Outcomes of the 2019 EMDataResource model challenge: validation of cryo-EM models at near-atomic resolution

Catherine L. Lawson[1]*, Andriy Kryshtafovych[2], Paul D. Adams[3,4], Pavel V. Afonine[3], Matthew L. Baker[5], Benjamin A. Barad[6], Paul Bond[7], Tom Burnley[8], Renzhi Cao[9], Jianlin Cheng[10], Grzegorz Chojnowski[11], Kevin Cowtan[7], Ken A. Dill[12], Frank DiMaio[13], Daniel P. Farrell[13], James S. Fraser[14] Mark A. Herzik Jr.[15],, Soon Wen Hoh[7], Jie Hou[16], Li-Wei Hung[17], Maxim Igaev[18], Agnel P. Joseph[8], Daisuke Kihara[19,20], Dilip Kumar[21], Sumit Mittal[22], Bohdan Monastyrskyy[2], Mateusz Olek[7], Colin M. Palmer[8], Ardan Patwardhan[23], Alberto Perez[24], Jonas Pfab[25], Grigore D. Pintilie[26], Jane S. Richardson[27], Peter B. Rosenthal[28], Daipayan Sarkar[19,22], Luisa U. Schäfer[29], Michael F. Schmid[30], Gunnar F. Schröder[29,31], Mrinal Shekhar[22,32], Dong Si[25], Abishek Singharoy[22], Genki Terashi[18], Thomas C. Terwilliger[33], Andrea Vaiana[18], Liguo Wang[34], Zhe Wang[23], Stephanie A. Wankowicz[14,35], Christopher J. Williams[27], Martyn Winn[8], Tianqi Wu[36], Xiaodi Yu[37], Kaiming Zhang[26], Helen M. Berman[38,39], Wah Chiu[26,30]*

[1]Institute for Quantitative Biomedicine and Research Collaboratory for Structural Bioinformatics, Rutgers, The State University of New Jersey, Piscataway, New Jersey 08854, USA

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

[3]Molecular Biophysics & Integrated Bioimaging Division, Lawrence Berkeley National Laboratory, Berkeley, CA, 94720, USA

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

[5]The University of Texas Health Science Center at Houston, Department of Biochemistry and Molecular Biology, Houston, TX 77030

[6]The Scripps Research Institute, Department of Integrated Computational Structural Biology, La Jolla, CA 92103

[7]York Structural Biology Laboratory, Department of Chemistry, University of York, York YO10 5DD, England, U.K.

[8]Scientific Computing Department, UKRI Science and Technology Facilities Council, Research Complex at Harwell, Didcot OX11 0FA, UK

[9]Department of Computer Science, Pacific Lutheran University, Tacoma, WA 98447, USA

[10]Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO 65211, USA

[11]European Molecular Biology Laboratory, c/o DESY, Notkestrasse 85, 22607 Hamburg, Germany

[12]Laufer Center, Stony Brook University, Stony Brook, New York, 11794, USA

[13]Department of Biochemistry and Institute for Protein Design, University of Washington, Seattle, WA 98195, USA

[14]Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, CA 94158, USA

[15]Department of Chemistry and Biochemistry, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA

[16]Department of Computer Science, Saint Louis University, St. Louis, MO, 63103, USA

[17]Los Alamos National Laboratory, Los Alamos, NM 87545, USA

[18]Theoretical and Computational Biophysics, Max Planck Institute for Biophysical Chemistry, D-37077 Göttingen, Germany

[19]Department of Biological Sciences, Purdue University, West Lafayette, IN, 47907, USA

[20]Department of Computer Science, Purdue University, West Lafayette, IN, 47907, USA

[21]Verna and Marrs McLean Department of Biochemistry and Molecular Biology, Baylor College of Medicine, Houston, TX 77030, USA

[22]Biodesign Institute, Arizona State University, Tempe, AZ, 85201, USA

[23]The European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridgeshire, United Kingdom

[24]Department of Chemistry, University of Florida, Gainesville, Florida, 32611, USA

[25]Division of Computing & Software Systems, University of Washington, Bothell, WA, 98011, USA

[26]Department of Bioengineering, Stanford University, Stanford, California 94305, USA

[27]Department of Biochemistry, Duke University, Durham NC 27710, USA

[28]Structural Biology of Cells and Viruses Laboratory, Francis Crick Institute, London, UK

[29]Institute of Biological Information Processing (IBI-7: Structural Biochemistry) and Jülich Centre for Structural Biology (JuStruct), Forschungszentrum Jülich, Jülich, Germany

[30]Division of CryoEM and Biomaging, SSRL, SLAC National Accelerator Laboratory, Stanford University, Menlo Park, California 94025, USA

[31]Physics Department, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

[32]Center for Development of Therapeutics, Broad Institute of MIT and Harvard, Cambridge, MA, 02141, USA

[33]New Mexico Consortium, Los Alamos NM 87544, USA

[34]Department of Biological Structure, University of Washington, Seattle, WA 98195, USA

[35]Biophysics Graduate Program, University of California, San Francisco, CA 94158, USA

[36]Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO 65211, USA

[37]SMPS, Janssen Research and Development, 1400 McKean Rd, Spring House, PA, 19477, USA

[38]Department of Chemistry and Chemical Biology and Research Collaboratory for Structural Bioinformatics, Rutgers, The State University of New Jersey, Piscataway, New Jersey 08854, USA

[39]Department of Biological Sciences and Bridge Institute, University of Southern California, Los Angeles, California 90089, USA

*Corresponding authors: cathy.lawson@rutgers.edu, wahc@stanford.edu

Abstract

This paper describes outcomes of the 2019 Cryo-EM Map-based Model Metrics Challenge sponsored by EMDataResource (www.emdataresource.org). The goals of this challenge were (1) to assess the quality of models that can be produced using current modeling software, (2) to check the reproducibility of modeling results from different software developers and users, and (3) compare the performance of current metrics used for evaluation of models. The focus was on near-atomic resolution maps with an innovative twist: three of four target maps formed a resolution series (1.8 to 3.1 Å) from the same specimen and imaging experiment. Tools developed in previous challenges were expanded for managing, visualizing and analyzing the 63 submitted coordinate models, and several novel metrics were introduced. The results permit specific recommendations to be made about validating near-atomic cryo-EM structures both in the context of individual laboratory experiments and holdings of structure data archives such as the Protein Data Bank. Our findings demonstrate the relatively high accuracy and reproducibility of cryo-EM models derived from these benchmark maps by 13 participating teams, representing both widely used and novel modeling approaches. We also evaluate the pros and cons of the commonly used metrics to assess model quality and recommend the adoption of multiple scoring parameters to provide full and objective annotation and assessment of the model, reflective of the observed density in the cryo-EM map.

Introduction

Electron cryo-microscopy (cryo-EM) has emerged as a key method to visualize and model a wide variety of biologically important macromolecules and cellular machines. Researchers can now routinely produce structures at near-atomic resolution, yielding new mechanistic insights into cellular processes and providing support for drug discovery[1-3]. Many academic institutions and pharmaceutical companies have invested in modern cryo-EM facilities, and multi-user resources are opening up worldwide[4].

The recent explosion of cryo-EM structures raises important questions. What are the limits of interpretability given the quality of the maps and resulting models? How do we quantify model accuracy and reliability under the simultaneous constraints of map density and chemical rules?

The EMDataResource Project (EMDR) was formed in 2006 as a collaboration between scientists in the UK (EMDataBank at the European Bioinformatics Institute) and the US (the Research Collaboratory for Structural Bioinformatics and the National Center for Macromolecular Imaging). Part of EMDR's mission is to derive validation methods and standards for cryo-EM maps and models through community consensus[5]. We created an EM Validation Task Force[6] analogous to those derived for X-ray crystallographic and NMR structures[7,8] and have sponsored Challenges, workshops and virtual conferences to engage cryo-EM experts, modellers, and end-users[5,9-13]. During this period, cryo-EM has evolved rapidly (Figure 1).

This paper describes outcomes of EMDR's most recent Challenge, the 2019 Model "Metrics" Challenge. The goals were three-fold: (1) to assess the quality of models that can be produced using established as well as newly implemented modeling software, (2) to check the reproducibility of modeling results from different software developers and users, and (3) to compare the performance of model evaluation metrics, particularly fit-to-map metrics. Map targets were selected in the near-atomic resolution regime (1.8-3.1 Å) with an innovative twist: three form a resolution series from the same specimen/imaging experiment (Figure 2). The results lead to several specific recommendations for validating near-atomic cryo-EM structures directed towards both individual researchers and the Protein Data Bank (PDB) structure data archive.

Results

We describe here the pipeline and outcomes of the EMDR 2019 Model Metrics Challenge (Figure 3). Four maps representing the state-of-the-art in cryo-EM single particle reconstruction were selected as the Challenge targets (Figures 2, 3a). Three maps of human heavy-chain apoferritin (APOF), a 500 kDa octahedral complex of 24 α-helix-rich subunits, formed a resolution series differing only in the number of particles used in reconstruction (EMDB entries EMD-20026, EMD-20027, EMD-20028)[14]. The fourth map was horse liver alcohol dehydrogenase (ADH), an 80 kDa α/β homodimer with NAD and Zn ligands (EMD-0406)[15].

A key criterion of target selection was availability of high quality experimentally determined model coordinates to serve as references. A 1.5 Å X-ray structure[16] (PDB id 3ajo) served as the reference for all three APOF maps, since no cryoEM-based model was available at the time. The X-ray model provides an excellent fit to each map, though not a fully optimized fit, owing to method/sample differences. The ADH reference was the model deposited by the original authors of the cryo-EM study (PDB id 6nbb)[15].

Thirteen teams from the US and Europe submitted 63 models in total, yielding 15-17 submissions per target (Figure 3b, Table I). The vast majority (51) were created *ab initio*, sometimes supported by additional manual steps, while others (12) were optimizations of publicly available models.

