nature research

Peer Review Information

Journal: Nature Methods
Manuscript Title: Residue-Wise Local Quality Estimation for Protein Models from Cryo-EM Maps
Corresponding author name(s): Daisuke Kihara

Reviewer Comments & Decisions:

Decision Letter, initial version:

Dear Daisuke,

Your Article entitled "Residue-Wise Local Quality Estimation for Protein Models from Cryo-EM Maps" has now been seen by three reviewers, whose comments are attached. While they find your work of some potential interest, they have raised concerns which in our view are sufficiently important that they preclude publication of the work in Nature Methods.

We will consider looking at a revised manuscript only if further experimental data allow you to address all the major criticisms of the reviewers (unless, of course, something similar has by then been accepted at Nature Methods or appeared elsewhere). This includes submission or publication of a portion of this work somewhere else.

A successful revision would address all the technical concerns, clarify the strengths of the tool relative to similar tools mentioned by the reviewers, remove any circular experiments that reinforce expectations, and make a stronger case that the method actually helps improve maps, not just find errors. Revisions for clarity and presentation will also be expected.

Please keep in mind, we cannot make a decision regarding sending the manuscript out for review again until we've seen the revision.

If you are interested in revising this manuscript for submission to Nature Methods in the future, please contact me to discuss your appeal before making any revisions. Otherwise, we hope that you find the reviewers' comments helpful when preparing your paper for submission elsewhere.

Sincerely,
Rita

Rita Strack, Ph.D.
Senior Editor
Nature Methods

Reviewers' Comments:

Reviewer #1:
Remarks to the Author:
A. Summary of the key results

The authors present a new method to evaluate the fit of a protein model to cryo-EM map, called DAQ score, or Deep-learning Amino acid-wise model Quality score. The method uses a neural network to characterize what the local environment of each type of residue looks like, based on a large number of maps and models. When applied to a given voxel in a map, the neural network then gives number which indicates to what degree that voxel represents either a secondary structure (SSscore), given amino acid type (AAscore) and Ca atom (ATOMscore). To evaluate a protein model, each score is calculated at the nearest voxel to the Ca atom of each residue. A positive value indicates the residue is likely ok, whereas a negative residue is flagged as possibly incorrect. The authors apply the scores to 15 structures in the PDB with updated structures, and show that the scores go up for the newer structures for residues that moved more than 2Å. They also apply it to 399 pairs of models in the PDB which have high sequence similarity but are different by more than 1Å RMSD, and a further data set of 4,485 models.

B. Originality and significance: if not novel, please include reference

This method is novel to my knowledge, although there are some similarities to Deeptracer (referenced) and Haruspex (not referenced – Mostosi et. Al., Angewandte Chemie, 2020).

C. Data & methodology: validity of approach, quality of data, quality of presentation

The approach seems valid, though there are limitations not adequately described, e.g. effect of resolution, step size, etc. More details and questions in the points below. The data used, coming from

the PDB/EMDB is of good quality. The presentation is also mostly ok, clearly described overall, thought again some further clarifications are needed.

D. Appropriate use of statistics and treatment of uncertainties

Statistics are used appropriately, for example to show that DAQ score goes up for improved regions of a model. Uncertainties are not addressed properly however, for example, the possibility that the map is not resolved in some regions is mostly ignored. The effect of resolution is also ignored; it would be very informative to know how the score can be applied and what to expect at different resolutions.

E. Conclusions: robustness, validity, reliability

The robustness seems reasonable, i.e. higher scores are obtained for improved model. However the validity and reliability at the per-residue level is not clear, since sliding windows; illustrations of actual scores for single residues in more cases would be useful to demonstrate the reliability of the score in different scenarios. Because the output is directly related to the training data, if the training data does not include all valid conformations of a residue, then the output could be incorrect. This limitation need to be mentioned. Based on looking at the provided result in the github repo, the results do not seem robust (details below).

F. Suggested improvements: experiments, data for possible revision

A potentially necessary improvement would be to illustrate the score for at least a few types of residues at different resolutions. Otherwise, it is relied upon general statistics to demonstrate. Other improvements would come mostly from being more clear on the limitations of the method and being more realistic with the conclusions.

G. References: appropriate credit to previous work?

Yes, but could be improved. They may consider to reference the Haruspex method. It would be appropriate to also reference other per-residue scoring methods such as TEMPY (CCPEM), and Phenix.

H. Clarity and context: lucidity of abstract/summary, appropriateness of abstract, introduction and conclusions

Descriptions are mostly clear, though further clarifications may be needed as indicated below. The abstract is a bit ambitious for what is actually presented, and the conclusions do not adequately indicate limitations.

Specific Points:

Abstract: It is stated that the method can identify "potential misassignment of residues in the map, including residue shifts". Later it is stated that none of the scores can tell the difference between misaligned or mispositioned residues (page 13). Perhaps these can be better consolidated and more clearly/realistically stated.

Page 3 bottom. For map-model scores, atom inclusion score should also be mentioned (reference Lagerstedt et al, JSB, 2013). It should be mentioned that it is currently used in the EMDB/PDB validation report.

Page 4 top: Molprobity score also includes bond lengths, and angles between 3 connected atoms.

Page 4 bottom paragraph. Suggest to add 'protein' in 'new approach for cryo-EM model validation'.

Page 5 top paragraph. To say that 'local structural features' are used instead of 'local density values' seems contradictory/misleading because ultimately the neural network is fed these local density values.

Page 5, bottom paragraph. It is stated that comparison of DAQ score to other scores showed a 'clear advantage' for the DAQ score. This may be too strong/vague/general of a statement. There are reasons why this is not so, for example, the training set does not include maps with higher than ~2.6Å resolution, so it's not clear what the method would do there. Also, the method can only be applied to protein residues. Lastly, more thorough tests and comparisons to other methods are needed before such a statement should be made.

Page 6 middle paragraph. 'The probabilities for a grid point …'. Not clear, is a 11^3 box used centered at the grid point?

Page 7 middle. 'atom type score of amino acid residue i'. The ATOMscore tries to indicate if a voxel appears to indicate a given type of atom, but only only the Ca atom is considered here. Atom type usually is meant to reflect whether an atom is C, O, N, etc. Indicating this score has something to do with

atom type is misleading. Perhaps this can be revised somehow to be more specific towards identifying the Ca atom.

Page 7, bottom paragraph. The accuracies (77.3% for secondary structures, 51.5% for Ca atoms, and 28.2% for amino acid type) seem low. It is mentioned that the accuracies do not need to be perfect for model validation, as long as the 'correct' residue is preferred for a position in the map. It is not described what is meant by 'preferred', and it seems quite likely that even this method could potentially give a higher score to the incorrect residue, i.e. it is not shown otherwise, which would be hard to do; unless it can be done, this statement needs to be revised).

Page 9. The two classes misaligned or mispositioned; misaligned have to do with shifted residues, but is the simple criteria that a residue is within 2Å of another residue always indicate a shift? They are a bit hard to comprehend from the text, could they be illustrated perhaps?

Page 9 bottom paragraph, Figure 2. Average DAQ scores of first and revised version are plotted. Figure 2a shows they are correlated when considering all residues, but what does this correlation mean - the higher the DAQ score of the initial model, the higher the DAQ of the revised as well; this seems reasonable and is a good 'sanity check'. Figure 2b – when considering inconsistent residues, there is almost an inverse correlation – the lower the DAQ score of the initial version, the higher it becomes in the revised version. The conclusion is somewhat circular, i.e. the improvement is in the inconsistent regions, but this is somewhat obvious it would be so, i.e. they would not have changed unless it was for some sort of improvement. But any other map-model score would likely show the same relationship. The DAQ score is higher in the inconsistent regions for only ~89% of the residues. Does this mean the DAQ score has a 11% error rate? Or is it because some inconsistent residues moved from a favourable spot to a less favourable spot, for the benefit of other residues perhaps. Or perhaphs that is for noisy parts of the map.

Page 10 second paragraph. The AUC-ROC could be described a bit further here for clarity. Having to use such a large window size of 19 residues is concerning, i.e. why is the score not reliable at a single residue level? Again the score seems to show that residues that move more tend to have better scores in the revised version, which is to be expected.

Page 12 middle paragraph. The sentence 'Comparison of AA-score with three other scores …' has incorrect syntax – it is not the comparison that demonstrates. 'The other scores only …' is also incorrect, since it is then mentioned that Q-scores also shows drops in all regions; the 'largely missed the problems in region (iii)' part is not clear and may be misleading.

Page 13. "AAscore is unique in that …'. It is not clear how it is unique, especially since it can't detect the difference in the two cases. It was mentioned in the previous paragraph that Q-scores also could

potentially detect the same issues. The rest of the sentences in this paragraph are also not too convincing in my opinion and are not sufficiently demonstrated.

Page 15 bottom paragraph. '… identifies mismodelled residues by absolute quality but not by…' not clear what is meant here. AUC-ROC and AUC Precision-Recall should be described as to how they are applied and what they indicate; it is hard to evaluate the results without a clear explanation.

Page 17 last paragraph. Last sentence again seems redundant, i.e. 'higher-scored models have higher scores'; this seems obvious and it is not clear why this should imply anything.

Page 18, Figure 4. The panels are very small and cryo-EM map density is not shown too well. It could be more informative to show less examples and more clear density as illustration, perhaps adding some to supplementary.

Page 19 first paragraph. The text is somewhat fragmented and hard to follow, i.e. from the sentence 'Positions in this model truncated as alanine…', and in the next sentence '… hydrophobic side chains …' which do not apply to the same 'truncated alanine' residues.

Page 19 bottom paragraph. Molprobity does not report map-model correlation.

Page 20 top paragraph. In a 4.4Å map, it is hard to talk about side chains with much confidence, so it not clear what the DAQ score is providing here. Other than reflecting approximate fit-to-map score, it is unreasonable to talk about truly validating such residues based on map densities alone.

Page 20 middle paragraph. Again in maps at ~3.8, 3.6Å resolutions, it can be hard to be fully sure of side chain placements and identities without knowing the sequence, and not clear how the DAQ score is showing here.

Page 20 bottom paragraph. In the last example, going from 3.43Å to 2.97Å of course should lead to a better model and better map-model scores.

Page 21 middle paragraph. The paragraph starting with 'Fig 5a' may be a bit too speculative. It seems a bit harsh to make the claim that maps and models are mismodeled in general without sufficient evidence. When dealing with maps at 3-5Å resolution, where side chains may not be clearly resolved, and furthermore with potentially unresolved regions, claiming such regions mismodeled may be the same as claiming they are unresolved; i.e. the map itself, the only data we have, may not support either claim.

Page 22. The density is Figure 5c is not clearly shown, and the paragraph on page 22 basically illustrates the point above, i.e. it seems that residues in question are not resolved in the density, and in that case is it fair (or is there enough evidence in the first place) to say that they are mis-assigned? It would be much more fair to find, for example, cases in 2.2Å maps where the residues are miss-assigned, in which case, we would know for sure that is the case because the density would clearly shows the correct assignment.

Page 24.
'… are too sensitive…' not clear what this sentence means to say.
'Many of the mistakes made would have been discovered by well-trained…' – This sentence is not true. Even a well-trained structural biologist may not be able build every residue with confidence in some maps, e.g. 3-4Å, especially in less-resolved regions. Such statement unfortunately underscores and does a bit of disservice to the efforts that go into model-building in the field of cryo-EM in general.
'… it is urgent…'. In my opinion and experience, existing tools that use local map-model correlations (or Q-scores) can also robustly indicate the same issues that the DAQ score would. (E.g. see Pintilie and Chiu, Acta. Cryst. D., 2021, Figure 6). It is true however that such tools should be more widely used and incorporated into validation reports.

Page 25.
'… trained on maps of 2.67 to 4.5Å' – why not also use higher resolution maps? Does this mean it would not be too meaningful to apply to maps at higher resolutions?
'…pretty good performance…' perhaps could be made a bit more concrete.
'…resolutions between 2.5 and 5.0Å..' why is it different than the training set resolutions.

Page 26.
'… collected voxels with voxels of size 11^3…' not clear – a voxel is a single grid point. Perhaps boxes with size 11^3? Is this box large enough for ARG residues which can be ~10 from Ca to end atom if the box is centered on the Ca?

Page 30. '… estimated Q-score…' would be better phrased '…expected Q-score…'. Also this is not really 'normalization' but more of an offset, making Q-scores above the expected Q-score positive and those below negative.

Code: I was able to install and run the method from the github account. First I did so on a Macbook laptop, but the code did not run as it requires GPU and CUDA.

Then I tried on our unix cluster; the code installed and ran but output some error messages, which appear related to CUDA:

mrc.c: In function ‘readmrc’:
mrc.c:193:11: error: ‘for’ loop initial declarations are only allowed in C99 mode
for(int i = 0; i<mrc->NumVoxels; ++i)
…
sh: /scratch/g/gregp/daq/DAQ/process_map/gen_trimmap/TrimMapAtom: No such file or directory

I have submitted an issue on github to see how it would be resolved.

Perhaps I missed it but there are no instructions how to run it in Google Colab.

Questions:
1. The default sliding ‘half window’ is 1, does this mean it averages the nearest 1 residue on each side of a given residue, so in total 3 residues?
2. What is the effect of the –stride parameter on the score and running times? Does it make it less accurate?

I looked at the result folder in the github folder, showing EMD:2566 and PDB:3j6b.
1. It is unclear what scores are being output in the pdb files. What is the raw score and the scores in the _w9? Are the SSScore, AAScore, ATOMScore combined?
2. What is unexpected is that the score appears widely different for 3 helices which are similarly resolved (image attached). Perhaps the actual residues are different, leading to different scores. However this makes me doubtful about the usefulness of the score, because it doesn’t seem to take into account whether the backbone is resolved or not.

Review by Greg Pintilie

Reviewer #2:
Remarks to the Author:
In this manuscript, Tersahi et al. propose a new validation metric to help detect errors in atomic models (PDB coordinates) that have been derived from cryoEM maps.

The metric is essentially a probability, for each amino acid position in the model, that the local cryoEM map feature indeed suggests the modeled amino acid (as opposed to one of the other 19 possibilities). The score is normalized to be easily interpretable: positive values indicating good agreement with the map, negative values suggesting likely errors in the model.

The task of algorithmically labeling parts of atomic models that are not well supported by experimental data is not trivial, and the authors claim that their method significantly outperforms existing metrics in that regard. If true, this could make this a very valuable tool for the cryoEM community - both investigators interpreting cryoEM results and repositories and databanks tasked with validating deposited results.

I found this manuscript difficult, at times painful, to review, largely because I found the text did not clearly explain what the authors had done, or the language was confusing in enough places that the intended meaning was lost. I will give specific examples of this later. However, despite all this, and following a few read-throughs, I am now of the mind that the authors' claims are for the most part substantiated by the evidence and that their validation metric indeed may well be a notable advance on the current state of the art.

For this reason, I would recommend that the authors resubmit a revised version of the manuscript for consideration. Assuming my concerns were addressed, and the main claims still stood intact, I would then be supportive of publication.

Below my signature is a list of issues I encountered when reading the manuscript, roughly sorted by importance.

Alexis Rohou, Genentech
January 2022

Obstacles to comprehension:

Near the beginning of Results, the authors go straight to a discussion of how their DAQ scores, but have not explained at all what Emap2sec+ is. Thus, this whole section is very confusing on first read, because you have never explained that Emap2sec+ is in fact a neural network whcih you have trained on a selected subset of the corpus of deposited maps & models. I strongly suggest you add a few sentences to first explain what Emap2Sec+ is (without details), and only then explain how you use the scores it outputs to compute DAQ scores. I would also suggest that you add to Figure 1 an overview panel with flowchart / algorithm design, showing (a) the training of the neural net (select training set, select test set, train on training set, etc), (b) the use of the trained neural net (iterate over each Calpha position, for each position use NN to compute probabilty, etc)

Claims that may not be supported by evidence:

- abstract:
-- "indicating a great need for improved validation tools to help achieve the most correct structures possible"; Does the evidence really support this claim? Many (most? see other comments) of these could be explained by errors in modeling caused by noise low-resolution regions of maps; in those cases, in my view, improved validation such as provided by DAQ or other tools would not have helped achieve more accurate models. At best they help flag a problem. I guess what I'm saying is even with improved validation tools telling them something is wrong with a model in a poorly-resolved part of the map, authors may still deposit
- result
-- "DAQ(AA)-score showed the highest values in the two types of metrics, which means DAQ(AA)-score identifies mismodelled residues by absolute quality but not by comparing with other residues in the same model." I do not follow this sentence. I don't understand the claim, and therefore cannot evaluate whether the results really support the claim. Please state your claim more explicitely / clearly and explain how the fact that DAQ(AA) the two metrics
-- "which implies that the inconsistent residues in lower scored models are probably misplaced." I do not follow.
--- What does misplaced mean in this context? Don't you mean misaligned?
--- Please clarify your meaning or remove. Your argument border on the circular, I fear. You first selected pairs of models with differences. You then ran your score and classified models according to whether their DAQ scores for inconsistent residues were higher or lower than the other pair member. Then you marvel at the fact that the residues in the lower-scored models have lower DAQ scores. I fail to see how this is remarkable. Maybe the only conclusion I would draw from this is that the depositors of the lower-scoring models tended to make mistakes all over, not just in one region.
-- Suppl Fig 1: "all the scores have positive values for the residues, suggesting that the scores can identify the residue specific
density features properly from background." I'm not sure this is correct. If I understand correctly, the score would identify a residue correctly only if that residue's score were higher than any other, which is a more stringent test than just having a postive score value. I wouldn't be surprised if several residues would score positive at a given position. This claim should be qualified or better supported by evidence. This may impact the main text results, near the sentence "correctly assigned residues all have positive values in Supplementary Fig. 1,"

Modifications needed to support claims:
- "Fig. 2c shows the Area Under the Curve of Receiver Operator Characteristic (AUC-ROC) with a window size of 19." Could the authors please add in the supplementary a pictoral and a text description of the what the AUC and ROC are? How exactly are they computed from the list of residues sorted by their DAQ scores? I am not familiar enough with these tools to intuit from the text how exactly they are calculated. Since AUC-ROC backs one of the most important claims of the paper (that DAQ scores

outperform existing validation metrics in identifying modeling errors, Table 1), this point should be explained in detail in suppl. I'm sure other validation metric developers will want to check these points carefully and reproduce these calculations.

