


Accurate flexible refinement for atomic-level protein structure using cryo-EM density maps and deep learning

Biao Zhang, Dong Liu, Yang Zhang, Hong-Bin Shen and Gui-Jun Zhang 

Corresponding authors: Gui-Jun Zhang, College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China. E-mail: zgj@zjut.edu.cn; Hong-Bin Shen, Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, and Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai, China; Department of Computer Science, Shanghai Jiao Tong University, Shanghai, China. E-mail: hbshen@sjtu.edu.cn

Abstract

With the rapid progress of deep learning in cryo-electron microscopy and protein structure prediction, improving the accuracy of the protein structure model by using a density map and predicted contact/distance map through deep learning has become an urgent need for robust methods. Thus, designing an effective protein structure optimization strategy based on the density map and predicted contact/distance map is critical to improving the accuracy of structure refinement. In this article, a protein structure optimization method based on the density map and predicted contact/distance map by deep-learning technology was proposed in accordance with the result of matching between the density map and the initial model. Physics- and knowledge-based energy functions, integrated with Cryo-EM density map data and deep-learning data, were used to optimize the protein structure in the simulation. The dynamic confidence score was introduced to the iterative process for choosing whether it is a density map or a contact/distance map to dominate the movement in the simulation to improve the accuracy of refinement. The protocol was tested on a large set of 224 non-homologous membrane proteins and generated 214 structural models with correct folds, where 4.5% of structural models were generated from structural models with incorrect folds. Compared with other state-of-the-art methods, the major advantage of the proposed methods lies in the skills for using density map and contact/distance map in the simulation, as well as the new energy function in the re-assembly simulations. Overall, the results demonstrated that this strategy is a valuable approach and ready to use for atomic-level structure refinement using cryo-EM density map and predicted contact/distance map.

Keywords: protein structure refinement, cryo-EM density map, contact/distance map, energy function

Introduction

With the development of cryogenic electron microscopy (cryo-EM) technology and deep learning, structural biology has experienced rapid progress in recent years, especially in creating and refining the high-accuracy of atomic-level structural models by using cryo-EM density maps and contact map [1–4]. In general, the high-resolution protein structures are important to understand the molecular interactions that give insights to their biological functions [5]. For example, membrane proteins are functional proteins that transport small molecules from outside of the membrane to the inside of the membrane or receive signals from outside the

cell and active an intracellular process. In the electron microscopy data resource (EMDR), 16 759 density map entries were released but only 8737 PDB coordinate entries were released, illustrating that ~48% of the EMDR density map does not have a corresponding atomic structure [6]. Therefore, robust computational methods have raised an urgent need to create and refine atomic structure models.

For decades, considerable efforts have been made to address this challenge, including cryo-EM technology and deep-learning technology. In cryo-EM technology, Cryo-EM has been developed for many years and has shown great progress in recent years due to the advancement

Biao Zhang is an assistant professor in the College of Information Engineering, Zhejiang University of Technology. His research interests include bioinformatics, cryo-EM density map and optimization theory.

Dong Liu is a MS candidate in the College of Information Engineering, Zhejiang University of Technology. His research interests include bioinformatics, intelligent information processing and optimization theory.

Yang Zhang is a professor in the Department of Computational Medicine and Bioinformatics, University of Michigan. His research interests include bioinformatics, protein structure prediction and protein design.

Hong-Bin Shen is a professor in the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, and Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai, China. His research interests include bioinformatics, intelligent information processing and image processing.

Guijun Zhang is a professor in the College of Information Engineering, Zhejiang University of Technology. His research interests include bioinformatics, intelligent information processing and optimization theory.

Received: November 5, 2021. **Revised:** December 26, 2021. **Accepted:** January 17, 2022