Submitted models were evaluated as in the previous Challenge[12] with multiple metrics in each of four tracks: Fit-to-Map, Coordinates-only, Comparison-to-Reference, and Comparison-among-Models (Figure 3c, Table II). The selected metrics include many already in common use, as well as several introduced via this Challenge.

Metrics to evaluate global Fit-to-Map included Map-Model Fourier Shell Correlation (FSC)[17] as encoded in Phenix[18], Refmac FSC average[19], EMDB atom inclusion[20], EMRinger[21], multiple Map vs. Model density-based correlation scores from TEMPy[22-25], Phenix[18], and the recently introduced Q-score to assess atom resolvability[14].

Metrics to evaluate overall Coordinates-only quality included Clashscore, Rotamer outliers, and Ramachandran outliers from MolProbity[26], as well as standard geometry measures (bond, bond angle, chirality, planarity, and dihedral angle RMSDs) from Phenix[27]. PDB currently uses each of these validation measures, based on community recommendations[6-8]. New in this round was MolProbity CaBLAM, which evaluates protein backbone conformation across multiple residues using novel virtual dihedral angle definitions[28].

Metrics assessing the similarity of a model to a reference structure included Global Distance Test total score[29], Local Difference Distance Test[30], CaRMSD from OpenStructure/QS[31], and Contact Area Difference[32]. Davis-QA was used to measure similarity among submitted models[33]. All of these measures are widely used in CASP competitions[33].

Several metrics were also evaluated at the per-residue level: Fit-to-Map: EMRinger, Q-score, EMDB atom inclusion, TEMPy SMOC, and Phenix CCbox; Coordinates-only: Clashes, Ramachandran outliers, and CaBLAM.

Evaluated metrics are tabulated with brief definitions in Table II; extended descriptions are provided in Online Methods.

An evaluation system website with interactive tables, plots and tools (Figure 3d) was established in order to organize and enable analysis of the Challenge results and to make the results accessible to all participants (model-compare.emdataresource.org).

Overall and local quality of models

The vast majority of submitted models scored well, landing in “acceptable” regions for metrics in each of the evaluation tracks, and in many cases performing better than the associated reference structure which served as a control (Supplementary Figure 1). For teams that submitted *ab initio* models, additional manual adjustment was beneficial, particularly for models built into the two lower resolution targets. In general, the best scoring models were produced by well-established methods and experienced modeling practitioners.

Evaluation exposed four fairly frequent issues: mis-assignment of peptide-bond geometry, misorientation of peptides, local sequence misalignment, and failure to model associated ligands. Sidechain model quality was not specifically assessed in this round.

Two-thirds of the submitted models had one or more peptide-bond geometry errors (Supplementary Figure 2).

At resolutions near 3 Å or in weak local density, the carbonyl O protrusion disappears into the tube of backbone density (Figure 2), and *trans* peptide bonds are more readily modeled in the wrong orientation. If ϕ,ψ values are explicitly refined, adjacent side chains can be pushed further in the wrong direction instead of fixing the underlying problem. Such cases are not flagged as Ramachandran outliers but they are still recognized by CaBLAM[34].

Sequence misthreadings misplace specific chemical groups over very large distances. The misalignment can be recognized by local Fit-to-Map criteria, with ends flagged by CaBLAM, bad geometry, *cis*-nonPro peptides, and clashes (Supplementary Figure 3).

The ADH map contains tightly bound ligands: an NADH cofactor as well as two zinc ions per subunit, with one zinc in the active site and the other in a spatially separate site where the metal coordinates with four cysteine residues[15]. A number of models lacking these ligands had considerable local modeling errors, sometimes even mistracing the backbone (Supplementary Figure 4).

Although there was evidence for ordered water in the higher resolution APOF maps[14], only two groups elected to model water oxygen atoms in their submissions. Model submissions were also split approximately 50:50 for the following practices: (1) inclusion of predicted hydrogen atom positions and (2) refinement of isotropic B-factors. Although near-atomic cryo-EM maps do not have a sufficient level of detail to directly identify hydrogen atom positions, inclusion of predicted H-atom positions can be useful for identifying model steric properties such as H-bonds or clashes[26]. Where provided, refined B-factors modestly improved Fit-to-Map scores against the highest resolution map target (APOF 1.8 Å) but had little to no benefit against lower resolution map targets.

Evaluating Metrics: Fit-to-Map

Fit-to-Map metrics (Table II, red section) were systematically compared using score distributions of the submitted models (Figure 4a-d). For APOF targets, subunit models were evaluated against masked subunit maps, whereas for the ADH target, dimeric models were evaluated against the full sharpened cryo-EM map (Figure 2d). To control for the impact of hydrogen atom inclusion or isotropic B-factor refinement on different subsets of Fit-to-Map metrics, all evaluated scores were produced with hydrogen atoms removed and with B-factors set to zero.

Score distributions were first evaluated *for all 63 models across all four Challenge targets*. Unexpectedly, a wide diversity in performance was observed, with poor correlations between most pairs of metrics (Figure 4a). This means that a model that scored well relative to all 62 others using one metric may have a much poorer ranking using another metric. A hierarchical cluster analysis identified three distinct clusters of similarly performing metrics (Figure 4a, boxes 1-3).

The observed sparse correlations and clustering of the Fit-to-Map metrics can be readily understood by considering their per target score distribution ranges, which differ substantially

from each other (Figure 4c). The three clusters identify sets of metrics that share similar trends (Fig. 4c, panels 1-3).

Cluster 1 metrics (Figure 4c, panel 1) share the trend of decreasing score values with increasing map target resolution. The cluster consists of six correlation measures, three from TEMPy[22-25] and three from Phenix[18]. Each evaluates a model's fit to the map in a similar way: by correlating a calculated model-map density with the experimental map density. In most cases (5 of 6), correlation is performed following model-based masking of the experimental map. The observed trend arises at least in part because as map resolution increases, the level of detail that a model-map must faithfully replicate in order to achieve a high correlation score must also increase.

Cluster 2 metrics (Figure 4c, panel 2) share the inverse trend: score values improve with increasing map target resolution. Cluster 2 metrics consist of Phenix Map-Model FSC=0.5[18], Qscore[14], and EMRinger[21]. The observed trend is expected: by definition each metric assesses a model's fit to the experimental map in a manner that is sensitive to map resolution.

Cluster 3 metrics (Figure 4c, panel 3) share a different trend: score values are significantly lower for ADH relative to APOF map targets. These measures include three unmasked correlation functions from TEMPy[22-25], Refmac FSCavg[19], EMDB Atom Inclusion[20] and TEMPy ENV[22]. All of these measures consider the full experimental map without masking, so can therefore be sensitive to background noise. Background noise was substantial in the unmasked ADH map and minimal in the masked APOF maps (Figure 2d).

Score distributions were also evaluated for how similarly they performed *per target*, and in this case most metrics were strongly correlated with each other (Figure 4b). This means that within any single target, a model that scored well relative to all others using one metric also fared well using nearly every other metric. This situation is illustrated by comparing scores for two different metrics, CCbox from Cluster 1 and Q-score from Cluster 2 (Figure 4d). The plot's four diagonal lines demonstrate that the scores are tightly correlated with each other within each map target. But as described above in the analyses of Clusters 1 and 2, the two metrics each have different sensitivities to map-specific factors. It is these different sensitivities that give rise to the separate and parallel spacings of the four diagonal lines, indicating score ranges on different relative scales for each target.

One Fit-to-Map metric showed poor correlation with all others in the per target analysis: TEMPy ENV (Figure 4b). ENV scores were poorly distributed with most models very close to the maximum possible value (1.0). ENV evaluates atom positions relative to a density threshold that is

determined from the sample molecular weight. At near-atomic resolution this threshold is overly generous and tends to include all modelled atoms. TEMPy Mutual Information (MI) and EMRinger also diverged somewhat from the other metrics (Figure 4b). Within each target, all MI scores were essentially identical to each other. This behavior may reflect a strong influence of background noise, since MI_OV, MI's masked version, yielded distributed scores that correlated well with other measures. As noted previously[21], EMRinger follows similar trends with other measures but yields distinct distributions owing to its focus on backbone placement.

Collectively these results reveal that multiple factors such as experimental map resolution, presence of background noise, and density threshold selection can strongly impact Fit-to-Map score values, depending on the chosen metric.

Evaluating metrics: Coordinates-only and vs-Reference

Metrics to assess model quality based on Coordinates-only (Table II, blue section), as well as Comparison-to-Reference and Comparison-among-Models (Table II, green and grey sections) were also evaluated and compared (Figure 4e-f).

Most of the Coordinates-only metrics were poorly correlated with each other (Figure 4e), with the exception of bond, bond angle, and chirality RMSD, which form a small cluster. Interestingly, Ramachandran outlier score, which is widely used to assess protein backbone conformation, was poorly correlated with all other Coordinate-only measures, including the novel CaBLAM scores[28]. Score distributions explain this in part: more than half (33) of submitted models had zero Ramachandran outliers, while only four had zero CaBLAM Conformation outliers (we note that Ramachandran statistics are increasingly used as restraints[35,36]). These results support the concept of CaBLAM as a new informative score for validating backbone conformation[28].

The CaBLAM Conformation and C-alpha measures, while orthogonal to other Coordinate-only measures, were unexpectedly found to perform very similarly to Comparison-to-Reference metrics; several Fit-to-Map metrics also performed somewhat similarly to Comparison-to-Reference metrics (Figure 4f). The similarity likely arises because the worst modeling errors in this Challenge were sequence and backbone conformation mis-assignments. These errors were equally flagged by CaBLAM, which compares models against statistics of high-quality structures from the PDB, and the Comparison-to-Reference metrics, which compare models directly against a high-quality reference. To a somewhat lesser extent these modeling errors were also flagged by Fit-to-Map metrics.