Suggested changes:
- intro:
-- "unlike existing map-model scores or (...) likely to be incorrect". Model-coordinate scores by definitely would not be able to detect mismatch with the experimental map, so should not be mentioned here, in my opinion. Also, can you specify which map-model scores you are thinking of here? I would have expected that local correlation-based metrics should be able to do this (at least before seeing your Table 1) - is that not a claim developers of those methods have made?
- results
-- In the text you refer to your scores as "DAQ scores" or "DAQ(AA)... how about calling them DAQ_SS or DAQ(SS), DAQ_AA or DAQ(AA), ..., in the equations?
-- Ref(ss): I assume this is the average of the probability from Emap2sec+, not from SPOT1D. If so, why not have the denominator written as SUM_j(P_ss(j)/N)) or something like that (where j iterates over all atom positions)?
-- on the topic of this denominator - I don't understand why you refer to it as "reference". For example, in the text when you say "if the predicted probability values for the local position are higher than the reference", why not just "higher than average"?

Confusing language:
- abstract:
-- "identified via deep learning" - is it the likelihood that is identified? I don't think a likelihood can be identified, can it? It can be computed, estimated, etc...
-- "reconstructed model" - what is this? I have heard of map 3D reconstruction
-- "how well the amino acids (...) agree with that likelihood" - amino acids cannot agree to anything; did you mean something like "the a.a. assignment is consistent with..."?
-- "scores that correlate the local density grandient in the map" - that just doesn't make sense. They correlate it with what? What kind of gradient are you talking about? Gradient of what with respect to what?
- results:
-- "indicating that Emap2sec+ detected each structural feature specific density features and distinguished them well from background." - This sentence looks mangled to me. I can't work out what you are trying to say.
- results:
-- "For both metrics" - I think "For each metric" would be better here
-- "two types of values are shown, one that averaged over values computed first for each PDB entry separately and the other that considered all the PDB entries together." This sentence threw me the first

4 or 5 times I tried to read it. Could the authors reformulate to make it clearer? I think I understood it eventually, but it's way too much work.
-- The section beginning "Fig 4C shows for examples" and ending "In this region, 6ZR3 clearly a clearly better DAQ(AA)-score than 6L54" was much better written than the results thus far. Despite the clear editing error in that last sentence ("clearly a clearly")
- I thought methods were clearly written also, they were easy to follow

Suggestions for topics to cover in the discussion:
- I'm curious - how would the DAQ score be affected by DL-based methods that aim to "clean up" / make maps look better (or more like your typical map). Now that the machines are learning what good maps look like, could one machine (e.g. DeepEMhancer, Sanchez-Garcia etal, Comms Biol 4:874(2021)) be used to fool another (DAQ)? What if a tool was trained like Emap2sec+ but instead of outputing a score it would output idealized version of each local map feature? In general I'm curious about the impact DL-based methods used during map polishing could have on DL-based methods used for validation. I suspect the authors' expert opinions would be of value.
- In the results: "Particularly, the regions Val661-Gln709 and Leu758-Lys778 have incorrect assignments as evidenced by implausible steric collision between backbone atoms. The density for these regions also does not support the PDB model."
I took a quick look. This is a low-resolution part of that cryoEM map. I am pretty sure other validation metrics, and the depositing authors, would have picked up on the fact that this was likely not an accurate region of the model.

I think this brings up the point that when selecting subsets of the PDB, the authors are using overall resolution value for each entry, but variations in local resolution will mean that many regions of maps and models from this dataset will be expected to have pretty bad models.

I was hoping that the authors would spell this out in the manuscript, perhaps in the discussion: that there will be times when despite best efforts by the model builder, locally a map doesn't have enough SNR to allow reliable model building. In those circumstances, low DAQ scores will be expected but not necessarily actionable.

On a similar point, in the discussion the authors state "Many of the mistakes made in the above examples would likely have been discovered by well-trained structural biologists before their publication". Discovered, yes. But fixed? I doubt it in many cases, since the problems are probably in many cases due to poor local quality.

I don't think DAQ (or any other validation metric) in itself solves the problem of depositing the correct model, or even an improved one. It flags the problems and then leaves it to the depositor to somehow

arrive at a better model, which may or may not be possible, depending on what the problem was due to in the first place.

- Would it be desirable / possible to remove from the training set regions of maps with bad local resolution?

Minor fixes:
- intro
-- "58% of the maps have a resolution", not "had"
-- "scores are naturally affected by the resolution of the local density at the position of a residue", not "of"
- results
-- "DAQ(AA)-scores in the lower-scored model centered at 0.0". According to Fig 4A, the residue scores from the lower scored models (red histogram) were NOT centered around 0.0. You may want to say "approximately" and get rid of the decimal place.

Reviewer #3:
Remarks to the Author:
The authors present a very interesting approach to CryoEM map and atomic model validation. Their approach is based on a neural network that extends part of their previous work. The method is technically sound and they apply it to many examples showing its validity. This reviewer recommends publication in the current form subject to some minor changes:

- Page 26. It currently says "we collected
voxels of a size of 113 Å3 by scanning across a map along three axes with a stride of". Probably "we collected BOXES ..." would be more appropriate.

- The number of parameters of the network should be reported. This is important to be able to compare it to the training dataset size.

- It would be interesting to report in the supplementary material the test and validation loss curves during the training. In this way, the reader can figure out if there has been some overfitting or the network is appropriately learning.

** For Nature Research Group general information and news for authors, see http://npg.nature.com/authors.

Author Rebuttal to Initial comments

Reviewers' Comments:

Reviewer #1:

Remarks to the Author:

A. Summary of the key results

The authors present a new method to evaluate the fit of a protein model to cryo-EM map, called DAQ score, or Deep-learning Amino acid-wise model Quality score. The method uses a neural network to characterize what the local environment of each type of residue looks like, based on a large number of maps and models. When applied to a given voxel in a map, the neural network then gives number which indicates to what degree that voxel represents either a secondary structure (SSscore), given amino acid type (AAscore) and Ca atom (ATOMscore). To evaluate a protein model, each score is calculated at the nearest voxel to the Ca atom of each residue. A positive value indicates the residue is likely ok, whereas a negative residue is flagged as possibly incorrect. The authors apply the scores to 15 structures in the PDB with updated structures, and show that the scores go up for the newer structures for residues that moved more than 2Å. They also apply it to 399 pairs of models in the PDB

which have high sequence similarity but are different by more than 1Å RMSD, and a further data set of 4,485 models.

B. Originality and significance: if not novel, please include reference

This method is novel to my knowledge, although there are some similarities to Deeptracer (referenced) and Haruspex (not referenced – Mostosi et. Al., Angewandte Chemie, 2020).

We will cite the two papers.

C. Data & methodology: validity of approach, quality of data, quality of presentation

The approach seems valid, though there are limitations not adequately described, e.g. effect of resolution, step size, etc. More details and questions in the points below. The data used, coming from the PDB/EMDB is of good quality. The presentation is also mostly ok, clearly described overall, thought again some further clarifications are needed.

D. Appropriate use of statistics and treatment of uncertainties

Statistics are used appropriately, for example to show that DAQ score goes up for improved regions of a model. Uncertainties are not addressed properly however, for example, the possibility that the map is not resolved in some regions is mostly ignored. The effect of resolution is also ignored; it would be very informative to know how the score can be applied and what to expect at different resolutions.

[New Data] We will provide data of average DAQ score for each amino acids in structures of different resolutions.

Our current plan is to provide the data for the same cryo-EM maps of Suppl Fig.6. of this work:

https://static-content.springer.com/esm/art%3A10.1038%2Fs41467-020-20295-w/MediaObjects/41467_2020_20295_MOESM1_ESM.pdf.

Since the table in this figure provides values of two scores including Q-score (the reviewer signed his name), the one developed by Reviewer 1, we can compare the score distribution of our score with the existing scores.

E. Conclusions: robustness, validity, reliability

The robustness seems reasonable, i.e. higher scores are obtained for improved model. However the validity and reliability at the per-residue level is not clear, since sliding windows; illustrations of actual scores for single residues in more cases would be useful to demonstrate the reliability of the score in different scenarios. Because the output is directly related to the training data, if the training data does not include all valid conformations of a residue, then the output could be incorrect. This limitation need to be mentioned. Based on looking at the provided result in the github repo, the results do not seem robust (details below).

In the discussion section, we will mention this limitation.

The reviewer mentioned the data we showed with a sliding window (Table 1), but this is not a specific trend of our score. The other existing score, including Q-score, the one developed by the reviewer, has the same trend where using a longer sliding window size shows better accuracy.

F. Suggested improvements: experiments, data for possible revision

A potentially necessary improvement would be to illustrate the score for at least a few types of residues at different resolutions. Otherwise, it is relied upon general statistics to demonstrate. Other improvements would come mostly from being more clear on the limitations of the method and being more realistic with the conclusions.

As responded above, we will provide the data for maps of different resolutions. We will address the limitation of the method in Discussion.

G. References: appropriate credit to previous work?

Yes, but could be improved. They may consider to reference the Haruspex method. It would be appropriate to also reference other per-residue scoring methods such as TEMPY (CCPEM), and Phenix.

We will cite these two papers.

H. Clarity and context: lucidity of abstract/summary, appropriateness of abstract, introduction and conclusions

Descriptions are mostly clear, though further clarifications may be needed as indicated below. The abstract is a bit ambitious for what is actually presented, and the conclusions do not adequately indicate limitations.

We will modify the abstract and discussion.

Specific Points:

Abstract: It is stated that the method can identify "potential misassignment of residues in the map, including residue shifts". Later it is stated that none of the scores can tell the difference between misaligned or mispositioned residues (page 13). Perhaps these can be better consolidated and more clearly/realistically stated.

We will rephrase the sentence.

Page 3 bottom. For map-model scores, atom inclusion score should also be mentioned (reference Lagerstedt et al, JSB, 2013). It should be mentioned that it is currently used in the EMDB/PDB validation report.

We will mention it.

Page 4 top: Molprobity score also includes bond lengths, and angles between 3 connected atoms.

We will correct it.

Page 4 bottom paragraph. Suggest to add 'protein' in 'new approach for cryo-EM model validation'.

We will correct it.

Page 5 top paragraph. To say that 'local structural features' are used instead of 'local density values' seems contradictory/misleading because ultimately the neural network is fed these local density values.

We will rephrase it.

Page 5, bottom paragraph. It is stated that comparison of DAQ score to other scores showed a 'clear advantage' for the DAQ score. This may be too strong/vague/general of a statement. There are reasons why this is not so, for example, the training set does not include maps with higher than ~2.6Å resolution, so it's not clear what the method would do there. Also, the method can only be applied to protein residues. Lastly, more thorough tests and comparisons to other methods are needed before such a statement should be made.

We will rephrase the statement. It is a good point that our method is only for protein models.

Page 6 middle paragraph. 'The probabilities for a grid point …'. Not clear, is a 11^3 box used centered at the grid point?

We will rephrase it to clarify the sentence.

Page 7 middle. 'atom type score of amino acid residue i'. The ATOMscore tries to indicate if a voxel appears to indicate a given type of atom, but only the Ca atom is considered here. Atom type usually is meant to reflect whether an atom is C, O, N, etc. Indicating this score has something to do with atom type is misleading. Perhaps this can be revised somehow to be more specific towards identifying the Ca atom.

We will rephrase it.

Page 7, bottom paragraph. The accuracies (77.3% for secondary structures, 51.5% for Ca atoms, and 28.2% for amino acid type) seem low. It is mentioned that the accuracies do not need to be perfect for model validation, as long as the 'correct' residue is preferred for a position in the map. It is not described what is meant by 'preferred', and it seems quite likely that even this method could potentially give a higher score to the incorrect residue, i.e. it is not shown otherwise, which would be hard to do; unless it can be done, this statement needs to be revised).

We will rephrase the statement.

Page 9. The two classes misaligned or mispositioned; misaligned have to do with shifted residues, but is the simple criteria that a residue is within 2Å of another residue always indicate a shift? They are a bit hard to comprehend from the text, could they be illustrated perhaps?

We will draw illustrations and put the image to Supplementary Information.

Page 9 bottom paragraph, Figure 2. Average DAQ scores of first and revised version are plotted. Figure 2a shows they are correlated when considering all residues, but what does this correlation mean - the higher the DAQ score of the initial model, the higher the DAQ of the revised as well; this seems reasonable and is a good 'sanity check'.

The reviewer's understanding is correct.

Figure 2b – when considering inconsistent residues, there is almost an inverse correlation – the lower the DAQ score of the initial version, the higher it becomes in the revised version. The conclusion is somewhat circular, i.e. the improvement is in the inconsistent regions, but this is somewhat obvious it would be so, i.e. they would not have changed unless it was for some sort of improvement. But any other map-model score would likely show the same relationship.

The logic actually not circular. Some clarification of the procedure is needed. We will rephrase the paragraph to clarify it.

The DAQ score is higher in the inconsistent regions for only ~89% of the residues. Does this mean the DAQ score has a 11% error rate? Or is it because some inconsistent residues moved from a favourable spot to a less favourable spot, for the benefit of other residues perhaps. Or perhaps that is for noisy parts of the map.

[New data] We will show a couple of types of examples among the 11% of the cases where DAQ score did not detect inconsistent residues.

Page 10 second paragraph. The AUC-ROC could be described a bit further here for clarity. Having to use such a large window size of 19 residues is concerning, i.e. why is the score not reliable at a single residue level? Again the score seems to show that residues that move more tend to have better scores in the revised version, which is to be expected.

We will explain it. As I responded one of the question above, it is general trend of all the scores that they perform better by using a large window size.

Page 12 middle paragraph. The sentence 'Comparison of AA-score with three other scores …' has incorrect syntax – it is not the comparison that demonstrates. 'The other scores only …' is also incorrect, since it is then mentioned that Q-scores also shows drops in all regions; the 'largely missed the problems in region (iii)' part is not clear and may be misleading.

We will rephrase it.

Page 13. "AAscore is unique in that …'. It is not clear how it is unique, especially since it can't detect the difference in the two cases. It was mentioned in the previous paragraph that Q-scores also could potentially detect the same issues. The rest of the sentences in this paragraph are also not too convincing in my opinion and are not sufficiently demonstrated.

We will rephrase it.

Page 15 bottom paragraph. '… identifies mismodelled residues by absolute quality but not by…' not clear what is meant here. AUC-ROC and AUC Precision-Recall should be described as to how they are applied and what they indicate; it is hard to evaluate the results without a clear explanation.

We will add the definition of the AUC.

Page 17 last paragraph. Last sentence again seems redundant, i.e. 'higher-scored models have higher scores'; this seems obvious and it is not clear why this should imply anything.

We will rephrase it.

Page 18, Figure 4. The panels are very small and cryo-EM map density is not shown too well. It could be more informative to show less examples and more clear density as illustration, perhaps adding some to supplementary.

We will revise the figure.

Page 19 first paragraph. The text is somewhat fragmented and hard to follow, i.e. from the sentence 'Positions in this model truncated as alanine…', and in the next sentence '… hydrophobic side chains …' which do not apply to the same 'truncated alanine' residues.

We will rephrase it.

Page 19 bottom paragraph. Molprobity does not report map-model correlation.

The Molprobity report for 5H64 indicated that there are map-model correlation problems in 5% of residues, while 6U62 chain A does not have any correlation problem

Molprobity does report map-model correlation. We will clarify where we see the map-model correlation data.

Page 20 top paragraph. In a 4.4Å map, it is hard to talk about side chains with much confidence, so it not clear what the DAQ score is providing here. Other than reflecting approximate fit-to-map score, it is unreasonable to talk about truly validating such residues based on map densities alone.

We will add more explanation.

Page 20 middle paragraph. Again in maps at ~3.8, 3.6Å resolutions, it can be hard to be fully sure of side chain placements and identities without knowing the sequence, and not clear how the DAQ score is showing here.

We will add more explanation.

Page 20 bottom paragraph. In the last example, going from 3.43Å to 2.97Å of course should lead to a better model and better map-model scores.

We agree.

Page 21 middle paragraph. The paragraph starting with 'Fig 5a' may be a bit too speculative. It seems a bit harsh to make the claim that maps and models are mismodeled in general without sufficient evidence. When dealing with maps at 3-5Å resolution, where side chains may not be clearly resolved, and furthermore with potentially unresolved regions, claiming such regions mismodeled may be the same as claiming they are unresolved; i.e. the map itself, the only data we have, may not support either claim.

