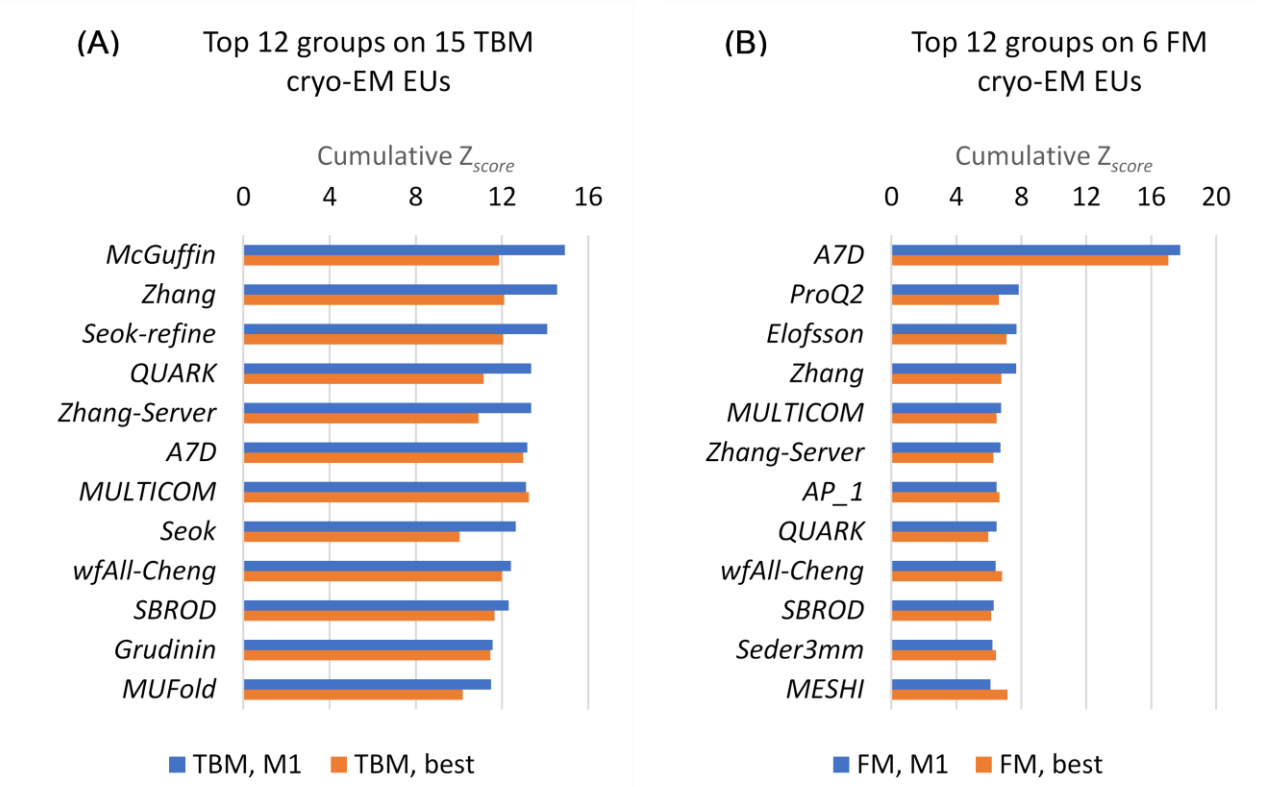

Figure 1.