Evaluating metrics: local scoring

As part of the evaluation pipeline, residue-level scores were calculated in addition to overall scores. Five Fit-to-Map metrics either considered masked density for both map and model around the evaluated residue (Phenix CCbox[18], TEMPy SMOC[24]), density profiles at non-hydrogen atom positions (Qscore[14]), density profiles of non-branched residue Cɣ-atom ringpaths (EMRinger[21]), or density values at non-hydrogen atom positions relative to a chosen threshold (EMDB Atom Inclusion[20]). In two of the five, residue-level scores were obtained as sliding-window averages over multiple contiguous residues (SMOC: 9-residues; EMRinger: 21-residues).

Residue-level correlation analyses similar to those described above showed that local fit-to-map scores diverged more than their corresponding global scores. Residue-level scoring was most similar across the evaluated metrics for high resolution maps. This observation suggests that the choice of method for scoring residue-level fit becomes less critical at higher resolution, where maps tend to have stronger density/contrast around atom positions.

A case study of a local modeling error in one of the APOF 2.3 Å models (Supplementary Figure 3) showed that EMDB Atom Inclusion[20], Phenix CCbox[18], and Qscore[14] measures produced significantly lower (worse) scores within a 4-residue α-helical misthread relative to correctly assigned flanking residues. In contrast, the two sliding-window-based metrics were largely insensitive (a more recent version of TEMPy offers single residue analysis (SMOCd) and adjustable window analysis based on map resolution (SMOCf)[37]). At near-atomic resolution, single residue fit-to-map evaluation methods are likely to be more useful than windowing methods for identifying local modelling issues.

Residue-level Coordinate-only metrics (Supplementary Figure 3), Comparison-to-Reference metrics and Comparison-among-Models metrics (not shown) were also evaluated for the same modeling error. The MolProbity server[26,28] flagged the problematic 4-residue misthread via CaBLAM, cis-Peptide, clashscore, bond, and angle scores, but all Ramachandran scores were either favored or allowed. The Comparison-to-Reference LDDT and LGA local scores and the Davis-QA model consensus score also strongly flagged this error. The example demonstrates the value of combining multiple orthogonal measures to identify geometry issues, and further highlights the value of CaBLAM as a novel, orthogonal measure for validation of backbone conformation.

Group performance

Group performance was examined by modeling category and target by combining Z-scores from metrics determined to be meaningful in the analyses described above (see Methods and Supplementary Figure 5).

For *ab initio* modeling, lower resolution targets were more challenging for some groups. For the higher resolution APOF 1.8 Å and 2.3 Å targets, six groups (10, 28, 35, 41, 73, 82, see Table I ids) did very well ($Z \geq 0.3$), and a seventh (54, models 2) was a runner-up. For the lower resolution APOF 3.1 Å and ADH 2.9 Å targets, a slightly different six groups (10, 27, 28, 35, 73, 82) did very well and another two (41, 90) were runners-up. A wide variety of map density features and algorithms to produce a model, and most were quite successful yet allowing a few mistakes, often in different places (see Supplementary Figures 2-4). For practitioners, it might be beneficial to compare/combine models from several *ab initio* methods to come up with a better initial model for subsequent refinement. Note that the performance results are specific to the Challenge task and may not be directly applicable to other modeling scenarios.

As for optimization-based modeling, all made improvements, but sample size was too small to produce a rating.

Discussion

This 3rd Model Challenge round has demonstrated that cryo-EM maps with resolution ≤ 3 Å and from samples with limited conformational flexibility, have excellent information content, and automated methods are able to generate fairly complete models from such maps, needing only small amounts of manual intervention to be finalized (but some is always needed). Modeling could readily be accomplished within a month, the time-period of this challenge. This outcome represents a great advance over the previous challenges.

Inclusion of three maps in a resolution series enabled controlled evaluation of metrics by resolution. Inclusion of a completely different map as the fourth target provided a useful additional control. These target selections enabled observation of important trends that otherwise could have been missed. In a recent evaluation of predicted models against several ~3 Å cryo-EM maps in the CASP13 competition, TEMPy and Phenix Fit-to-Map correlation measures performed very similarly[37]. In this Challenge, because the chosen map targets covered a wider resolution range and had more variability in background noise, the same measures were found to have distinctive, map target feature-sensitive performance profiles.

The majority of submitted models were overall either equivalent to or better than the reference model in terms of the fit of their atomic coordinates to the target map. This achievement reflects significant advances in the development of modeling tools relative to the state presented a decade ago in our first Model Challenge[9]. However, several factors beyond atom positions that become important for accurate modelling at near-atomic resolution were not uniformly addressed: only half of the submitted models included refinement of atomic displacement factors (B-factors), and a minority of modellers attempted to fit water or bound ligands.

Fit-to-Map measures were found to be sensitive to different physical properties of the map, including experimental map resolution and background noise level, as well as input parameters such as density threshold. Coordinates-only measures were found to be largely orthogonal to each other, while Comparison-to-Reference measures were generally well correlated with each other.

The cryo-EM modeling community as represented by the Challenge participants have introduced a number of metrics to evaluate cryo-EM models with sound biophysical basis. We find that some of them are correlated to each other and to the resolution of the map, while others are not. Based on our careful analyses of these metrics and their relationships, we make four recommendations regarding validation practices for cryo-EM models of proteins determined at near-atomic resolution as studied here between 3.1 Å and 1.8 Å, a rising trend for cryo-EM (Figure 1).

Recommendation 1: For researchers optimizing a model against a single map, nearly any of the evaluated global fit-to-map metrics (Table II) can be used to evaluate progress because they are all largely equivalent in performance. Exception: TEMPy ENV is more appropriate for medium to low resolution (>4 Å).

Recommendation 2: To flag issues with local (per residue) fit to a map, metrics that evaluate single residues such as CCbox, Qscore, and EMDB Atom Inclusion are more suitable than those using sliding window averages over multiple residues.

Recommendation 3: The ideal Fit-to-Map metric for *archive-wide ranking* will be insensitive to map background noise (appropriate masking or alternative data processing can help), will not require input of estimated parameters that affect score value (*e.g.*, resolution limit, threshold). and will yield overall better scores for maps with trustworthy higher-resolution features. The three Cluster 2 metrics identified in this Challenge (Figure 4a) meet these criteria.

- Map-Model FSC[17,18] is already in common use [13], and can be compared with the experimental map's independent half-map FSC curve.

- Global EMRinger score[21] can assess non-branched protein side chains.
- Q-scoreis a relatively new correlation metric that can be used both globally and locally for validating non-hydrogen-atom x,y,z positions.[14].

Other Fit-to-map metrics may be rendered suitable for archive-wide comparisons through conversion of raw scores to Z-scores over narrow resolution bins, as is currently done by the PDB for some X-ray-based metrics[7,38].

Recommendation 4: CaBLAM statistical measures and MolProbity cis-peptide detection[28] are useful to detect protein backbone conformation issues. These are valuable new tools for cryo-EM protein structure validation, particularly since maps at typical resolutions (2.5 - 4.0 Å; Figure 1) may not resolve backbone carbonyl oxygens (Figure 2).

The 2019 Model “Metrics” Challenge was more successful than previous challenges because more time could be devoted to analysis and because infrastructure for model collection, processing and assessment is now established. EMDR plans to sponsor additional model challenges in order to continue promoting development and testing of cryo-EM modeling and validation methods. Future challenge topics are likely to cover medium resolution (3 to 4 Å), particle heterogeneity, membrane proteins, ligand modeling, nucleic acids, and models derived from tomograms.

Online Methods

Challenge Process and Organization

Informed by previous Challenges[9,10,12], the 2019 Model Challenge process was significantly streamlined in this round. In March, a panel of advisors with expertise in cryo-EM methods, modeling, and/or model assessment was recruited. The panel worked with EMDR team members to develop the challenge guidelines, identify suitable map targets from EMDB and reference models from PDB, and recommend the metrics to be calculated for each submitted model.

The Challenge rules and guidance were as follows: (1) *Ab initio* modeling is encouraged but not required. For optimization studies, any publicly available coordinate set can be used as the starting model. (2) Regardless of the modeling method used, submitted models should be as complete and as accurate as possible (*i.e.*, equivalent to publication-ready). (3) For each target, a separate modeling process should be used. (4) Fitting to either the unsharpened/unmasked map or one of the half-maps is strongly encouraged. (5) Submission in mmCIF format is strongly encouraged.

Members of cryo-EM and modeling communities were invited to participate in mid-April 2019; details were posted on the challenges website (challenges.emdataresource.org). Models were submitted by participant teams between May 1 and May 28, 2019. For apoferritin (APOF) targets, coordinate models were submitted as single subunits at the position of a provided segmented density consisting of a single subunit. Alcohol dehydrogenase (ADH) models were submitted as dimers. For each submitted model, metadata describing the full modeling workflow were collected via a Drupal webform, and coordinates were uploaded and converted to PDBx/mmCIF format using PDBextract[39]. Model coordinates were then processed for atom/residue ordering and nomenclature consistency using PDB annotation software (Feng Z., https://swtools.rcsb.org/apps/MAXIT) and additionally checked for sequence consistency and correct position relative to the designated target map. Models were then evaluated as described below (Model Evaluation System).

In early June, models, workflows, and initial calculated scores were made available to all participants for evaluation, blinded to modeler team identity and software used. A 2.5-day workshop was held in mid-June at Stanford/SLAC to review the results, with panel members attending in person. All modeling participants were invited to attend remotely and present overviews of their modeling processes and/or assessment strategies. Recommendations were made for additional evaluations of the submitted models as well as for future challenges. Modeler

teams and software were unblinded at the end of the workshop. In September, a virtual follow-up meeting with all participants provided an overview of the final evaluation system after implementation of recommended updates.