We will rephrase the sentence.

Page 22. The density is Figure 5c is not clearly shown, and the paragraph on page 22 basically illustrates the point above, i.e. it seems that residues in question are not resolved in the density, and in that case is it fair (or is there enough evidence in the first place) to say that they are mis-assigned? It would be much more fair to find, for example, cases in 2.2Å maps where the residues are miss-assigned, in which case, we would know for sure that is the case because the density would clearly shows the correct assignment.

[New Data] We will probably use a different example.

Page 24.

'… are too sensitive…' not clear what this sentence means to say.

'Many of the mistakes made would have been discovered by well-trained…' – This sentence is not true. Even a well-trained structural biologist may not be able build every residue with confidence in some maps, e.g. 3-4Å, especially in less-resolved regions. Such statement unfortunately underscores and does a bit of disservice to the efforts that go into model-building in the field of cryo-EM in general.

'… it is urgent…'. In my opinion and experience, existing tools that use local map-model correlations (or Q-scores) can also robustly indicate the same issues that the DAQ score would. (E.g. see Pintilie and Chiu, Acta. Cryst. D., 2021, Figure 6). It is true however that such tools should be more widely used and incorporated into validation reports.

We will rephrase the sentences and add more explanation.

Page 25.

'… trained on maps of 2.67 to 4.5Å' – why not also use higher resolution maps? Does this mean it would not be too meaningful to apply to maps at higher resolutions?

We will add more explanation.

'…pretty good performance…' perhaps could be made a bit more concrete.

'…resolutions between 2.5 and 5.0Å..' why is it different than the training set resolutions.

We will rephrase the sentences.

Page 26.

'… collected voxels with voxels of size 11^3…' not clear – a voxel is a single grid point. Perhaps boxes with size 11^3? Is this box large enough for ARG residues which can be ~10 from Ca to end atom if the box is centered on the Ca?

We will revise and add more explanation.

Page 30. '… estimated Q-score…' would be better phrased '…expected Q-score…'. Also this is not really 'normalization' but more of an offset, making Q-scores above the expected Q-score positive and those below negative.

We will rephrase it.

Code: I was able to install and run the method from the github account. First I did so on a Macbook laptop, but the code did not run as it requires GPU and CUDA.

Then I tried on our unix cluster; the code installed and ran but output some error messages, which appear related to CUDA:

mrc.c: In function 'readmrc':

mrc.c:193:11: error: 'for' loop initial declarations are only allowed in C99 mode

for(int i = 0; i<mrc->NumVoxels; ++i)

…

sh: /scratch/g/gregp/daq/DAQ/process_map/gen_trimmap/TrimMapAtom: No such file or directory

I have submitted an issue on github to see how it would be resolved.

We have responded to this comment on Github when it was posted. The problem occurred for Macbook.

Perhaps I missed it but there are no instructions how to run it in Google Colab.

The Google Colab site has sufficient instruction we believe but there was no explanation on the Github page about the Google Colab page. We will add some explanation.

Questions:

1. The default sliding 'half window' is 1, does this mean it averages the nearest 1 residue on each side of a given residue, so in total 3 residues?

Yes.

2. What is the effect of the –stride parameter on the score and running times? Does it make it less accurate?

We will explain it. The short answer is No.

I looked at the result folder in the github folder, showing EMD:2566 and PDB:3j6b.

1. It is unclear what scores are being output in the pdb files. What is the raw score and the scores in the _w9? Are the SSScore, AAScore, ATOMScore combined?

It is AAscore. We will add more explanation.

2. What is unexpected is that the score appears widely different for 3 helices which are similarly resolved (image attached). Perhaps the actual residues are different, leading to different scores. However this makes me doubtful about the usefulness of the score, because it doesn't seem to take into account whether the backbone is resolved or not.

We will add more explanation.

Review by Greg Pintilie

Reviewer #2:

Remarks to the Author:

In this manuscript, Tersahi et al. propose a new validation metric to help detect errors in atomic models (PDB coordinates) that have been derived from cryoEM maps.

The metric is essentially a probability, for each amino acid position in the model, that the local cryoEM map feature indeed suggests the modeled amino acid (as opposed to one of the other 19 possibilities). The score is normalized to be easily interpretable: positive values indicating good agreement with the map, negative values suggesting likely errors in the model.

The task of algorithmically labeling parts of atomic models that are not well supported by experimental data is not trivial, and the authors claim that their method significantly outperforms existing metrics in that regard. If true, this could make this a very valuable tool for the cryoEM community - both investigators interpreting cryoEM results and repositories and databanks tasked with validating deposited results.

I found this manuscript difficult, at times painful, to review, largely because I found the text did not clearly explain what the authors had done, or the language was confusing in enough places that the intended meaning was lost. I will give specific examples of this later. However, despite all this, and following a few read-throughs, I am now of the mind that the authors' claims are for the most part

substantiated by the evidence and that their validation metric indeed may well be a notable advance on the current state of the art.

For this reason, I would recommend that the authors resubmit a revised version of the manuscript for consideration. Assuming my concerns were addressed, and the main claims still stood intact, I would then be supportive of publication.

Below my signature is a list of issues I encountered when reading the manuscript, roughly sorted by importance.

Alexis Rohou, Genentech

January 2022

Obstacles to comprehension:

Near the beginning of Results, the authors go straight to a discussion of how their DAQ scores, but have not explained at all what Emap2sec+ is. Thus, this whole section is very confusing on first read, because you have never explained that Emap2sec+ is in fact a neural network whcih you have trained on a selected subset of the corpus of deposited maps & models. I strongly suggest you add a few sentences to first explain what Emap2Sec+ is (without details), and only then explain how you use the scores it outputs to compute DAQ scores.

We will add more explanation.

I would also suggest that you add to Figure 1 an overview panel with flowchart / algorithm design, showing (a) the training of the neural net (select training set, select test set, train on training set, etc), (b) the use of the trained neural net (iterate over each Calpha position, for each position use NN to compute probabilty, etc)

We will revise Figure 1.

Claims that may not be supported by evidence:

- abstract:

-- "indicating a great need for improved validation tools to help achieve the most correct structures possible"; Does the evidence really support this claim? Many (most? see other comments) of these could be explained by errors in modeling caused by noise low-resolution regions of maps; in those cases, in my view, improved validation such as provided by DAQ or other tools would not have helped achieve more accurate models. At best they help flag a problem. I guess what I'm saying is even with improved validation tools telling them something is wrong with a model in a poorly-resolved part of the map, authors may still deposit

We basically agree. We will rephrase it.

- result

-- "DAQ(AA)-score showed the highest values in the two types of metrics, which means DAQ(AA)-score identifies mismodelled residues by absolute quality but not by comparing with other residues in the same model." I do not follow this sentence. I don't understand the claim, and therefore cannot evaluate whether the results really support the claim. Please state your claim more explicitly / clearly and explain how the fact that DAQ(AA) the two metrics

We will rephrase it.

-- "which implies that the inconsistent residues in lower scored models are probably misplaced." I do not follow.

--- What does misplaced mean in this context? Don't you mean misaligned?

We will rephrase these sentences for clarification.

--- Please clarify your meaning or remove. Your argument border on the circular, I fear. You first selected pairs of models with differences. You then ran your score and classified models according to whether their DAQ scores for inconsistent residues were higher or lower than the other pair member. Then you marvel at the fact that the residues in the lower-scored models have lower DAQ scores. I fail to see how this is remarkable. Maybe the only conclusion I would draw from this is that the depositors of the lower-scoring models tended to make mistakes all over, not just in one region.

It is not a circular logic. We will add more explanation.

-- Suppl Fig 1: "all the scores have positive values for the residues, suggesting that the scores can identify the residue specific density features properly from background." I'm not sure this is correct. If I understand correctly, the score would identify a residue correctly only if that residue's score were higher than any other, which is a more stringent test than just having a postive score value. I wouldn't be surprised if several residues would score positive at a given position. This claim should be qualified or better supported by evidence. This may impact the main text results, near the sentence "correctly assigned residues all have positive values in Supplementary Fig. 1,"

We will add more explanation for Supplementary Fig. 1.

Modifications needed to support claims:

- "Fig. 2c shows the Area Under the Curve of Receiver Operator Characteristic (AUC-ROC) with a window size of 19." Could the authors please add in the supplementary a pictoral and a text description of the what the AUC and ROC are? How exactly are they computed from the list of residues sorted by their DAQ scores? I am not familiar enough with these tools to intuit from the text how exactly they are calculated. Since AUC-ROC backs one of the most important claims of the paper (that DAQ scores outperform existing validation metrics in identifying modeling errors, Table 1), this point should be explained in detail in suppl. I'm sure other validation metric developers will want to check these points carefully and reproduce these calculations.

We will add a description of AUC-ROC and AUC-Precision-Recall.

Suggested changes:

- intro:

-- "unlike existing map-model scores or (...) likely to be incorrect". Model-coordinate scores by definitely would not be able to detect mismatch with the experimental map, so should not be mentioned here, in my opinion. Also, can you specify which map-model scores you are thinking of here? I would have expected that local correlation-based metrics should be able to do this (at least before seeing your Table 1) - is that not a claim developers of those methods have made?

We will add more explanation.

- results

-- In the text you refer to your scores as "DAQ scores" or "DAQ(AA)... how about calling them DAQ_SS or DAQ(SS), DAQ_AA or DAQ(AA), ..., in the equations?

Ok, we will rephrase them.

-- Ref(ss): I assume this is the average of the probability from Emap2sec+, not from SPOT1D. If so, why not have the denominator written as SUM_j(P_ss(j)/N)) or something like that (where j iterates over all atom positions)?

Yes, the reviewer's understanding is correct. We will rephrase it.

-- on the topic of this denominator - I don't understand why you refer to it as "reference". For example, in the text when you say "if the predicted probability values for the local position are higher than the reference", why not just "higher than average"?

We will rephrase it.

Confusing language:

- abstract:

-- "identified via deep learning" - is it the likelihood that is identified? I don't think a likelihood can be identified, can it? It can be computed, estimated, etc…

We will rephrase it.

-- "reconstructed model" - what is this? I have heard of map 3D reconstruction

We will rephrase it.

-- "how well the amino acids (...) agree with that likelihood" - amino acids cannot agree to anything; did you mean something like "the a.a. assignment is consistent with..."?

Yes. We will rephrase it.

-- "scores that correlate the local density grandient in the map" - that just doesn't make sense. They correlate it with what? What kind of gradient are you talking about? Gradient of what with respect to what?

We will rephrase it.

- results:

-- "indicating that Emap2sec+ detected each structural feature specific density features and distinguished them well from background." - This sentence looks mangled to me. I can't work out what you are trying to say.

We will rephrase it.

- results:

-- "For both metrics" - I think "For each metric" would be better here

We will rephrase it.

-- "two types of values are shown, one that averaged over values computed first for each PDB entry separately and the other that considered all the PDB entries together." This sentence threw me the first 4 or 5 times I tried to read it. Could the authors reformulate to make it clearer? I think I understood it eventually, but it's way too much work.

We will rephrase it.

-- The section beginning "Fig 4C shows for examples" and ending "In this region, 6ZR3 clearly a clearly better DAQ(AA)-score than 6L54" was much better written than the results thus far. Despite the clear editing error in that last sentence ("clearly a clearly")

We will correct the sentence.

- I thought methods were clearly written also, they were easy to follow

Thank you.

Suggestions for topics to cover in the discussion:

- I'm curious - how would the DAQ score be affected by DL-based methods that aim to "clean up" / make maps look better (or more like your typical map). Now that the machines are learning what good maps look like, could one machine (e.g. DeepEMhancer, Sanchez-Garcia etal, Comms Biol 4:874(2021)) be used to fool another (DAQ)? What if a tool was trained like Emap2sec+ but instead of outputing a score it would output idealized version of each local map feature? In general I'm curious about the impact DL-based methods used during map polishing could have on DL-based methods used for validation. I suspect the authors' expert opinions would be of value.

An interesting question. We will discuss it in the discussion.

- In the results: "Particularly, the regions Val661-Gln709 and Leu758-Lys778 have incorrect assignments as evidenced by implausible steric collision between backbone atoms. The density for these regions also does not support the PDB model."

I took a quick look. This is a low-resolution part of that cryoEM map. I am pretty sure other validation metrics, and the depositing authors, would have picked up on the fact that this was likely not an accurate region of the model.

Basically we agree. Add more explanation.

I think this brings up the point that when selecting subsets of the PDB, the authors are using overall resolution value for each entry, but variations in local resolution will mean that many regions of maps and models from this dataset will be expected to have pretty bad models.

I was hoping that the authors would spell this out in the manuscript, perhaps in the discussion: that there will be times when despite best efforts by the model builder, locally a map doesn't have enough SNR to allow reliable model building. In those circumstances, low DAQ scores will be expected but not necessarily actionable.

We will discuss more in the discussion.

On a similar point, in the discussion the authors state "Many of the mistakes made in the above examples would likely have been discovered by well-trained structural biologists before their publication". Discovered, yes. But fixed? I doubt it in many cases, since the problems are probably in many cases due to poor local quality.

We basically agree. Will rephrase it.

I don't think DAQ (or any other validation metric) in itself solves the problem of depositing the correct model, or even an improved one. It flags the problems and then leaves it to the depositor to somehow arrive at a better model, which may or may not be possible, depending on what the problem was due to in the first place.

We basically agree. Will rephrase it.

- Would it be desirable / possible to remove from the training set regions of maps with bad local resolution?

We will add explanation. We will explain pros and cons of the idea.

Minor fixes:

- intro

-- "58% of the maps have a resolution", not "had"

We will rephrase it.

-- "scores are naturally affected by the resolution of the local density at the position of a residue", not "of" - results

We will rephrase it.

-- "DAQ(AA)-scores in the lower-scored model centered at 0.0". According to Fig 4A, the residue scores from the lower scored models (red histogram) were NOT centered around 0.0. You may want to say "approximately" and get rid of the decimal place.

We will rephrase it.

Reviewer #3:

Remarks to the Author:

The authors present a very interesting approach to CryoEM map and atomic model validation. Their approach is based on a neural network that extends part of their previous work. The method is technically sound and they apply it to many examples showing its validity. This reviewer recommends publication in the current form subject to some minor changes:

- Page 26. It currently says "we collected

voxels of a size of 113 Å3 by scanning across a map along three axes with a stride of". Probably "we collected BOXES ..." would be more appropriate.

We will rephrase it.

- The number of parameters of the network should be reported. This is important to be able to compare it to the training dataset size.

[New Data] We will add the number of parameters.

- It would be interesting to report in the supplementary material the test and validation loss curves during the training. In this way, the reader can figure out if there has been some overfitting or the network is appropriately learning.

[New Data] We will add the loss curves from training.

(End of the reviewers' comments)

Decision Letter, first revision:

Dear Daisuke,

Thank you for your letter asking us to reconsider our decision on your Article, "Residue-Wise Local Quality Estimation for Protein Models from Cryo-EM Maps". After careful consideration we have decided that we are willing to consider a revised version of your manuscript that is updated as you've described.

We wanted to note that in some places you did not provide us sufficient detail in your responses for us to fully assess whether the revisions are likely to satisfy our editorial concerns and those of the referees. For this reason, we cannot guarantee we will send the paper for review until we have carefully considered the revised version.

When revising your paper:

* include a point-by-point response to our referees and to any editorial suggestions

* please underline/highlight any additions to the text or areas with other significant changes to facilitate review of the revised manuscript

* address the points listed described below to conform to our open science requirements

* ensure it complies with our general format requirements as set out in our guide to authors at www.nature.com/naturemethods

* resubmit all the necessary files electronically by using the link below to access your home page

[Redacted] This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

We hope to receive your revised paper within XX weeks [**ED TO CUSTOMIZE AS NEEDED**]. If you cannot send it within this time, please let us know. In this event, we will still be happy to reconsider your paper at a later date so long as nothing similar has been accepted for publication at Nature Methods or published elsewhere.

OPEN SCIENCE REQUIREMENTS

REPORTING SUMMARY AND EDITORIAL POLICY CHECKLISTS

When revising your manuscript, please submit reporting summary and editorial policy checklists.

Reporting summary: https://www.nature.com/documents/nr-reporting-summary.zip

Editorial policy checklist: https://www.nature.com/documents/nr-editorial-policy-checklist.zip

If your paper includes custom software, we also ask you to complete a supplemental reporting summary.

Software supplement: https://www.nature.com/documents/nr-software-policy.pdf

Please submit these with your revised manuscript. They will be available to reviewers to aid in their evaluation if the paper is re-reviewed. If you have any questions about the checklist, please see http://www.nature.com/authors/policies/availability.html or contact me.

Please note that these forms are dynamic 'smart pdfs' and must therefore be downloaded and completed in Adobe Reader. We will then flatten them for ease of use by the reviewers. If you would like to reference the guidance text as you complete the template, please access these flattened versions at http://www.nature.com/authors/policies/availability.html.

DATA AVAILABILITY

Please include a "Data availability" subsection in the Online Methods. This section should inform readers about the availability of the data used to support the conclusions of your study, including accession codes to public repositories, references to source data that may be published alongside the paper, unique identifiers such as URLs to data repository entries, or data set DOIs, and any other statement about data availability. At a minimum, you should include the following statement: "The data that support the findings of this study are available from the corresponding author upon request", describing which data is available upon request and mentioning any restrictions on availability. If DOIs are provided, please include these in the Reference list (authors, title, publisher (repository name), identifier, year). For more guidance on how to write this section please see: http://www.nature.com/authors/policies/data/data-availability-statements-data-citations.pdf

CODE AVAILABILITY

Please include a "Code Availability" subsection in the Online Methods which details how your custom code is made available. Only in rare cases (where code is not central to the main conclusions of the paper) is the statement "available upon request" allowed (and reasons should be specified).

