

2 **Supplementary Information for**

3 **Protein Complex Structure Modeling by Cross-Modal Alignment between Cryo-EM Maps and**

4 **Protein Sequences**

5 **Sheng Chen, Sheng Zhang, Xiaoyu Fang, Liang Lin, Huiying Zhao, and Yuedong Yang**

6 **Yuedong Yang**

7 **E-mail: yangyd25@mail.sysu.edu.cn**

8 **This PDF file includes:**

- 9 Supplementary Note
- 10 Figs. S1 to S18
- 11 Tables S1 to S2
- 12 SI References

1. Modeling for maps that have no deposited PDB structure

The modeling performance of EModelX has been validated in the main text. Here, we attempted to apply it to cryo-EM maps that have no fitted PDB structure. We first gathered a dataset comprising 126 cryo-EM maps released in EMDB from 2023/1/1 to 2023/12/11. All these maps and their XML files were downloaded from EMDB(1). These maps were filtered by released date (2023/1/1 to 2023/12/11), resolution (2 to 4Å), and no fitted PDB structure. The complete data list of these 126 cryo-EM maps is available in Supplementary Data 3. Subsequently, we implemented EModelX to build models without protein sequences, namely EModelX(w/o seq). As shown in Algorithm 1, EModelX(w/o seq) was implemented as a greedy algorithm. As an ablation for sequence input, EModelX(w/o seq) was first applied to our test set of 99 maps and compared with EModelX. As shown in Fig. S4, the average TM-score of EModelX(w/o seq) was 0.433, lower than that of EModelX (0.808) and EModelX(init) (0.628). It underscores the importance of C α -sequence alignment that uses the sequence input. EModelX(w/o seq) was then applied to the 126 cryo-EM maps that have no fitted PDB structure. Several maps were further tested by EModelX and EModelX(+AF) using sequence information extracted from XML files. The modeling results are provided in Supplementary Data 5.

Algorithm 1: EModelX without sequence

Input: $\mathcal{C}$ as C α candidates, $\mathcal{H}$ as the estimated C α neighbor connection likelihood

```

1  $\mathcal{V} \leftarrow$  edges of C $\alpha$  candidates whose distance in (2,6)Å;
2  $\mathcal{V} \leftarrow$  sorted  $\mathcal{V}$  by  $\mathcal{H}$  from lower to higher;
3 for  $v$  in  $\mathcal{V}$  do
4    $(n_1, n_2) \leftarrow v$ ;
5   if  $n_1.degree > 2$  or  $n_2.degree > 2$  then
6      $\mathcal{V}.remove(v)$ ;
7   end
8 end
9  $\mathcal{T} \leftarrow (\mathcal{C}, \mathcal{V})$  // traces and loops;
10  $\mathcal{T} \leftarrow \mathcal{T}$  after loops disconnecting;
Output: Integrate at most 62 traces in  $\mathcal{T}$  as the built model
```

Fig. S5 exhibits three cases of maps with no deposited PDB structure. A 3.22Å map (EMD-29509) of the circadian clock protein KaiC-RS-S413E/S414E dodecamer has been released in the study of Pitsawong et al. (2) However, no structure has been deposited in PDB to fit this map before the date this paper is announced. As depicted in Fig. S5a, the model built by EModelX(w/o seq) was well-fitted into the cryo-EM map and obtained a satisfactory CC_{box} of 0.737. Nevertheless, due to the absence of sequence guidance, EModelX (w/o seq) failed to achieve symmetric dodecamer modeling. Therefore, we extracted the protein sequence from the XML file released in EMDB and consequently applied EModelX to model this map. As depicted in the lower portion of Fig. S5a, the dodecamer built by EModelX exhibited D6 symmetry. The CC_{box} was also improved to 0.778. Similar trends were also found in another case. McCarthy et al. have reconstructed multiple cryo-EM maps of the I3-01 (3). Among them, a 3.9Å cryo-EM map (EMD-28028) has no deposited structure in the PDB. As shown in Fig. S5b, the 62-mer assembled by EModelX (w/o seq) exhibited satisfactory I4 symmetry, achieving a CC_{box} of 0.816. We further employed EModelX for sequence-guided model construction, which improved the CC_{box} to 0.836. The last case is also a nanocage cage_T3_5 designed by Huddy et al(4). Both models built by EModelX(w/o) and EModelX did not exhibit the T symmetry described in their study. We then attempted to build a complex model using EModelX(+AF). As illustrated in Fig. S5c, EModelX(+AF) successfully built a dodecamer model with T symmetry. Its all-atom model exhibited good fitness to the density map, which contributed to its high CC_{box} of 0.915.

While some EMDB entries do not have an associated PDB file, many of those could be additional maps to another entry with an associated PDB file. Therefore, we performed additional analyses using the most related PDB structure (in the same released paper but was not deposited for this map) to evaluate our modeling results. First, corresponding to EMD-29509 in Fig. S5a and EMD-28028 in Fig. S5b, we observed that EModelX achieved very high TM-scores of 0.992 and 0.996 with related pdb (8fw1 for EMD-29505 (2) and 8ed3 for EMD-28027(3)), respectively. However, when we applied EModelX to another map EMD-27890 of mouse cold-sensing TRPM8 channel protein, only pdb structures with different conformational states (EMD-27895: Class I, C0 state; EMD-27890: Class III putative desensitized state) can be found in Yin et al.'s study (5), and a lower TM-score of 0.865 was obtained. Focusing on the voltage-sensor-like domain, as shown in Supplementary Fig. S6c-d, we observed that 8e4p exhibited a straight conformation in S5 helix, consistent with the description provided by its authors(5). In contrast, the S5 helix modeled by EModelX for EMD-27890 appeared to be bent. This suggested that the putative desensitized state may share a similar S5 helix conformation with the C1 state (5). It indicated that comparing the structural differences between the EModelX model and the relevant PDB structures may reveal the dynamic changes in molecular conformations across different maps.

2. Implementation of Other Methods

- Phenix** We use Phenix (phenix-1.20.1-4487, released date: Jan 20, 2022) to benchmark our test dataset following the tutorial from https://www.phenix-online.org/documentation/reference/map_to_model.html and http://www.phenix-online.org/documentation/tutorials/cryo_em_structure_solution_model_building.html: **i.** get symmetry automatically by `phenix.map_symmetry denmod_map.ccp4 resolution=2 symmetry_out=denmod_map.ncs_spec`. **ii.** build a model into the map with `map_to_model: phenix.map_to_model denmod_map.ccp4 resolution=2 nproc=4 ncs_file=denmod_map.ncs_spec pdb_out=map_to_model.pdb seq_file=seq.dat`. In a half hour or so you should get the model. **iii.** refine the build_in model either against the original map: `phenix.real_space_refine map_to_model.pdb denmod_map.ccp4 resolution=2 ignore_symmetry_conflicts=True`
- MAINMAST**: We run MAINMAST (version 1.0, released date: Mar 1, 2017) following the tutorial from <http://kih.aralab.org/mainmast/Tutorials.html>: **i.** Generate density map in situs format and predict secondary structure from SPIDER2: `map2map input.mrc input.situs SPIDER2/misc/run_local.sh input.seq`. **ii.** Generate main chain traces from density map: `MAINMAST -m input.situs -filter 0.3 -Rlocal 10 > path.pdb`. **iii.** Thread the amino acid sequence on the traces: `ThreadCA -i path.pdb -a 20AA.param -spd input.fasta.spd33 -fw 1.4 -Ab 3.4 -Wb 0.9 -r`
- DeepTracer**: A web service is globally accessible at <https://deeptimizer.uw.edu>. We run DeepTracer in this web service.
- ModelAngelo**: We run ModelAngelo (version 1.0.12, released date: Nov 29, 2023) following the tutorial <https://github.com/3dem/model-angelo>: `model_angelo build -v map.mrc -pf prot.fasta -o output`
- AlphaFold**: We run AlphaFold following the tutorial <https://github.com/google-deepmind/alphafold>. The default setting of the monomer version is used: `python3 docker/run_docker.py --fasta_paths=*** --max_template_date=2020-05-14 --model_preset=monomer --db_preset=full_dbs`.

