nature portfolio

https://doi.org/10.1038/s41592-022-01574-4

Supplementary information

Residue-wise local quality estimation for protein models from cryo-EM maps

In the format provided by the authors and unedited

Supplementary Information for

Residue-Wise Local Quality Estimation for Protein Models from Cryo-EM Maps

Genki Terashi[1,†], Xiao Wang[2,†], Sai Raghavendra Maddhuri Venkata Subramaniya[2], John J. G. Tesmer[1,3], and Daisuke Kihara[1,2,3*]

[1] Department of Biological Sciences, Purdue University, West Lafayette, Indiana, 47907, USA

[2] Department of Computer Science, Purdue University, West Lafayette, Indiana, 47907, USA

[3] Department of Medicinal Chemistry and Molecular Pharmacology, Purdue University, West Lafayette, IN, 47907, USA

† Equal contribution

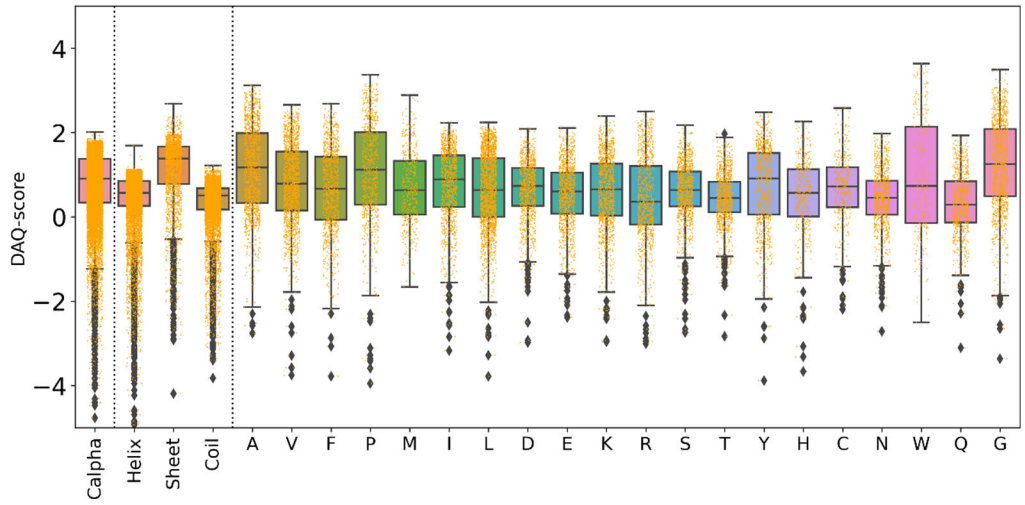
* Corresponding author:

E-mail: dkihara@purdue.edu

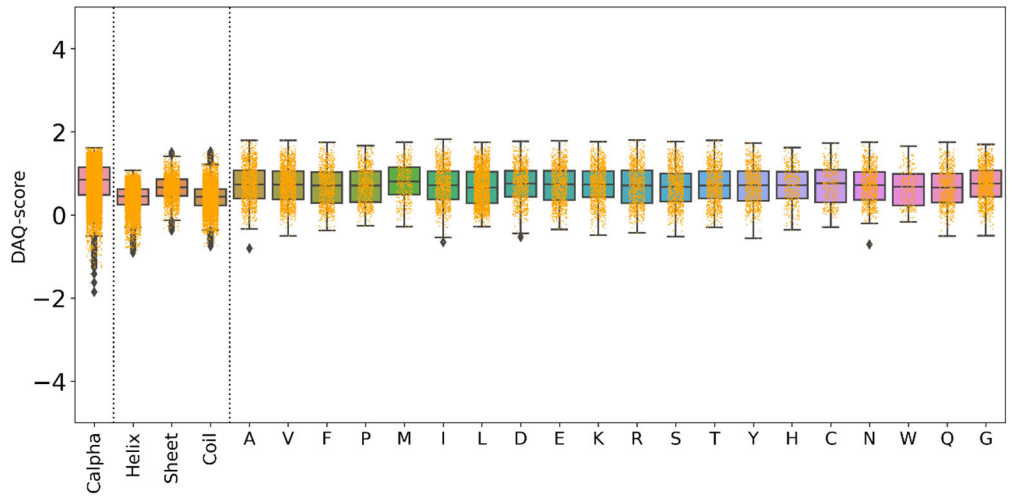
Supplementary Table 1. (separate Excel file)

Structure feature detection accuracy in the 40 cryo-EM maps in the testing set of Emap2sec+ (Methods). The third column, SS_acc, shows the accuracies of detecting secondary structure. The fourth column, AA_acc, shows the average accuracy of 20 amino acids for each map. The fifth column, Atom(Cα)_acc, shows the accuracy of Cα atom detection. The average accuracy is shown at the bottom of the table.

a. Individual residues (window size of 1).