© The Author(s) 2022. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

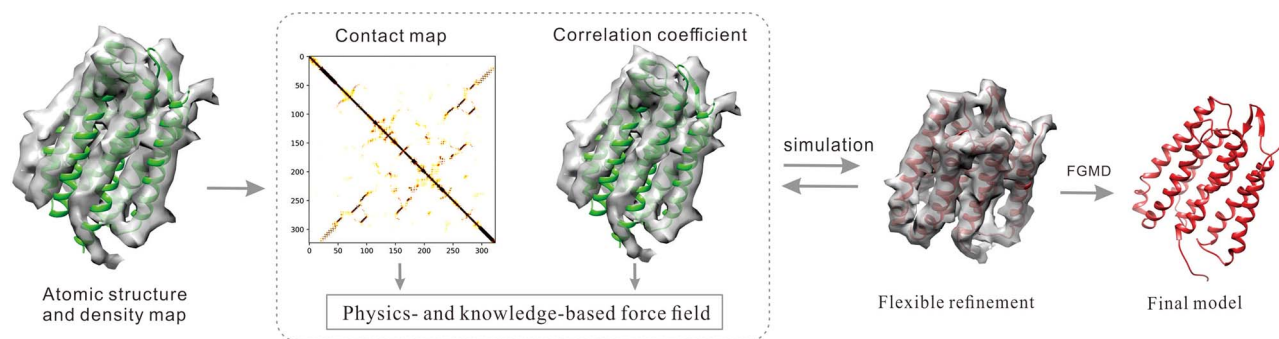


Figure 1. Flowchart of EMCMR, which refines predicted protein structures generated by C-I-TASSER using cryo-EM density map and contact/distance map.

in electron detector technology and image-processing techniques [7]. The resolution of the density map has increased due to the cryo-EM technology, making it possible to be solved at near-atomic resolution. However, some parts of density maps still stayed in the intermediate range (5–10 Å). How to derive a high-resolution atomic structure from the cryo-EM is still an urgent problem that needs to be solved in the field. Some computational methods have been developed for this task, including elastic network models-based refinement [8, 9], fragment assemble-based refinement [10, 11] and Bayesian-based refinement [12]. Even though these methods have promoted the development of structural construction and refinement on the basis of cryo-EM density map, some challenges still need further efforts, such as the absence of long-range correlation between the density map and predicted model, insufficient sampling of the conformational space, and etc. The absence of long-range correlation limits the flexible structure refinement simulations because the movement is strongly guided by the correlation coefficient (CC) score. Protein structure prediction has shown great progress due to the recent deep-learning technology. The predicted contact/distance map is driven by deep-learning technology and enhances protein structure prediction, especially shown by the results in the CASP14 [13]. Many approaches, such as TripletRes [14], RaptorX [15] and DeepContact [16], were developed to address the protein inter-residue contact/distance predicted. In the contact/distance map, a long-range correlation was provided among the residues. This correlation is absent in the CC score. The development of cryo-EM and contact/distance map prediction has played a role in promoting the development of structural biology, respectively, but the joint use of the two technologies to improve the accuracy of released proteins is still an urgent problem to be solved. Therefore, how to effectively use density map and predicted contact/distance map and design a new folding strategy is crucial to improve the refinement accuracy in this paper.

In this work, a protein structure refinement protocol (EMCMR, the cryo-EM density map and Contact Map for protein structure Refinement) to membrane protein was proposed, where the pipeline is depicted in Figure 1. By using a new energy function based on a combination

of the cryo-EM density map and predicted contact distance map (CM), most of the predicted protein structures were further refined. In the protocol, a strategy to use density map and contact/distance map for different refinement situations in accordance with the confidence score was designed. The density map could dominate the movement and guide protein folding in the simulation when the density map and initial atomic model have a higher confidence score. Otherwise, the predicted contact/distance map could dominate the movement and guide protein folding in the simulation. For examination of the preference of this method, a large-scale benchmark test was performed on 218 non-redundant membrane proteins with density maps created from both noise-free simulations and collected from cryo-EM experiments. The results demonstrated the advantages of this method for the refinement of predicted protein structure on the basis of the cryo-EM density map and contact/distance map.

Materials and methods

The CC of each residue between the atomic structure and density map

The CC of each residue between the atomic structure and density map is very important to the simulation. It is the key point in the calculation of the confidence of protein fragments and locating the bad matching area. The formula of CC of each residue could be defined as follows:

$$CC(res) = \frac{\sum_{y \in Res(l)} (\rho_o(y) - \bar{\rho}_o) (\rho_c(y) - \bar{\rho}_c)}{\sqrt{\sum_{y \in Res(l)} (\rho_o(y) - \bar{\rho}_o)^2} \sqrt{\sum_{y \in Res(l)} (\rho_c(y) - \bar{\rho}_c)^2}}$$

where $Res(l)$ represents the grid point set in diffraction space for each residue; $\bar{\rho}_c$ and $\bar{\rho}_o$ are the average values of calculated density map and experimental density map; y represents the corresponding grid point and $\rho_c(y)$ is the calculated density value for given atomic structure and could be calculated as,

$$\rho_c(y) = \sum_{x_i \in N} C \cdot \exp(-k \cdot \|x_i - y\|^2),$$

where $k = (\pi/(2.4 + 0.8R_0))^2$ and $C = a \bullet (k/\pi)^{1.5}$ are the parameters to describe the shape of Gaussian kernel; a is the mass of an atom in the residue and R_0 is the resolution of the density map and x_i is the atom coordinate [17].