It should be noted that for MAINMAST and Phenix, the input maps are manually sharpened by PDB structure (following MAINMAST (6) and EMBuild (7)) since they are not automated for chain separation. However, for DeepTracer, ModelAngelo, and EModelX, the input maps are raw maps.

3. Pruning Search for Sequence-Guiding C α Threading

We employed a strategy of pruning search to accelerate the sequence-guiding C α threading, which is a modified version of our previous work(8). Here, the pruning search algorithm is described by Algorithm 2 and a schematic flowchart in Fig. S8. The breadth-first searching (BFS) is performed iteratively to expand traces to their neighbor C α s until the expanded traces reach both the sequential and structural endpoints. During iteration, when the number of traces exceeds a threshold, all traces would be clustered by structural similarity. Only the top-ranked trace in each cluster could be preserved, and the scoring function to rank traces has been detailed in Equation 28 of the main text.

Algorithm 2: Pruning Search Algorithm of Sequence-Guiding C α Threading

Input: $\mathcal{C}$ as C α candidates, $\mathcal{F}$ as the scoring function detailed in the main text, N as the max number of searched traces

```
1 while not reach the end or failed do
2    $\mathcal{T} \leftarrow \text{BFS\_STEP}(\mathcal{T}, \mathcal{C})$ ;
3   if  $\text{len}(\mathcal{T}) > N$  then
4      $\mathcal{T}' \leftarrow$  empty list;
5      $\mathcal{T} \leftarrow$  sorted  $\mathcal{T}$  by  $\mathcal{F}$  from higher to lower;
6     for  $t$  in  $\mathcal{T}$  do
7       if  $t$  not similar with traces in  $\mathcal{T}'$  then
8          $\mathcal{T}'.\text{append}(t)$ ;
9       end
10    end
11     $\mathcal{T} \leftarrow \mathcal{T}'$ ;
12  end
13 end
```

Output: Return the top scoring t in $\mathcal{T}$ if reaching the end or return None if failed

4. Secondary structure recall

Similar to the sequence recall, the secondary structure recall is defined as the proportion of the PDB residues that is neighboring (C α distance $\leq 3\text{\AA}$) to a modeled residue with the same type of secondary structure. DSSP program(9) was used to classify the secondary structures into eight categories based on the atomic models built by each method. As illustrated in Fig. S9 EModelX(+AF) achieved an average secondary structure recall of 67.0%, surpassing EModelX's 61.4%, ModelAngelo's 63.7%, DeepTracer's 50.6%, MAINMAST's 14.1%, and Phenix's 19.7%.

91 5. Resolution cutoff up to which the model and map Fourier coefficients (FSC > 0.5)

92 d_FSC_model_05 is defined as the resolution cutoff up to which the model and map Fourier coefficients are similar (FSC >
93 0.5). As depicted in Fig. S10, EModelX(+AF) achieved the best average estimated resolution of 4.13Å, close to the PDB's
94 3.61Å, better than 4.65Å for EModelX, 5.63Å for ModelAngelo, 6.07Å for DeepTracer, 14.34Å for MAINMAST, and 15.09Å
95 for Phenix.

96 6. Abilities to fold single chains compared to AlphaFold

97 Here we discussed how EModelX and other methods performed compared to AlphaFold in each single chain from complex
98 structures of our test set. TM-scores were computed for AlphaFold, EModelX, and other methods on 660 single chains split
99 from 99 complex structures in our test set. There are 28 models built by Phenix, 28 models by MAINMAST, 116 models by
100 DeepTracer, and 167 models by ModelAngelo that obtained higher TM-scores than the corresponding AlphaFold model. In
101 contrast, EModelX successfully built 273 models that achieved higher TM-scores than the corresponding AlphaFold model.
102 Corresponding to the case study in Fig. 5d-f of the main text, Fig. S11 visualized models built by all compared methods for
103 a 3.76Å Cryo-EM map of SARS-CoV-2 Nsp2 (EMD-23970), which was fitted by a single chain structure (PDB ID:7msw).
104 The model built by EModelX exhibited higher similarity to the PDB structure (TM-score: 0.734), outperforming compared
105 methods (Phenix: 0.106, MAINMAST: 0.447, DeepTracer: 0.407, ModelAngelo: 0.580).

106 7. Applying EModelX to maps of lower (>4 Å) resolution

107 We additionally applied EModelX to 8 maps of lower (>4 Å) resolution. These 8 maps satisfied the same protein sequence
108 de-homogenous filtering (to our training set) as the test set. As shown in Supplementary Table S2, EModelX (+AF) achieved
109 an average TM-score of 0.767, higher than the 0.392 obtained by EModelX. This is reasonable because introducing AlphaFold
110 structures is akin to template-based modeling, which is commonly employed by biologists in low-resolution structures where de
111 novo modeling is hard to solve.

112 8. Time consumption by EModelX and EModelX(+AF) to build C α model

113 Fig. S12 illustrated the running time costs of EModelX and EModelX(+AF) in different intervals of residue number in the
114 PDB structure. The cryo-EM maps were divided into four ranges based on the number of residues ([473,1086]→[1086,1521]→
115 [1521,2536]→[2536,8533]). Within these ranges, the median time consumption by EModelX (excluding post-processing of atomic
116 refinement and AlphaFold prediction) showed an increasing trend (98→156→205→816 seconds). Similarly, EModelX(+AF)
117 also exhibited an increasing trend in median time consumption (261→460→552→1291 seconds). EModelX(+AF) cost more
118 time than EModelX, which is reasonable since EModelX(+AF) additionally scored and aligned the sampled fragments with
119 AlphaFold predicted structure. In the range with a larger number of residues ([1521,2536]), EModelX and EModelX(+AF)
120 were able to complete the modeling within median value of 4 minutes and 10 minutes, respectively. Even in the range with the
121 highest number of residues ([2536,8533]), the median time consumption for EModelX and EModelX(+AF) was approximately
122 14 minutes and 22 minutes, respectively.

123 9. Atomic level evaluation by molprobity scores, clash scores, Ramachandran outliers, and rotamer scores

124 We have computed the molprobity scores, clash scores, Ramachandran outliers, and rotamer scores for EModelX and compared
125 them with other methods. The molprobity score was computed by *phenix.molprobity*. As shown in Fig.S13, EModelX(+AF)
126 achieved an average molprobity score of 2.61, which is the lowest in all compared methods and is nearest to that of PDB
127 structures (1.80). The clash score was computed by *phenix.clashscore*. As shown in Fig.S14, EModelX achieved an average
128 clash score of 25.7 (lower is better). Among the compared methods, only Phenix had a lower clash score (21.4) than EModelX.
129 However, Phenix had a lower modeling coverage (31.6%, far lower than EModelX's 83.0%), and higher modeling coverage might
130 lead to more inter-residue clashes. The Ramachandran outlier rate was computed by *phenix.ramalyze*. As depicted in Fig.S15,
131 EModelX and EModelX(+AF) achieved average Ramachandran outlier rates of 3.30% and 3.06%, respectively, which both were
132 lower than Phenix, MAINMAST, and DeepTracer (4.36%, 34.28%, and 29.82%) and were comparable with ModelAngelo's rate
133 of 2.79%. The rotamer outlier rate was computed by *phenix.rotalyze*. As illustrated in Fig.S16 EModelX and EModelX(+AF)
134 achieved average rates of 1.47% and 1.49%, respectively, outperforming Phenix, MAINMAST, and ModelAngelo with rates of
135 2.69%, 28.28%, and 9.33%. DeepTracer, as it only models main-chain atoms, was not included in the comparison.

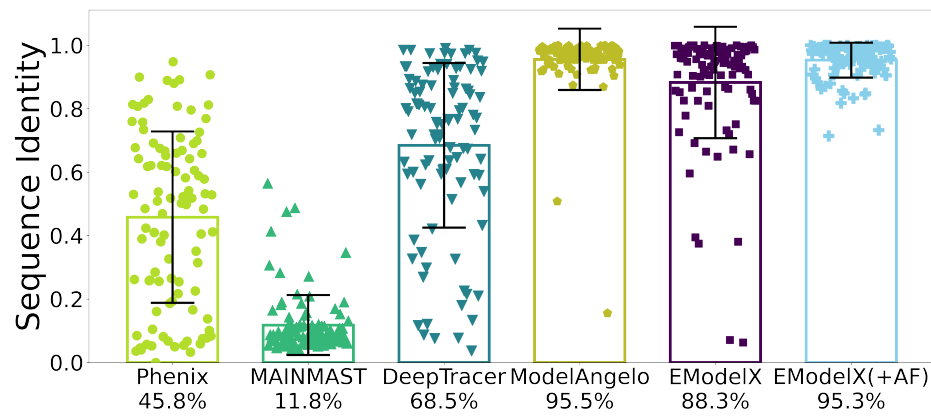


Fig. S1. The average sequence identity on the test cases of 99 Cryo-EM maps. Error bars indicate ± 1.0 standard deviation. The sequence identity was computed by MMalign(10).