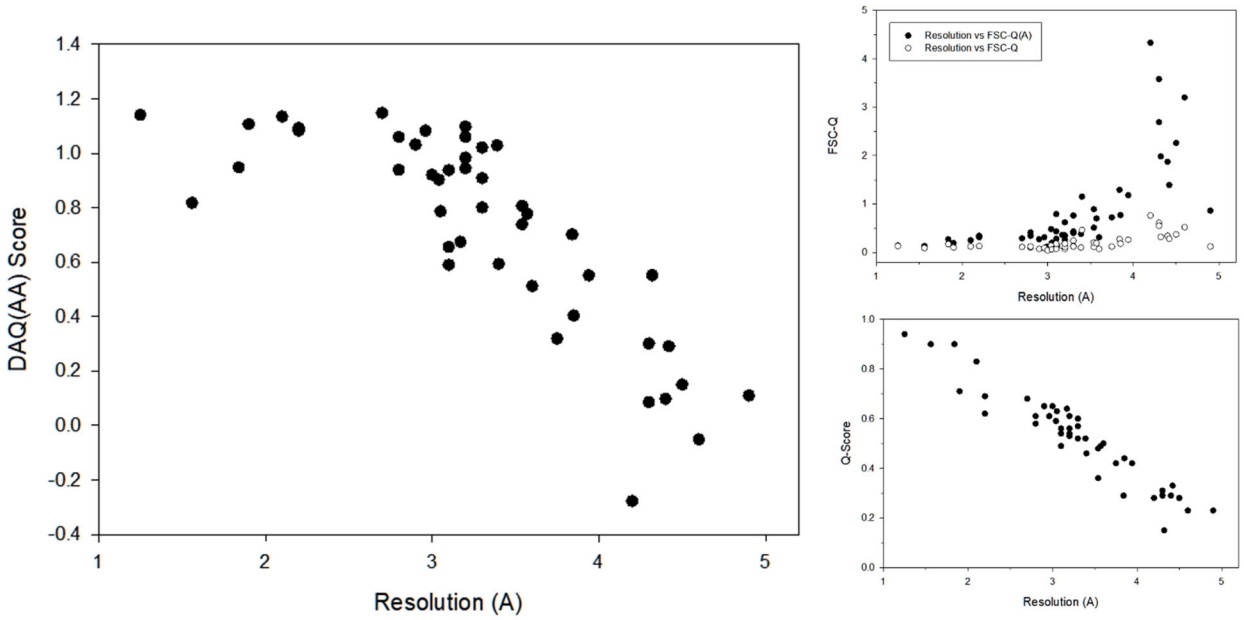
b. The average with a window size of 19.



Supplementary Fig 1. Distribution of DAQ(Cα), DAQS(SS), and DAQ(AA) of residues in the 35 revised PDB models of the PDB2Ver dataset used for comparison of mis-built structures in the main text (Figs. 2 and 3). We used the revised PDB models, because they should in theory have more correctly modeled residues than the originally deposited structures. The graph includes DAQ(Cα) of Cα atoms (n=18,670), DAQ(SS) of residues in helices (n=8,761), β strands (n=2,065), and others (n=7,844), and DAQS(AA) of 20 amino acids (n= 1,305, 1,493, 1,027, 758, 442, 1,049, 2,047, 966, 1,075, 1,127, 901, 1,052, 1,011, 823, 386, 324, 796, 243, 716 and 1,129 residues for amino-acid type A, V, F, P, M, I, L, D, E, K, R, S, T, Y, H, C, N, W, Q, and G, respectively). In the box plots, the middle line in a box represents the median, and the two ends of each box show quartiles, the upper and lower horizontal bars show the third quartile + 1.5*(the interquartile range) and the first quartile + 1.5*(the interquartile range), respectively. Outliers are shown as individual dots. **a**, scores of individual residues are shown. **b**, scores were averaged within a window of 19 residues centering each target residue. Supplementary Table 2 provides the fraction of residues that have positive scores.

Supplementary Table 2. (separate Excel file).

The fraction of residues that have a positive DAQ-score in the revised PDB models in the PDB2Ver dataset. This table corresponds to Supplementary Figure 1. A window of 19 resdiues was used. When individual residues were considered, for DAQ(SS), on average over the three secondary structures 84.6 (93.4)% of residues have positive values. 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.



Supplementary Fig. 2. Average DAQ(AA) score for maps determined at different resolutions. 50 maps 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) are used. For these maps, we applied a Gaussian filter of 0.9. There are two data points that have negative values. They are for EMD-10290 (resolution: 4.2 Å) and EMD-10294 (resolution: 4.6 Å). These two maps are determined at a relatively low resolution, and over 88.8% and 94.1% of residues, respectively, have no amino acid type assigned (In the PDB file, they are UNK). For comparison, FSC-Q and FSC-Q(A) scores as well as Q-score provided in the data table of Supplementary Fig. 6 of the paper by Ramirez-Aportela et al.[1] are also plotted.

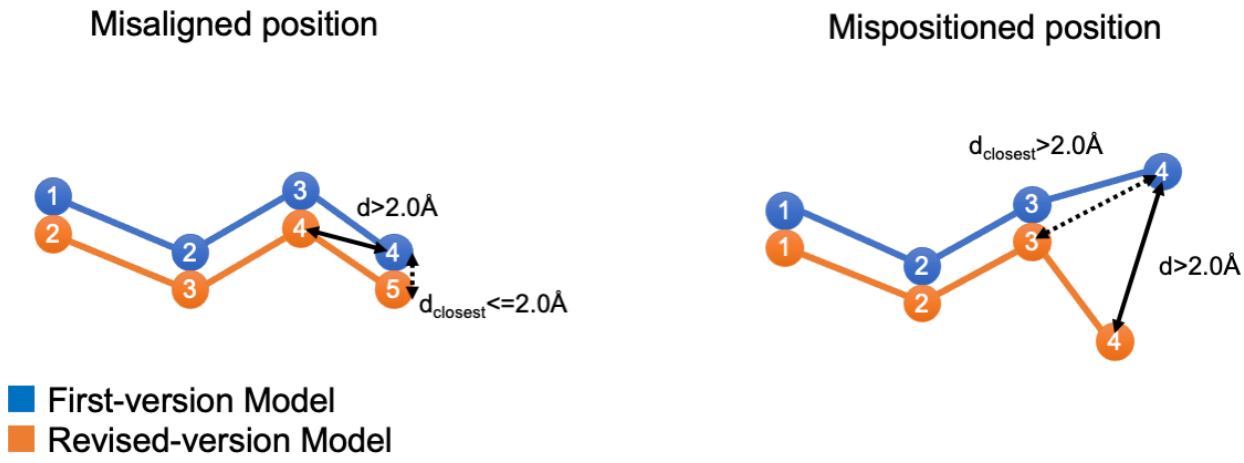
Supplementary Table 3. Examples of DAQ(AA) score of individual residues in EM maps of apoferritin at different resolutions.