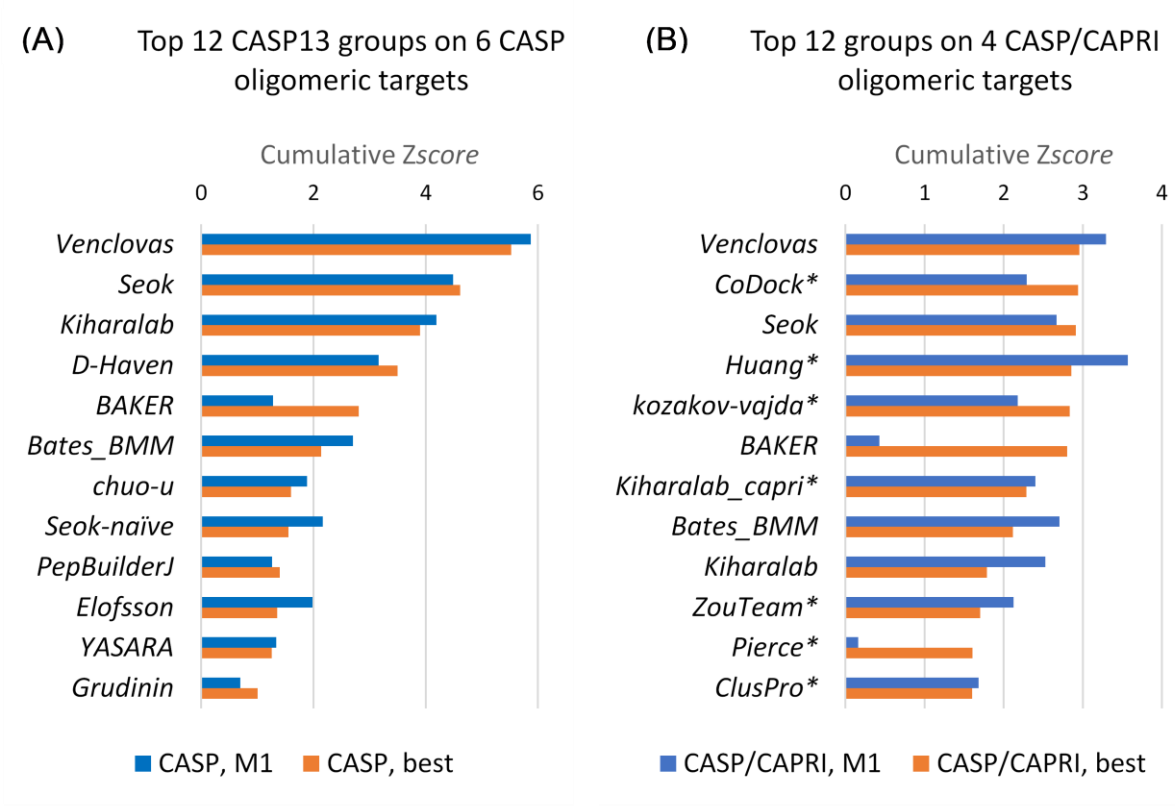

Figure 2.

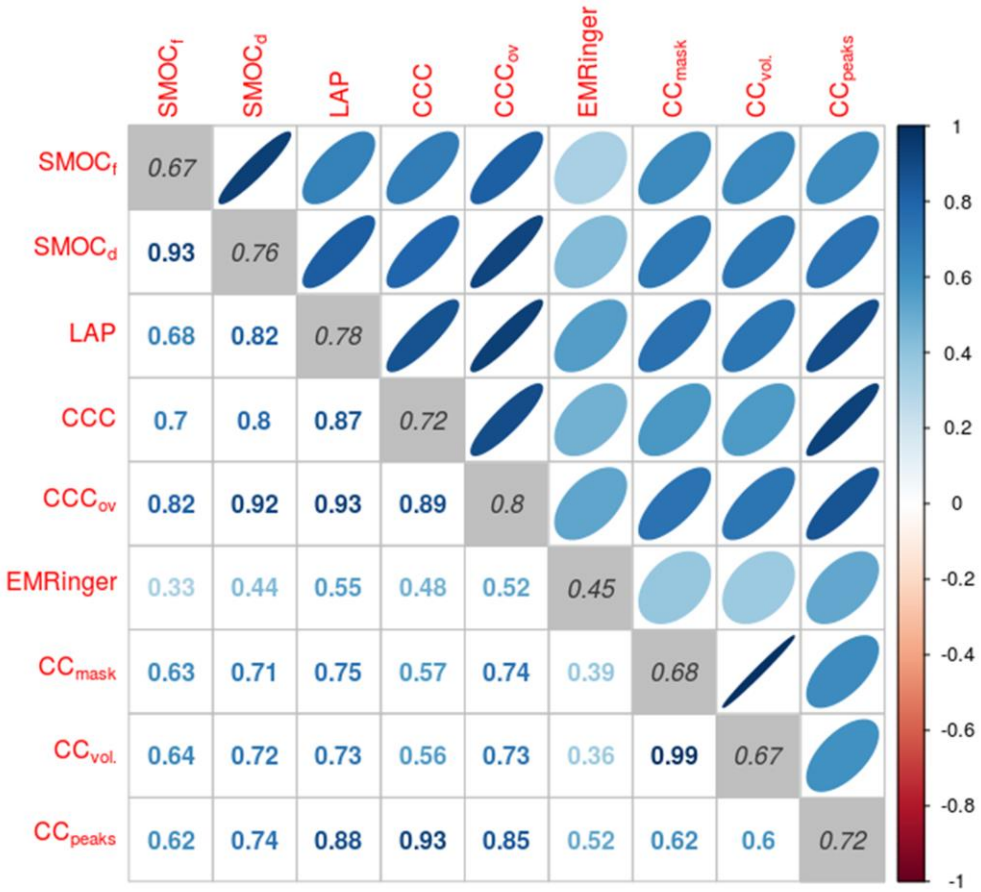

Figure 3.