Contact/distance potential

The predicted inter-residue distance provides abundant spatial constraint information for protein folding. It could guide protein folding when the CC has a lower value (<0.05) between the density map and predicted atomic structure in the simulation stage. In this study, the distance is constructed as a potential to improve the accuracy of refinement. The specific formula could be defined as follows:

$$E_{cp} = \sum_i^N \log \left((\|x_i - y_i\|_2 - \mu)^2 + 1 \right) / \sigma, \quad (1)$$

where x_i and y_i are the atomic coordinates of the i th pair of residues in the simulated stage; N is the number of effective inter-residue distances; μ is the mean value obtained by Gaussian fitting of the i th pair of residues distance distribution and σ is the standard deviation for the i th pair of residues distance distribution [18].

Confidence score of fragments to guide protein folding

Estimating the confidence score of fragments between the density map and superposed protein structure in the refinement is important. The confidence score of fragments could guide protein folding in different situations. The confidence score of fragments includes two parts. One part is S_{cc} , which measures the confidence score between the density map and the protein structure. The density map could guide the movement when the high confidence score S_{cc} exists in the density map and the selected fragment of protein structure. The CC could also play a major role in the energy function and dominate the protein folding in the simulation. The second part is S_{cp} , which represents the confidence score of the predicted contact/distance map by deep-learning technology. The predicted contact/distance map could guide protein folding when the high confidence score S_{cp} exists in the protein structure. The process of protein folding could be adjusted by the confidence score. The confidence score could contribute to the success rate of movement when the initial structure has a lower TM-score (<0.5). The main reason is that the quality of the fragment of the model is proportional to the confidence score of the protein structure. The specific calculation formula of confidence score is as follows:

$$S_{cc} = \left(\frac{pnum - ncg}{pnum} \right) \quad (1)$$

$$S_{cp} = \left(\frac{pnum}{pnum - ncg} \right) \quad (2)$$

where $pnum$ is the number of the selected fragment in the simulation and ncg is the number of the residue that has $CC(i) < 0$ between the density map and selected fragment in the simulation.

Conformational sampling

Although fragment assembly has greatly promoted conformational sampling and reduced the sampling space, conformational sampling remains a challenging problem. It easily generated insufficient or ineffective sampling in the matching region between the density map and atomic structure at the iteration process. Crossover is one of the sources of power for a population in nature. The sampling process of the crossover was introduced to increase the effectiveness and sufficient of sampling in our program. Crossover sampling was the fragment selection process in accordance with CC scores of residues at the entire sequence in our protocol. The fragment of atomic structure with low CC scores and high CC scores were cross-selected in the fragment sampling process. For example, the poor local region and matching region were determined in accordance with CC scores of residues at first in the iteration process. Then the poor local region was sampled after each N ($N=6$) matching region sampling at the refinement (cross-sampling). This cross-sampling method could not only improve the sufficient or effective refinement in the poor local region but also prevent over-sampling in the poor local region. Moreover, it also prevents the entire structure from falling into the local optimum at the refinement.

EMCMR force field

The energy function of EMCMR consists of seven energy terms for protein structure refinement. Physics- and knowledge-based energy functions, integrated with Cryo-EM density map data and predicted contact/distance map through deep learning, were used to optimize the protein structure in the simulation. The dynamic confidence score was introduced to the iterative process for choosing whether it is a density map or a contact/distance map to dominate the movement in the simulation to improve the accuracy of refinement. The specific energy function can be found in the EMCMR force field from the Supplementary Data available online.