EMID	Res (Å)	Ile-20	Asn-25	Leu-26	Tyr-29	Pro-127
20026	1.75	1.70	0.86	1.29	1.77	3.27
20027	2.3	1.62	1.85	1.23	1.87	3.16
20028	3.1	1.86	1.78	1.10	1.92	2.52
0140	3.9	0.91	0.00	0.23	1.09	2.34

Res, reported map resolution.

Supplementary Table 4. (separate Excel file)

15 EMDB entries and 35 protein chains in the PDB2Ver dataset. Sixth and seventh columns show the file names of the deposited models in the wwPDB database (ftp-versioned.wwpdb.org). Eighth column shows Cα–RMSD values between the first and revised versions of the deposited PDB models.



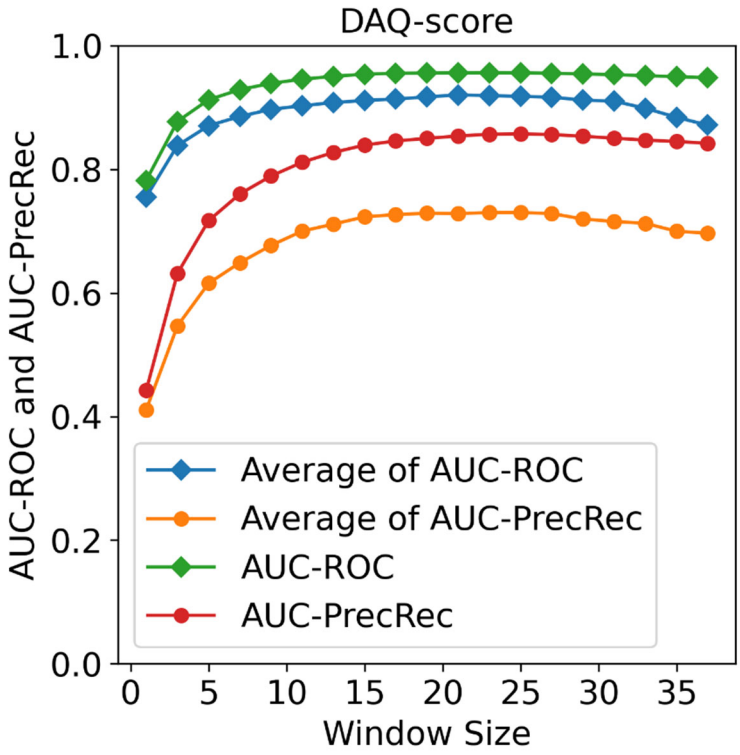
Supplementary Fig 3. Illustration of two types of inconsistent residues: misaligned and mispositioned. A misaligned residue (Cα atom) in the first-version model is the one that is within 2.0 Å of a different amino acid in the revised-version model. A mispositioned residue is denoted if a Cα atom in the first version is not within 2.0 Å of any other Cα atom in the revised-version model. d is the distance between the same amino acids in the first and revised models. $d_{closest}$ is the distance between a residue in the first model and a different closest residue in the space in the revised model.

Supplementary Table 5. AUC-ROC and AUC-PrecRoc values for detecting inconsistent residues in individual first version models of the 35 chains of the PDB2Ver dataset.

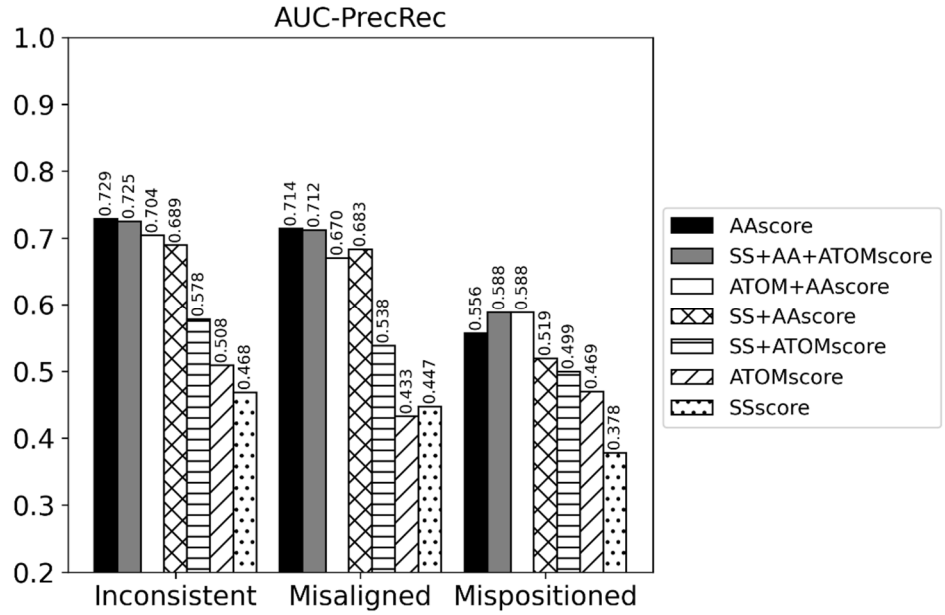
EMID	PDB-ID	AUC-ROC	AUC-PrecRoc
6865	5z10A	0.959	0.945
6865	5z10B	0.939	0.912
6865	5z10C	0.939	0.917
7546	6cp3C	1.000	1.000
7546	6cp3Y	0.928	0.949
7548	6cp6C	1.000	1.000
7548	6cp6Y	0.994	0.993
9906	6k1hC	0.927	0.755
9906	6k1hF	0.947	0.848
9906	6k1hZ	0.947	0.799
30039	6m17E	0.740	0.124
30039	6m17F	0.822	0.202
30127	6m71A	0.913	0.196
10573	6tt7J	0.965	0.889
10573	6tt7K	0.780	0.322
10573	6tt7O	0.561	0.128
10573	6tt7R	0.974	0.943
10573	6tt7S	0.772	0.545
20655	6u5tA	0.903	0.787
20655	6u5tG	0.925	0.390
11127	6za9K	0.997	0.993
11127	6za9O	0.689	0.170
11127	6za9R	0.990	0.972
11127	6za9S	0.825	0.619
30178	7btfA	1.000	1.000
30209	7bv1A	1.000	1.000
30210	7bv2A	0.914	0.721
30226	7bw4A	0.967	0.304