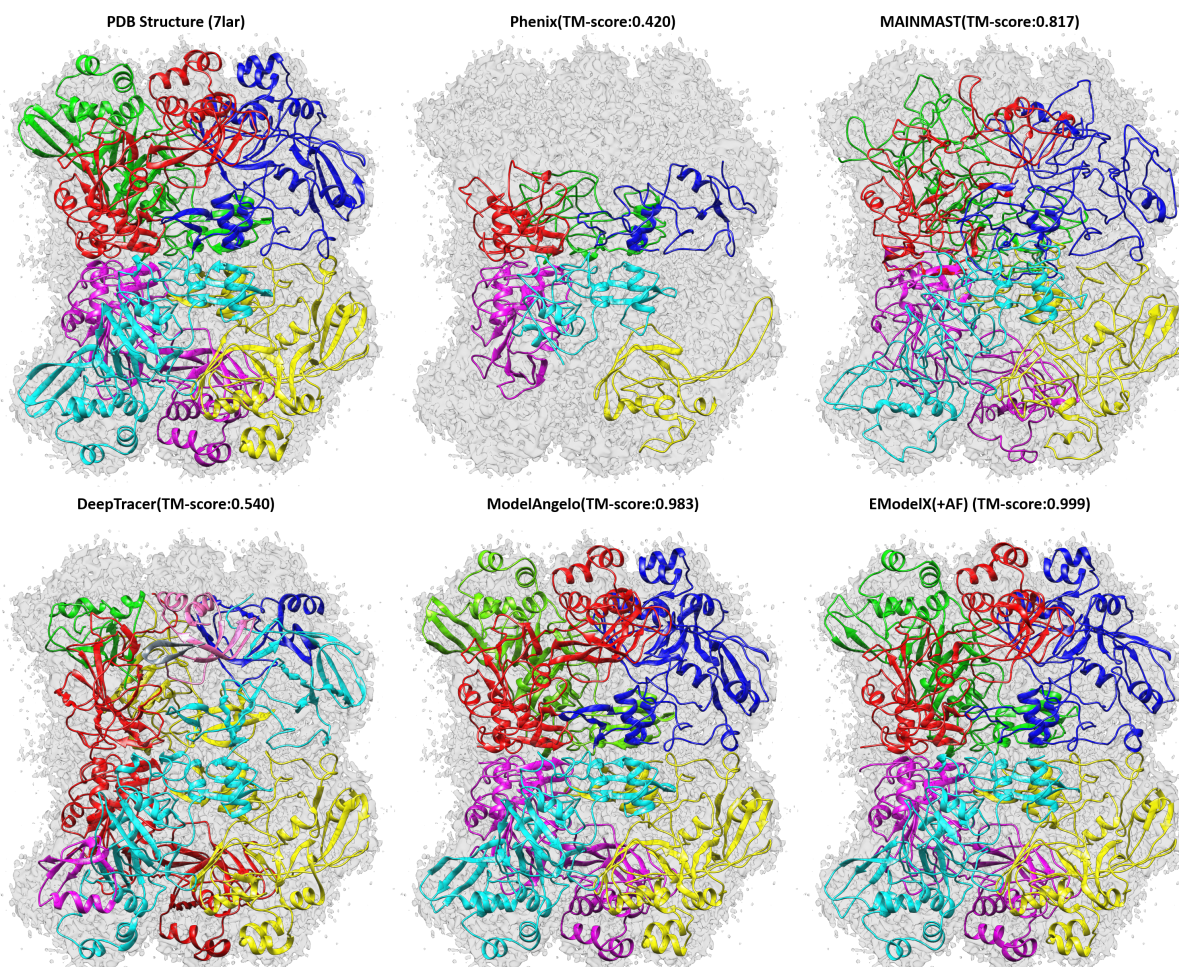


Fig. S2. The deposited PDB structure and models built by compared methods for Map EMD-24101 of SARS-CoV-2 Nsp15 endoribonuclease post-cleavage state at the resolution of 2.2Å.

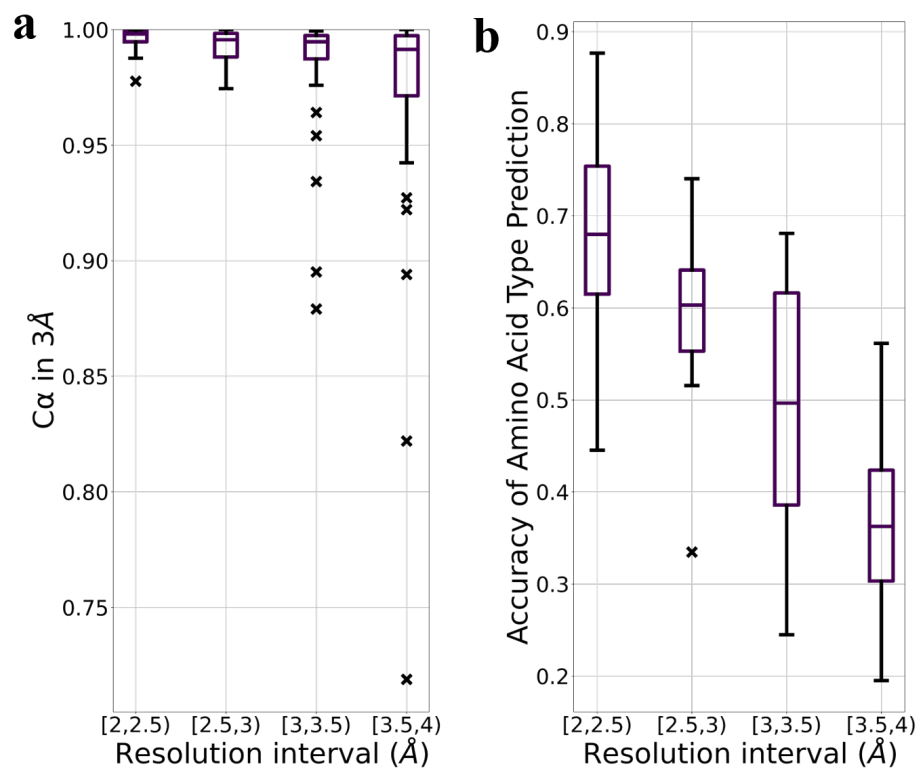


Fig. S3. The boxplot of $C\alpha$ in 3Å (a) and accuracy of amino acid type prediction (b) at different resolution ranges for the neural network trained for EModelX. The box extends from the first quartile (Q1) to the third quartile (Q3), with a line at the median. The whiskers extend from the box to the farthest data point lying within $1.5 \times (Q3 - Q1)$ from the box. Flier points are those past the end of the whiskers.