Modeling Software

Modelling teams created *ab initio* models or optimized previously known models available from the PDB. *Ab initio* software included ARP/wARP[40], Buccaneer[41,42], Cascaded-CNN[43], Mainmast[44, Terashi 2020], Pathwalker[45], and Rosetta[46]. Optimization software included CDMD[47], CNS[48], DireX[49], Phenix[27], REFMAC[19], MELD[50], MDFF[51], and reMDFF[52]. Participants made use of VMD[53], Chimera[54], and COOT[35] for visual evaluation and/or manual model improvement of map-model fit. See Table I for software used by each modeling team.

Model Evaluation System

The evaluation system for 2019 Challenge (model-compare.emdataresource.org) was built on the basis of the 2016/2017 Model Challenge system[12], updated with several new evaluation measures and analysis tools. Submitted models were evaluated for >70 individual metrics in four tracks: Fit-to-Map, Coordinates-only, Comparison-to-Reference, and Comparison-among-Models. A detailed description of the updated infrastructure and each calculated metric is provided as a help document on the model evaluation system website.

For brevity, a representative subset of metrics from the evaluation website are discussed in this paper. The selected metrics are listed in Table II, and are further described below. All scores were calculated according to package instructions using default parameters.

Fit-to-Map

The evaluated metrics included several ways to measure the correlation between map and model density[55], as implemented in TEMPy[22-25] v.1.1 (CCC, CCC_OV, SMOC, LAP, MI, MI_OV) and the Phenix[27] v.1.15.2 map_model_cc module[18] (CCbox, CCpeaks, CCmask). These methods compare the experimental map with a model map produced on the same voxel grid, integrated either over the full map or over selected masked regions. The model-derived map is generated to a specified resolution limit by inverting Fourier terms calculated from coordinates, B-factors, and atomic scattering factors. Some measures compare density-derived functions instead of density (Mutual Information, Laplacian[22]).

The newly introduced Q-score (MAPQ v1.2[14] plugin for UCSF Chimera[54] v.1.11) uses a real-space correlation approach to assess the resolvability of each model atom in the map. Experimental

map density is compared to a Gaussian placed at each atom position, omitting regions that overlap with other atoms. The score is calibrated by the reference Gaussian, which is formulated so that a highest score of 1 would be given to a well-resolved atom in a map at ~1.5 Å resolution. Lower scores (down to -1) are given to atoms as their resolvability and the resolution of the map decreases. The overall Q-score is the average value for all model atoms.

Measures based on Map-Model FSC curve, atom inclusion, and protein side chain rotamers were also compared. Phenix Map-Model FSC is calculated using a soft mask and is evaluated at FSC=0.5[18]. REFMAC FSCavg[19] (module of CCPEM[56] v1.4.1) integrates the area under the Map-Model FSC curve to a specified resolution limit[19]. EMDB Atom Inclusion determines the percentage of atoms inside the map at a specified density threshold[20]. TEMPy ENV is also threshold-based and penalizes unmodeled regions[22]. EMRinger (module of Phenix) evaluates backbone positioning by measuring the peak positions of unbranched protein C_γ atom positions versus map density in ring-paths around C_α-C_β bonds[21].

Coordinates-only

Standard measures assessed local configuration (bonds, bond angles, chirality, planarity, dihedral angles; Phenix model statistics module), protein backbone (MolProbity Ramachandran outliers[26]; Phenix molprobity module) and side-chain conformations, and clashes (MolProbity rotamers outliers and clashscore[26]; Phenix molprobity module).

New in this Challenge round is CaBLAM[28] (part of MolProbity and as Phenix cablam module), which employs two novel procedures to evaluate protein backbone conformation. In both cases, virtual dihedral pairs are evaluated for each protein residue *i* using C_α positions *i*-2 through *i*+2. To define CaBLAM outliers, the third virtual dihedral is between the CO groups flanking residue *i*. To define Calpha-geometry outliers, the third parameter is the Calpha virtual angle at *i*. The residue is then scored according to virtual triplet frequency in a large set of high-quality models from PDB[28].

Comparison-to-Reference and Comparison-among-Models

Assessing the similarity of the model to a reference structure and similarity among submitted models, we used metrics based on atom superposition (LGA GDT-TS and GDC scores[29] v.04.2019), interatomic distances (LDDT score[30] v.1.2), and contact area differences (CAD[32] v.1646). HBPLUS[57] was used to calculate nonlocal hydrogen bond precision, defined as the fraction of correctly placed hydrogen bonds in residue pairs with > 6 separation in sequence.

DAVIS-QA determines for each model the average of pairwise GDT-TS scores among all other models[33].

Local (per residue) Scores

Residue-level visualization tools for comparing the submitted models were also provided for the following metrics. Fit-to-Map: Phenix CCbox, TEMPy SMOC, Qscore, EMRinger, EMDB Atom Inclusion; Comparison-to-Reference: LGA and LDDT; Comparison-among-Models: DAVIS-QA.

Metric Score Pairwise Correlations and Distributions

For pairwise comparisons of metrics, Pearson correlation coefficients (P) were calculated for all model scores and targets (N=63). For average per-target pairwise comparisons of metrics, P values were determined for each target and then averaged. Metrics were clustered according to the similarity score (1-|P|) using a hierarchical algorithm with complete linkage. At the beginning, each metric was placed into a cluster of its own. Clusters were then sequentially combined into larger clusters, with the optimal number of clusters determined by manual inspection. In the fit-to-map evaluation track, the procedure was stopped after three divergent score clusters were formed for the all-model correlation data (Figure 4a), and after two divergent clusters were formed for the average per-target clustering (Figure 4b).

Score distributions are represented in box-and-whisker format in Figure 4c. Each box represents the interquartile range (IQR) and is drawn between Q1 (25th percentile) and Q3 (75th percentile) values. The inner horizontal line represents the median value (excluding outliers). Whisker lines extend out to the highest and lowest measured scores that are within 1.5*IQR of each box end. Scores falling outside the 1.5*IQR limits are considered outliers and are separately plotted as dots.

Controlling for Model Systematic Differences

As initially calculated, some Fit-to-Map scores had unexpected distributions, owing to differences in modeling practices among participating teams. For models submitted with all atom occupancies set to zero, occupancies were reset to one and rescored. In addition, model submissions were split approximately 50:50 for each of the following practices: (1) inclusion of hydrogen atom positions and (2) inclusion of refined B-factors. For affected fit-to-map metrics, modified scores were produced excluding hydrogen atoms and/or setting B-factors to zero. Both original and modified scores are provided in the web interface. Only modified scores were used in the pairwise metric comparisons described here.

Evaluation of Group Performance

Rating of group performance was done using the Model Compare Pipeline/Comparative Analyses/Model Ranks (per target) tool on the Challenge evaluation website. The tool permits users, for a specified target and for all or a subcategory of models (e.g., *ab initio*), to calculate Z-scores for each individual model, using any combination of 47 of the evaluated metrics with any desired relative weightings. The Z-scores for each metric are calculated from all submitted models for that target. The metrics (weights) used to generate individual-model Z-scores were as follows:

Coordinates-only: CaBLAM outliers (0.5), Calpha-geometry outliers (0.3), and Clashscore (0.2). CaBLAM outliers and Calpha-geometry outliers had the best correlation with match-to-target parameters (Figure 5b), and clashscore is an orthogonal measure. Ramachandran and rotamer criteria were excluded since they are often restrained in refinement and are zero for many models.

Fit-to-Map: EMringer (0.3) and Q-score (0.3), Atom Inclusion-backbone (0.2), and SMOC (0.2). EMringer and Q-score were among the most promising model-to-map metrics, and the other two provide distinct measures.

Comparison-to-Reference: LDDT (0.9) and GDC_all (0.9) and HBPR>6 (0.2). LDDT is superposition-independent and local, while GDC_all requires superposition; H-bonding is distinct. Metrics in this category are weighted higher, because although the target models are not perfect, they are a reasonable estimate of the right answer.

Individual Z-scores for each model were then averaged across each group's models on a given target, and further averaged across T1+T2 and across T3+T4, yielding overall Z-scores for high and low resolutions. The scores consistently showed 3 quite separate clusters: a good cluster at Z>0.3, an unacceptable cluster at Z<-0.3, and a small cluster near Z=0 (see Supplementary Figure 5). Other choices of metrics were tried, with very little effect on clustering.

Group 54 models were rated separately because they used different methods, their 2nd model versions were much better. Group 73's second model on target T4 was not rated because the metrics are not set up to meaningfully evaluate an ensemble.

Molecular Graphics

Molecular graphics images were generated using UCSF Chimera[54] (Figure 2 and Supplementary Figure 3: maps with superimposed models) and KiNG[58] (Supplementary Figures 2 and 4: maps with superimposed models and validation flags).

Acknowledgements

EMDataResource is supported by the US National Institutes of Health (NIH)/National Institute of General Medical Science, R01GM079429.

The following additional grants are acknowledged for participant support.

JSR and CW: NIH/ R35GM131883, NIH/P01GM063210.

AS: National Science Foundation (NSF)/MCB-1942763 (CAREER), NIH/R01GM095583. The Singharoy team used supercomputing resources of the OLCF at the Oak Ridge National Laboratory, which is supported by the Office of Science at DOE under Contract No. DE-AC05-00OR22725.

DKihara: NIH/ R01GM123055, NSF/DMS1614777, NSF/CMMI1825941, NSF/MCB1925643, Purdue Institute of Drug Discovery/DBI2003635.

JSF: NIH/R01GM123159.

MI: Max Planck Society German Research Foundation/IG 109/1-1.

ACV: Max Planck Society German Research Foundation /FOR-1805.