22458	7jsnA	0.989	0.965
22458	7jsnB	0.994	0.987
22458	7jsnC	0.993	0.998
22458	7jsnD	0.983	0.995
23020	7ksmA	0.931	0.520
23020	7ksmC	0.992	0.880
23020	7ksmD	0.907	0.747
Average		0.917	0.729

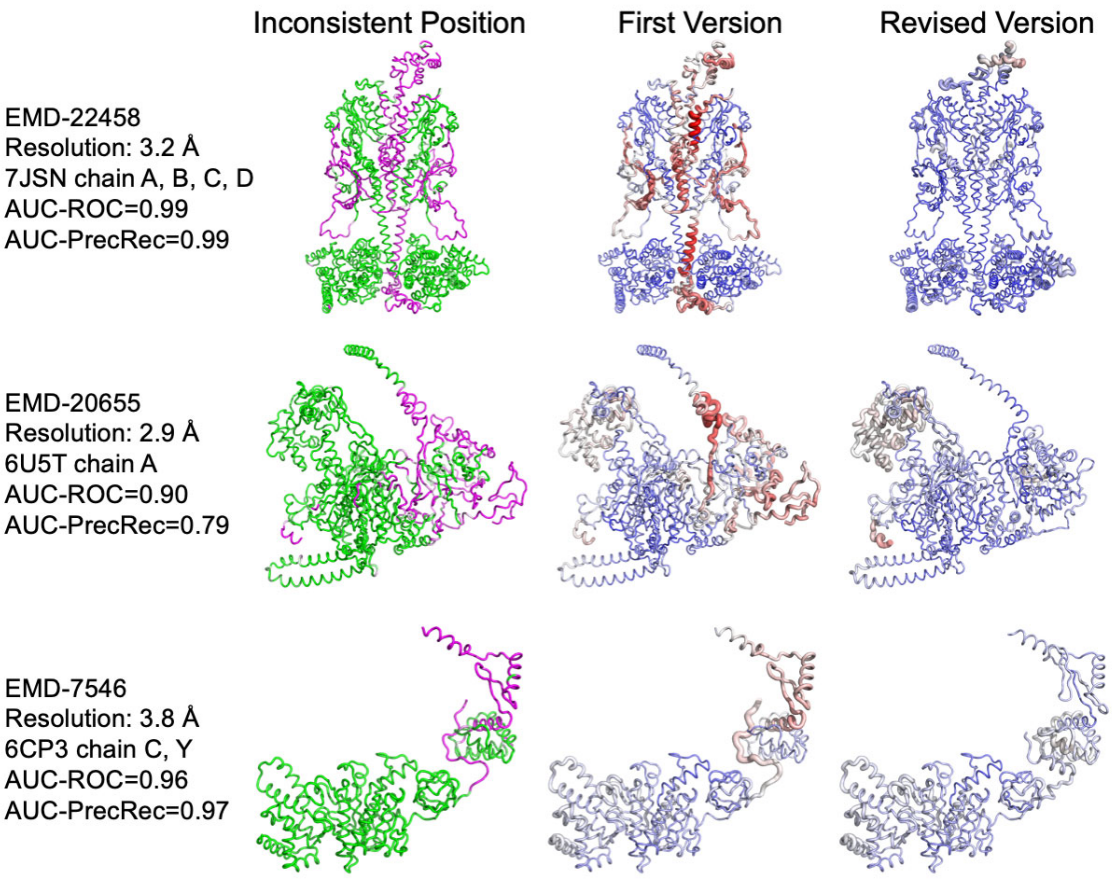
DAQ(AA) was used. The sliding window size was set to 19.