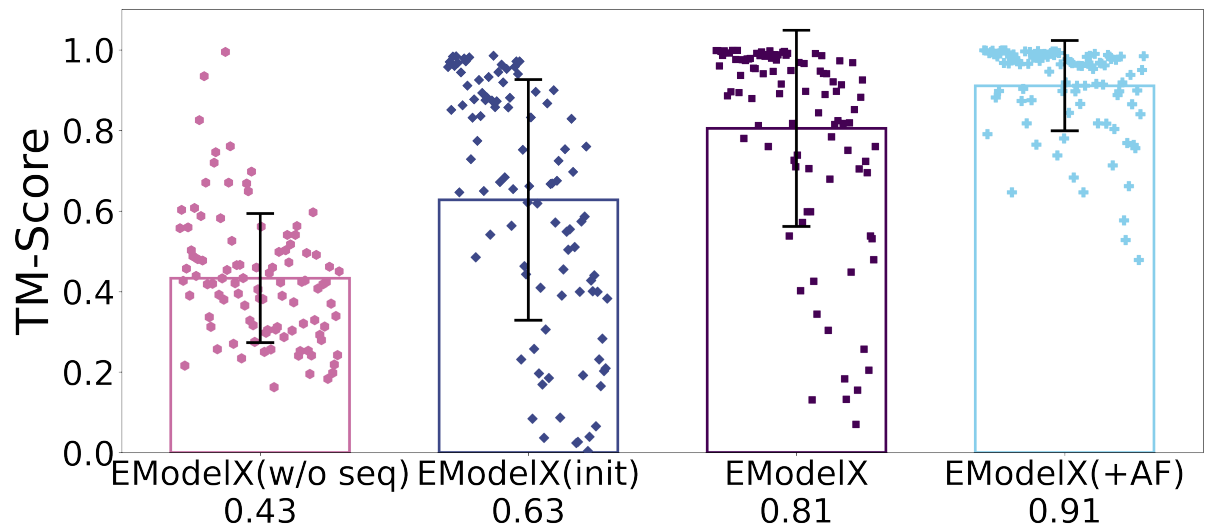


Fig. S4. Ablation study of EModelX on TM-score by removing sequence input (EModelX(w/o seq)), removing-sequence guiding C α threading (EModelX(init)), and adding AlphaFold module (EModelX(+AF)).

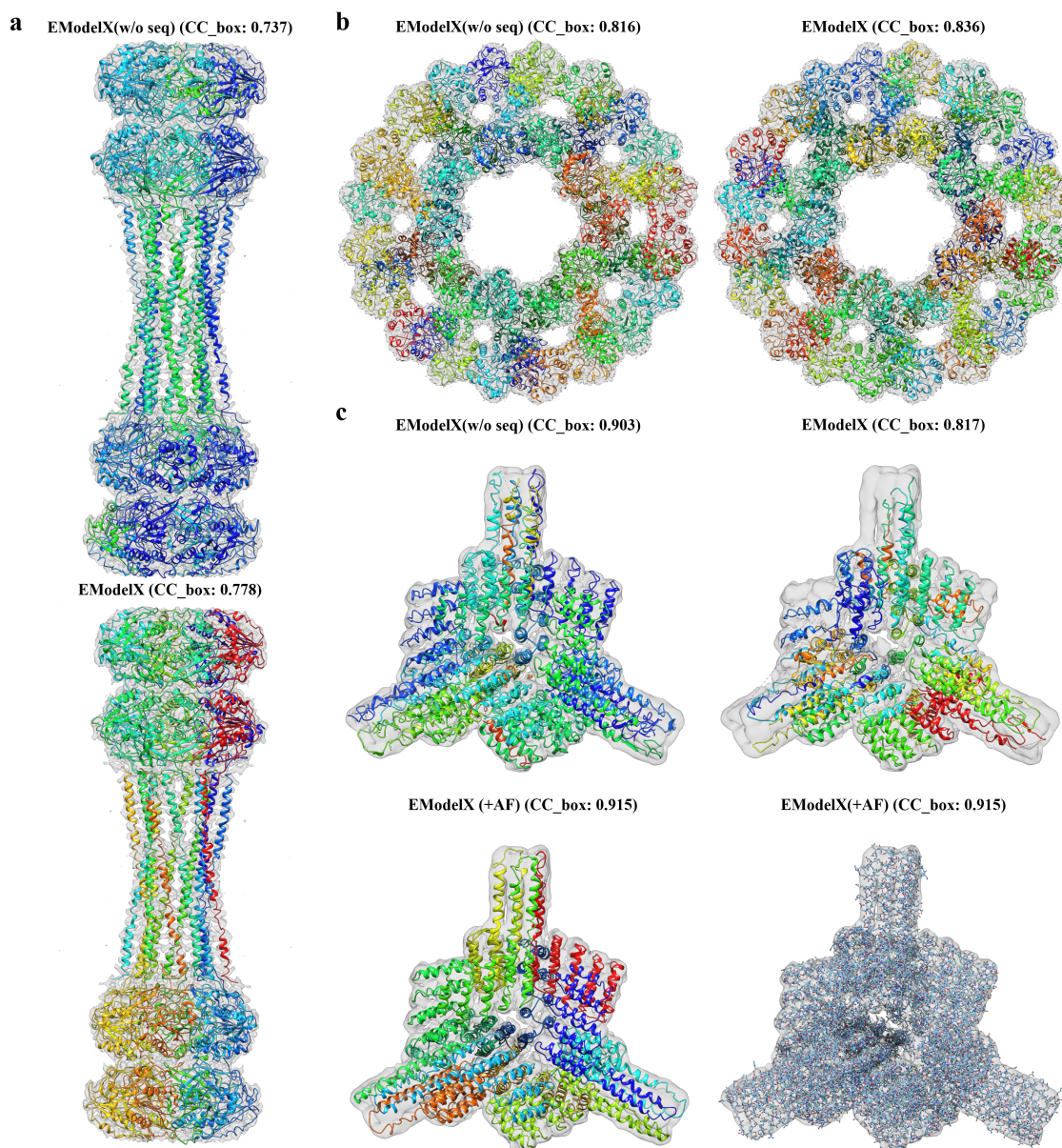


Fig. S5. Applying EModelX to cryo-EM maps that have no fitted PDB structure. **a** EMD-29509 at 3.22Å resolution for dodecameric circadian clock protein KaiC-RS-S413E/S414E. **b** EMD-28028 at 3.9Å resolution for the extended nanocage I3-01 with I4 symmetry. **c** EMD-40074 at 3.6Å resolution for synthetic cage_T3_5 reconstructed with T symmetry. The Cryo-EM density maps are colored in transparent gray. Each chain in a model is rendered by a unique color.