DKumar: NIH/R37AI36040 and Welch Foundation/Q1279 (PI: BVV Prasad).

DS: NSF/DBI2030381.

TB, CMP, MW: Medical Research Council MR/N009614/1.

APJ, MW: Wellcome Trust 208398/Z/17/Z.

KC: Biotechnology and Biological Sciences Research Council / BB/P000517/1.

MW: Biotechnology and Biological Sciences Research Council BB/P000975/1.

Author Contributions

PDA, PVA, JF, FDM, JSR, PBR, HMB, WC, AK, CLL, GDP, MFS: expert panel that selected targets, reference models and assessment metrics, set challenge rules; attended face-to-face results review workshop. KZ generated the APF maps for the challenge. MAH provided the published ADH map. CLL: designed and implemented the challenge model submission pipeline, drafted initial manuscript. Authors as listed in Table I: Built and submitted models; presented modeling strategies at review workshop. AK: designed and implemented evaluation pipeline and website, calculated scores. AK, CLL, BM, MAH, JSR, CJW, PVA, JF: analyzed models and model scores. AP, ZW, MT, ADJ, GDP, PVA, CJW: contributed software, advice on use and scores interpretation. CLL, AK, GDP, JSR: drafted figures. AK, HMB, GDP, WC, MFS, MAH, JSR: contributed to manuscript writing. All authors: reviewed and approved final manuscript.

Competing Interests

The authors declare no competing interests.

References

1 Merk, A. *et al.* Breaking Cryo-EM Resolution Barriers to Facilitate Drug Discovery. *Cell* **165**, 1698-1707, https://doi.org/10.1016/j.cell.2016.05.040 (2016).

2 Venien-Bryan, C., Li, Z., Vuillard, L. & Boutin, J. A. Cryo-electron microscopy and X-ray crystallography: complementary approaches to structural biology and drug discovery. *Acta Cryst F* **73**, 174-183, https://doi.org/10.1107/S2053230X17003740 (2017).

3 Mitra, A. K. Visualization of biological macromolecules at near-atomic resolution: cryo-electron microscopy comes of age. *Acta Cryst F* **75**, 3-11, https://doi.org/10.1107/S2053230X18015133 (2019).

4 Alewijnse, B. *et al.* Best practices for managing large CryoEM facilities. *J Struct Biol* **199**, 225-236, https://doi.org/10.1016/j.jsb.2017.07.011 (2017).

5 Lawson, C. L., Berman, H. M. & Chiu, W. Evolving data standards for cryo-EM structures. *Struct Dyn* **7**, 014701, https://doi.org/10.1063/1.5138589 (2020).

6 Henderson, R. *et al.* Outcome of the first electron microscopy validation task force meeting. *Structure* **20**, 205-214, https://doi.org/10.1016/j.str.2011.12.014 (2012).

7 Read, R. J. *et al.* A new generation of crystallographic validation tools for the Protein Data Bank. *Structure* **19**, 1395-1412, https://doi.org/10.1016/j.str.2011.08.006 (2011).

8 Montelione, G. T. *et al.* Recommendations of the wwPDB NMR Validation Task Force. *Structure* **21**, 1563-1570, https://doi.org/10.1016/j.str.2013.07.021 (2013).

9 Ludtke, S. J., Lawson, C. L., Kleywegt, G. J., Berman, H. & Chiu, W. The 2010 cryo-em modeling challenge. *Biopolymers* **97**, 651-654, https://doi.org/10.1002/bip.22081 (2012).

10 Lawson, C. L. & Chiu, W. Comparing cryo-EM structures. *J Struct Biol* **204**, 523-526, https://doi.org/10.1016/j.jsb.2018.10.004 (2018).

11 Heymann, J. B. *et al.* The first single particle analysis Map Challenge: A summary of the assessments. *J Struct Biol* **204**, 291-300, https://doi.org/10.1016/j.jsb.2018.08.010 (2018).

12 Kryshtafovych, A., Adams, P. D., Lawson, C. L. & Chiu, W. Evaluation system and web infrastructure for the second cryo-EM model challenge. *J Struct Biol* **204**, 96-108, https://doi.org/10.1016/j.jsb.2018.07.006 (2018).

13 Editorial. Challenges for cryo-EM. *Nat Methods* **15**, 985, https://doi.org/10.1038/s41592-018-0256-z (2018).

14 Pintilie, G. *et al.* Measurement of atom resolvability in cryo-EM maps with Q-scores. *Nat Methods*, https://doi.org/10.1038/s41592-020-0731-1 (2020).

15 Herzik, M. A., Jr., Wu, M. & Lander, G. C. High-resolution structure determination of sub-100 kDa complexes using conventional cryo-EM. *Nat Commun* **10**, 1032, https://doi.org/10.1038/s41467-019-08991-8 (2019).

16 Masuda, T., Goto, F., Yoshihara, T. & Mikami, B. The universal mechanism for iron translocation to the ferroxidase site in ferritin, which is mediated by the well conserved transit site. *Biochem Biophys Res Commun* **400**, 94-99, https://doi.org/10.1016/j.bbrc.2010.08.017 (2010).

17 Rosenthal, P. B. & Henderson, R. Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. *J Mol Biol* **333**, 721-745, https://doi.org/10.1016/j.jmb.2003.07.013 (2003).

18 Afonine, P. V. *et al.* New tools for the analysis and validation of cryo-EM maps and atomic models. *Acta Cryst D* **74**, 814-840, https://doi.org/10.1107/S2059798318009324 (2018).

19 Brown, A. *et al.* Tools for macromolecular model building and refinement into electron cryo-microscopy reconstructions. *Acta Cryst D* **71**, 136-153, https://doi.org/10.1107/S1399004714021683 (2015).

20 Lagerstedt, I. *et al.* Web-based visualisation and analysis of 3D electron-microscopy data from EMDB and PDB. *J Struct Biol* **184**, 173-181, https://doi.org/10.1016/j.jsb.2013.09.021 (2013).

21 Barad, B. A. *et al.* EMRinger: side chain-directed model and map validation for 3D cryo-electron microscopy. *Nat Methods* **12**, 943-946, https://doi.org/10.1038/nmeth.3541 (2015).

22 Vasishtan, D. & Topf, M. Scoring functions for cryoEM density fitting. *J Struct Biol* **174**, 333-343, https://doi.org/10.1016/j.jsb.2011.01.012 (2011).

23 Farabella, I. *et al.* TEMPy: a Python library for assessment of three-dimensional electron microscopy density fits. *J Appl Crystallogr* **48**, 1314-1323, https://doi.org/10.1107/S1600576715010092 (2015).

24 Joseph, A. P. *et al.* Refinement of atomic models in high resolution EM reconstructions using Flex-EM and local assessment. *Methods* **100**, 42-49, https://doi.org/10.1016/j.ymeth.2016.03.007 (2016).

25 Joseph, A. P., Lagerstedt, I., Patwardhan, A., Topf, M. & Winn, M. Improved metrics for comparing structures of macromolecular assemblies determined by 3D electron-microscopy. *J Struct Biol* **199**, 12-26, https://doi.org/10.1016/j.jsb.2017.05.007 (2017).

26 Chen, V. B. *et al.* MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Cryst D* **66**, 12-21, https://doi.org/10.1107/S0907444909042073 (2010).

27 Liebschner, D. *et al.* Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Cryst D* **75**, 861-877, https://doi.org/10.1107/S2059798319011471 (2019).

28 Williams, C. J. *et al.* MolProbity: More and better reference data for improved all-atom structure validation. *Protein Sci.* **27**, 293-315, https://doi.org/10.1002/pro.3330 (2018).

29 Zemla, A. LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Res* **31**, 3370-3374, https://doi.org/10.1093/nar/gkg571 (2003).

30 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* **29**, 2722-2728, https://doi.org/10.1093/bioinformatics/btt473 (2013).

31 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* **7**, 10480, https://doi.org/10.1038/s41598-017-09654-8 (2017).

32 Olechnovic, K., Kulberkyte, E. & Venclovas, C. CAD-score: a new contact area difference-based function for evaluation of protein structural models. *Proteins* **81**, 149-162, https://doi.org/10.1002/prot.24172 (2013).

33 Kryshtafovych, A., Monastyrskyy, B. & Fidelis, K. CASP prediction center infrastructure and evaluation measures in CASP10 and CASP ROLL. *Proteins* **82 Suppl 2**, 7-13, https://doi.org/10.1002/prot.24399 (2014).

34 Prisant, M. G., Williams, C. J., Chen, V. B., Richardson, J. S. & Richardson, D. C. New tools in MolProbity validation: CaBLAM for CryoEM backbone, UnDowser to rethink "waters," and NGL Viewer to recapture online 3D graphics. *Protein Sci.* **29**, 315-329, https://doi.org/10.1002/pro.3786 (2020).

35 Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Cryst D* **66**, 486-501, https://doi.org/10.1107/S0907444910007493 (2010).

36 Headd, J. J. *et al.* Use of knowledge-based restraints in phenix.refine to improve macromolecular refinement at low resolution. *Acta Cryst D* **68**, 381-390, https://doi.org/10.1107/S0907444911047834 (2012).

37 Kryshtafovych, A. *et al.* Cryo-electron microscopy targets in CASP13: Overview and evaluation of results. *Proteins* **87**, 1128-1140, https://doi.org/10.1002/prot.25817 (2019).

38 Gore, S. *et al.* Validation of Structures in the Protein Data Bank. *Structure* **25**, 1916-1927, https://doi.org/10.1016/j.str.2017.10.009 (2017).

39 Yang, H. *et al.* Automated and accurate deposition of structures solved by X-ray diffraction to the Protein Data Bank. *Acta Cryst D* **60**, 1833-1839, https://doi.org/10.1107/S0907444904019419 (2004).