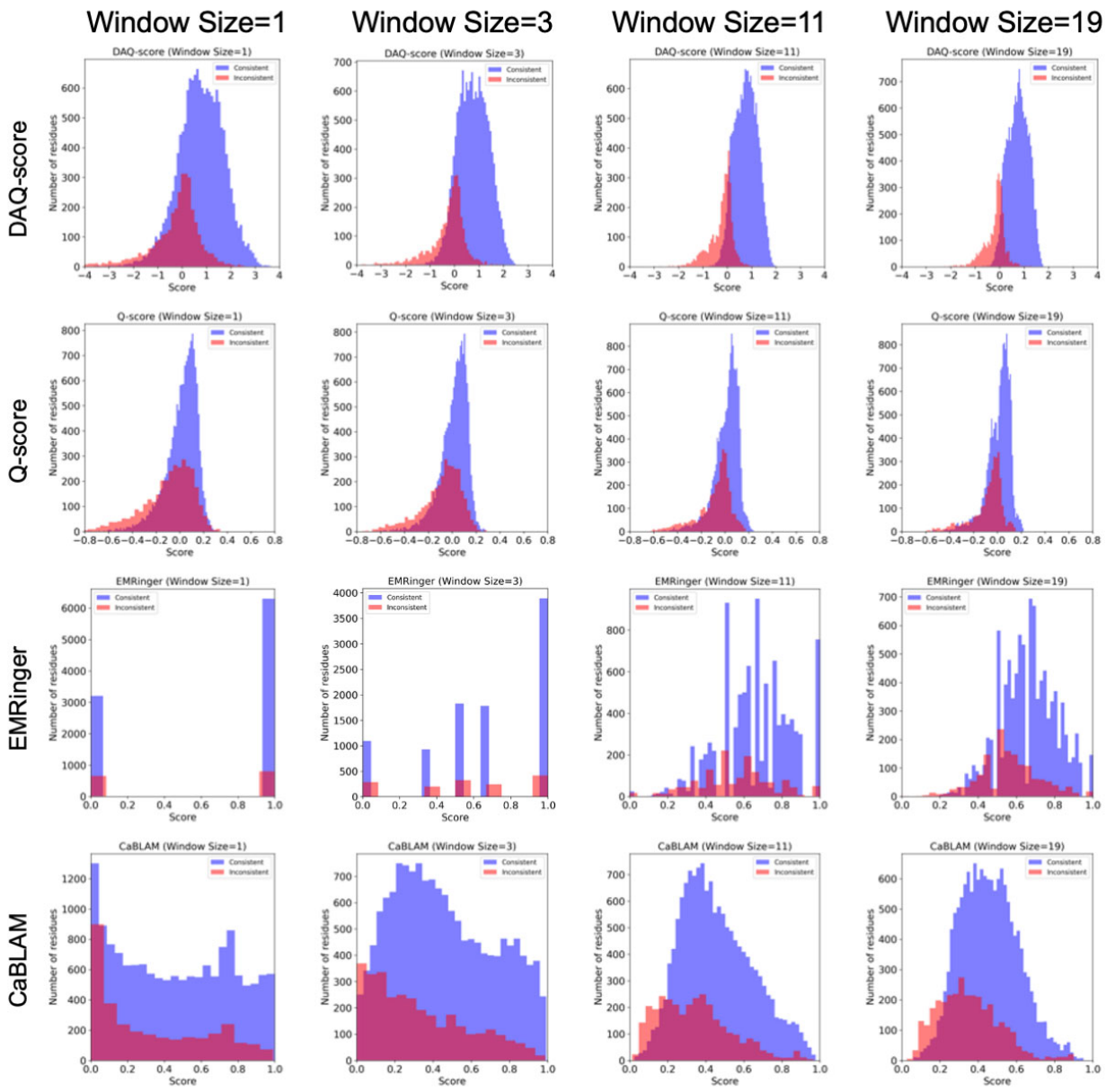
Supplementary Fig. 4. AUC-ROC and Area Under the Curve of Precision-Recall curve (AUC-PrecRoc) values for detecting inconsistent residue positions in first version models using DAQ(AA) with different sliding window sizes. Average of AUC-ROC (blue) and Average of AUC-PrecRec (orange), AUC-ROC and AUC-PrecRoc were first computed for individual models separately, then averaged over all the models. AUC-ROC (green) and AUC-PrecRec (red), all the models were combined.



Supplementary Fig. 5. AUC-PrecRec corresponding to Fig. 2c in the main text. The detection of inconsistent, misaligned, and mispositioned residues in the first-version models of the 35 protein chains was evaluated.



Supplementary Fig. 6. Examples of DAQ score applied to first and revised protein models (relates to Fig. 3). Left, the first version of the model colored according to the deviation of the Cα atom positions from the revised version. Deviation is colored from green (deviation < 1.0 Å) to magenta (deviation > 4.0 Å). Models in the middle and right columns correspond to the first and the revised version of the deposited PDB models, respectively. The models are colored by DAQ(AA) score from red (DAQ(AA) < -2.0) to blue (DAQ(AA) > 2.0). The radius of the tube representation reflects DAQ score when it is a negative value. AUC-ROC and AUC-PrecRec values for detecting inconsistent residues in the first-version models are also shown. Top row: four chains (A-D) of cGMP phosphodiesterase 6 (PDB entry 7JSN; EMD-22458); in Fig. 3a, we only showed chain B. Middle row: fatty acid synthase subunit α (PDB entry 6U5T; EMD-20655). Bottom row: ATP synthase subunit α (chain C) and protein 8 (chain Y) (PDB entry 6CP3; EMD-7546).



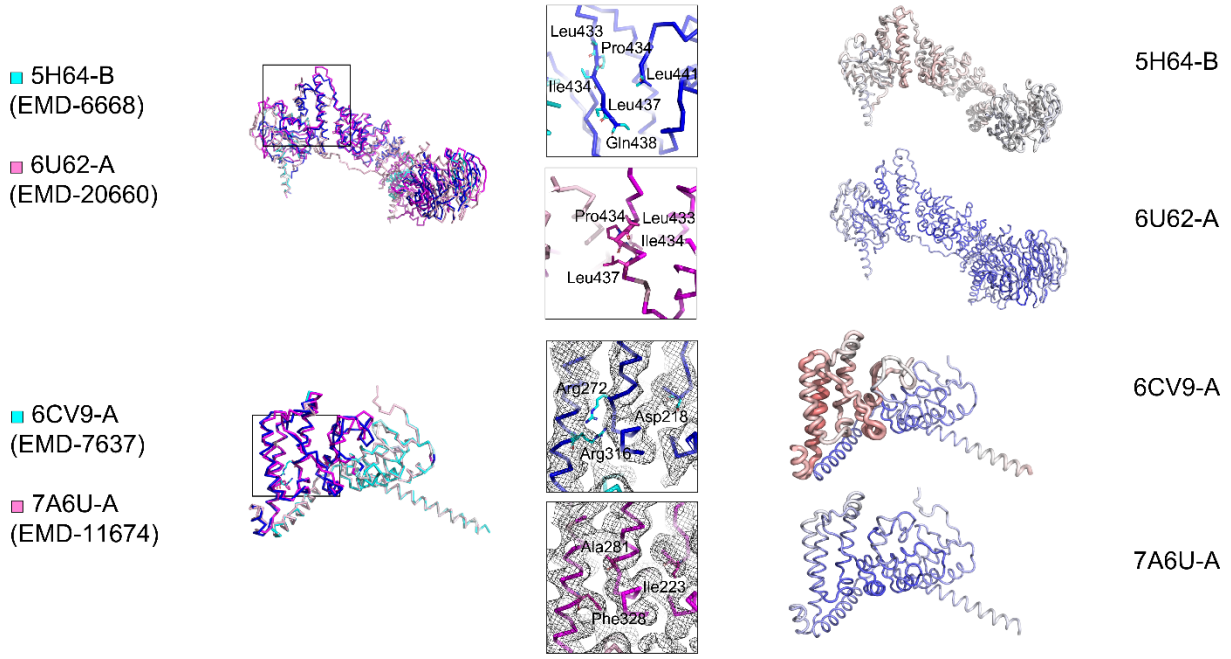
Supplementary Fig. 7. Score distributions for consistent and inconsistent positions in the first-version models of the 35 protein chains used for comparisons in the PDB2ver dataset. The same dataset as in Figure 2 was used. For the four scores, DAQ(AA) score, Q-score, EMRinger, and CaBLAM, four averaging window sizes (1, 3, 11, and 19 residues) were used. Blue and red bars represent consistent and inconsistent positions, respectively.