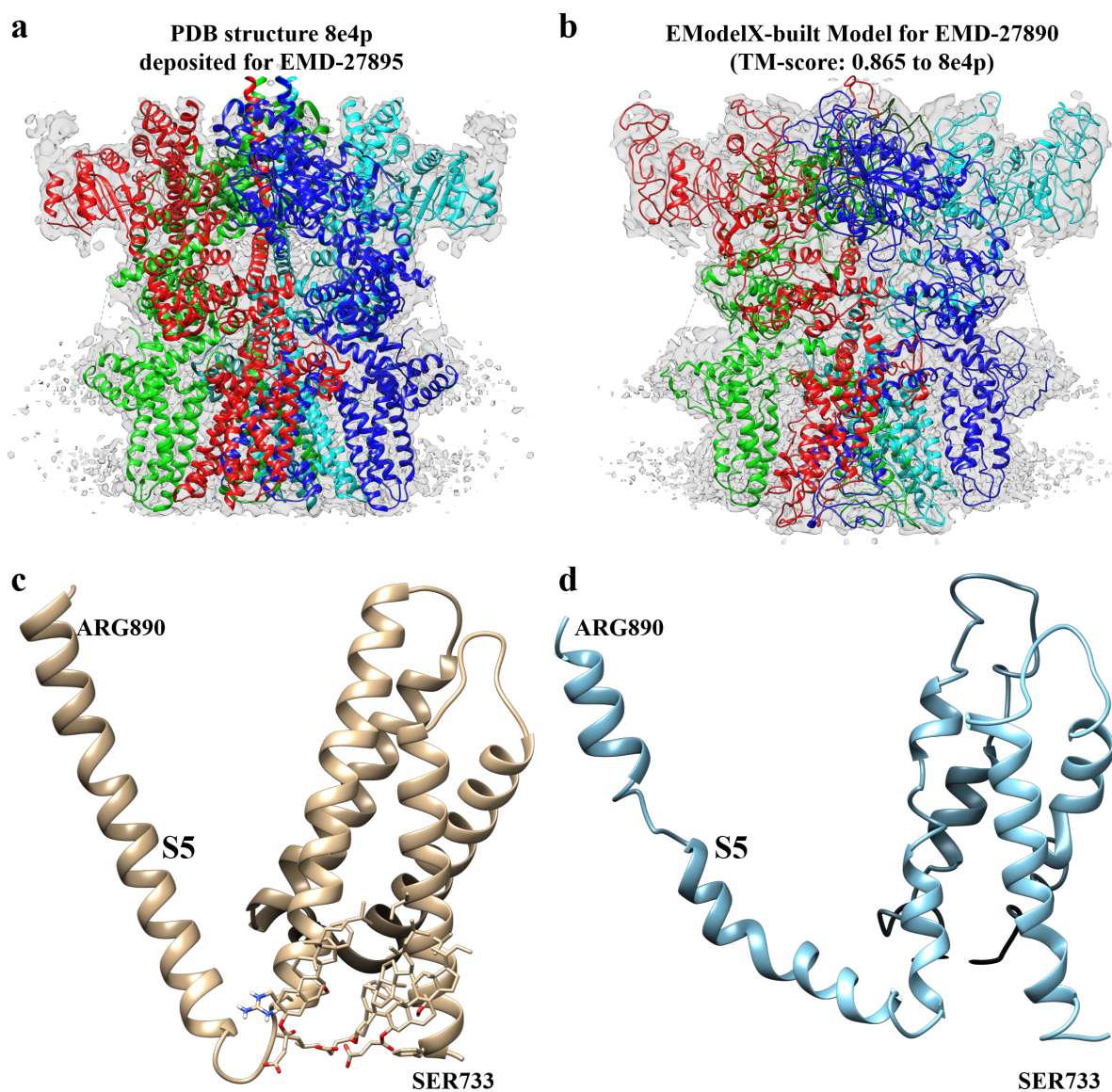


Fig. S6. **a.** The pdb structure 8e4p deposited for EMD-27895 of Mouse TRPM8 structure determined in the ligand- and PI(4 5)P2-free condition Class I C0 state with a resolution of 3.59Å. **b.** EModelX-built Model for EMD-27890 mouse transient receptor potential cation channel subfamily M member 8 with a resolution of 3.43Å. The voltage-sensor-like domain conformation of **(c)** the pdb structure 8e4p deposited for EMD-27895 and **(d)** the EModelX-built model for EMD-27890.

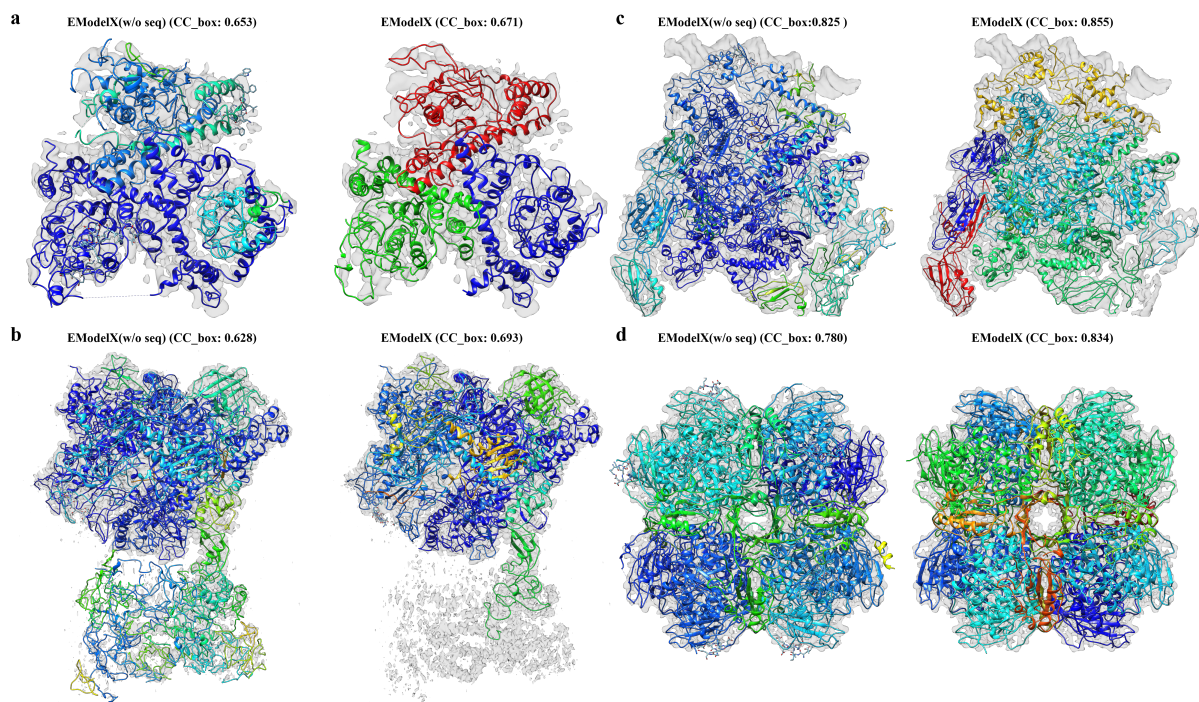


Fig. S7. Applying EModelX to cryo-EM maps that have no fitted PDB structure. **a** EMD-29545(11) at 3.39Å resolution for GltPh RSMR mutant bound with Na and Asp. **b** EMD-16828(12) at 3.04Å resolution for The Pol II-SPT6 transcription elongation complex. **c** EMD-28783(13) at 2.6Å resolution for the EsN-dhsU36mm2. **d** EMD-35166(14) at 2.2Å resolution for Ribulose 1,5-bisphosphate carboxylase/oxygenase from *Phaeodactylum tricornutum*. The Cryo-EM density maps are colored in transparent gray. Each chain in a model was rendered by a unique color.

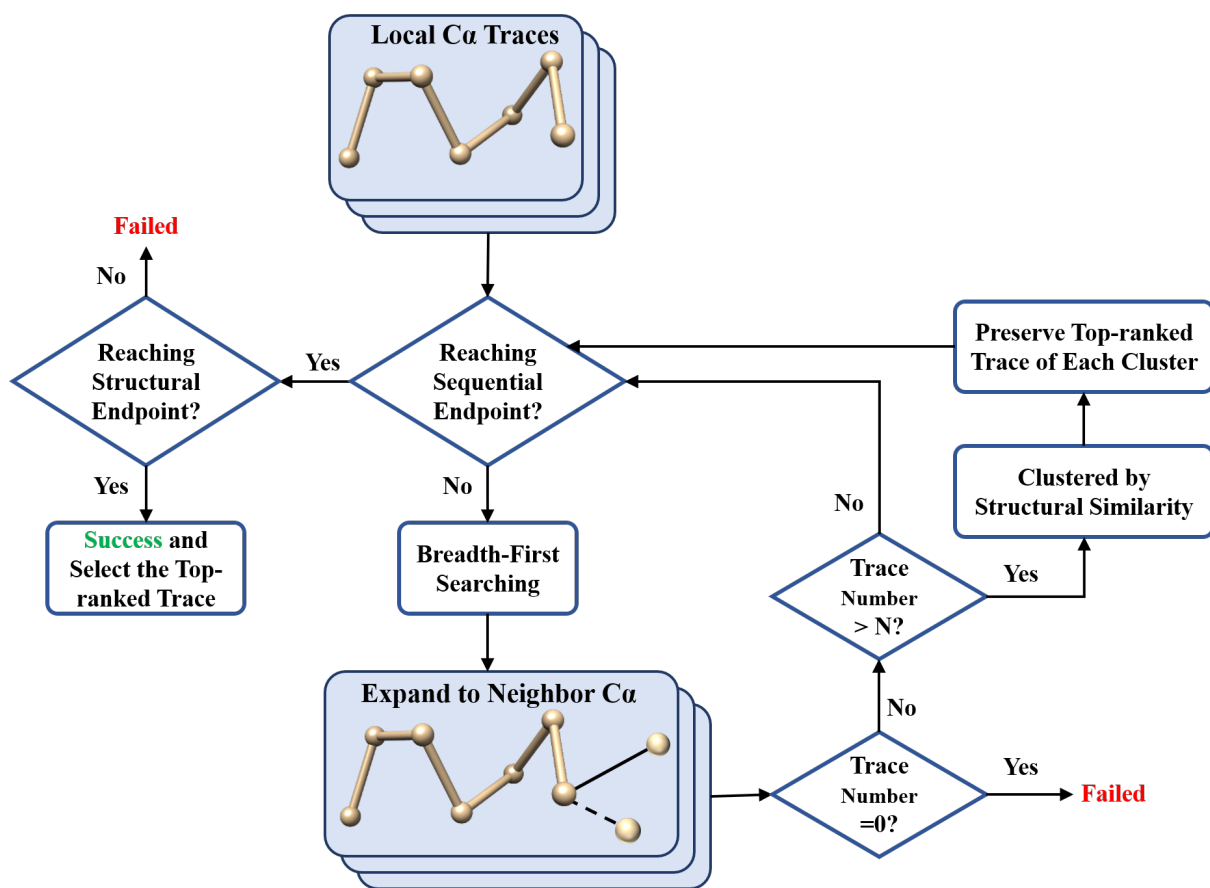


Fig. S8. The schematic flowchart of the pruning search algorithm in sequence-guiding C α threading.