Result

Benchmark results using simulated density maps and experimental density maps

Benchmark dataset building

A comprehensive set of benchmark membrane proteins from the OPM (orientations of proteins in membranes) database was collected to examine the preference of EMCMR [19]. A total of 218 single-chain proteins with sequence lengths ranging from 100 residues to 620 residues and pair-wise sequence identities $<30\%$ were selected from 4231 membrane proteins. Simulated

noise-free density maps were generated from the target structures by using EMAN2 (pdb2mrc), where the resolution for each membrane protein varied from 3.0 to 10.0 Å [20]. Here, the resolution of density maps ranging from 3.0 to 10.0 Å was matched to the recent progress in the electron microscopy data resource (EMDR; [6]). In Table S1 from the Supplementary Data, the resolution parameters for all membrane proteins are listed, where a histogram is given in Figure S1, Supplementary Data available online. One reason for choosing the range of 3–10 Å is that it represents a typical resolution range of the released experimental structure in the EMDR data bank in recent years. Another reason is that structural modeling tools are most needed, as high-resolution structure determination was difficult to create from the density maps alone, although the average resolution of the cryo-EM data in the community has kept improving in the last years [6].

Since the proposed protocol focuses on protein structure refinement and relies on the initial model, the initial structure models were obtained by C-I-TASSER [21] (a hierarchical approach to protein structure prediction), for which all homologous templates with sequence identities $\geq 30\%$ to the query were excluded using CD-HIT [22]. Initial membrane protein models were generated with the average TM-score of 0.732, where 198 out of the 218 targets had the correct folding with TM-scores > 0.5 and 20 targets had incorrect folding with TM-scores < 0.5 [23]. Here, TM-score is a metric used to measure the similarity between two protein structures [24]. The value of TM-score falls in the range (0,1], where a TM-score of 1 indicates a perfect match between two structures and a value ≥ 0.5 indicates that two structures share the same fold [23]. A more detailed TM-score distribution of each target is displayed in Figure S2, Supplementary Data available online. Here, the assessments of the initial and final model quality were mainly based on TM-score [24], since it is more sensitive than RMSD to the topological similarity of protein structures.

Parameter setting

For estimation of the efficiency of the proposed strategy and excluding the effect of superposition on refinement, the superposed structure was used as input to the program which was also used in Rosetta refinement based on the density map [17]. The density map and its resolution are necessary to the program. The format of the contact/distance map file that included residue id, the average value and the standard value of distance is also necessary to the program (see Figure S3 in the Supplementary Data available online). In our protocol, the contact/distance map was predicted by trossetta [25]. The residue pairs of movement fragments were selected to use in the energy function calculation for each iteration step of the refinement. However, considering there were some potential errors in the predicted contact/distance map, the residue pairs combined with the density map

Table 1. Summary of modelling results by I-TASSER structure prediction and the follow-up EM density map refinement methods on 218 test proteins

Methods	TM-score (med*)	RMSD (Å)	#TM-score < 0.5	P-value
C-I-TASSER	0.732 (0.770)	7.59	20	$6.71 * 10^{-44}$
Rosetta*	0.794 (0.840)	5.60	14	$5.06 * 10^{-6}$
EM-Refiner	0.812 (0.856)	5.31	14	$3.0 * 10^{-2}$
EMCMR	0.818 (0.870)	5.04	10	–

The P-values were calculated using two-tail Student's t-tests between the TM-scores produced by EMCMR and the other programs. Med*: the medium value. Rosetta*: Rosetta refinement using density map.

were integrated into the energy function to determine whether it was residue pairs of contact/distance map or a density map to dominate the movement in the simulation. EMCMR is a protein structure optimization program that is based on a rapid replica-exchange Monte Carlo (REMC) simulation. In the REMC simulation, 30 replicas (decoys) were performed, and each replica involves 50 temperatures ranging from $KT = 3.0$ to $KT = 0.01$ with 500 MC movements attempted at each temperature. The model was finally selected from the conformation with the lowest energy in the REMC simulation.

Performance of EMCMR on structure refinement

The average TM-score and RMSD of EMCMR were 0.818 and 5.04 Å in 218 test proteins, respectively. The TM-score and RMSD of EMCMR in 218 test proteins were shown in Figure S4, Supplementary Data available online and Figure 2. The TM-score of EMCMR was higher than that of the initial C-I-TASSER model (0.732), the difference of which corresponded to a P-value of $6.71E-44$, as determined by a two-tailed Student's t-test. The average RMSD of EMCMR also decreased by 33.60% from 7.59 Å (C-I-TASSER) to 5.04 Å (EMCMR), corresponding to a P-value of $1.57E-3$. The specific statistic results are summarized in Table 1. Figure 2 showed the TM-score distribution of structure models obtained by EMCMR (blue), EM-Refiner (green), Rosetta (red) and C-I-TASSER (black) on 218 test proteins, which had the median TM-score of 0.865, 0.856, 0.840, 0.768, respectively. These results demonstrated the performance of the proposed method on the basis of the density map and contact/distance map.