Supplementary Table 6. (separate Excel file)

399 protein structure model pairs in the PDBNR90 dataset. Fourth and ninth columns show misaligned positions in the first and second PDB-IDs, respectively. The columns of DAQ (AA) score show the total value at the misaligned positions.

Supplementary Table 7. (separate Excel file)

4,485 non-redundant PDB-chain models in the PDBNR1Å dataset. Fifth and sixth columns show Cross Correlation and Overlap between EM-maps and PDB-chain models computed by the TEMPy package.



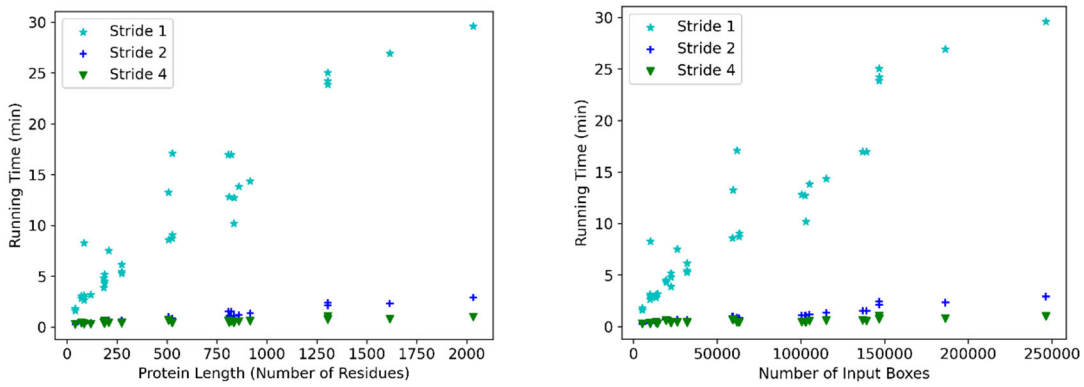
Supplementary Fig. 8. Two additional examples of protein model pairs in the PDBNR90 dataset that have a large score difference. The left column shows the superposition of the pairs (cyan and pink). Inconsistent Cα atom positions (deviation > 4.0 Å) between the two models are indicated in blue and magenta in the cyan and pink models, respectively. Models in the middle and right columns correspond to the lower and higher scored PDB models. Surface meshes represent the EM map at author recommended contour levels. In models in the right column, DAQ(AA) score is indicated in colors from red (DAQ(AA) < -2.0) to blue (DAQ(AA) > 2.0) and by the radius of the tube (thicker being more negative). The first example is a structure of regulatory-associated protein of mTOR[2,3]. Chain B of PDB entry 5H64, derived from the 4.4 Å EM map EMD-6668 is compared to chain A of 6U62, derived from the 3.2 Å map EMD-20660. Although they have 96.2% sequence identity, the sequence assignment does not agree with each other in almost all regions. The full wwPDB EM validation 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 in its report. One example is the residue span Gln432-Leu441. In 5H64 chain B, these are modeled as a β-strand, with residues Leu433, Ile435, and Leu437 presenting their hydrophobic side chains to solvent. Pro434, an unusual residue for a β strand, is central to the strand. In 6U62 chain B, these residues instead form an α helix, with Pro434 capping the N-terminus of the helix, a preferred position for this amino acid. Moreover, there is a missing loop connecting positions Gln438 to Leu441 in 5H64 chain B (not modeled), but two residues are not enough to make the connection. For these reasons, 6U62 chain A is a better fit for the EM map. Our local quality score also suggests that the residues Gln432-Leu441 in 6U62 chain A agree well with the EM map, whereas the DAQ(AA) score suggests that 5H64 chain B does not fill well to EMD-6668. The example on the bottom is structure models of short transient receptor potential channel 6[4]. The sequence identity between 6CV9 chain A (EDM-7637, 3.8Å resolution) and 7A6U chain A (EMD-11674, 3.62 Å resolution) is 97.4%. 6CV9 residues are out of register with respect to 7A6U from Glu187 through Asn337. For example, the 6CV9 Asp218 side chain packs in hydrophobic core with no charge compensation. At this position, 7A6U Ile223 would be more appropriate choice. Also, 6CV9 Arg272 and Arg318 were both

modeled with side chains packing into the core of the protein and, in fact, adjacent to each other to create a highly unfavorable buried electrostatic repulsion. The electron density for the 6CV9 map also does not support the modeling for these residues, and they are more likely Ala281 and Phe328, respectively, as modeled in 7A6U. DAQ(AA) score also indicates that 7A6U has better fit to the density than 6CV9 in the helical regions.

Supplementary Table 8. PDB chains with low-scoring residues according to DAQ(AA) score and fraction cutoff in the PDBNR1Å dataset.