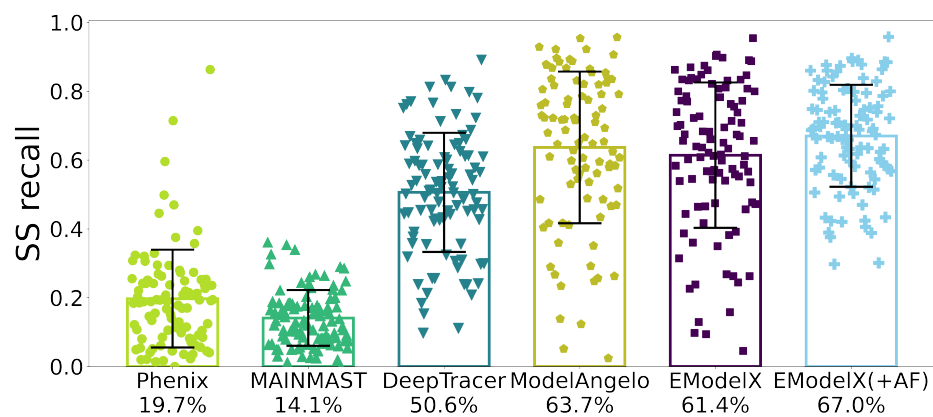


Fig. S9. The average secondary structure recall on the test cases of 99 Cryo-EM maps. Error bars indicate ± 1.0 standard deviation.

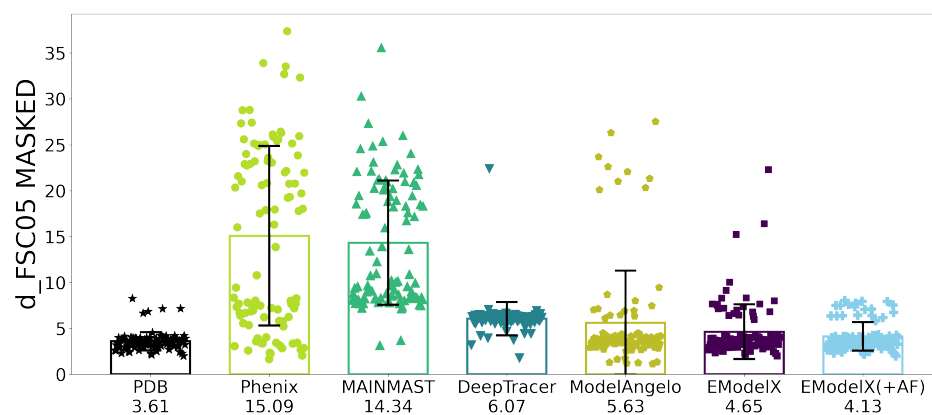


Fig. S10. The average FSC05 on the test cases of 99 Cryo-EM maps. Error bars indicate ± 1.0 standard deviation. The FSC05 value was calculated by *phenix.mtriage*(15).

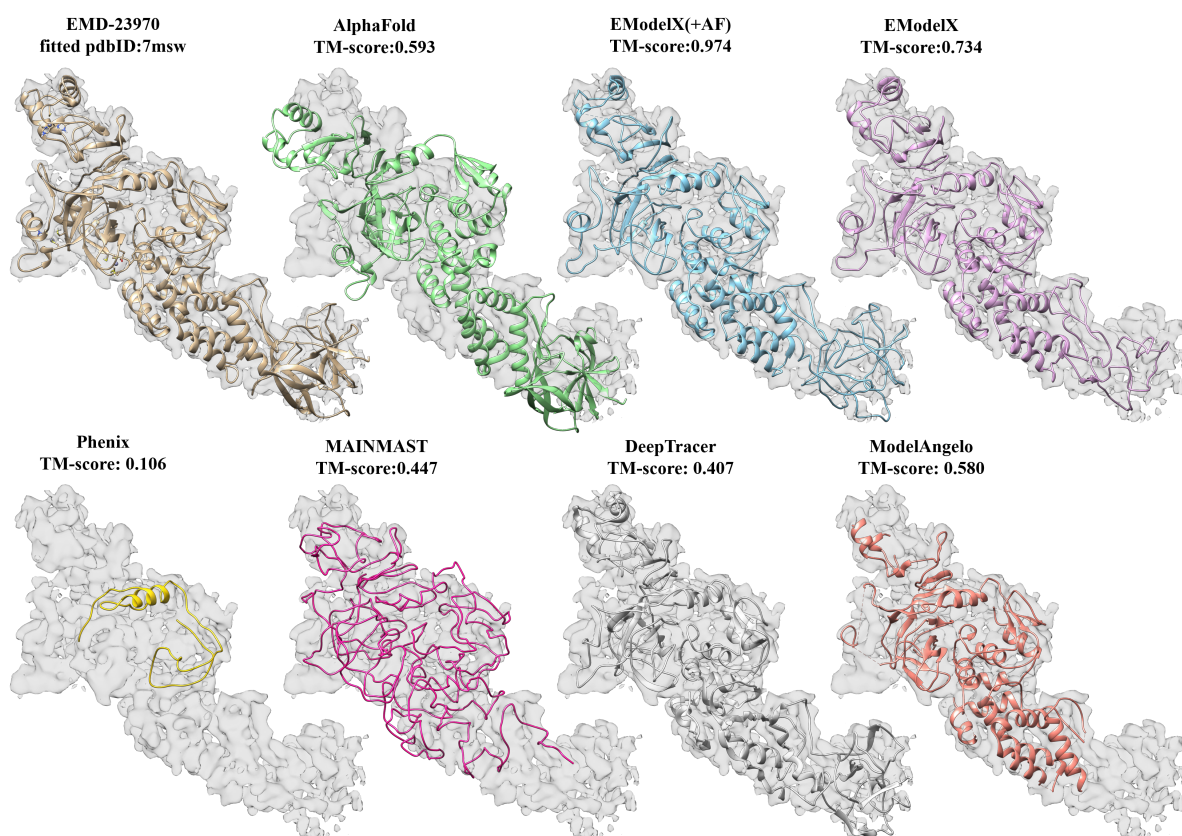


Fig. S11. Map EMD-23970 of Full length SARS-CoV-2 Nsp2 at the resolution of 3.8Å with the deposited pdb structure (pdbid: 7msw) and models built by EModelX and compared methods.