40 Chojnowski, G., Pereira, J. & Lamzin, V. S. Sequence assignment for low-resolution modelling of protein crystal structures. *Acta Cryst D* **75**, 753-763, https://doi.org/10.1107/S2059798319009392 (2019).

41 Cowtan, K. The Buccaneer software for automated model building. 1. Tracing protein chains. *Acta Cryst D* **62**, 1002-1011, https://doi.org/10.1107/S0907444906022116 (2006).

42 Hoh, S., Burnley, T. & Cowtan, K. Current approaches for automated model building into cryo-EM maps using Buccaneer with CCP-EM. *Acta Cryst D* **76**, 531-541, https://doi.org/10.1107/S2059798320005513 (2020).

43 Si, D. *et al.* Deep Learning to Predict Protein Backbone Structure from High-Resolution Cryo-EM Density Maps. *Sci Rep* **10**, 4282, https://doi.org/10.1038/s41598-020-60598-y (2020).

44 Terashi, G. & Kihara, D. De novo main-chain modeling for EM maps using MAINMAST. *Nat Commun* **9**, 1618, https://doi.org/10.1038/s41467-018-04053-7 (2018).

45 Chen, M. & Baker, M. L. Automation and assessment of de novo modeling with Pathwalking in near atomic resolution cryoEM density maps. *J Struct Biol* **204**, 555-563, https://doi.org/10.1016/j.jsb.2018.09.005 (2018).

46 Frenz, B., Walls, A. C., Egelman, E. H., Veesler, D. & DiMaio, F. RosettaES: a sampling strategy enabling automated interpretation of difficult cryo-EM maps. *Nat Methods* **14**, 797-800, https://doi.org/10.1038/nmeth.4340 (2017).

47 Igaev, M., Kutzner, C., Bock, L. V., Vaiana, A. C. & Grubmuller, H. Automated cryo-EM structure refinement using correlation-driven molecular dynamics. *Elife* **8**, https://doi.org/10.7554/eLife.43542 (2019).

48 Brünger, A. T. Version 1.2 of the Crystallography and NMR system. *Nature protocols* **2**, 2728-2733, https://doi.org/10.1038/nprot.2007.406 (2007).

49 Wang, Z. & Schröder, G. F. Real-space refinement with DireX: from global fitting to side-chain improvements. *Biopolymers* **97**, 687-697, https://doi.org/10.1002/bip.22046 (2012).

50 MacCallum, J. L., Perez, A. & Dill, K. A. Determining protein structures by combining semireliable data with atomistic physical models by Bayesian inference. *Proc Natl Acad Sci U S A* **112**, 6985-6990, https://doi.org/10.1073/pnas.1506788112 (2015).

51 Trabuco, L. G., Villa, E., Mitra, K., Frank, J. & Schulten, K. Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics. *Structure* **16**, 673-683, https://doi.org/10.1016/j.str.2008.03.005 (2008).

52 Singharoy, A. *et al.* Molecular dynamics-based refinement and validation for sub-5 A cryo-electron microscopy maps. *Elife* **5**, https://doi.org/10.7554/eLife.16105 (2016).

53 Hsin, J., Arkhipov, A., Yin, Y., Stone, J. E. & Schulten, K. Using VMD: an introductory tutorial. *Curr Protoc Bioinformatics* **Chapter 5**, Unit 5 7, https://doi.org/10.1002/0471250953.bi0507s24 (2008).

54 Pettersen, E. F. *et al.* UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem* **25**, 1605-1612, https://doi.org/10.1002/jcc.20084 (2004).

55 Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved Methods for Building Protein Models in Electron-Density Maps and the Location of Errors in These Models. *Acta Cryst A* **47**, 110-119, https://doi.org/10.1107/S0108767390010224 (1991).

56 Burnley, T., Palmer, C. M. & Winn, M. Recent developments in the CCP-EM software suite. *Acta Cryst D* **73**, 469-477, https://doi.org/10.1107/S2059798317007859 (2017).

57 McDonald, I. K. & Thornton, J. M. Satisfying hydrogen bonding potential in proteins. *J Mol Biol* **238**, 777-793, https://doi.org/10.1006/jmbi.1994.1334 (1994).

58 Chen, V. B., Davis, I. W. & Richardson, D. C. KING (Kinemage, Next Generation): a versatile interactive molecular and scientific visualization program. *Protein Sci.* **18**, 2403-2409, https://doi.org/10.1002/pro.250 (2009).

59 wwPDB Consortium. Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res* **47**, D520-D528, https://doi.org/10.1093/nar/gky949 (2019).

60 Kuhlbrandt, W. Biochemistry. The resolution revolution. *Science* **343**, 1443-1444, https://doi.org/10.1126/science.1251652 (2014).

Figure Legends/Figures

Figure 1. Single particle cryo-EM models in PDB

The availability of models in the Protein Data Bank[59] derived from single particle cryo-EM maps has increased dramatically since the "resolution revolution" circa 2014[60]. (a) Plot shows that the steepest increase is for structures with reported resolution in the 3-4 Å range (orange bars). Higher resolution structures (blue bars), the topic of the Challenge presented here, are also beginning to trend upward. (b) EMDataResource Challenge activities timeline.

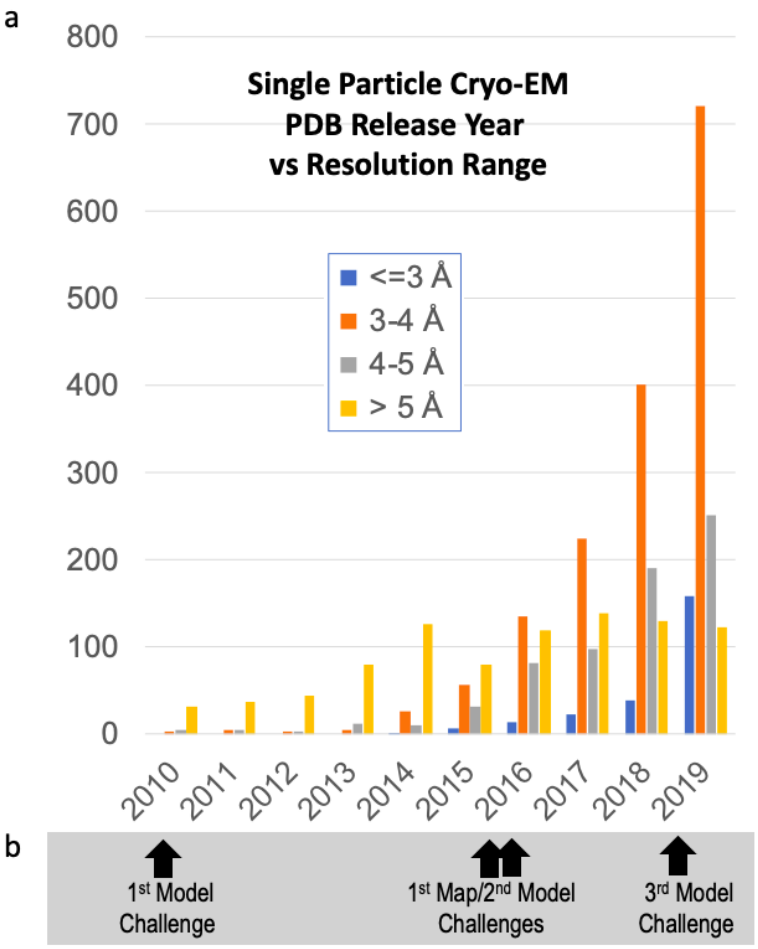

Figure 2. Challenge targets: cryo-EM maps at near-atomic resolution

Three density maps of α-helix rich apoferritin (APOF) at 1.8, 2.3, and 3.1 Å form a resolution series differing only in the number of particles averaged (EM Data Bank entries EMD-20026, EMD-20027, EMD-20028). The fourth density map, alcohol dehydrogenase (ADH) at 2.9 Å, contains both α-helices and β-sheets, as well as ligands (yellow density bound to blue and red subunits is NAD; EMDB entry EMD-0406). (a) Full map for each target. (b,c) Representative secondary structural elements (APOF: residues 14-42; ADH: residues 34-45) with masked density for protein backbone atoms only (b), and for all protein atoms (c).

Over the 1.8-3.1 Å resolution range represented by the four target maps, visible map features transition from near-atomic to secondary structure-dominated. At 1.8 Å (APOF), most protein atom positions are well defined by the map density: backbone protrusions delineate carbonyl oxygen positions and holes appear inside aromatic rings. At 2.3 Å (APOF), most protein atom positions are within the contoured map density; carbonyl oxygen atom “bumps” in the map help to define direction of backbone trace. At 2.9 Å (ADH) and 3.1 Å (APOF), secondary structure features begin to predominate. Notably, carbonyl oxygen atoms “bumps” are absent in the map, making it harder to identify the direction of backbone trace.

(d) EMDB maps used in model Fit-to-Map analysis (APOF targets: masked single subunits; ADH: unmasked sharpened map). The molecular boundary is shown in red (EMDB recommended contour level), background noise is represented in grey (1/3rd of EMDB recommended contour level), and the full map extent is indicated by the black outline.