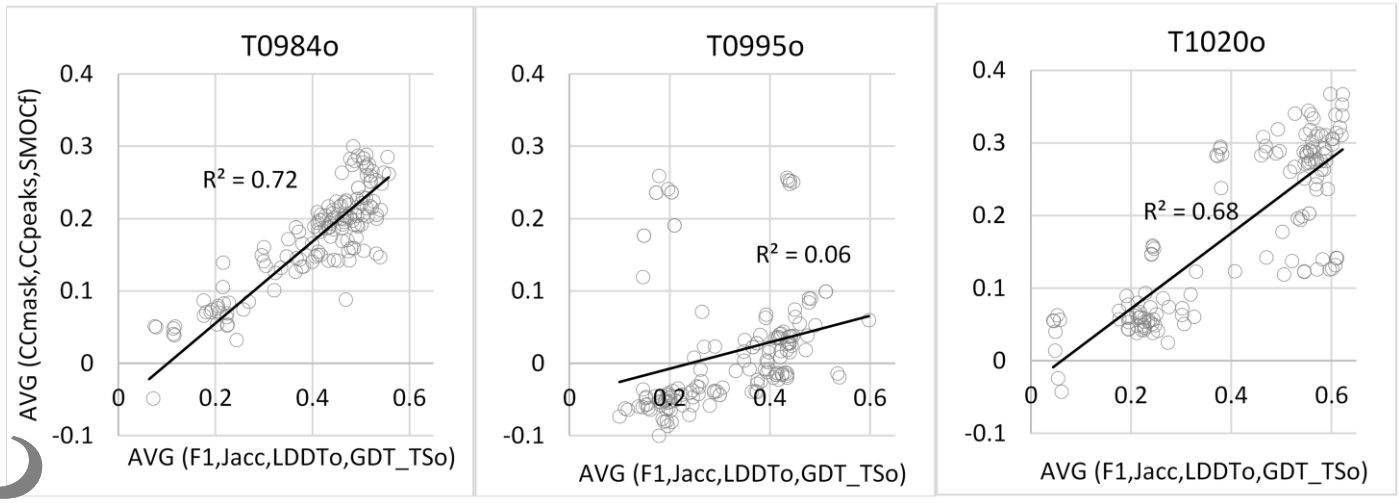

Figure 4.

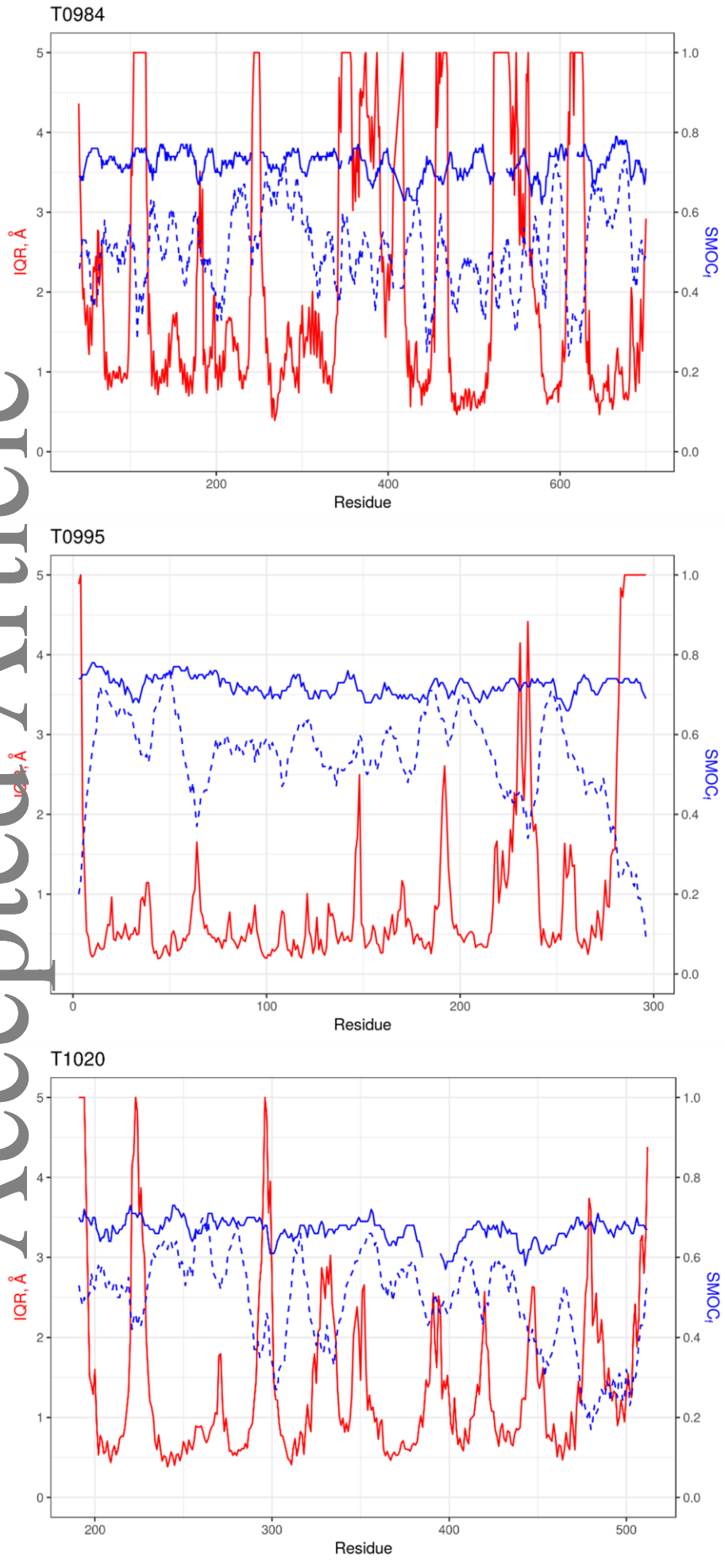

Figure 5