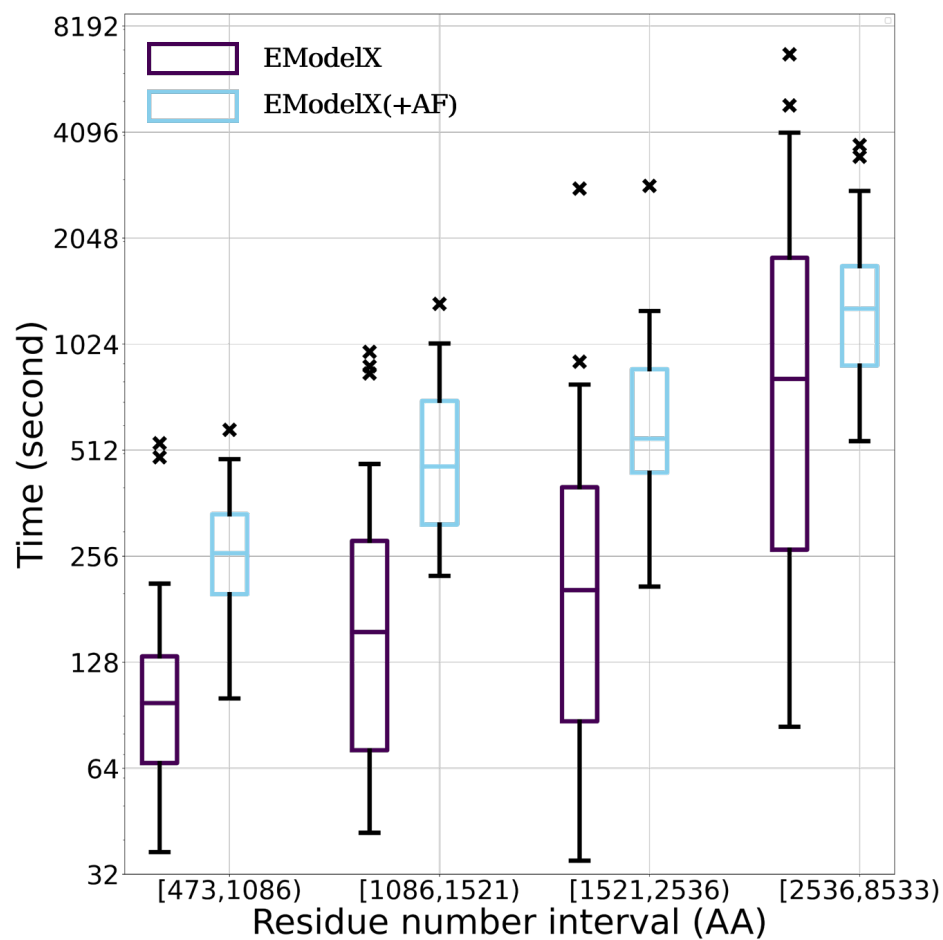


Fig. S12. The boxplot of the time cost by EModelX computation in different residue number ranges. The box extends from the first quartile (Q1) to the third quartile (Q3), with a line at the median. The whiskers extend from the box to the farthest data point lying within $1.5 \times (Q3 - Q1)$ from the box. Flier points are those past the end of the whiskers.

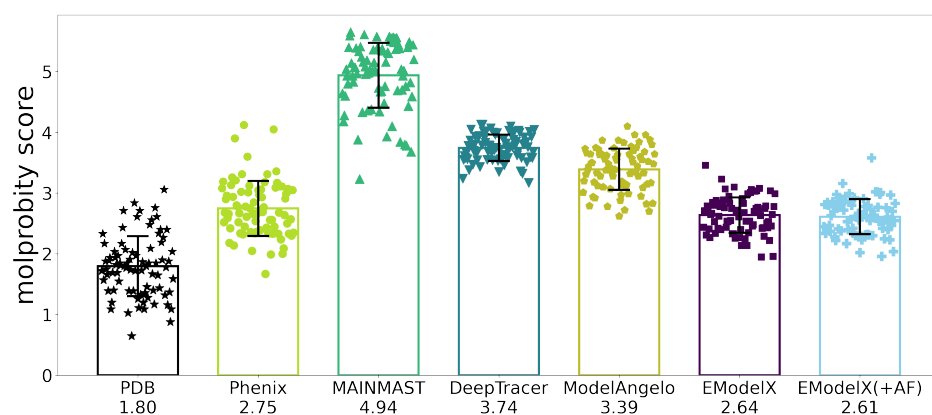


Fig. S13. The average molprobability scores on the test cases of 80 Cryo-EM maps (19 maps were excluded since their MAINMAST models failed to compute the molprobability scores). Error bars indicate ± 1.0 standard deviation. The molprobability scores were computed by *phenix.molprobability*(15).

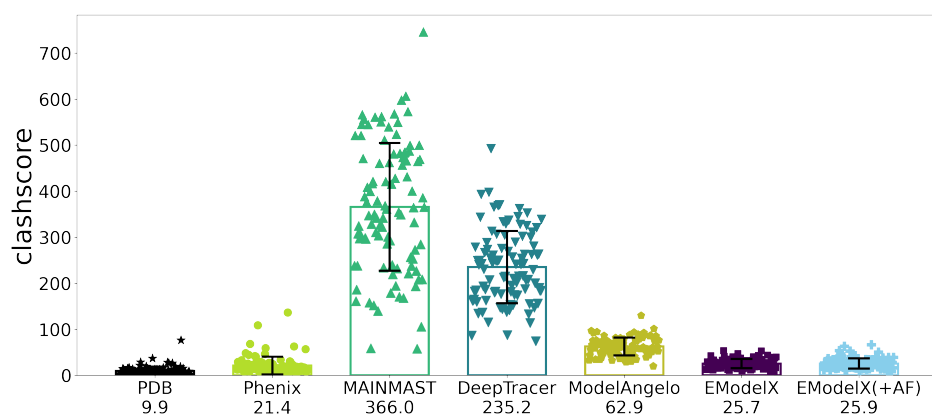


Fig. S14. The average clash scores on the test cases of 99 Cryo-EM maps. Error bars indicate ± 1.0 standard deviation. The clash scores were computed by *phenix.clashscore*(15).

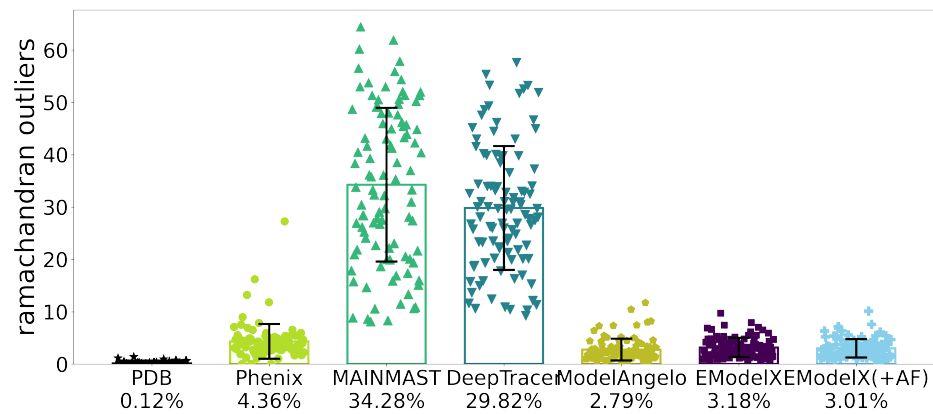


Fig. S15. The average rates of ramachandran outliers on the test cases of 99 Cryo-EM maps. Error bars indicate ± 1.0 standard deviation. The ramachandran outliers were computed by *phenix.ramalyze*(15).

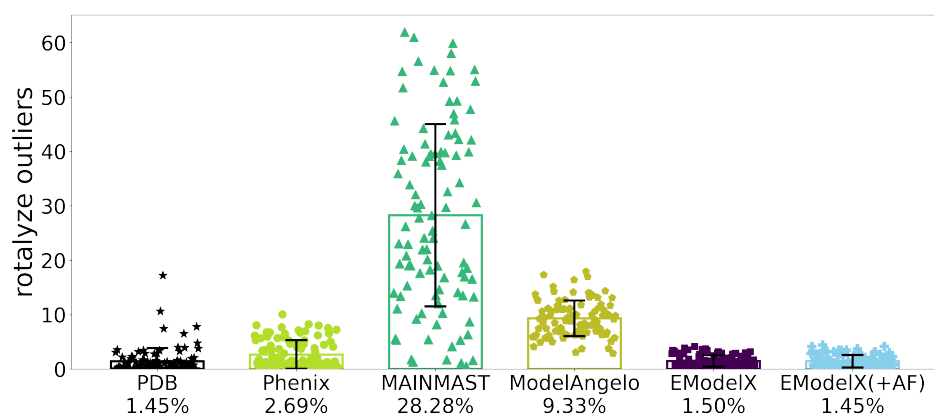


Fig. S16. The average rates of rotamer outliers on the test cases of 99 Cryo-EM maps. Error bars indicate ± 1.0 standard deviation. The rotamer outliers were computed by *phenix.rotalyze*(15).