DAQ(AA) score cutoff	Fraction cutoff	Number of chains	% Low scoring models
-1.0	0.1	6	0.13
-1.0	0.2	4	0.09
-1.0	0.3	1	0.02
-1.0	0.4	0	0.0
-1.0	0.5	0	0.0
-0.5	0.1	89	1.98
-0.5	0.2	28	0.62
-0.5	0.3	8	0.18
-0.5	0.4	4	0.09
-0.5	0.5	2	0.04
0.0	0.1	1642	36.61
0.0	0.2	1074	23.95
0.0	0.3	744	16.59
0.0	0.4	494	11.01
0.0	0.5	335	7.47

Number of chains, the number of models that have low-scoring residues cutoff more than the fraction cutoff. To define low-scoring residues, three cutoff values were used, -1.0, -0.5, and 0.0. Fourth column is the percentage of the low-scoring models among the PDBNR1Å dataset (4,485 PDB-chain models).

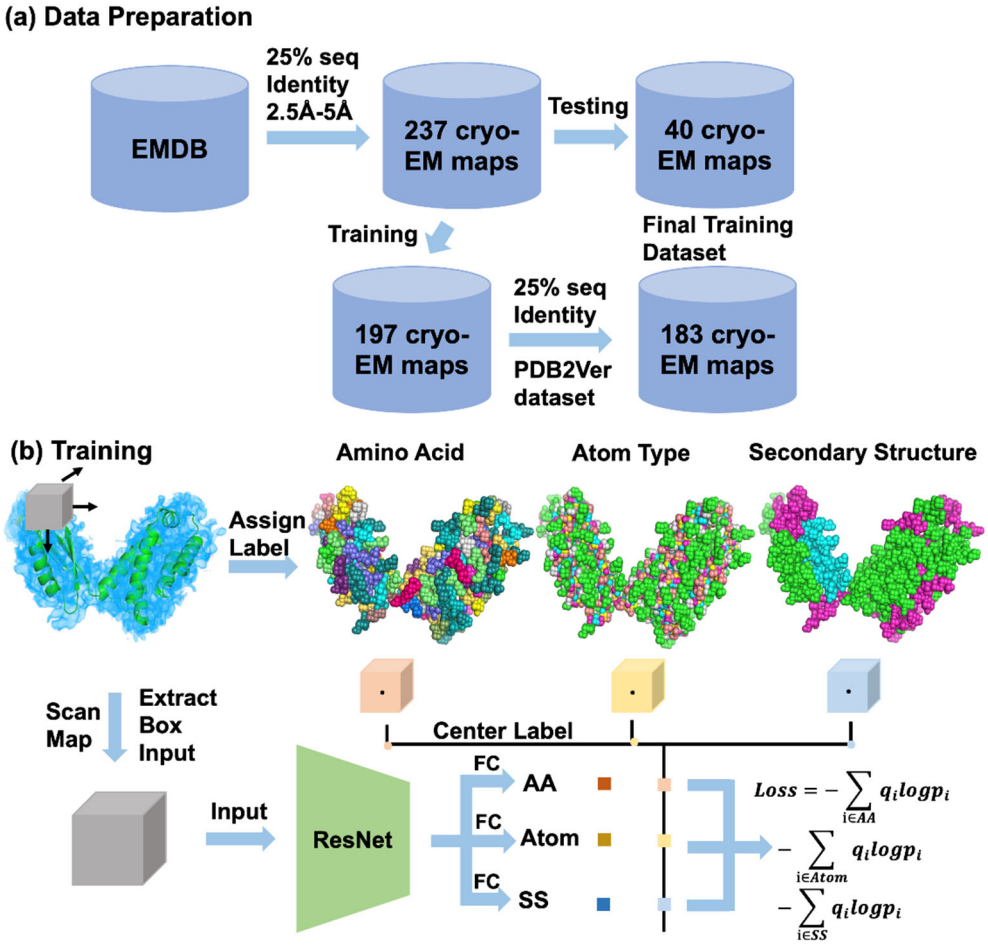


Supplementary Fig. 9. The computational time of DAQ versus the size of the input maps. Three stride values, 1 (default), 2 and 4 were used with 35 maps. The times were measured on a machine with 4 Intel(R) Xeon(R) 3.60GHz CPU cores, a NVIDIA RTX 2080Ti GPU and 64GB memory. The size of input data was quantified in two metrics, the length of the proteins and the number of input boxes from the map.

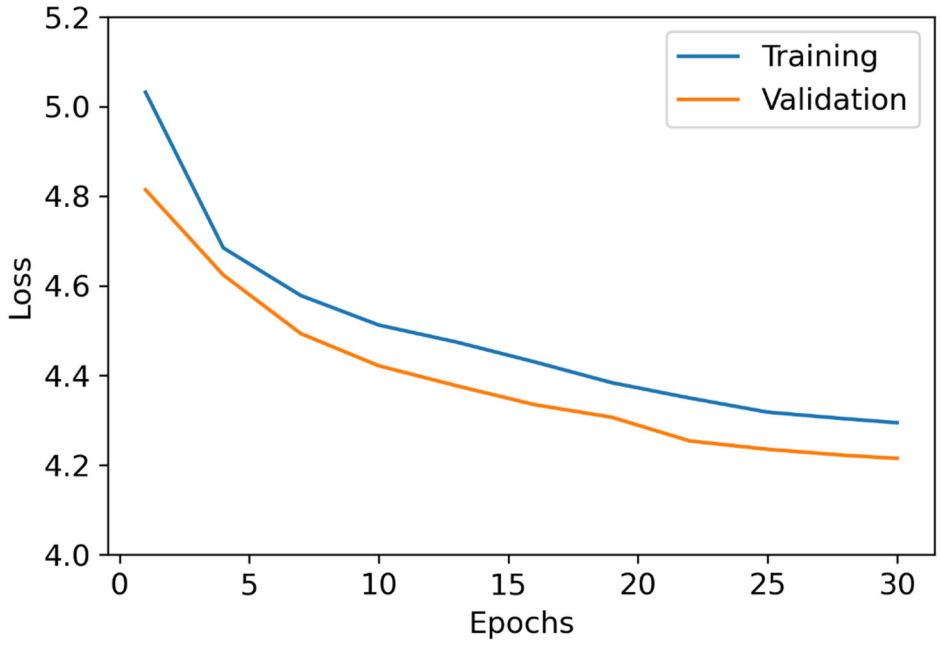
Supplementary Table 9. Accuracy of identifying inconsistent positions in the 35 protein models using different stride sizes (relates to Table 1).