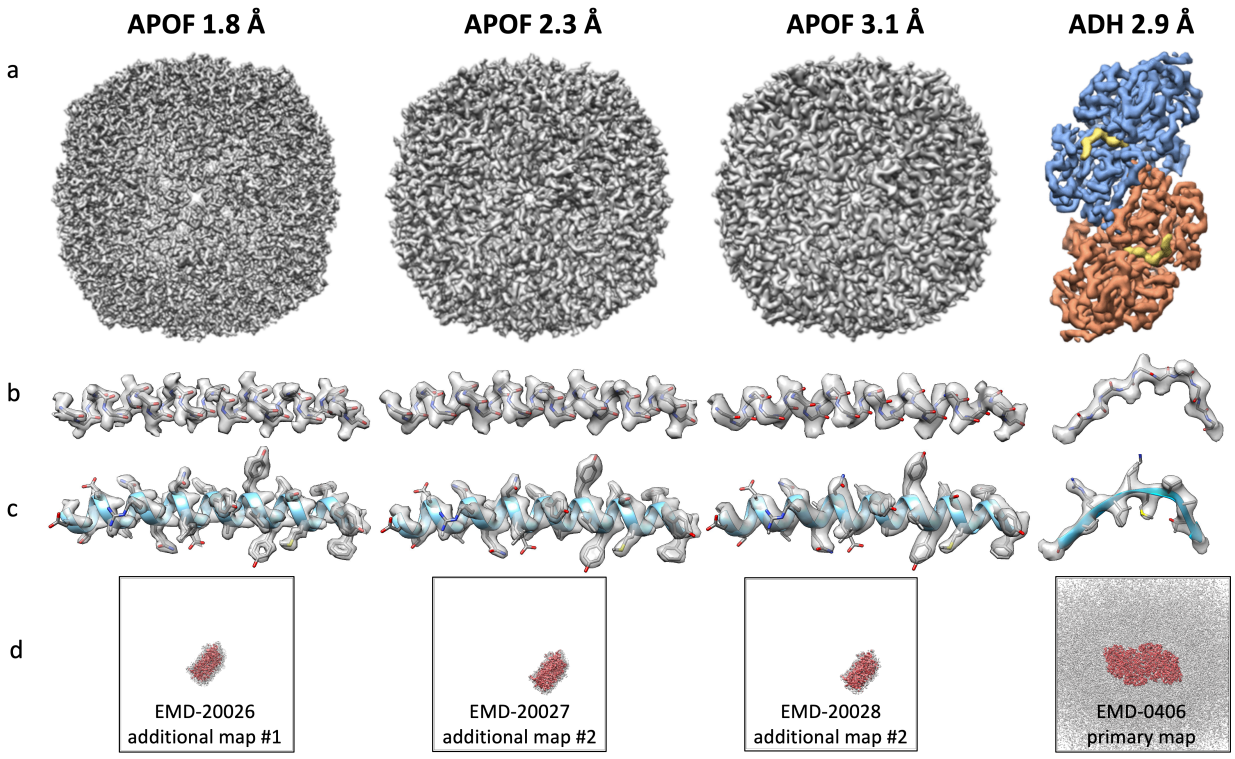

Figure 3. Challenge pipeline

Setup (a): A panel of experts selected four target maps in the 1.8-3.1 Å resolution range as well as reference models.

Submissions (b): Methods involving *ab initio* and optimization, both with and without additional manual steps, were represented by 63 models submitted by 16 modeling teams.

Evaluation (c): Building on the previous model challenge round[12], the evaluation system was organized in four tracks (coordinates only, fit-to-map, comparison to reference model, and comparison among models), each with its own set of software tools for generating scores.

Scores Comparison (d): Multiple interactive tabular and graphical displays enable comparative evaluations on the model-compare website (model-compare.emdataresource.org).

Top: Map-Model Fourier Shell Correlation (FSC) curves for models submitted against the APOF 1.8 Å target (random light colors), versus curve for the reference model (bold cherry red). Map-Model FSC measures the overall agreement of the experimental density map with a density map derived from the coordinate model (model map)[17]. Curves are calculated from the Fourier coefficients of the two maps and plotted vs. frequency ($resolution^{-1}$). Visualization of the full curve is useful to ensure that it follows the expected sigmoidal shape with the tail decaying exponentially. The resolution value corresponding to FSC=0.5 (black horizontal line) is typically reported, with smaller values indicating better fit. In this example, some overlaid curves indicate poor agreement between submitted model and the APOF 1.8 Å experimental map, but the majority indicate equivalence to or even improvement over the X-ray reference model (PDB entry 3ajo was rigid-body fitted to the cryoEM target map without further refinement).

Bottom: scores comparison tool for ADH 2.9 Å target models. Interactive score distribution sliders reveal at a glance how well the submitted models performed relative to each other. Parallel lanes display score distributions for each evaluated metric. They are conceptually similar to the graphical display for key metrics used in wwPDB validation reports[7,38]. Displayed here are the score distributions for all models submitted against the ADH target for four representative metrics, one from each evaluation track, and scores for two individual models are also highlighted. Model scores are plotted horizontally (semi-transparent diamonds) with color-coding to indicate worse (left, orange) and better (right, green) values. Darker, opaque diamonds indicate multiple overlapping scores. The interactive display enables scores for individual models to be identified and compared (red and blue triangles). The model indicated by red triangles scored better than the model indicated by blue triangles for Fit-to-Map: Q-score, Comparison-to-Reference: LDDT,

and Comparison-among-Models: Davis-QA, but the same model scored lower for Coordinates-only: CaBLAM Conformation. The example demonstrates the importance of evaluating models in each track: a high score in any one metric/track does not guarantee full optimization as judged by another metric.

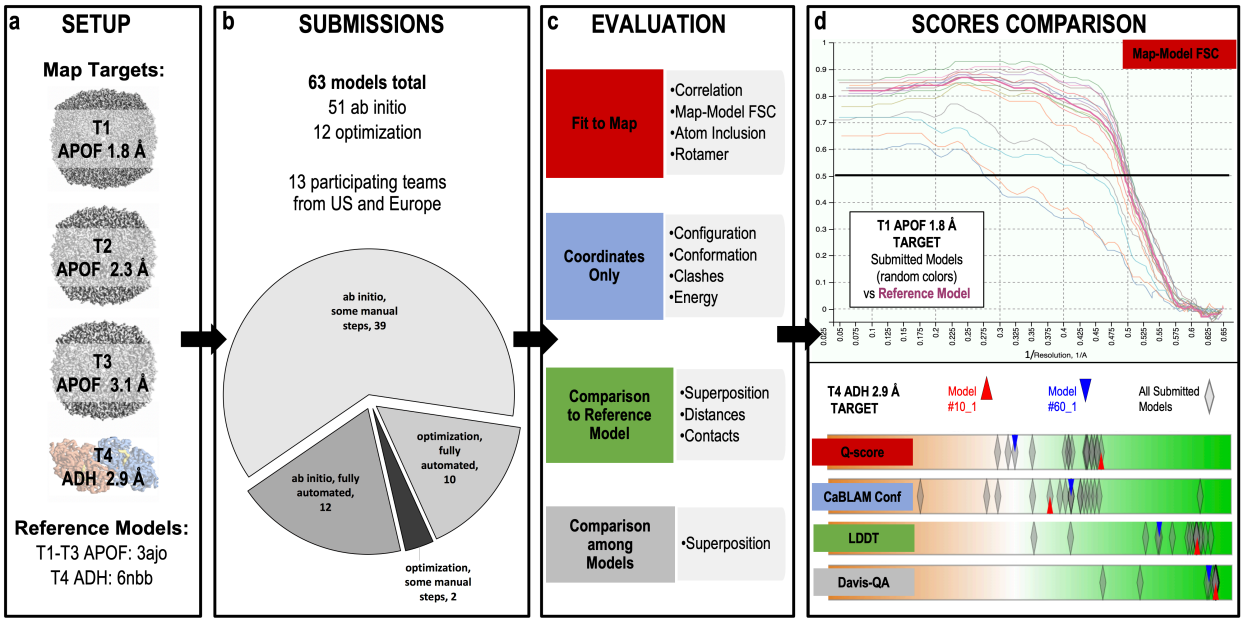

Figure 4. Evaluation of metrics

Model metrics (Table II) were compared with each other to assess how similarly they performed in scoring the Challenge models.

(a-d) Fit-to-Map metrics were compared with each other to assess how similarly they performed in scoring the Challenge models. Their similarity was evaluated in two ways: (a) Pairwise correlation coefficients were calculated for all models across all map targets (n=63); (b) Average correlation per target was calculated (separate correlation coefficients for each map target were averaged). In both, pairs of metrics that are strongly similar in performance are indicated by dark shading (high correlation/black: 0.85-1.0; moderately high correlation/grey: 0.7-0.84). Those that perform very differently (poor correlation) are indicated with no shading. Correlation-based metrics are identified by bold labels.

The ordering of metrics in (a) is based on hierarchical cluster analysis of all-model correlation values (see Methods). Three red-outlined boxes along the table diagonal correspond to identified clusters (#1-3); labels at left are also boxed in red according to these clusters. For ease of comparison, the ordering of metrics in (b) is identical to (a). The red-outlined box in (b) identifies the single cluster identified by hierarchical cluster analysis of the per-target correlation averages. This one cluster includes all metrics but ENV.

In (c), representative score distributions for nine metrics, three from each cluster in (a), are plotted for each map target. These plots illustrate the systematic differences in scoring per map target that are responsible for the division of the evaluated metrics into the three clusters. In Cluster 1, score distributions are lowest for the highest resolution target and increase as map resolution decreases. Cluster 2 metrics have the opposite trend: score distributions are highest for the highest resolution target and decrease as map resolution decreases. Cluster 3 metrics have a mixed trend with respect to map resolution, but uniformly have lower score distributions for ADH map target models relative to all three APOF map target models. See the main text for discussion of the most likely factors behind these trends.

In (d), scores for one representative pair of metrics that belong to different clusters in (a) but to the same cluster in (b) are plotted against each other. As highlighted by the diagonal lines representing linear fits to scores for map target, both metrics (CCbox from Cluster 1 and Qscore from Cluster 2) perform very similarly to each other within any one map target. Different sensitivities of the two metrics to map-specific factors give rise to the separate and parallel spacings of the four diagonal lines, with score ranges on different relative scales for each target.