We request that you deposit code in a DOI-minting repository such as Zenodo, Gigantum or Code Ocean and cite the DOI in the Reference list. We also request that you use code versioning and provide a license.

For more information on our code sharing policy and requirements, please see:

https://www.nature.com/nature-research/editorial-policies/reporting-standards#availability-of-computer-code

ORCID

Nature Methods is committed to improving transparency in authorship. As part of our efforts in this direction, we are now requesting that all authors identified as 'corresponding author' on published papers create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System (MTS), prior to acceptance. This applies to primary research papers only. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the MTS by clicking on 'Modify my Springer Nature account'. For more information please visit please visit www.springernature.com/orcid.

We look forward to hearing from you soon.

Sincerely,

Rita

Rita Strack, Ph.D.

Senior Editor

Nature Methods

Author Rebuttal, first revision:

Responses to Comments by Reviewer 1

Remarks to the Author:

A. Summary of the key results

The authors present a new method to evaluate the fit of a protein model to cryo-EM map, called DAQ score, or Deep-learning Amino acid-wise model Quality score. The method uses a neural network to characterize what the local environment of each type of residue looks like, based on a large number of maps and models. When applied to a given voxel in a map, the neural network then gives number which indicates to what degree that voxel represents either a secondary structure (SSscore), given amino acid

type (AAscore) and Ca atom (ATOMscore). To evaluate a protein model, each score is calculated at the nearest voxel to the Ca atom of each residue. A positive value indicates the residue is likely ok, whereas a negative residue is flagged as possibly incorrect. The authors apply the scores to 15 structures in the PDB with updated structures, and show that the scores go up for the newer structures for residues that moved more than 2Å. They also apply it to 399 pairs of models in the PDB which have high sequence similarity but are different by more than 1Å RMSD, and a further data set of 4,485 models.

B. Originality and significance: if not novel, please include reference

This method is novel to my knowledge, although there are some similarities to Deeptracer (referenced) and Haruspex (not referenced – Mostosi et al., Angewandte Chemie, 2020).

We now cited the Harupsex paper in the Introduction (page 5).

C. Data & methodology: validity of approach, quality of data, quality of presentation

The approach seems valid, though there are limitations not adequately described, e.g. effect of resolution, step size, etc. More details and questions in the points below. The data used, coming from the PDB/EMDB is of good quality. The presentation is also mostly ok, clearly described overall, thought again some further clarifications are needed.

Thank you; we reply each questions below in detail.

D. Appropriate use of statistics and treatment of uncertainties

Statistics are used appropriately, for example to show that DAQ score goes up for improved regions of a model. Uncertainties are not addressed properly however, for example, the possibility that the map is not resolved in some regions is mostly ignored. The effect of resolution is also ignored; it would be very informative to know how the score can be applied and what to expect at different resolutions.

In the new Supplementary Fig 2, we plotted DAQ(AA) scores on a dataset of 50 maps taken from Supplementary Fig. 6 of “FSC-Q: A CryoEM map-to-atomic model quality validation based on the local Fourier Shell Correlation” Ramirez-Aportela et al., Nature Commn. 12: 42 (2021) (PDF: https://static-content.springer.com/esm/art%3A10.1038%2Fs41467-020-20295-w/MediaObjects/41467_2020_20295_MOESM1_ESM.pdf). We followed the way the authors of this paper plotted their scores (FSC-Q scores). We also added DAQ(AA) scores of a few residues in apoferritin maps of different resolutions in Supplementary Table 3. The result is discussed on page 8.

E. Conclusions: robustness, validity, reliability

The robustness seems reasonable, i.e. higher scores are obtained for improved model. However the validity and reliability at the per-residue level is not clear, since sliding windows; illustrations of actual scores for single residues in more cases would be useful to demonstrate the reliability of the score in different scenarios. Because the output is directly related to the training data, if the training data does not include all valid conformations of a residue, then the output could be incorrect. This limitation need to be mentioned. Based on looking at the provided result in the github repo, the results do not seem robust (details below).

We added a new paragraph in Discussion to discuss limitations (page 27-28). The paragraph reads:

"There are limitations of the DAQ score. This score currently only provides scores to protein models and cannot handle other molecules, such as DNA/RNA, stereochemical compounds, and water molecules. DAQ is a residue-wise score, and thus does not provide scores for individual atoms. Also, the targeted map resolution for DAQ score is
2.5 Å to 5.0 Å because this was the resolution range of maps in the training set, although DAQ seem to work fine for maps at a higher resolution (Supplementary Fig. 2). For a map with a worse resolution (> 5.0 Å), we observe cases that local density features are not sufficient to distinguish individual amino acids, and DAQ score values go down around 0 (neutral). Each of these characteristics of DAQ score is complementary to atom level scores, such as Q-score, therefore, we recommend to use a proper validation score depending on the needs. As we have shown above, DAQ is very efficient at identifying misalignments."

Regarding the use of a sliding window, in Table 1 we showed the accuracy of detecting potentially wrong amino acids in the first model in PDB entries. There, we showed results using no window (i.e. window = 1), and windows of 3, 11, 19. As shown, using no window showed a sufficiently high accuracy in 4 evaluation metrics. Indeed, using no window, DAQ still performed the best among the four scores compared in Table 1. We just used 19 residue long window in the analysis because 19 residues showed the highest accuracy. The superior performance by using 19 residue window is not specific for DAQ. It is true for all other scores, too, Q-score, EM-Ringer, and CaBLAM.

Regarding the output example in the github repo, we now added more explanation in a readme file in the github folder: https://github.com/kiharalab/DAQ/tree/main/result. The example in the github repo is actually an example we explained in Fig. 4c. In the main text. (please also see below where we answered this question).

F. Suggested improvements: experiments, data for possible revision

A potentially necessary improvement would be to illustrate the score for at least a few types of residues at different resolutions. Otherwise, it is relied upon general statistics to demonstrate. Other improvements would come mostly from being more clear on the limitations of the method and being more realistic with the conclusions.

As we noted above, we added a new paragraph explaining limitation of this method in Discussion section in page 26. Regarding the residues with different resolutions, we added Supplementary Table 3. We also added Supplementary Fig. 2 that shows the influence of the map resolution to the DAQ score. It is a good suggestion, thank you for suggesting it.

G. References: appropriate credit to previous work?

Yes, but could be improved. They may consider to reference the Haruspex method. It would be appropriate to also reference other per-residue scoring methods such as TEMPY (CCPEM), and Phenix.

TEMPy, CCP-EM and Phenix are now cited in page 3. (Reference numbers are 11, 12, and 13, respectively).

H. Clarity and context: lucidity of abstract/summary, appropriateness of abstract, introduction and conclusions

Descriptions are mostly clear, though further clarifications may be needed as indicated below. The abstract is a bit ambitious for what is actually presented, and the conclusions do not adequately indicate limitations.

In Abstract and Discussion, several lines are deleted which sounded ambitious. Also, Discussion was substantially revised. As mentioned above, we added a new paragraph to address limitations.

Specific Points:

Abstract: It is stated that the method can identify “potential misassignment of residues in the map, including residue shifts”. Later it is stated that none of the scores can tell the difference between misaligned or mispositioned residues (page 13). Perhaps these can be better consolidated and more clearly/realistically stated.

I agree that the description in page 15 was confusing. We wrote the paragraph to clarify different nature of the three component scores, DAQ(AA), DAQ(Atom), and DAQ(SS). It now reads:

To summarize, the three scores in DAQ have complementary nature. DAQ(Ca) identifies Ca atoms that are likely modeled incorrectly, whereas DAQ(SS) can identify inconsistencies in assigned secondary structure. Thus, these two scores are adept at detecting mispositioned residues. In contrast, DAQ(AA) examines characteristic local density features of amino acid residues, thus able to detect misaligned residues along otherwise correctly modelled backbone structure.

Page 3 bottom. For map-model scores, atom inclusion score should also be mentioned (reference Lagerstedt et al, JSB, 2013). It should be mentioned that it is currently used in the EMDB/PDB validation report.

We cited the paper at the bottom of page 3.

Page 4 top: Molprobity score also includes bond lengths, and angles between 3 connected atoms.

We rephrase the sentence to include these terms. According to the article “MolProbity: More and better reference data for improved all-atom structure validation” Protein Science (2017) https://onlinelibrary.wiley.com/doi/full/10.1002/pro.3330#pro3330-bib-0013, “At the frequent request of users, we developed the MolProbity score as a combined single indicator of model quality.13 It uses a weighted function of clashes, Ramachandran favored, and rotamer outliers, scaled and normalized so that its value approximates the resolution at which that score would be average.”, which we originally meant and put it as “…in a single score of overall model quality”. But as pointed out by the reviewer,

the Molprobity package includes evaluation of other terms. Thus, we revised the sentence. Thank you for pointing it out.

Page 4 bottom paragraph. Suggest to add ‘protein’ in ‘new approach for cryo-EM model validation’.
Thank you, we added ‘protein’.
Page 5 top paragraph. To say that ‘local structural features’ are used instead of ‘local density values’ seems contradictory/misleading because ultimately the neural network is fed these local density values.

Thank you for pointing it out. We changed it to “local density”.

Page 5, bottom paragraph. It is stated that comparison of DAQ score to other scores showed a 'clear advantage' for the DAQ score. This may be too strong/vague/general of a statement. There are reasons why this is not so, for example, the training set does not include maps with higher than ~2.6Å resolution, so it's not clear what the method would do there. Also, the method can only be applied to protein residues. Lastly, more thorough tests and comparisons to other methods are needed before such a statement should be made.

We deleted the sentence "Comparison of the DAQ score with Q-score, EMRinger, and CaBLAM showed a clear advantage for the DAQ score in identifying errors in cryo-EM models." Also, in Discussion, we addressed the limitation of the DAQ score including that it is only for proteins and its target resolution range is 2.5 to 5.0 A. In discussion, we also suggested to also use other scores that use different approaches.

Page 6 middle paragraph. 'The probabilities for a grid point …'. Not clear, is a 11^3 box used centered at the grid point?

Yes, that is correct. To clarify, we rephrased the sentence to "The probabilities for the center grid point of a box of a 113 Å3 size are computed…" in page 6.

Page 7 middle. 'atom type score of amino acid residue i'. The ATOMscore tries to indicate if a voxel appears to indicate a given type of atom, but only the Ca atom is considered here. Atom type usually is meant to reflect whether an atom is C, O, N, etc. Indicating this score has something to do with atom type is misleading. Perhaps this can be revised somehow to be more specific towards identifying the Ca atom.

We agree. We now noted the atom score as DAQ(Cα) throughout the manuscript.

Page 7, bottom paragraph. The accuracies (77.3% for secondary structures, 51.5% for Ca atoms, and 28.2% for amino acid type) seem low. It is mentioned that the accuracies do not need to be perfect for model validation, as long as the 'correct' residue is preferred for a position in the map. It is not described what is meant by 'preferred', and it seems quite likely that even this method could potentially give a higher score to the incorrect residue, i.e. it is not shown otherwise, which would be hard to do; unless it can be done, this statement needs to be revised).

DAQ score is not for specifying a residue at each position in a map to guide structure modeling. Rather, DAQ identifies incompatible residues that do not fit to the local density. Thus, as long as the correct

residue has a positive score, the validation will work. It is possible that multiple residues have positive values. But we realized that we have not provided the statistics of correct amino acids that have positive DAQ score. Now we added new data in revised Supplementary Fig. 1 and Supplementary Table 2 (Excel). With these new data, the corresponding section in page 7 is revised.

Page 9. The two classes misaligned or mispositioned; misaligned have to do with shifted residues, but is the simple criteria that a residue is within 2 Å of another residue always indicate a shift? They are a bit hard to comprehend from the text, could they be illustrated perhaps?

We made a figure for the misaligned and mispositioned residues as Supplementary Figure 3 and referenced in page 9.

Page 9 bottom paragraph, Figure 2. Average DAQ scores of first and revised version are plotted. Figure 2a shows they are correlated when considering all residues, but what does this correlation mean - the higher the DAQ score of the initial model, the higher the DAQ of the revised as well; this seems reasonable and is a good 'sanity check'.

Yes, the first model and revised model have many residues in common with difference in only some local regions, overall the scores correlate; i.e. if a first model has a high score, then the revised model also has a high score.

Figure 2b – when considering inconsistent residues, there is almost an inverse correlation – the lower the DAQ score of the initial version, the higher it becomes in the revised version. The conclusion is somewhat circular, i.e. the improvement is in the inconsistent regions, but this is somewhat obvious it would be so, i.e. they would not have changed unless it was for some sort of improvement. But any other map-model score would likely show the same relationship.

The logic is not circular.
The story is, when a model in PDB was revised, for most of the cases, the revised model has a higher DAQ score. Thus, the DAQ score agrees and supports how the model was revised. Note that the revised regions in two models were detected not by the DAQ score but by structure comparison between the first-version and revised version models. We analyzed the revised, and thus, inconsistent regions between two versions of the models by DAQ.

To clarify, we revised the summary sentences in the paragraph (page 11). The sentences read:

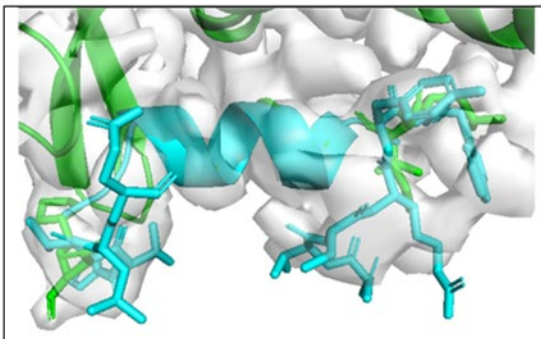
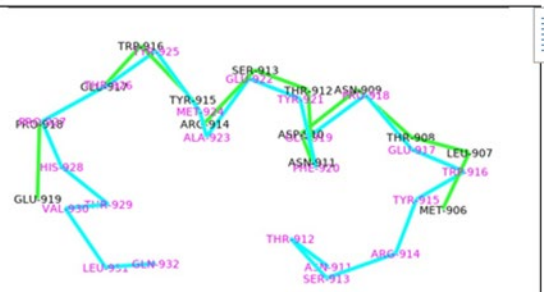
“To summarize, for the majority of the cases where a structure model for a PDB entry was revised, the revised model showed a higher DAQ score, which indicates that the DAQ score agrees with and supports the way in which the models were revised.”

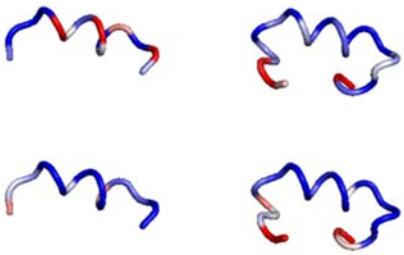
Also, to provide more insights of the data of Fig 2b, we plotted the same data differently in Fig. 2c and 2d: Here, the x-axis is the average DAQ score of consistent residues (common regions in the two versions of the models), and the y-axis is the score change of the inconsistent regions from the first-version (Fig. 2c) to the revised models (Fig. 2d). As shown, in the first models, inconsistent regions have a lower score than the other parts. And after the model revision (Fig. 2d), the same regions now has a higher score than before for most of the cases, to a similar level as the other regions (close to the y=x line). This indicates that those regions had a poor quality before and revised to have the similar quality as the other parts of the model.

The DAQ score is higher in the inconsistent regions for only ~89% of the residues. Does this mean the DAQ score has a 11% error rate? Or is it because some inconsistent residues moved from a favourable spot to a less favourable spot, for the benefit of other residues perhaps. Or perhaps that is for noisy parts of the map.

The 11% corresponds to 4 maps, EMD-30127 (resolution: 2.9 A; PDB: 6M71-A), EMD-11127 (resolution: 3.76 A; PDB: 6za9-O), EMD-20655 (resolution: 2.9A; PDB: 6u5t-G), and EMD-30226 (resolution: 3.7A; 7bw4-A). Overall, as you speculated, the revised regions of these 4 maps are in noisy part of the maps and revised models have more exposed side-chains than the first model. When we used Resmap, local resolution of the regions tend to be low, too. In the main text, we added this information on page 11. Below, we discuss individual cases.

EMD-30127 PDB 6m71A
The inconsistent regions are residue 906-919. In the revised model, the authors have shifted residues by 9 residues and slightly changed the CA positions. We found a clear DAQ score improvement in these regions with revised residue assignment and CA positions. However, the authors added additional residues in the model at the two ends of the helix, residue assignment 911-914, 929-932 (numbering in the revised model), which are noisy regions with a lower resolution by Resmap (Quantifying the local resolution of cryo-EM density maps, Kucukelbir et al., Nature Methods 11: 63– 65 (2014). The DAQ of these added regions are low, which lead to the average DAQ score drop from the first version model to the revised model.

	
Green, the first model; cyan, the revised model. The revised Model has more exposed side-chains.	Superimposition of the inconsistent regions in the two models. green, the first model; cyan, the revised model.