Figure S4 (Supplementary Data available online) presents a head-to-head comparison of the TM-score and RMSD value between the models produced by C-I-TASSER and EMCMR. For the initial model with correct folds [23] (which largely corresponds to models with TM-score > 0.5), the average TM-score increased by 10.40% from 0.770 (C-I-TASSER) to 0.853 (EMCMR) on 198 membrane protein models, corresponding to a P-value of $2.54E-47$. For initial models with incorrect folds, the improvement was significant, with a P-value = $2.25E-3$ between the EMCMR and C-I-TASSER models, in which the average TM-score increased by 30.48% from 0.361 (C-I-TASSER) to 0.471 (EMCMR) on the 20 membrane protein models. This finding was mainly due to a new energy function of EMCMR that combines with the density map

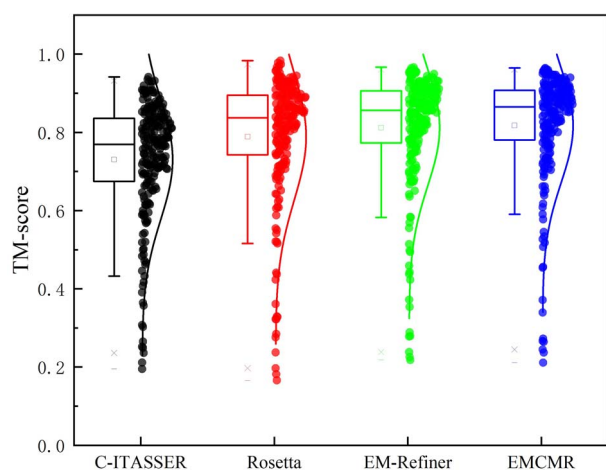


Figure 2. The TM-score distribution of structure models was obtained by EMCMR (blue), EM-Refiner (green), Rosetta (red) and C-I-TASSER (black) on 218 test proteins.

and the predicted contact/distance map. When initial models have incorrect folds, the density map could not guide the atomic protein to fold correctly due to the unreliable energy field between the density map and initial models. The confidence score (Methods section) could be calculated in accordance with the CC between the density map and atomic structure and the predicted contact/distance map could dominate the protein folding in this situation. Overall, these results suggested that the inherent knowledge-based force field of EMCMR is capable of refining the structural model for some small proteins, when coupled with cryo-EM density map and predicted contact/distance map.

Control results compared with Rosetta

As a control, the other commonly used density map-based refinement method, Rosetta [17], which is a version of the Rosetta refinement on the basis of the density map, was compared with EMCMR. In the stages of the Rosetta refinement, the same density map potential as EMCMR was used to guide the fragment folding together with the other energy function. The final model was generated by 300 independent trajectories with the parameter `-default_max_cycles = 300` and refinement protocol file and other parameters were set up following the tutorial instructions. The same C-I-TASSER models as a starting point were applied to EMCMR and Rosetta to ensure a fair comparison. The refinement results of EMCMR and Rosetta on the benchmark dataset are summarized in Table 1.

The average TM-score and RMSD of Rosetta were 0.794 and 5.59 Å, respectively, as shown in Figure S5, Supplementary Data available online and Figure 2. Compared with that of Rosetta, the average RMSD of the model produced by EMCMR decreased by 10.00% from 5.60 Å (Rosetta) to 5.04 Å (EMCMR), corresponding to a P -value of $5.7E-4$. The average TM-score was increased by 3.02% from 0.794 (Rosetta) to 0.818 (EMCMR), corresponding to a P -value of $1.65E-5$, as shown in Table 1. Ten out of

the 20 targets with incorrect folding (TM-scores < 0.5) were refined and they generated the correct folding (TM-scores > 0.5) by EMCMR, corresponding to a 50.0% growth rate. However, 6 out of the 20 targets that had incorrect folding were refined and they generated the correct folding by Rosetta, corresponding to a 30.0% growth rate, as shown in Table 1. This finding illustrated that EMCMR has better performance than Rosetta in hard targets with TM-scores < 0.5 . The main reason was that EMCMR regarded the contact/distance map as a constraint and adjusted the confidence score between the density map and the contact/distance map in the iterative process to ensure protein folding at the correct time. Overall, these results suggested that EMCMR could outperform Rosetta on the benchmark dataset.