(e) Coordinates-only and (f) Comparison-to-Reference metrics comparisons. Analogously to (a) and (b), the Pearson correlation coefficient was determined for each metric pair for all submitted model scores (n=63) across all map targets. In (f), Comparison-to-reference metrics are contrasted with each other as well as with several Fit-to-Map and Coordinates-only metrics.

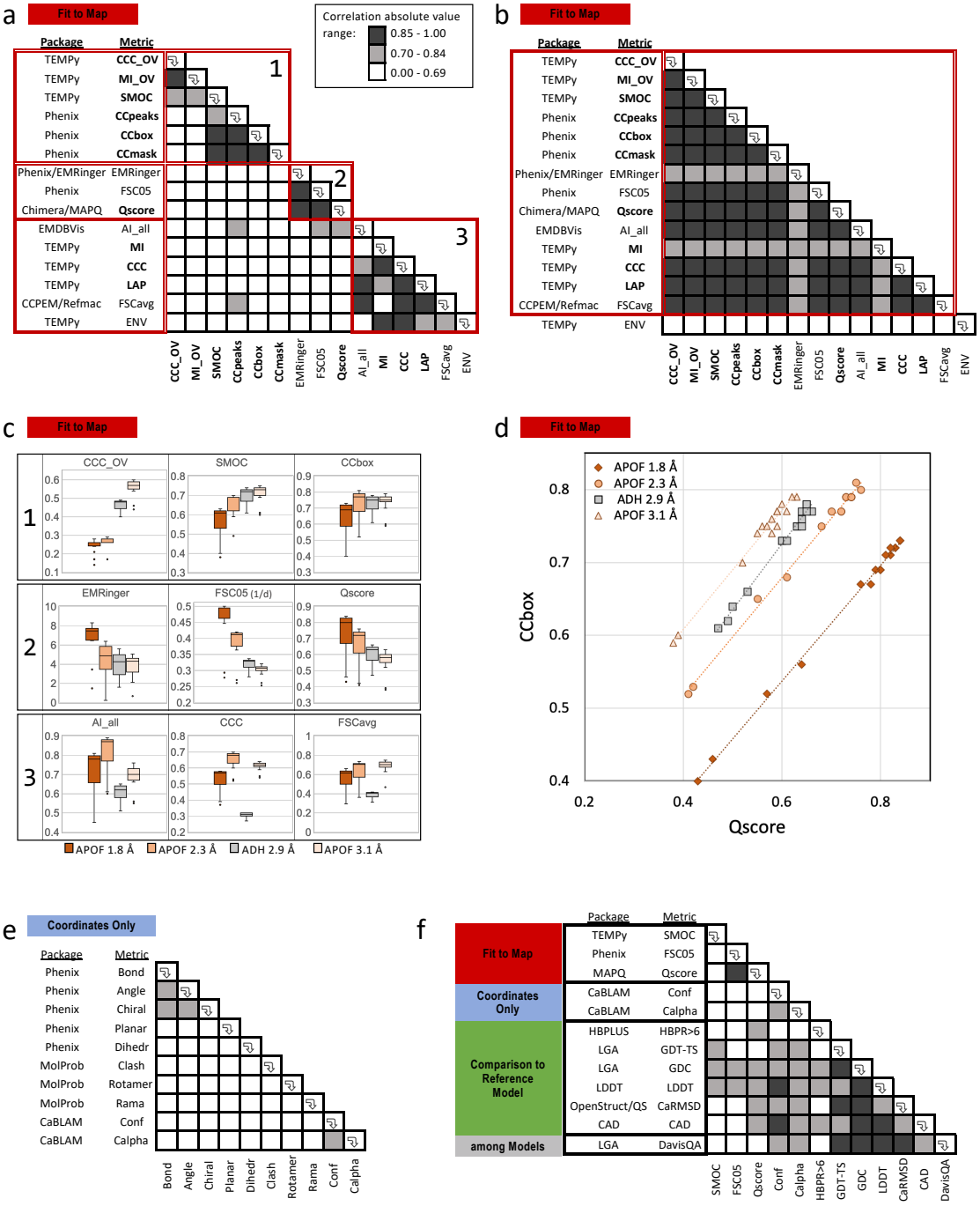

Table I. Participating Modeling Teams

Team ID, name	Team Members	# of Submitted Models	Effort Type(s)	Software
10 yu	Xiaodi Yu	4	ab initio+manual	Phenix, Buccaneer, Chimera, Coot, Pymol
25 cdmd	Maxim Igaev, Andrea Vaiana	4	optimization automated	CDMD
27 kumar	Dilip Kumar	1	ab initio+manual	Phenix, Rosetta, Buccaneer, ARP/wARP, Coot
28 ccpem	Soon Wen Hoh, Kevin Cowtan, Agnel Praveen Joseph, Colin Palmer, Martyn Winn, Tom Burnley, Mateusz Olek, Paul Bond, Eleanor Dodson	4	ab initio+manual	CCP-EM, Refmac, Buccaneer, Coot, TEMPy
35 phenix	Pavel Afonine, Tom Terwilliger, Li-Wei Hung	4	ab initio+manual	Phenix, Coot
38 fzjuelich	Gunnar Schroeder, Luisa Schaefer	3	optimization automated	Phenix, Chimera, DireX, MDFF, CNS, Gromacs
41 arpwarp	Grzegorz Chojnowski	8	ab initio automated, ab initio+manual	Refmac, ARP/wARP, Coot
54 kihara	Daisuke Kihara, Genki Terashi	8	ab initio+manual	Rosetta, Mainmast, MDFF, Chimera
60 deeptracer	Liguo Wang, Dong Si, Renzhi Cao, Jianlin Cheng, Spencer A. Moritz, Jonas Pfab, Tianqi Wu, Jie Hou	10	ab initio automated, ab initio+manual	Cascaded-CNN, Chimera
73 singharoy	Mrinal Shekhar, Genki Terashi, Sumit Mittal, Daipayan Sarkar, Daisuke Kihara, Ken Dill, Alberto Perez, Abishek Singharoy	5	ab initio+manual, optimization automated	reMDFF, MELD, VMD, Chimera, Mainmast
82 rosetta	Frank DiMaio, Dan Farrell	8	ab initio automated, ab initio+manual	Rosetta, Chimera
90 mbaker	Matt Baker	2	ab initio+manual	Pathwalker, Phenix, Chimera, Coot
91 chiu	Greg Pintilie, Wah Chiu	2	optimization+manual	Phenix, Chimera, Coot

Table II. Evaluated Metrics

	Metric Class	Package Metric Definition
Fit-to-Map	Correlation Coefficient, all voxels	Phenix CCbox full grid map vs model-map density correlation coefficient[18] TEMPy CCC full grid map vs model-map density correlation coefficient[23]
	Correlation Coefficient, selected voxels	Phenix CCmask map vs model-map density, only modelled regions[18] Phenix CCpeaks map vs model-map density, only high-density map and model regions[18] TEMPy CCC_OV map vs model-map density, overlapping map and model regions[25] TEMPy SMOC Segment Manders' Overlap, map vs model-map density, only modelled regions[25]
	Correlation Coefficient, other density function	TEMPy LAP map vs model-map Laplacian filtered density (partial 2nd derivative)[22] TEMPy MI map vs model-map Mutual Information entropy-based function[22] TEMPy MI_OV map vs model-map Mutual Information, only modelled regions[25]
	Correlation Coefficient, atom positions	Chimera/MAPQ Qscore map density at each modeled atom vs reference Gaussian density function[14]
	Fourier Shell Correlation	Phenix FSC05 Resolution (distance) of Map-Model FSC curve read at point FSC=0.5[18] CCPEM/Refmac FSCavg FSC curve area integrated to map resolution limit[19,56]
	Atom Inclusion	EMDB/VisualAnalysis AI_all Atom Inclusion, percentage of all atoms inside depositor-provided density threshold[20] TEMPy ENV Atom Inclusion in envelope corresponding to sample MW; penalizes unmodeled regions[22]
	Side Chain Density	Phenix EMRinger evaluates backbone positioning by sampling map density around Cγ-atom ring-paths for non-branched residues[21]
Coordinates-only	Configuration	Phenix Bond Root-mean-square deviation (RMSD) of bond lengths[27] Phenix Angle RMSD of bond angles[27] Phenix Chiral RMSD of chiral centers[27] Phenix Planar RMSD of planar group planarity[27] Phenix Dihedral RMSD of dihedral angles[27]
	Clashes	MolProbity Clashscore Number of steric overlaps ≥0.4Å per 1000 atoms [26]
	Conformation	MolProbity Rotamer sidechain conformation outliers[26] MolProbity Rama Ramachandran ϕ,ψ mainchain conformation outliers[26] MolProbity CaBLAM 3D outliers of CO-CO virtual dihedral incompatible with 2 surrounding Ca-Ca virtual dihedrals[28] MolProbity Calpha 3D outliers of 2 Cα virtual mainchain dihedrals plus Calpha virtual bond angle[28]
vs. Reference Model(s)	Atom Superposition	LGA GDT-TS Global Distance Test Total Score, average % of model Cα that superimpose with reference Cα, multiple distance cutoffs[29] LGA GDC Global Distance Calculation, average % of all model atoms that superimpose with reference, multiple distance cutoffs[29] OpenStructure/QS CaRMSD Root-Mean-Square Deviation of Cα atoms[31]
	Interatomic Distances	LDDT LDDT Local Difference Distance Test, superposition-free comparison of all-atom distance maps between model and reference[30]
	Contact Area	CAD CAD Contact Area Difference, superposition-free measure of differences in interatom contacts[32] HBPLUS[57] HBPRC>6, Hydrogen bond precision, nonlocal. fraction of correctly placed hydrogen bonds in residue pairs with > 6 separation in sequence.
	Atom Superposition, Multiple	DAVIS QA average of pairwise LGA GDT-TS scores among submitted models[33]