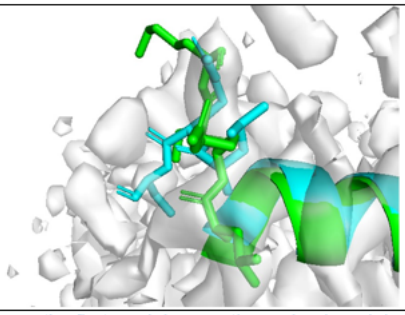
The top row, DAQ(AA)-score of the first-version (left) and the revised model (right). In the revised model, the additional residues in red have DAQ scores of -2.33 to -0.27(average: -1.33).

The bottom row, local resolution computed by Resmap. The first-version (left) and the revised model (right). The local resolution of the additional residues in red in the revised model was 3.5 to 3.6Å, while the surrounding regions have a resolution around 3.1 to 3.2 Å.

EMD-11127 PDB 6za9O

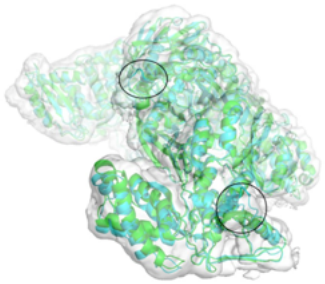
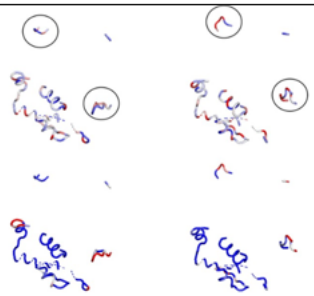
For this example, the inconsistent region of the revised models has a negative DAQ score. At the inconsistent regions from residue 47 to 49, the authors changed the residue assignment positions for

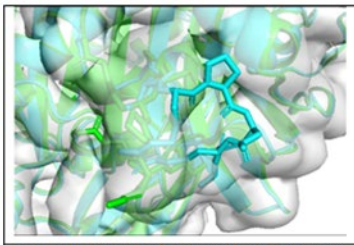
3 residues and added 1 more residue, 50LYS in the revised model. We found the average DAQ score decrease from 0.065 to -0.507 in this region in the revised model.

	
Green, the first model; cyan, the revised model. The revised Model has more exposed side-chains.	Superimposition of the inconsistent regions in the two models. green, the first model; cyan, the revised model.

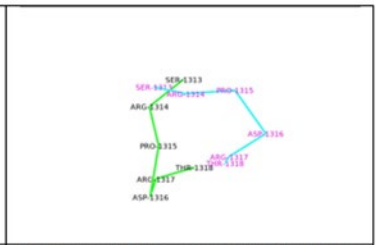
EMD-20655 PDB 6u5tG

There are two inconsistent regions in the models for this map, residue 1313-1318 (the top region) and 1739-1748 (right bottom region in the figure below), where the DAQ In the right panel below, low DAQ scores are shown in red. The DAQ score dropped from 0.127 to -0.808 (residues 1313-1318), and 0.015 to -0.372 (residues 1739-1748). These regions have a lower local resolution when analyzed by Resmap. The entire map is reported to have a 2.9Å resolution, while the local resolution of the 3 regions are 3.4-3.5 Å and 3.4-3.6 Å, respectively.

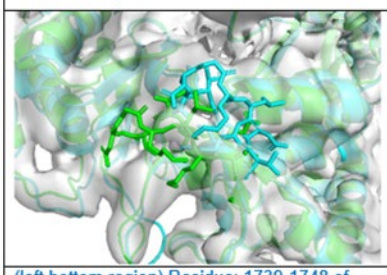
	
Green, the first model; cyan, the revised model.	The top row, DAQ(AA)-score of the first-version (left) and the revised model (right). The bottom row, local resolution computed by Resmap. The first-version (left) and the revised model (right).



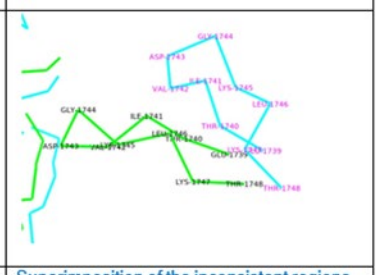
(top region) Residue 1313-1318 of 6u5t-G. Green, the first model; cyan, the revised model. The chain exposed in the revised model.



Superimposition of the inconsistent regions in the two models



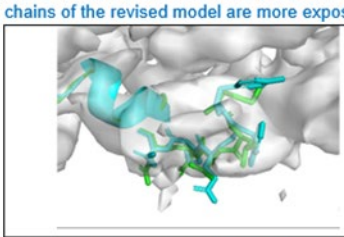
(left bottom region) Residue: 1739-1748 of 6u5t-G chain. The chain more exposed in the revised mode (cyan).



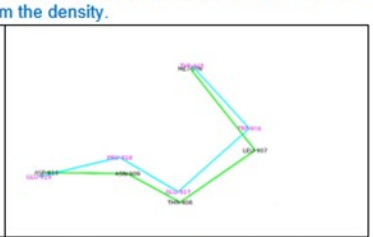
Superimposition of the inconsistent regions in the two models. green, the first model; cyan, the revised model.

EMD-30226 PDB 7bw4A

For this example, the inconsistent region dropped from the positive DAQ score, 0.394 to negative score, -0.086 in the revised model. The inconsistent regions are residue 906 to 910. In the revised model, sequence assignment was shifted by nine residues (Asp910 to Glu919, Asn909 to Pro918) and some residue positions and types were updated (Met906 to Trp915, Leu-907 to Trp916) .Side-chains of the revised model are more exposed from the density.



Green, the first model; cyan, the revised model. The revised Model has more exposed side-chains.



Superimposition of the inconsistent regions in the two models. green, the first model; cyan, the revised model.

Page 10 second paragraph. The AUC-ROC could be described a bit further here for clarity. Having to use such a large window size of 19 residues is concerning, i.e. why is the score not reliable at a single residue level? Again the score seems to show that residues that move more tend to have better scores in the revised version, which is to be expected.

In page 34 in the Methods section, we added a new paragraph to explain how AUC-ROC was computed. Supplementary Fig. 12 shows an example of ROC curve.

Regarding the window size of 19, later in Table 1 (page 18) we compared DAQ and three other scores with different window sizes. As shown, for all the scores improved their performance by using a large window size. Thus, not only for DAQ, but also for Q-score, EM-Ringer, CaBLAM, using an averaging window increases AUC-ROC because using a window allows detecting regions which are consistently in a low (or high) quality in its local sequential neighbors.
To explain we added a sentence "Using a single amino acid (i.e. a window of 1) showed a high AUC- ROC of 0.78. It further increased by using a larger window size because using a window allows detecting regions which are consistently in a low quality in its local sequential neighbors.".

Using DAQ at a single residue level is not unreliable as shown in Table 1. DAQ showed the highest performance among the four scores by sing a single residue only (window size of 1). We used a window of 19 because it further improved the accuracy of detecting inconsistent residues.

Page 12 middle paragraph. The sentence 'Comparison of AA-score with three other scores …' has incorrect syntax – it is not the comparison that demonstrates. 'The other scores only …' is also incorrect, since it is then mentioned that Q-scores also shows drops in all regions; the 'largely missed the problems in region (iii)' part is not clear and may be misleading.

We deleted the phrase that Q-score "but largely missed the problems in region (iii).".

Page 13. "AAscore is unique in that …'. It is not clear how it is unique, especially since it can't detect the difference in the two cases. It was mentioned in the previous paragraph that Q-scores also could potentially detect the same issues. The rest of the sentences in this paragraph are also not too convincing in my opinion and are not sufficiently demonstrated.

We deleted the sentence "Aascore is unique …" and rewrote the whole section, which now starts "To summarize, the three scores in DAQ have complementary nature. .."

Page 15 bottom paragraph. '… identifies mismodelled residues by absolute quality but not by…' not clear what is meant here. AUC-ROC and AUC Precision-Recall should be described as to how they are applied and what they indicate; it is hard to evaluate the results without a clear explanation.

In page 34-35 in Methods, we explained AUC-ROC and AUC precision-recall. In Supplementary Fig. 12, we showed an example of the ROC, which was used to compute AUC-ROC. In both AUC-ROC and AUC precision-recall, residues in structure models are sorted by a score (e.g. DAQ) and checked if inconsistent residues locates high in the ranking. A more general explanation can be found, for example, https://en.wikipedia.org/wiki/Receiver_operating_characteristic.

Page 17 last paragraph. Last sentence again seems redundant, i.e. 'higher-scored models have higher scores'; this seems obvious and it is not clear why this should imply anything.

Agreed; we deleted the sentence.

Page 18, Figure 4. The panels are very small and cryo-EM map density is not shown too well. It could be more informative to show less examples and more clear density as illustration, perhaps adding some to supplementary.

As suggested, we moved two examples, the second and the third examples in the original Figure 4, to Supplementary file (Supplementary Fig. 8).

Page 19 first paragraph. The text is somewhat fragmented and hard to follow, i.e. from the sentence 'Positions in this model truncated as alanine…', and in the next sentence '… hydrophobic side chains …' which do not apply to the same 'truncated alanine' residues.

We rewrote the sentences to "His227, Leu229, Asn231, Asn234-Lys241 are truncated as alanine, which underscore the lack of interpretability in this region."
Also, in the next sentence, we rewrote a list of example residues as "(3J6B His227, which aligns with a hydrophobic core residue in 5MRF, Leu229; 3J6B Ser228, which aligns with 5MRF Val230; and 3J6B Ser219, which aligns with 5MRF Leu220)."

Page 19 bottom paragraph. Molprobity does not report map-model correlation.
"The Molprobity report for 5H64 indicated that there are map-model correlation problems in 5% of residues, while 6U62 chain A does not have any correlation problem"

It was a mistake. It is "Full wwPDB EM Validation Report" which associated with this entry, EMD- 6668, PDB ID: 5H64, not Molprobity report. The file is
https://files.rcsb.org/pub/pdb/validation_reports/h6/5h64/5h64_full_validation.pdf

The second page of this report has a summary table of geometric issues found in each chain and chain B is reported to have 5% of outliers shown in red bar. On the other hand, the corresponding chain 6U62 chain A does not have any geometric issue reported in its report. We corrected it. (This example is moved to Supplementary Fig. 8, following another suggestion by this reviewer).

Page 20 top paragraph. In a 4.4Å map, it is hard to talk about side chains with much confidence, so it not clear what the DAQ score is providing here. Other than reflecting approximate fit-to-map score, it is unreasonable to talk about truly validating such residues based on map densities alone.

Page 20 middle paragraph. Again in maps at ~3.8, 3.6Å resolutions, it can be hard to be fully sure of side chain placements and identities without knowing the sequence, and not clear how the DAQ score is showing here.

Following another suggestion above the reviewer 1, these two examples are moved to Supplementary Fig. 8.
These are examples that protein model pairs that have a large DAQ(AA) score difference. We agree that the density is poor and by human eyes side-chain density is not clearly evident. Therefore, in the description we did not say that one is “mismodelled”, “misaligned” or “wrong”. However, when we looked at the model with a significant lower score, we found that there are obvious problems with modeling, such as exposure of hydrophobic amino acids, in the lower scoring model that would support why DAQ favored one structure over the other. For those problematic regions, DAQ showed a negative low score, which means that the fit of the amino acid at that position is way below average of all the positions in the entire map, which seems to make sense.

Page 20 bottom paragraph. In the last example, going from 3.43Å to 2.97Å of course should lead to a better model and better map-model scores.

Agreed. The models from the two maps have residue shift, and DAQ score as well as visual inspection suggest that the one from the 2.97A map is a better fit to the map.

Page 21 middle paragraph. The paragraph starting with ‘Fig 5a’ may be a bit too speculative. It seems a bit harsh to make the claim that maps and models are mismodeled in general without sufficient evidence. When dealing with maps at 3-5Å resolution, where side chains may not be clearly resolved, and furthermore with potentially unresolved regions, claiming such regions mismodeled may be the same as claiming they are unresolved; i.e. the map itself, the only data we have, may not support either claim.

I agree that the term “potential misaligned region” is strong. We wrote the sentence to just say the distribution of DAQ score have significantly low values (below -0.5, -1), which correspond to inconsistent residues in the PDB2Ver dataset.

Page 22. The density is Figure 5c is not clearly shown, and the paragraph on page 22 basically illustrates the point above, i.e. it seems that residues in question are not resolved in the density, and in that case is it fair (or is there enough evidence in the first place) to say that they are mis-assigned? It would be much more fair to find, for example, cases in 2.2Å maps where the residues are miss- assigned, in which case, we would know for sure that is the case because the density would clearly shows the correct assignment.

We agree that the example of Fig. 5c is in a low resolution and it is difficult to show correct assignment. Therefore, we revised the text and removed words such as “miss assignment” or “wrong”. We have chosen this model because the potential problem is obvious because it has entanglement of chains and have serious steric clashes.
But to further address this point, we now provide two examples of models (Fig. 5c and 5d) built into high resolution maps (<3.0 Å) where potentially mis-assigned residues are in regions where the density is of high quality (see pages 23-24 of resubmission). We also fit these regions with models generated by AlphaFold2, which agreed with our assessment.

Page 24.
‘… are too sensitive…’ not clear what this sentence means to say.
‘Many of the mistakes made would have been discovered by well-trained…’ – This sentence is not true. Even a well-trained structural biologist may not be able build every residue with confidence in some maps, e.g. 3-4Å, especially in less-resolved regions. Such statement unfortunately underscores and does a bit of disservice to the efforts that go into model-building in the field of cryo-EM in general.

We deleted the sentence starting with “Many of the mistake ..”.

‘… it is urgent…’. In my opinion and experience, existing tools that use local map-model correlations (or Q-scores) can also robustly indicate the same issues that the DAQ score would. (E.g. see Pintilie and Chiu, Acta. Cryst. D., 2021, Figure 6). It is true however that such tools should be more widely used and incorporated into validation reports.

We deleted the two phrases, ‘… are too sensitive …” and ‘Many of the mistakes…. by well-trained ..”. In a sentence, I put DAQ, Q-score, and EM-Ringer in parallel. “Validation tools including DAQ, Q- score, EMRinger..”

Page 25.
‘… trained on maps of 2.67 to 4.5Å’ – why not also use higher resolution maps? Does this mean it would not be too meaningful to apply to maps at higher resolutions?

In this work, we focused on maps of 2.5 to 5.0A resolution because models from maps in this resolution may have errors, and many of them can be reasonably corrected. We thought maps better than 2.5A resolution usually have correct models and the DAQ score may not be needed. On the other hand, maps at worse than 5.0A resolution may have many errors. But at that resolution, in many cases residue types are not confidently detectable and alternative models may not be suggested. To explain it, we added the sentence: “In this work we focused on this range of resolution because maps at higher resolution would

not have many modeling errors and maps at a resolution lower than 5.0 Å may have errors but correct residue types may not be detected from the density features."

'…pretty good performance…' perhaps could be made a bit more concrete. '…resolutions between 2.5 and 5.0Å..' why is it different than the training set resolutions.

We removed the phrase, "pretty good performance".
Regarding 2.5 and 5.0 A, we applied the same criteria but the criteria yielded maps from 2.67A to 4.5A. To avoid confusion, we unified it to 2.5 and 5.0 A.

Page 26.
'… collected voxels with voxels of size 11^3…' not clear – a voxel is a single grid point. Perhaps boxes with size 11^3? Is this box large enough for ARG residues which can be ~10 from Ca to end atom if the box is centered on the Ca?

We changed voxels to boxes throughout the manuscript. Regarding the question about ARG, we have not checked all the ARGs in all the dataset, but according to Supplementary Figure 1, ARG does not show any particularly different score distributions.

Page 30. '… estimated Q-score…' would be better phrased '…expected Q-score…'. Also this is not really 'normalization' but more of an offset, making Q-scores above the expected Q-score positive and those below negative.

Thank you, we corrected it.

Code: I was able to install and run the method from the github account. First I did so on a Macbook laptop, but the code did not run as it requires GPU and CUDA.

Then I tried on our unix cluster; the code installed and ran but output some error messages, which appear related to CUDA:
mrc.c: In function 'readmrc':
mrc.c:193:11: error: 'for' loop initial declarations are only allowed in C99 mode for(int i = 0; i<mrc->NumVoxels; ++i)
…
sh: /scratch/g/gregp/daq/DAQ/process_map/gen_trimmap/TrimMapAtom: No such file or directory I have submitted an issue on github to see how it would be resolved.
Perhaps I missed it but there are no instructions how to run it in Google Colab.

We noticed the issue posted, and answered it on https://github.com/kiharalab/DAQ/blob/main/QA.md. We are very happy to help the installation process. If you still have problem, let us know.
The instruction about Google Colab is provided in the Google Colab page. It is here: https://colab.research.google.com/drive/1Q-Dj42QjVO8TCOLXMQBJlvm1zInxPkOu#scrollTo=lok4D8YC4lOb

To clarify on the Github page https://github.com/kiharalab/DAQ , we added the following sentence, right after the link to the Google Colab: “All the functions in this github are available here. Related instructions are included in the Colab website.”

Questions:

1. The default sliding ‘half window’ is 1, does this mean it averages the nearest 1 residue on each side of a given residue, so in total 3 residues?

Yes, that is correct. A half window 1 at i-th position considers i+1, i and i -1 th residues, thus in total three residues. But as we showed in the Running Example (https://github.com/kiharalab/DAQ#running-example) , the recommended half window size is 9 (thus the total length is 19, as we mainly used in the manuscript).