	Average AUC-ROC				**Average Ave-Precision**				**AUC-ROC**				**Ave-Precision**			
Stride size	1	3	11	19	1	3	11	19	1	3	11	19	1	3	11	19
stride=1	0.76	0.84	0.90	0.92	0.41	0.55	0.70	0.73	0.78	0.88	0.95	0.96	0.44	0.63	0.81	0.85
stride=2	0.76	0.84	0.90	0.92	0.41	0.55	0.70	0.73	0.78	0.88	0.95	0.96	0.44	0.63	0.81	0.85
stride=4	0.67	0.75	0.84	0.88	0.32	0.40	0.55	0.61	0.68	0.78	0.89	0.91	0.31	0.44	0.65	0.71

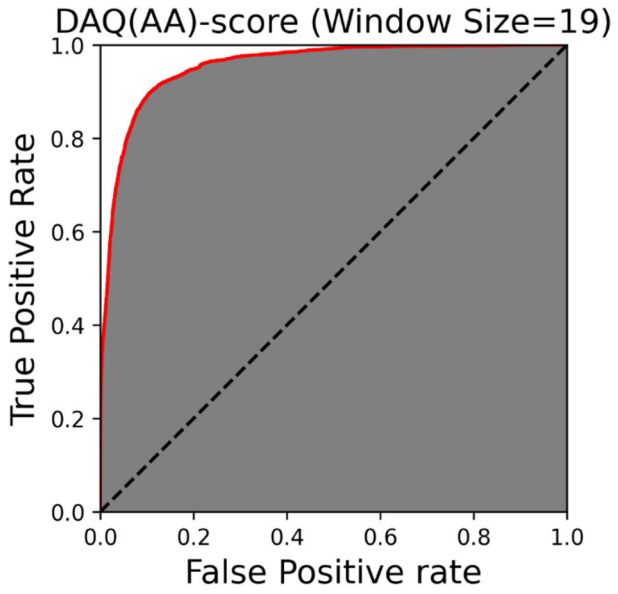
This table corresponds to Table 1. The stride determines the amount of shifts of sampling boxes from an EM map. Stride of 1 (default), 2, 4, are used.



Supplementary Fig. 10. Illustration of the dataset construction and network training process. **a.** The training and testing dataset collection process. From EMDB, we collected 237 independent maps with protein structures by applying a 65% map-structure cross-correlation and a 25% sequence identity cutoff. After considering sequence identity with PDB2Ver dataset, we further reduced to 183 maps for training and 40 maps for testing. **b.** Overview of the training process. We collected boxes of a size of 11^3 Å^3 by scanning across a map along three axes. For each box, 3 labels were assigned, the amino acid type, the atom type and the secondary structure type by checking the closest residue located within 1 Å to the center of the box. The ResNet network was trained by considering a multi-task loss, which is a sum of cross entropy loss of amino acid type (AA), Atom type, and the secondary structure (SS) classification tasks. Note that six types of atoms, Cα, Cβ, main-chain N, C, and O, and other heavy atom types, were used during the training. But in the inference, we only used probability of Cα to compute the DAQ(Cα) score.



Supplementary Fig. 11. Training and validation loss. The training was performed on 146 maps and the validation on 37 maps. The loss considers the sum of the loss values from the amino acid probabilities, atom probabilities, and secondary structure probabilities as shown in Supplementary Fig. 7. The learning rate of 0.002 with LR regularization parameter of $1e^{-5}$ were used for this plot. The loss values were recorded after every 3 epochs. Training was stopped at the 30th epoch because the difference of the loss on the training set was small (0.015) between the 30th epoch and the 27th epoch. The loss of both training and validation were decreasing up to the 30th epoch, showing no trace of overfitting to the training set. The loss of the training was higher than validation in this case, partly because we rotated input boxes randomly during training to achieve better generalization. In image processing, augmenting training images often lead do higher training loss than validation loss. For example, see a paper by Yun et al.[5]



Supplementary Fig. 12. Example of a Receiver Operator Characteristic (ROC) curve. This is the ROC curve of detecting Inconsistent residues using DAQ(AA) score in the first models of the 35 protein chains in the PDB2Ver dataset. This curve corresponds to the left-most black bar in Fig. 2c (AUC-ROC: 0.917).

References

1 Ramirez-Aportela, E. *et al.* FSC-Q: a CryoEM map-to-atomic model quality validation based on the local Fourier shell correlation. *Nature communications* **12**, 42, doi:10.1038/s41467-020-20295-w (2021).

2 Yang, H. *et al.* 4.4 A Resolution Cryo-EM structure of human mTOR Complex 1. *Protein Cell* **7**, 878-887, doi:10.1007/s13238-016-0346-6 (2016).

3 Rogala, K. B. *et al.* Structural basis for the docking of mTORC1 on the lysosomal surface. *Science* **366**, 468-475, doi:10.1126/science.aay0166 (2019).

4 Azumaya, C. M., Sierra-Valdez, F., Cordero-Morales, J. F. & Nakagawa, T. Cryo-EM structure of the cytoplasmic domain of murine transient receptor potential cation channel subfamily C member 6 (TRPC6). *J Biol Chem* **293**, 10381-10391, doi:10.1074/jbc.RA118.003183 (2018).

5 Yun, S. *et al.* CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 6023-6032 (2019).