Comparison with EM-Refiner

EM-Refiner is a Monte Carlo-based method for protein structure refinement and determination using Cryo-EM density map [20]. During the refinement simulations, the backbone structures kept flexible movement and were guided by a composite of physics- and knowledge-based force fields, integrated with Cryo-EM density map data. The difference of EMCMR with EM-Refiner was the contact distance map, which was added into the energy function to guide protein folding and the concept of confidence score, which was introduced to determine whether it is a contact distance map or a density map to dominate the movement in the simulation. The introduction of the confidence score promoted the efficiency of movement in the simulation and improved the quality of the refinement structure, especially for the initial model with incorrect folds (Figure 4 and Figure S7 in the Supplementary Data available online). The comparison results of EMCMR and EM-Refiner on the benchmark dataset are summarized in Table 1.

The average TM-score and RMSD of EM-Refiner were 0.812 and 5.31 Å, respectively, as shown in Figure S6, Supplementary Data available online and Figure 2. Compared with that of EM-Refiner, the average RMSD of EMCMR was decreased by 5.08% from 5.31 Å (EM-Refiner) to 5.04 Å (EMCMR), corresponding to a P -value of $3.7E-2$. The average TM-score increased 0.7% from 0.812 (Rosetta) to 0.818 (EMCMR), corresponding to a P -value of $3.3E-2$. Six out of the 20 targets with incorrect folding were refined by EM-Refiner and they generated the correct folding, corresponding to a 30.0% growth rate, as shown in Table 1. A similar result was found in Rosetta (30.0% growth rate) and it was lower than that in EMCMR (50.0% growth rate). These results suggested that EMCMR could outperform EM-Refiner and has a better performance than EM-Refiner for an initial model with incorrect folds in the refinement (such as Rosetta).

On average, although both of the control methods had a good performance at refining the initial C-I-TASSER models, the improvement by EMCMR was larger than that by both programs, where the P -values between

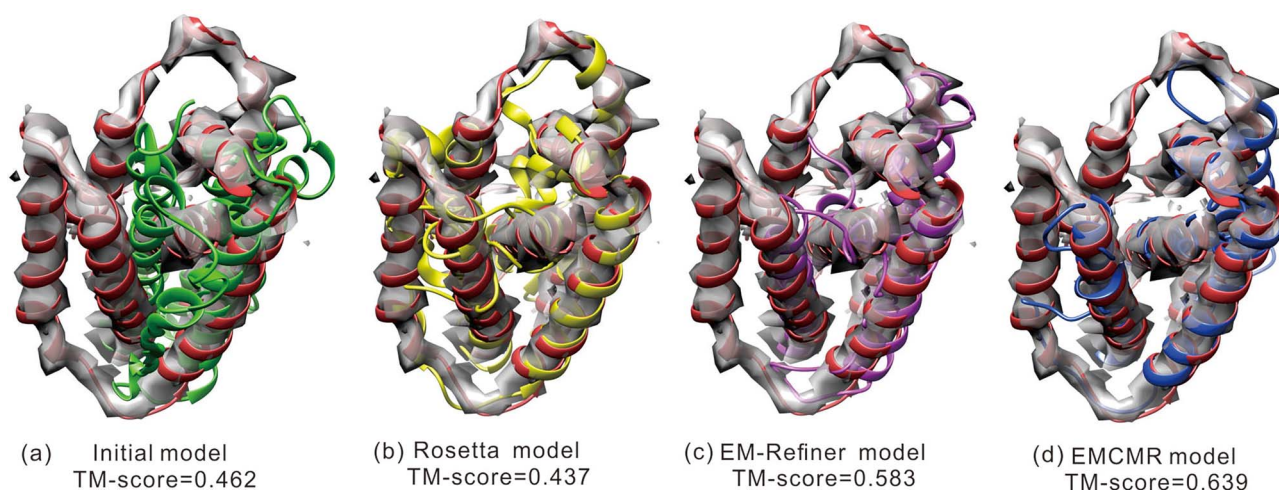


Figure 3. Illustrative example of refinement using simulated experimental data at a resolution of 6.39 Å for the chain A of 4PGR. Cartoons show the initial model from C-I-TASSER overlaid on the density map (A), Rosetta refined model overlaid on the density map (B), EM-Refiner refined model overlaid on the density map (C) and the EMCMR model overlaid on the density map (D). The EMCMR model had a TM-score of 0.639, an RMSD of 8.64 Å to the native structure (red) and a CC of 0.723 with the density map, compared with a TM-score of 0.437, an RMSD of 10.41 Å and a CC of 0.709 for the Rosetta model and a TM-score of 0.583, an RMSD of 9.51 Å and a CC of 0.723 for the EM-Refiner refined model.