2. What is the effect of the –stride parameter on the score and running times? Does it make it less accurate?

We provided a plot that shows computational time with using stride parameter of 1, 2, 4 in Supplementary Fig. 9. Also, the accuracy of detecting inconsistent residues (potentially wrong residue assignments in the first model in PDB entries, PDBVer2. Dataset) in Supplementary Table 9. These new results are described in a new section, Computational Time in the manuscript (page 25).
Compared to the default stride of 1, using 2 and 4 reduces the computational time drastically, because in principle the number of boxes to process will reduce by 2^3 = 8 folds if stride of 2 is used. In terms of accuracy, only stride 4 substantially deteriorated the results.

I looked at the result folder in the github folder, showing EMD:2566 and PDB:3j6b.

1. It is unclear what scores are being output in the pdb files. What is the raw score and the scores in the _w9? Are the SSScore, AAScore, ATOMScore combined?

In the result folder https://github.com/kiharalab/DAQ/tree/main/result , the raw score (dqa_raw_score.pdb) provides individual DAQ(AA) score without window average. In the daq_raw_w9.pdb file, the score is averaged using a half-window size of 9 (i.e. the window size of 19 in total). Thank you for pointing it out, we put an explanation in a readme file in the folder, https://github.com/kiharalab/DAQ/blob/main/result/README.md.

2. What is unexpected is that the score appears widely different for 3 helices which are similarly resolved (image attached). Perhaps the actual residues are different, leading to different scores. However this makes me doubtful about the usefulness of the score, because it doesn't seem to take into account whether the backbone is resolved or not.

This is an example discussed in the manuscript in Fig 2c. To clarify, in the Github folder of this example, we put the explanation in the README.md (https://github.com/kiharalab/DAQ/tree/main/result/README.md), based on the discussion we have in the manuscript. It reads:
"This protein is Chain 9 of the Ribonuclease III domain from PDB entries 3J6B. When 3J6B is compared with another PDB entry 5MRF, the sequence identity of these pairs is 99.5%. However, residues 227 to 241 are shifted between the two structures.
In the 3.2 Å resolution EM map (EMD-2566, 3J6B), density for the terminal helix region (His227-Lys241) is not of high enough quality on its own to confirm choices on residue or atom identity. Positions in this model truncated as alanine underscore the lack of interpretability in this region. There are, however, a few telltale signs that there are misalignments, such as hydrophobic side-chains exposed to solvent (3J6B Ile232, which aligns with 5MRF Asn234) and polar residues packed into the core of the fold (3J6B His227 vs. 5MRF Leu229, 3J6B

Ser228 vs. 5MRF Val230, and 3J6B Ser219 vs. 5MRF Leu220). At these positions, side chains in 5MRF fit the environment well. For these reasons, 5MRF-9 seems to be a better fit for both EM maps, consistent with the DAQ(AA) score, and the inconsistent regions in 3J6B (227 to 241), which correspond to the red regions in the associated image, may have errors. To have a more clear understanding of the region, we attached the comparison in visualization.png. The left panel shows the DAQ(AA) score for this structure (EMD-2566, 3J6B chain 9). The misalignment region is shown in red by DAQ, which indicates lower scores for this region. In the right panel, we compared the assignment of 3J6B-9 (shown in green) and 5MRF-9 (shown in cyan) in the helix region where DAQ outputs low scores. We can see the difference of the residue assignment in 3J6B-9 from 5MRF-9." Review by Greg Pintilie
We appreciate your careful and thorough comments.

Responses to Comments by Reviewer 2

Remarks to the Author:
In this manuscript, Tersahi et al. propose a new validation metric to help detect errors in atomic models (PDB coordinates) that have been derived from cryoEM maps.

The metric is essentially a probability, for each amino acid position in the model, that the local cryoEM map feature indeed suggests the modeled amino acid (as opposed to one of the other 19 possibilities). The score is normalized to be easily interpretable: positive values indicating good agreement with the map, negative values suggesting likely errors in the model.

The task of algorithmically labeling parts of atomic models that are not well supported by experimental data is not trivial, and the authors claim that their method significantly outperforms existing metrics in that regard. If true, this could make this a very valuable tool for the cryoEM community - both investigators interpreting cryoEM results and repositories and databanks tasked with validating deposited results.

I found this manuscript difficult, at times painful, to review, largely because I found the text did not clearly explain what the authors had done, or the language was confusing in enough places that the intended meaning was lost. I will give specific examples of this later. However, despite all this, and following a few read-throughs, I am now of the mind that the authors' claims are for the most part substantiated by the evidence and that their validation metric indeed may well be a notable advance on the current state of the art.

For this reason, I would recommend that the authors resubmit a revised version of the manuscript for consideration. Assuming my concerns were addressed, and the main claims still stood intact, I would then be supportive of publication.

Below my signature is a list of issues I encountered when reading the manuscript, roughly sorted by importance.

Alexis Rohou, Genentech January 2022

We appreciate your careful and thorough comments.

Obstacles to comprehension:

Near the beginning of Results, the authors go straight to a discussion of how their DAQ scores, but have not explained at all what Emap2sec+ is. Thus, this whole section is very confusing on first read, because you have never explained that Emap2sec+ is in fact a neural network whcih you have trained on a selected subset of the corpus of deposited maps & models. I strongly suggest you add a few sentences to first explain what Emap2Sec+ is (without details), and only then explain how you use the scores it outputs to compute DAQ scores.

We agree. We have now completely revised that section. Since the abrupt appearance of Emap2sec+ was a source of confusion, we explained it in page 8 by referring to Fig. 1b, which shows the network architecture. Figure 1 is completely revised to provide detailed steps of DAQ computation (panel a) and the network architecture (panel b).

I would also suggest that you add to Figure 1 an overview panel with flowchart / algorithm design, showing (a) the training of the neural net (select training set, select test set, train on training set, etc), (b) the use of the trained neural net (iterate over each Calpha position, for each position use NN to compute probabilty, etc).

Thank you for the suggestion. We revised the Figure 1 to include more detailed steps of the inference process (i.e. when the software is used for assessing the quality of maps) and the neural network architecture. We made another new figure, Supplementary Figure 10, which illustrates how the datasets were constructed and how the network was trained, and referred to it in page 27, where we explained the training process. We separated these two figures because it was hard to put everything in one figure.

Claims that may not be supported by evidence:
- abstract:
-- "indicating a great need for improved validation tools to help achieve the most correct structures possible"; Does the evidence really support this claim? Many (most? see other comments) of these could be explained by errors in modeling caused by noise low-resolution regions of maps; in those cases, in my view, improved validation such as provided by DAQ or other tools would not have helped achieve more accurate models. At best they help flag a problem. I guess what I'm saying is even with improved validation tools telling them something is wrong with a model in a poorly-resolved part of the map, authors may still deposit

We removed the phrase “indicating a great need for improved validation tools to help achieve the most correct structures possible".
Validation tools including DAQ flag potential errors in a model so that researchers pay attention to the flagged regions. Of course it is up to the users if they wish to fix the problem. The aim of validation tools is basically the same as structure validation report that is associated with each PDB entry. It is also meant to help others trying to understand the validity of the deposited models. We also answered to the similar comments given to the similar statement we originally had in Discussion.

- result
-- "DAQ(AA)-score showed the highest values in the two types of metrics, which means DAQ(AA)- score identifies mismodelled residues by absolute quality but not by comparing with other residues in the

same model." I do not follow this sentence. I don't understand the claim, and therefore cannot evaluate whether the results really support the claim. Please state your claim more explicitly / clearly and explain how the fact that DAQ(AA) the two metrics

As we answered to the same comments below, we have added a new paragraph in Methods (page 34) to explain. Also, the corresponding section in the main text in page 17 is largely rewritten.

-- "which implies that the inconsistent residues in lower scored models are probably misplaced." I do not follow. What does misplaced mean in this context? Don't you mean misaligned?

A low DAQ(AA) score occurs if the amino acid was misaligned or when the local region of the model is mismodelled and the residue is mispositioned. We changed the word “misplaced” to “misaligned or mispositioned” to make it consistent with the explanation given in page 12 and now illustrated in Supplementary Fig. 3.

--- Please clarify your meaning or remove. Your argument border on the circular, I fear. You first selected pairs of models with differences. You then ran your score and classified models according to whether their DAQ scores for inconsistent residues were higher or lower than the other pair member. Then you marvel at the fact that the residues in the lower-scored models have lower DAQ scores. I fail to see how this is remarkable. Maybe the only conclusion I would draw from this is that the depositors of the lower-scoring models tended to make mistakes all over, not just in one region.

I see the description is confusing (page 19-20). To clarify, we revised Fig. 4a and Fig. 4b and totally rewrote the section (page 19-20).
The story of the section of “Evaluating inconsistent protein model pairs of high sequence identity” is the following: The dataset in this section are pairs of proteins with high sequence identity (>90%) yet has difference in sequence assignment in local regions. Although these protein pairs have difference in sequence assignment, there is a slight chance that both models are correct because these two proteins do not have 100% identical sequences (although highly similar). If that is the case, both models should have similarly high DAQ(AA) score. However, when we computed DAQ score of the inconsistent regions in model pairs, for almost all the cases one model has a lower, negative DAQ scores, and the other model has a positive score. These results implies that one model maybe wrong.

-- Suppl Fig 1: "all the scores have positive values for the residues, suggesting that the scores can identify the residue specific density features properly from background." I'm not sure this is correct. If I understand correctly, the score would identify a residue correctly only if that residue's score were higher than any other, which is a more stringent test than just having a postive score value. I wouldn't be surprised if several residues would score positive at a given position. This claim should be qualified or

better supported by evidence. This may impact the main text results, near the sentence "correctly assigned residues all have positive values in Supplementary Fig. 1,"

The original Suppl Fig. 1 was showing the average DAQ scores, and I agree that it was not sufficient to say “all the scores have positive values…”. We should have said “all the average scores..” to be accurate, but we decided to provide more data.
In the revised Suppl. Fig 1, we now provided the distribution of DAQ scores of all the residues in the dataset. The distribution is shown in box plots with all the data points explicitly shown (orange data points). We provided two plots, a) DAQ score distribution of individual residues; b) DAQ scores when averaged over a window or 19 residues. We added b) because as we show later in the manuscript, we use the window average of 19 residues when analyzing protein structure models. As shown in the plot, not 100%, but the majority of the residues have positive DAQ scores. In the associated Supplementary Table 2, we showed that the % of residues that have positive scores for each category of DAQ scores. For DAQ(AA), on average over the 20 amino acids 78.3 (96.6)% of residues have positive values. In the parentheses shown are values when an average window of 19 was used. This new result support our decision of using 19 residue-long window, as the fraction of

residues with positive DAQ score increased substantially to 96.6% by using the window. This is described on pages 7-8.

Modifications needed to support claims:
- "Fig. 2c shows the Area Under the Curve of Receiver Operator Characteristic (AUC-ROC) with a window size of 19." Could the authors please add in the supplementary a pictoral and a text description of the what the AUC and ROC are? How exactly are they computed from the list of residues sorted by their DAQ scores? I am not familiar enough with these tools to intuit from the text how exactly they are calculated. Since AUC-ROC backs one of the most important claims of the paper (that DAQ scores outperform existing validation metrics in identifying modeling errors, Table 1), this point should be explained in detail in suppl. I'm sure other validation metric developers will want to check these points carefully and reproduce these calculations.

We added a new section in Methods (page 34-35) to explain it. To explain AUC-ROC, we added a new figure, Supplementary Fig. 12, which is a curve used to compute Fig. 2c.

Suggested changes:
- intro:
-- "unlike existing map-model scores or (...) likely to be incorrect". Model-coordinate scores by definitely would not be able to detect mismatch with the experimental map, so should not be mentioned here, in my opinion. Also, can you specify which map-model scores you are thinking of here? I would have

expected that local correlation-based metrics should be able to do this (at least before seeing your Table 1) - is that not a claim developers of those methods have made?

We rewrote the sentence just to indicate that the approach DAQ score is taking is different from existing map-model scores where DAQ uses deep learning to capture amino acid specific local density distribution. We revised the entire paragraph (page 4-5). Particularly, we removed the phrase of “Therefore, unlike existing map-model scores or model-coordinate scores,”.

- results
-- In the text you refer to your scores as "DAQ scores" or "DAQ(AA)... how about calling them DAQ_SS or DAQ(SS), DAQ_AA or DAQ(AA), ..., in the equations?

Throughout the manuscript and equations, we unified the notation to DAQ(SS), DAQ(AA), DAQ(C□).

-- Ref(ss): I assume this is the average of the probability from Emap2sec+, not from SPOT1D. If so, why not have the denominator written as SUM_j(P_ss(j)/N)) or something like that (where j iterates over all atom positions)?

Yes, your understanding is correct. Thank you for the suggestion. We revised the equations as suggested.

-- on the topic of this denominator - I don't understand why you refer to it as "reference". For example, in the text when you say "if the predicted probability values for the local position are higher than the reference", why not just "higher than average"?

Thank you. We rephrased it to “average”.

Confusing language:
- abstract:
-- "identified via deep learning" - is it the likelihood that is identified? I don't think a likelihood can be identified, can it? It can be computed, estimated, etc…

We changed it to “estimated”.

-- "reconstructed model" - what is this? I have heard of map 3D reconstruction We changed it to “atomic protein structure model”.

-- "how well the amino acids (...) agree with that likelihood" - amino acids cannot agree to anything; did you mean something like "the a.a. assignment is consistent with..."?

We added “assignment” and “is consistent with”.

-- "scores that correlate the local density grandient in the map" - that just doesn't make sense. They correlate it with what? What kind of gradient are you talking about? Gradient of what with respect to what?

We rephrased it to “that check the fit of a model structure with the density in the map”.

- results:
-- "indicating that Emap2sec+ detected each structural feature specific density features and distinguished them well from background." - This sentence looks mangled to me. I can't work out what you are trying to say.

We rewrote the phrase into a separate sentence. It reads “As these scores are log odds scores for an assigned secondary structure, amino acid, or Ca atom relative to background (Eqs. 1, 2, 3), the positive scores indicate that the deep learning saw the fit of the correct amino acids at their positions in the maps.“

- results:
-- "For both metrics" - I think "For each metric" would be better here We changed it to “For each metric”.
-- "two types of values are shown, one that averaged over values computed first for each PDB entry separately and the other that considered all the PDB entries together." This sentence threw me the first 4 or 5 times I tried to read it. Could the authors reformulate to make it clearer? I think I understood it eventually, but it's way too much work.

I agree that it needs more explanation. In page 34 in Methods, we added the following paragraph to explain the two types. The corresponding part in page 17 was also largely rewritten.

(page 34): Table 1, we computed two types of AUC-ROC, “Average AUC-ROC” and “AUC- ROC All”. The average AUC-ROC is the average of AUC-ROC computed for the 35 models. Thus, for each model, AUC-ROC was separately computed and then averaged over models. In contrast, the latter “AUC-ROC All” considered all the residues in the 35 models altogether, sorted them by the DAQ score and computed an AUC-ROC value once. Although they are similar, these two approaches have notable differences. The average AUC-ROC will be high

if a score is able to sort residues within each model by their quality and if all the models have a similar fraction of inconsistent residues. On the other hand, as AUC-ROC All sort all the residues from all the models, it tends to examine if absolute values of the score (but not relative within each model) can tell inconsistent residues or not.

-- The section beginning "Fig 4C shows for examples" and ending "In this region, 6ZR3 clearly a clearly better DAQ(AA)-score than 6L54" was much better written than the results thus far. Despite the clear editing error in that last sentence ("clearly a clearly")

- I thought methods were clearly written also, they were easy to follow We removed one “clearly”.

Suggestions for topics to cover in the discussion:

- I'm curious - how would the DAQ score be affected by DL-based methods that aim to "clean up" / make maps look better (or more like your typical map). Now that the machines are learning what good maps look like, could one machine (e.g. DeepEMhancer, Sanchez-Garcia etal, Comms Biol 4:874(2021)) be used to fool another (DAQ)? What if a tool was trained like Emap2sec+ but instead of outputing a score it would output idealized version of each local map feature? In general I'm curious about the impact DL-based methods used during map polishing could have on DL-based methods used for validation. I suspect the authors' expert opinions would be of value.

Thank you for asking, it is an interesting question.
DAQ reads local density values of the input cryo-EM map and detects specific features for amino acids, secondary structures and atoms, and compare them with the structure model. DAQ does not ask how the map has been generated. DAQ just compares the local map density and the model structure and examine if they are compatible. Thus, if the map has been modified by DeepEMhancer or other map sharpening tool, DAQ still simply checks the compatibility of the model with the modified map. If the modified map still looks “real” (i.e. still have density distribution which is similar to usual experimental maps) and fits well to the modelled amino acid, a high DAQ score will be assigned. On the other hand, if the modified map looks unusual for an experimental map, probably most modelled amino acids in the map will have a low DAQ score because such correspondence has not been seen in EM maps and their modelled protein structures in EMDB and PDB. Thus, preprocessing maps and structure model validation by DAQ are separate things. To clarify, we added the following sentence in Discussion (page 25): “Note that DAQ is aimed for comparing the local map density distribution and the model structure and examines if they are compatible regardless of how the map is generated or if the map is modified by a map sharpening tool or not.”