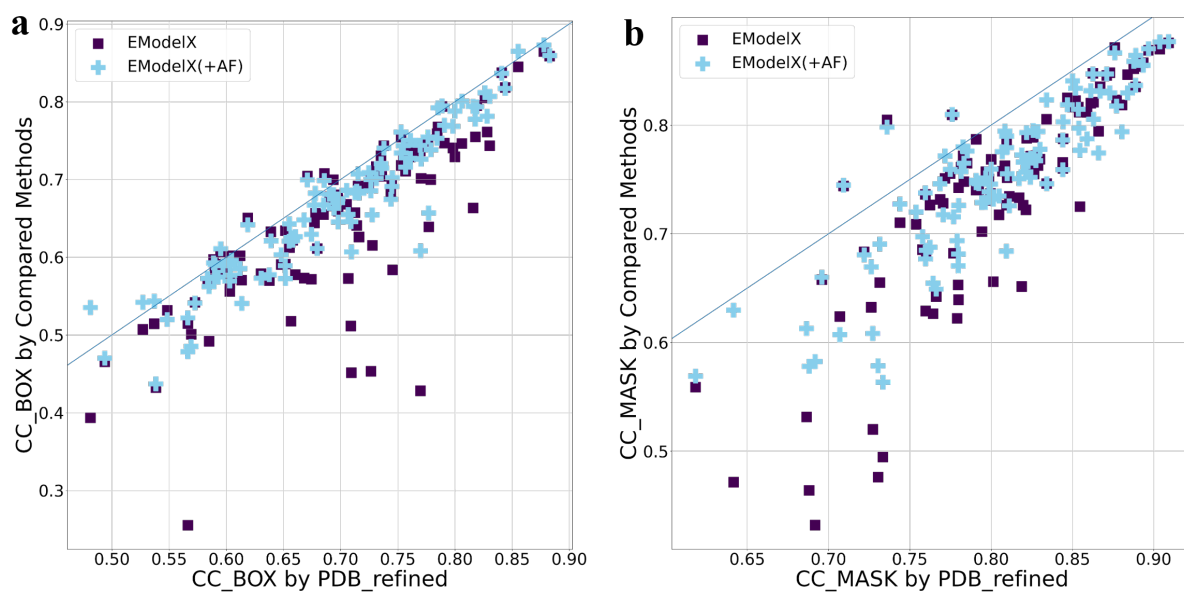


Fig. S17. Comparison of the (a) CC_box and (b) CC_mask between EModelX, EModelX(+AF), and the PDB structures re-refined by phenix.real_space_refine using the same restraints. Re-refining of the PDB structure improved the average CC_box (0.687→0.698) and CC_mask (0.780→0.803). EModelX and EModelX(+AF) still yielded comparable results with the PDB structures.

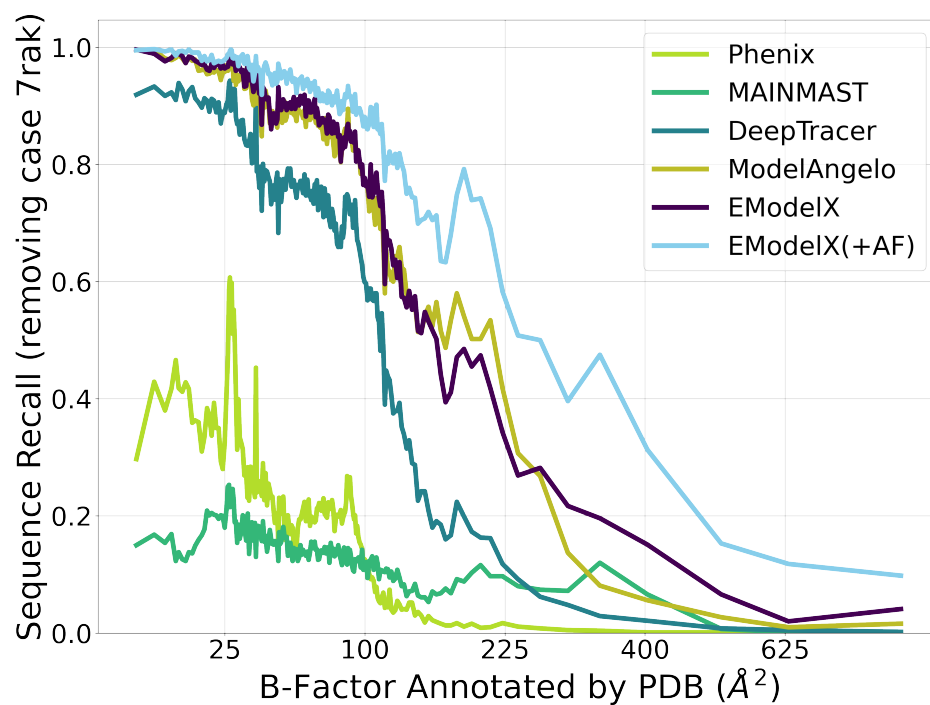


Fig. S18. The sequence recall in a subset (excluding 7rak) as a function of the B-factor. Each data point represents the average sequence recall on a B-factor interval that contains 1,000 residues. Compared to Fig. 2b in the main text, the sequence recall curves turn smoother by removing case 7rak.

Table S1. The average performance of neural network prediction

<i>Method</i>	<i>False Positive^a</i>	<i>Cα in 3Å^b</i>	<i>AA acc^c</i>
EModelX	10.7%	98.1%	48.4%

^apercentage of C α atoms in the predicted model that are more than 3Å away from any native C α atom.

^bpercentage of native C α atoms within 3Å of its closest predicted C α .

^camino acid type prediction accuracy on native structure

Table S2. Additional test on maps of lower ($>4\text{\AA}$) resolution

emid	resolution($\text{\AA}$)	pdbid	TM-score by EModelX	TM-score by EModelX(+AF)
11467	4	6zvp	0.255	0.692
31080	4	7eeq	0.845	0.953
11806	4	7ajq	0.070	0.469
11202	4.1	6zgd	0.818	0.919
24618	4.2	7rpo	0.143	0.770
13193	4.2	7p4i	0.616	0.925
31296	4.2	7et2	0.245	0.933
12698	4.8	7o24	0.144	0.473

References

1. CL Lawson, et al., Emdatabank unified data resource for 3dem. *Nucleic acids research* **44**, D396–D403 (2016).
2. W Pitsawong, et al., From primordial clocks to circadian oscillators. *Nature* **616**, 183–189 (2023).
3. S McCarthy, S Gonen, Improved interface packing and design opportunities revealed by cryoem analysis of a designed protein nanocage. *Heliyon* **8** (2022).
4. TF Huddy, et al., Blueprinting extendable nanomaterials with standardized protein blocks. *Nature* **627**, 898–904 (2024).
5. Y Yin, et al., Activation mechanism of the mouse cold-sensing trpm8 channel by cooling agonist and pip2. *Science* **378**, eadd1268 (2022).
6. G Terashi, D Kihara, De novo main-chain modeling for em maps using mainmast. *Nat. communications* **9**, 1–11 (2018).
7. J He, P Lin, J Chen, H Cao, SY Huang, Model building of protein complexes from intermediate-resolution cryo-em maps with deep learning-guided automatic assembly. *Nat. Commun.* **13**, 4066 (2022).
8. S Chen, S Zhang, X Li, Y Liu, Y Yang, Segem: a fast and accurate automated protein backbone structure modeling method for cryo-em in *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. (IEEE), pp. 24–31 (2021).
9. W Kabsch, C Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolym. Orig. Res. on Biomol.* **22**, 2577–2637 (1983).
10. S Mukherjee, Y Zhang, Mm-align: a quick algorithm for aligning multiple-chain protein complex structures using iterative dynamic programming. *Nucleic acids research* **37**, e83–e83 (2009).
11. Y Huang, et al., Environmentally ultrasensitive fluorine probe to resolve protein conformational ensembles by 19f nmr and cryo-em. *J. Am. Chem. Soc.* **145**, 8583–8592 (2023).
12. Y Chen, et al., Structure of the transcribing rna polymerase ii–elongin complex. *Nat. Struct. & Mol. Biol.* **30**, 1925–1935 (2023).
13. ZG Oh, et al., A linker protein from a red-type pyrenoid phase separates with rubisco via oligomerizing sticker motifs. *Proc. Natl. Acad. Sci.* **120**, e2304833120 (2023).
14. AU Mueller, et al., A general mechanism for transcription bubble nucleation in bacteria. *Proc. Natl. Acad. Sci.* **120**, e2220874120 (2023).
15. PV Afonine, et al., New tools for the analysis and validation of cryo-em maps and atomic models. *Acta Crystallogr. Sect. D: Struct. Biol.* **74**, 814–840 (2018).