- In the results: "Particularly, the regions Val661-Gln709 and Leu758-Lys778 have incorrect assignments as evidenced by implausible steric collision between backbone atoms. The density for these regions also does not support the PDB model."

I took a quick look. This is a low-resolution part of that cryoEM map. I am pretty sure other validation metrics, and the depositing authors, would have picked up on the fact that this was likely not an accurate region of the model.

I agree. We selected this example because we thought it is an obvious mis modeling, and readers would agree to our evaluation. But we now added two more examples in Fig 5, Fig 5c and 5d, which are models with a higher resolution (2.41 A and 2.7 A).

I think this brings up the point that when selecting subsets of the PDB, the authors are using overall resolution value for each entry, but variations in local resolution will mean that many regions of maps and models from this dataset will be expected to have pretty bad models.

We agree that low local resolution is a problem. As we show in Fig. 5 and discussed in page 24, there are potentially many bad models in the current PDB.

I was hoping that the authors would spell this out in the manuscript, perhaps in the discussion: that there will be times when despite best efforts by the model builder, locally a map doesn't have enough SNR to allow reliable model building. In those circumstances, low DAQ scores will be expected but not necessarily actionable.

In principle we agree with the reviewer. We added the following sentences to clarify the point in Discussion (page 25):
As shown in examples, generally, there are two cases where DAQ score is low: Obviously, DAQ score will be low, may be a large negative value, if a wrong amino acid is assigned to a position in a map of a high resolution, Also, DAQ score tend to be low when the local resolution of the map is low. In this case, usually DAQ tend to be close to 0, indicating that the fit of the residue at that position is not better than the average across all the positions in the map.

On a similar point, in the discussion the authors state "Many of the mistakes made in the above examples would likely have been discovered by well-trained structural biologists before their publication". Discovered, yes. But fixed? I doubt it in many cases, since the problems are probably in many cases due to poor local quality.

We agree that the phrase “Many of the mistakes …” is a bit strong, hence we deleted it.

I don't think DAQ (or any other validation metric) in itself solves the problem of depositing the correct model, or even an improved one. It flags the problems and then leaves it to the depositor to somehow

arrive at a better model, which may or may not be possible, depending on what the problem was due to in the first place.

We agree that validation tools do not solve the problem; it is not the tool that will automatically fix the problem in structure models. It would, however, help biologists to identify potential errors in a model and guide them to fix the problems, and allow endusers to assess whether or not the structural element they are looking at is reliable. Indeed, as we used as a dataset in this work, biologists often update their deposited models to PDB, making correction to the models. As we have shown, DAQ showed clear quality improvement in the updated model over the initial model for almost all the cases. DAQ and other validation tools can help that process. Model validation is very important also in crystal structures of biomolecules and as you know, currently each PDB entry has structure wwwPDB validation report, which indicates potential errors in a model. The validation metrics for cryo-EM, including DAQ, is basically for the same purpose. They tell inconsistencies between the density and the model, and suggest that researchers to check them before deposition or actioning experiments based on these regions.

To clarify that, we added the following sentence in discussion (page 25): Validation tools including DAQ, Q-score, EMRinger, do not fix the model itself, but help researchers to identify potential problems in a structure model and guide them to fix the problems.

- Would it be desirable / possible to remove from the training set regions of maps with bad local resolution?

We think it is an interesting idea, but it has advantages and disadvantages. An advantage of removing bad regions from maps is that deep learning will be trained on “cleaner”, higher resolution data only, which may be able to produce a neural network that performed better than the current one on high resolution, clean maps. On the other hand, a practical disadvantage of removing “bad local resolution regions” from the method development point of view is that we need to define “bad local resolution regions” of maps, which adds additional subjective choice and step (which users would not agree) in the development process. We need to decide how to (which methods or software to use) compute local resolution and decide the cutoff values and procedure to declare low resolution regions.
Another potential disadvantage is that using only high resolution regions in maps would make the dataset far from the reality, and resulting neural network may not be able to provide good results for locally low resolution regions in maps, because it has never seen such local density.
Therefore, we use the entire maps for training. Because maps with some low resolution regions are current reality.

Minor fixes:
- intro

-- "58% of the maps have a resolution", not "had" Thank you, we corrected it.
-- "scores are naturally affected by the resolution of the local density at the position of a residue", not "of"
We corrected it to “at”.

- results
-- "DAQ(AA)-scores in the lower-scored model centered at 0.0". According to Fig 4A, the residue scores from the lower scored models (red histogram) were NOT centered around 0.0. You may want to say "approximately" and get rid of the decimal place.

We put “approximately”.

Responses to Comments by Reviewer 3
Remarks to the Author:
The authors present a very interesting approach to CryoEM map and atomic model validation. Their approach is based on a neural network that extends part of their previous work. The method is technically sound and they apply it to many examples showing its validity. This reviewer recommends publication in the current form subject to some minor changes:

- Page 26. It currently says "we collected
voxels of a size of 113 Å3 by scanning across a map along three axes with a stride of". Probably "we collected BOXES ..." would be more appropriate.

We changed voxels to boxes as suggested throughout the manuscript.

- The number of parameters of the network should be reported. This is important to be able to compare it to the training dataset size.

We added the information in page 31. To summarize, the network has about 6.47 million parameters to train. The training dataset has in total 31.9 values to predict. Thus, the training dataset size is larger than the number of parameters.

Please also note that as described in page 29, we paid the best effort we could think to make the training set, the validation set, and the test set non redundant. Any pairs of cryo-EM maps from, for example, the training set and the test set, do not have even one protein pairs with over 25% sequence identity. Thus, cryo-EM maps in the validation set and the test set are not similar to any cryo-EM maps in the training set.

- It would be interesting to report in the supplementary material the test and validation loss curves during the training. In this way, the reader can figure out if there has been some overfitting or the network is appropriately learning.

We provided the training and validation curve as Supplementary Figure 11. It is mentioned in page 31. There is no trace of overfitting as both training and validation curve went down as the epochs progressed.

Decision Letter, second revision:

Dear Daisuke,

Your Article, "Residue-Wise Local Quality Estimation for Protein Models from Cryo-EM Maps", has now been seen again by two reviewers. Although they find the work greatly improved, referee 1 still has concerns. We are interested in the possibility of publishing your paper in Nature Methods, but would like to consider your response to these concerns before we reach a final decision on publication.

We therefore invite you to revise your manuscript to address these concerns, including new experiments, if necessary.

We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

When revising your paper:

* include a point-by-point response to the reviewers and to any editorial suggestions

* please underline/highlight any additions to the text or areas with other significant changes to facilitate review of the revised manuscript

* address the points listed described below to conform to our open science requirements

* ensure it complies with our general format requirements as set out in our guide to authors at www.nature.com/naturemethods

* resubmit all the necessary files electronically by using the link below to access your home page

[Redacted] This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

We hope to receive your revised paper within six weeks. If you cannot send it within this time, please let us know. In this event, we will still be happy to reconsider your paper at a later date so long as nothing similar has been accepted for publication at Nature Methods or published elsewhere.

OPEN SCIENCE REQUIREMENTS

REPORTING SUMMARY AND EDITORIAL POLICY CHECKLISTS
When revising your manuscript, please update your reporting summary and editorial policy checklists.

Reporting summary: https://www.nature.com/documents/nr-reporting-summary.zip
Editorial policy checklist: https://www.nature.com/documents/nr-editorial-policy-checklist.zip

If your paper includes custom software, we also ask you to complete a supplemental reporting summary.

Software supplement: https://www.nature.com/documents/nr-software-policy.pdf

Please submit these with your revised manuscript. They will be available to reviewers to aid in their evaluation if the paper is re-reviewed. If you have any questions about the checklist, please see http://www.nature.com/authors/policies/availability.html or contact me.

Please note that these forms are dynamic 'smart pdfs' and must therefore be downloaded and completed in Adobe Reader. We will then flatten them for ease of use by the reviewers. If you would like to reference the guidance text as you complete the template, please access these flattened versions at http://www.nature.com/authors/policies/availability.html.

DATA AVAILABILITY
We strongly encourage you to deposit all new data associated with the paper in a persistent repository where they can be freely and enduringly accessed. We recommend submitting the data to discipline-

specific and community-recognized repositories; a list of repositories is provided here: http://www.nature.com/sdata/policies/repositories

All novel DNA and RNA sequencing data, protein sequences, genetic polymorphisms, linked genotype and phenotype data, gene expression data, macromolecular structures, and proteomics data must be deposited in a publicly accessible database, and accession codes and associated hyperlinks must be provided in the "Data Availability" section.

Refer to our data policies here: https://www.nature.com/nature-research/editorial-policies/reporting-standards#availability-of-data

To further increase transparency, we encourage you to provide, in tabular form, the data underlying the graphical representations used in your figures. This is in addition to our data-deposition policy for specific types of experiments and large datasets. For readers, the source data will be made accessible directly from the figure legend. Spreadsheets can be submitted in .xls, .xlsx or .csv formats. Only one (1) file per figure is permitted: thus if there is a multi-paneled figure the source data for each panel should be clearly labeled in the csv/Excel file; alternately the data for a figure can be included in multiple, clearly labeled sheets in an Excel file. File sizes of up to 30 MB are permitted. When submitting source data files with your manuscript please select the Source Data file type and use the Title field in the File Description tab to indicate which figure the source data pertains to.

Please include a "Data availability" subsection in the Online Methods. This section should inform readers about the availability of the data used to support the conclusions of your study, including accession codes to public repositories, references to source data that may be published alongside the paper, unique identifiers such as URLs to data repository entries, or data set DOIs, and any other statement about data availability. At a minimum, you should include the following statement: "The data that support the findings of this study are available from the corresponding author upon request", describing which data is available upon request and mentioning any restrictions on availability. If DOIs are provided, please include these in the Reference list (authors, title, publisher (repository name), identifier, year). For more guidance on how to write this section please see: http://www.nature.com/authors/policies/data/data-availability-statements-data-citations.pdf

CODE AVAILABILITY

Please include a "Code Availability" subsection in the Online Methods which details how your custom code is made available. Only in rare cases (where code is not central to the main conclusions of the paper) is the statement "available upon request" allowed (and reasons should be specified).

We request that you deposit code in a DOI-minting repository such as Zenodo, Gigantum or Code Ocean and cite the DOI in the Reference list. We also request that you use code versioning and provide a license.

For more information on our code sharing policy and requirements, please see: https://www.nature.com/nature-research/editorial-policies/reporting-standards#availability-of-computer-code

MATERIALS AVAILABILITY

As a condition of publication in Nature Methods, authors are required to make unique materials promptly available to others without undue qualifications.

Authors reporting new chemical compounds must provide chemical structure, synthesis and characterization details. Authors reporting mutant strains and cell lines are strongly encouraged to use established public repositories.

More details about our materials availability policy can be found at https://www.nature.com/nature-portfolio/editorial-policies/reporting-standards#availability-of-materials

ORCID

Nature Methods is committed to improving transparency in authorship. As part of our efforts in this direction, we are now requesting that all authors identified as 'corresponding author' on published papers create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System (MTS), prior to acceptance. This applies to primary research papers only. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the MTS by clicking on 'Modify my Springer Nature account'. For more information please visit please visit www.springernature.com/orcid.

Please do not hesitate to contact me if you have any questions or would like to discuss these revisions further. We look forward to seeing the revised manuscript and thank you for the opportunity to consider your work.

Sincerely,

Rita

Rita Strack, Ph.D.
Senior Editor
Nature Methods

Reviewers' Comments:

Reviewer #1:
Remarks to the Author:

The authors have addressed my initial questions and comments adequately, which helped to clarify the method and results, also importantly the differences between the 3 score types. The new supplementary figures and tables are very good for further explaining the methods and results.

A few more questions and comments:

In regards to the use of a 11^3 Å^3 box centered at the grid point/Calpha position for example. There is a potential issue here that an Arg residue can measure 6.7Å from the Calpha to the terminal atom (NH1), and Lys 6.2Å from Calpha to NZ. In such a box, anything beyond 5.5Å (half the box side which extends form the Calpha position) would be ignored, and hence the full side chain may not be considered. Since Arg,Lys side chains are longer than other side chains, this may not be an issue, but for example the method may have trouble distinguishing between Lys and Arg. Trp can also extend 6.6Å between Calpha and atom CH2. In a 5.5Å box it may look a lot like a His.

• Page 7: Eq.2 “The density of probability is normalized by the average probability of amino acid type aa(i) averaged over all the atom positions in the protein model.”

So a residue has a high DAQ(AA) if the probability is higher at the given location vs. other locations in the same map/model? If all locations have high probabilities, then the DAQ(AA) would be close to 1. At the same time, if all locations have low probabilities, then DAQ(AA) would also (perhaps misleadingly) be close to 1? So does the score depend on the map/model having the majority of a given residue type well-defined in the map to work properly? This may be unlikely but it could be an issue in a small number of cases, e.g. in cases where the entire model is fitted poorly, or for residues that are not as common, e.g. Met.

• Page 7: “Note that these accuracies do not need to be perfect for model validation so long as the correct residue is preferred”.

‘so long as the correct residue is preferred’ seems to imply that the accuracy is important and that the residue should be preferred over other types of residues. In many cases the correct residue does not seem to be preferred: 100 – 28.2% or the accuracy for DAQ(aa). Perhaps what is meant is more what was given in a response that “as long as the correct residue has a positive score, the validation will work.” Although the term 'validation' may be too strong, given it is not ensured that the correct rescue is chosen, just that it has a positive score.

• Page 8: “The average DAQ(AA) score are not much influenced by the map resolution up to around 3 Å.”

Interesting to see a strong correlation to resolution. I wonder if it tapers off at 2-3Å because the training data was only up to 2.5Å?

• Page 15. “Comparison of DAQ(AA) with three other scores, Q-score, EMRinger (map-model scores), and CaBLAM (a model-coordinate score) demonstrates that DAQ(AA) captured all four inconsistent regions (the right panels in Fig. 3c) while the other scores only consistently detected region (i), which is mispositioned. ”

I can see that EMRinger and CaBLAM do not show dips in all regions, but it does look like Q-score shows dips (below the 0.0 line or expected Q-score), in all regions, even if they are lower in magnitude; so I’m not sure this statement is completely right. It may be noted too that CaBLAM is only based on model geometry and not map-model, but still interesting to compare.

• Page 15. “Q-score did show local minima in regions (ii) and (iv) with less magnitude than DAQ(AA) score.”

In the Fig. 2c, Q-score also shows significant dips below the 0-line in regions (i) and (iii) as well, so this statement is somewhat misleading. It could be a lot more clear and useful to compare the scores by normalizing (subtract minimum, divide by max-min) all of them and plotting them together on the same axes. For Q-score, it would be useful to include the 0-line (as in the current graph), or expected Q-score for reference as this is standard to do.

• Page 34 AUC-ROC, Supp Fig. 12

Great to see this explanation of AUC-ROC and how it is calculated in this case. Does the 'true positive' mean that the score for a residue is simply positive, not that it is higher for a given residue type rather than for other residue types? I suspect yes, since the latter is a harder task as mentioned above.

Review by Greg Pintilie

Reviewer #3:
None

Author Rebuttal, second revision:

Response to Comments by Reviewer #1:
Remarks to the Author:

The authors have addressed my initial questions and comments adequately, which helped to clarify the method and results, also importantly the differences between the 3 score types. The new supplementary figures and tables are very good for further explaining the methods and results.

A few more questions and comments:

In regards to the use of a 11^3 Å^3 box centered at the grid point/Calpha position for example. There is a potential issue here that an Arg residue can measure 6.7Å from the Calpha to the terminal atom (NH1), and Lys 6.2Å from Calpha to NZ. In such a box, anything beyond 5.5Å (half the box side which extends form the Calpha position) would be ignored, and hence the full side chain may not be considered. Since Arg,Lys side chains are longer than other side chains, this may not be an issue, but for example the method may have trouble distinguishing between Lys and Arg. Trp can also extend 6.6Å between Calpha and atom CH2. In a 5.5Å box it may look a lot like a His.

Thank you, it is an interesting point. Among 190 (=20x19/2) different amino acid pair combinations, Lys – Arg pair is among relatively frequently confused pairs (the 9th among the 190 pairs) but it is not the most confused pair. Other pairs, including Trp – Tyr, combinations among Ile, Val, Leu, and the Asp – Pro pair were more confused between them, for example. I agree that the box size is too small to fit fully-extended Lys and Arg if it is placed along the shortest distance in the box, and it could be one reason of

the confusion. But it is not so simple because using a larger box may include density from other neighboring residues, which will act as noise and would further confuse network, deteriorating the overall performance. We used 11^3 Å for the box in this work because it performed well in our previous similar works (e.g. Wang et al., Nature Communications, 12: 2302, https://pubmed.ncbi.nlm.nih.gov/33863902/). At the time of the development of the previous work, we tried a couple of different box size and concluded that 11^3 Å was the best among tested.

Nevertheless, I agree with the reviewer that the box size is one of the important parameters in the network. Besides the box size, using a different architecture of network, which may even not use a fixed boxed size, e.g. using graph representation etc, may further improve the performance. We added a new paragraph to discuss it in Discussion (page 27).

- Page 7: Eq.2 "The density of probability is normalized by the average probability of amino acid type aa(i) averaged over all the atom positions in the protein model."

So a residue has a high DAQ(AA) if the probability is higher at the given location vs. other locations in the same map/model? If all locations have high probabilities, then the DAQ(AA) would be close to 1. At the same time, if all locations have low probabilities, then DAQ(AA) would also (perhaps misleadingly) be close to 1? So does the score depend on the map/model having the majority of a given residue type well-defined in the map to work properly? This may be unlikely but it could be an issue in a small number of cases, e.g. in cases where the entire model is fitted poorly, or for residues that are not as common, e.g. Met.

Let me explain how DAQ(AA) score (Eq. 2) works with an example. DAQ(AA) consists of two parts in the logarithm, the numerator and the denominator. The numerator, Paa(i)(i), is the probability that the amino acid currently assigned at position i, is correct. For example, if Val is assigned at position i, the numerator is PVal(i), which is the probability that Val is at position i. The neural network has learned this probability from the training dataset. If Val is the correct assignment for position i, PVal(i) will be a high value. The denominator, is the average PVal(j) across all the atom positions j in the protein models in the map, i.e. the average probability of Val across all the positions.
Since for most of the cases protein(s) in a map has many amino acids other than Val, the numerator will usually be a smaller value. Thus, overall, the numerator is larger than the denominator if the assignment by the author is correct, which makes the DAQ(AA) a positive value after taking logarithm.
On the other hand, if an amino acid assignment is wrong at a certain position,
e.g. if Ile is assigned at position i, which is wrong, the numerator will clearly be a low value. And the denominator will be a higher value than the numerator usually, partly

because Ile is the correct assignment in a fraction of the places in protein(s) in the map, which contribute in making the average score of Ile larger than the numerator.

If a local density does not have a distinct pattern at that position for the assigned amino acid, DAQ(AA) score (Eq. 2) will be close to 0 (not 1), because the numerator and the denominator will be a similar value, and log(numerator/denominator) ~= log(1) = 0. (i.e. DAQ does not show clear support nor disagreement).
"If the entire model is fitted poorly": This is not a problem. If all the residue assignment is wrong, simply all the residues in the model will have a negative DAQ score, indicating that each residue in the model may be incorrectly assigned.
"For residues that are not as common, e.g. Met".: This is not a problem, particularly in the case when Met is the correct assignment for the position. If Met is the correct assignment for a position i in a map, in Equation (2), the numerator within logarithm, i.e. Paa(i)(i), which is the probability of Met at the position i, will be a high value. In contrast, for the denominator, when Met = Paa(i) is assigned to all the positions (position j's) in the proteins in the map, it will be a small value for most of j's because the correct amino acid for most of the position is not Met. Thus, DAQ(Met)(i) will be a high positive value, after taking log.
Now, in a rare case that the Met for position i is a wrong assignment and if there is only one Met in proteins in the map, the DAQ score of Met at i can be close to 0: Here is why it could happen. If Met at position i is a wrong assignment, the numerator, PMet(i) will be a small value. Then, the denominator, which is the average PMet(j) for all positions j in the proteins in the map, can be also a small value because none of the position j has Met as its correct assignment. Thus, log(numerator/denominator) may be close to 0 unless the incompatibility of Met at i (the numerator) is not much worse than the average. The result will also depend on how small the numerator, PMet(i) is (i.e. how incompatible Met is for the position i.)
I would also like to mention that the averaging window we use is to alleviate this scenario, because the window will highlight a local region that has consistently negative DAQ scores in many amino acids in the region.

To reflect this discussion to users, in the github and the Google Colab server we added a description of the DAQ score, and added above two examples in QA page.

They are at https://github.com/kiharalab/DAQ#overview-of-daq-score and https://github.com/kiharalab/DAQ/blob/main/QA.md at Github;
and at the Google Colab page: https://colab.research.google.com/drive/1Q-Dj42QjVO8TCOLXMQBJlvm1zInxPkOu#scrollTo=MdPRqudcZXu-&line=5&uniqifier=1

- Page 7: "Note that these accuracies do not need to be perfect for model validation so long as the correct residue is preferred".

'so long as the correct residue is preferred' seems to imply that the accuracy is important and that the residue should be preferred over other types of residues. In many cases the correct residue does not

seem to be preferred: 100 – 28.2% or the accuracy for DAQ(aa). Perhaps what is meant is more what was given in a response that “as long as the correct residue has a positive score, the validation will work.” Although the term 'validation' may be too strong, given it is not ensured that the correct rescue is chosen, just that it has a positive score.

The original sentence was “Note that these accuracies do not need to be perfect for model validation so long as the correct residue is preferred, i.e. a positive DAQ score is computed, for the position in the map.”
We now revised it to: “Note that these accuracies do not need to be perfect for model validation so long as the correct residue has a positive DAQ score for the position in the map.”

- Page 8: “The average DAQ(AA) score are not much influenced by the map resolution up to around 3 Å.”

Interesting to see a strong correlation to resolution. I wonder if it tapers off at 2-3Å because the training data was only up to 2.5Å?

Yes, I agree that it is a possible reason.

- Page 15. “Comparison of DAQ(AA) with three other scores, Q-score, EMRinger (map- model scores), and CaBLAM (a model-coordinate score) demonstrates that DAQ(AA)

captured all four inconsistent regions (the right panels in Fig. 3c) while the other scores only consistently detected region (i), which is mispositioned. ”
I can see that EMRinger and CaBLAM do not show dips in all regions, but (map-model scores), and CaBLAM i), which is mispositioned. ”
I can see that EMRinger and CaBLAM do not show dips in all regions, but it does look like Q-score shows dips (below the 0.0 line or expected Q-score), in all regions, even if they are lower in magnitude; so I’m not sure this statement is completely right. It may be noted too that CaBLAM is only based on model geometry and not map-model, but still interesting to compare.

- Page 15. “Q-score did show local minima in regions (ii) and (iv) with less magnitude than DAQ(AA) score.”

In the Fig. 2c, Q-score also shows significant dips below the 0-line in regions (i) and (iii) as well, so this statement is somewhat misleading. It could be a lot more clear and useful to compare the scores by normalizing (subtract minimum, divide by max-min) all of them and plotting them together on the same axes. For Q-score, it would be useful to include the 0-line (as in the current graph), or expected Q-score for reference as this is standard to do.

We now revised the Figure 3c, right panel, specifically, the Q-score and the CaBLAM panels. In the Q-score panel, as suggested, we put a baseline of the expected score for this map. Similarly, in the CaBLAM plot, we put two lines, 1% outlier cutoff and 5% disfavored cutoff, as the author discussed in their paper. We did not draw a line for the EMRinger as we did not find a recommended base line in their paper. For DAQ-score, the cutoff value is essentially 0.0, as described throughout the manuscript.
We agree that the sentence "Q-score did show local minima in regions (ii) and (iv) with less magnitude than DAQ(AA) score." is indeed misleading. Actually, the sentence was intended to mention that Q-score obviously identified region (i) and (iii), and (ii) and (iv) with a less magnitude, but somehow we dropped the first part by mistake.
In the revised manuscript, we now simply refer the panels of the other three scores, Q-score, CaBLAM, EMRinger as follows: "The right three panels in Fig. 3c show three other scores, Q-score, EMRinger (map-model scores), and CaBLAM (a

model-coordinate score), on the same target for reference." And removed sentences that mention our observation of individual scores.

- Page 34 AUC-ROC, Supp Fig. 12

Great to see this explanation of AUC-ROC and how it is calculated in this case. Does the 'true positive' mean that the score for a residue is simply positive, not that it is higher for a given residue type rather than for other residue types? I suspect yes, since the latter is a harder task as mentioned above.

In general for AUC-ROC, true positive is a data point that is correctly identified below (or above) the score cutoff value used. In our particular problem, true positive' means an inconsistent residue (which we want to detect) that are correctly detected by the cutoff value shown in the x-axis. I added a sentence to clarify it in page 34.
Review by Greg Pintilie

Response to Comments by Reviewer #3:
None

Thank you.

Decision Letter, third revision:

1st Jun 2022

Dear Daisuke,

Thank you for submitting your revised manuscript "Residue-Wise Local Quality Estimation for Protein Models from Cryo-EM Maps" (NMETH-A47664C). It has now been seen by the original referees and their comments are below. The reviewers find that the paper has improved in revision, and therefore we'll be happy in principle to publish it in Nature Methods, pending minor revisions to satisfy the referees' final requests and to comply with our editorial and formatting guidelines.

We are now performing detailed checks on your paper and will send you a checklist detailing our editorial and formatting requirements in about a week. Please do not upload the final materials and make any revisions until you receive this additional information from us.

TRANSPARENT PEER REVIEW
Nature Methods offers a transparent peer review option for new original research manuscripts submitted from 17th February 2021. We encourage increased transparency in peer review by publishing the reviewer comments, author rebuttal letters and editorial decision letters if the authors agree. Such peer review material is made available as a supplementary peer review file. Please state in the cover letter 'I wish to participate in transparent peer review' if you want to opt in, or 'I do not wish to participate in transparent peer review' if you don't. Failure to state your preference will result in delays in accepting your manuscript for publication.
Please note: we allow redactions to authors' rebuttal and reviewer comments in the interest of confidentiality. If you are concerned about the release of confidential data, please let us know specifically what information you would like to have removed. Please note that we cannot incorporate redactions for any other reasons. Reviewer names will be published in the peer review files if the reviewer signed the comments to authors, or if reviewers explicitly agree to release their name. For more information, please refer to our FAQ page.

Thank you again for your interest in Nature Methods Please do not hesitate to contact me if you have any questions.

Sincerely,
Rita

Rita Strack, Ph.D.
Senior Editor
Nature Methods

ORCID
IMPORTANT: Non-corresponding authors do not have to link their ORCIDs but are encouraged to do so. Please note that it will not be possible to add/modify ORCIDs at proof. Thus, please let your co-authors know that if they wish to have their ORCID added to the paper they must follow the procedure described in the following link prior to acceptance:
https://www.springernature.com/gp/researchers/orcid/orcid-for-nature-research

Reviewer #1 (Remarks to the Author):

I want to thank the authors again for taking the time to work on this important issue of map-model metrics and to answer my questions.

First point, ending with response on Page 2:
“Besides the box size, using a different architecture of network, which may even not use a fixed boxed size, e.g. using graph representation etc, may further improve the performance. We added a new paragraph to discuss it in Discussion (page 27).”

Thank you for this detailed break-down, it does make sense that the network may work better with this box size as explained, and perhaps the box does not need to encapsulate the entire residues (for the larger ones like TRP, ARG) for the network to be able to give them a positive score. I see the brief paragraph added on page 27, but it is a bit vague and does not fully address this issue. Regardless, it is a bit of a minor technical point perhaps. Being more clear here would however help others understand the method better and any future efforts in this direction.

To response starting on Page 2, for my comment
Page 7: Eq.2 “The density of probability is normalized by the average probability of amino acid type aa(i) averaged over all the atom positions in the protein model.”

I appreciate that detailed explanation regarding Eq. 2. The sentence above which is still the same in the text is vague, and I’m not sure what ‘density of probability’ refers to. The variable j should be explained, i.e. that it is every residue position regardless of type (not every atom position). It is great the authors have added the detailed explanation to the external QA section but if it can be clarified here too with a sentence or two I’m sure it would be very useful.

Further I am ok with the rest of the responses as they addressed my questions and suggestions.

Finally, I was able to run the method through the Google Colab, for which the tutorial was very useful. So I could replicate the plots and comparisons to Q-scores in Figure 3c. It is quite impressive the method is able to detect the misaligned residues in region (iii) in particular.

I would like to congratulate the authors on a very nice method and implementation.

Review by Greg Pintili

Author Rebuttal, third revision:

Responses to Comments by Reviewer #1
Reviewer #1:
Remarks to the Author:

I want to thank the authors again for taking the time to work on this important issue of map-model metrics and to answer my questions.

First point, ending with response on Page 2:
"Besides the box size, using a different architecture of network, which may even not use a fixed boxed size, e.g. using graph representation etc, may further improve the performance. We added a new paragraph to discuss it in Discussion (page 27)."

Thank you for this detailed break-down, it does make sense that the network may work better with this box size as explained, and perhaps the box does not need to encapsulate the entire residues (for the larger ones like TRP, ARG) for the network to be able to give them a positive score. I see the brief paragraph added on page 27, but it is a bit vague and does not fully address this issue. Regardless, it is a bit of a minor technical point perhaps. Being more clear here would however help others understand the method better and any future efforts in this direction.

In page 17, we added following phrases: "The current scanning box size is adopted from our previous work. It may be worthwhile to optimize the box size specifically for DAQ…"

To response starting on Page 2, for my comment
Page 7: Eq.2 "The density of probability is normalized by the average probability of amino acid type aa(i) averaged over all the atom positions in the protein model."

I appreciate that detailed explanation regarding Eq. 2. The sentence above which is still the same in the text is vague, and I'm not sure what 'density of probability' refers to. The variable j should be explained,

i.e. that it is every residue position regardless of type (not every atom position). It is great the authors have added the detailed explanation to the external QA section but if it can be clarified here too with a sentence or two I'm sure it would be very useful.

I agree that "the density of probability" is a strange expression. We fixed it to "probability". We also now added explicit explanation of the variable j in all three equations, Equations 1, 2, and 3.

Further I am ok with the rest of the responses as they addressed my questions and suggestions.

Finally, I was able to run the method through the Google Colab, for which the tutorial was very useful. So I could replicate the plots and comparisons to Q-scores in Figure 3c. It is quite impressive the method is able to detect the misaligned residues in region (iii) in particular.

I would like to congratulate the authors on a very nice method and implementation.

Review by Greg Pintilie

Thank you very much for all your constructive comments.

Final Decision Letter:

11th Jul 2022

Dear Daisuke,

I am pleased to inform you that your Article, "Residue-Wise Local Quality Estimation for Protein Models from Cryo-EM Maps", has now been accepted for publication in Nature Methods. Your paper is tentatively scheduled for publication in our August or September print issue, and will be published online prior to that. The received and accepted dates will be Nov 20, 2021 and July 11, 2022. This note is intended to let you know what to expect from us over the next month or so, and to let you know where to address any further questions.

In approximately 10 business days you will receive an email with a link to choose the appropriate publishing options for your paper and our Author Services team will be in touch regarding any additional information that may be required.

You will not receive your proofs until the publishing agreement has been received through our system.

Your paper will now be copyedited to ensure that it conforms to Nature Methods style. Once proofs are generated, they will be sent to you electronically and you will be asked to send a corrected version within 24 hours. It is extremely important that you let us know now whether you will be difficult to contact over the next month. If this is the case, we ask that you send us the contact information (email, phone and fax) of someone who will be able to check the proofs and deal with any last-minute problems.

If, when you receive your proof, you cannot meet the deadline, please inform us at rjsproduction@springernature.com immediately.

If you have any questions about our publishing options, costs, Open Access requirements, or our legal forms, please contact ASJournals@springernature.com

Once your manuscript is typeset and you have completed the appropriate grant of rights, you will receive a link to your electronic proof via email with a request to make any corrections within 48 hours. If, when you receive your proof, you cannot meet this deadline, please inform us at rjsproduction@springernature.com immediately.

Once your paper has been scheduled for online publication, the Nature press office will be in touch to confirm the details.

Content is published online weekly on Mondays and Thursdays, and the embargo is set at 16:00 London time (GMT)/11:00 am US Eastern time (EST) on the day of publication. If you need to know the exact publication date or when the news embargo will be lifted, please contact our press office after you have submitted your proof corrections. Now is the time to inform your Public Relations or Press Office about your paper, as they might be interested in promoting its publication. This will allow them time to prepare an accurate and satisfactory press release. Include your manuscript tracking number NMETH-A47664D and the name of the journal, which they will need when they contact our office.

About one week before your paper is published online, we shall be distributing a press release to news organizations worldwide, which may include details of your work. We are happy for your institution or funding agency to prepare its own press release, but it must mention the embargo date and Nature Methods. Our Press Office will contact you closer to the time of publication, but if you or your Press Office have any inquiries in the meantime, please contact press@nature.com.

If you are active on Twitter, please e-mail me your and your coauthors' Twitter handles so that we may tag you when the paper is published.

Please note that *Nature Methods* is a Transformative Journal (TJ). Authors may publish their research

with us through the traditional subscription access route or make their paper immediately open access through payment of an article-processing charge (APC). Authors will not be required to make a final decision about access to their article until it has been accepted. Find out more about Transformative Journals

Authors may need to take specific actions to achieve compliance with funder and institutional open access mandates. If your research is supported by a funder that requires immediate open access (e.g. according to Plan S principles) then you should select the gold OA route, and we will direct you to the compliant route where possible. For authors selecting the subscription publication route, the journal's standard licensing terms will need to be accepted, including self-archiving policies. Those licensing terms will supersede any other terms that the author or any third party may assert apply to any version of the manuscript.

If you have posted a preprint on any preprint server, please ensure that the preprint details are updated with a publication reference, including the DOI and a URL to the published version of the article on the journal website.

To assist our authors in disseminating their research to the broader community, our SharedIt initiative provides you with a unique shareable link that will allow anyone (with or without a subscription) to read the published article. Recipients of the link with a subscription will also be able to download and print the PDF. As soon as your article is published, you will receive an automated email with your shareable link.

Please note that you and your coauthors may order reprints and single copies of the issue containing your article through Nature Research Group's reprint website, which is located at http://www.nature.com/reprints/author-reprints.html. If there are any questions about reprints please send an email to author-reprints@nature.com and someone will assist you.

Please feel free to contact me if you have questions about any of these points.

Best regards,
Rita