EMCMR and the control methods were all below 0.05, showing that the differences were statistically significant (Table 1). Figure 3 shows an illustrative example of a single chain of *Bacillus Subtilis* at pH 8 from the Human Bax inhibitor (hBI-1; PDBID: 4pgrA) [26], where the resolution of its simulated density map was 6.39 Å. For this target, C-I-TASSER built an initial structure with an incorrect fold (TM-score = 0.462, RMSD = 10.58), as shown in Figure 3A. After the fragment adjustment, EMCMR created a significantly refined model with a TM-score of 0.639 and an RMSD of 8.64 Å, where the CC increased to 0.723, as shown in Figure 3D. The major improvements by EMCMR were in the regions with poor local CC scores, where the fragment adjustment procedure identified and correctly adjusted these regions into the density map in accordance with the density map and contact/distance map. For this same target, the Rosetta model had a TM-score of 0.437, a CC of 0.709 and an RMSD of 10.41 Å in Figure 3B, whereas the EM-Refiner model had a TM-score of 0.583, a CC of 0.719 and an RMSD of 9.51 Å in Figure 3C. In Figure 4, we present the TM-score and RMSD of the structural decoy conformations generated by EMCMR simulations (blue circle) versus EM-Refiner simulations (orange square) in 4pgrA protein. Although the TM-score and RMSD of the structural decoy conformations were both becoming better with the iteration step (as shown in Figure S7, Supplementary Data available online), EMCMR sufficiently outperformed EM-Refiner. This was mainly due to the contact/distance map that guided protein folding in EMCMR simulation, whereas it was absent in EM-Refiner simulation. Overall, this example highlighted the importance and effectiveness of the refinement procedure utilized by EMCMR, which helps achieve a more significant model improvement based on the EM-density data and contact/distance map over the control methods.

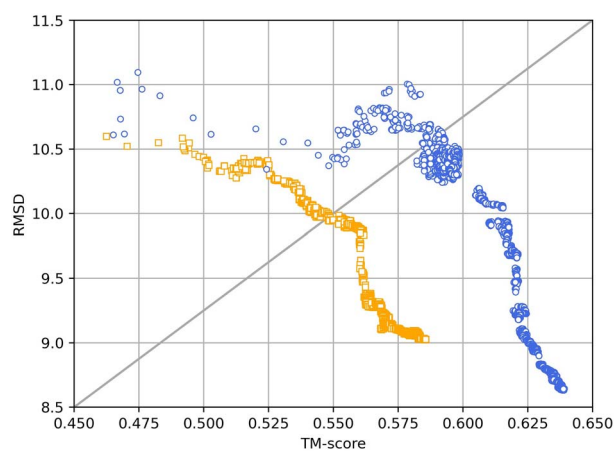


Figure 4. The TM-score and RMSD of the structural decoy conformations generated by EMCMR simulations (blue circle) versus EM-Refiner simulations (orange square) in the Human Bax inhibitor (hBI-1) (PDBID: 4pgrA). EMCMR refined model with a TM-score of 0.639 and an RMSD of 8.64 Å is better than the EM-Refiner refined model with a TM-score of 0.583 and an RMSD of 9.51 Å.

A comparison experiment was added to evaluate the robustness of EMCMR using a real contact/distance map and predicted contact/distance map on the simulated density maps. The actual contact/distance map could be produced from the actual structure with the distance among the residue pairs <20 Å, and the predicted contact/distance map can be produced by trossetta [25]. The average TM-score and RMSD of EMCMR with the actual contact/distance map were 0.862 and 3.91 Å, respectively, which was significantly higher than EMCMR with the predicted contact/distance map with the P-values = 9.62E-36. Figure 5 showed the comparison results of EMCMR with the predicted contact/distance map or the actual contact/distance map. The result illustrated that a highly accurate contact/distance map would contribute to building the high-accurate

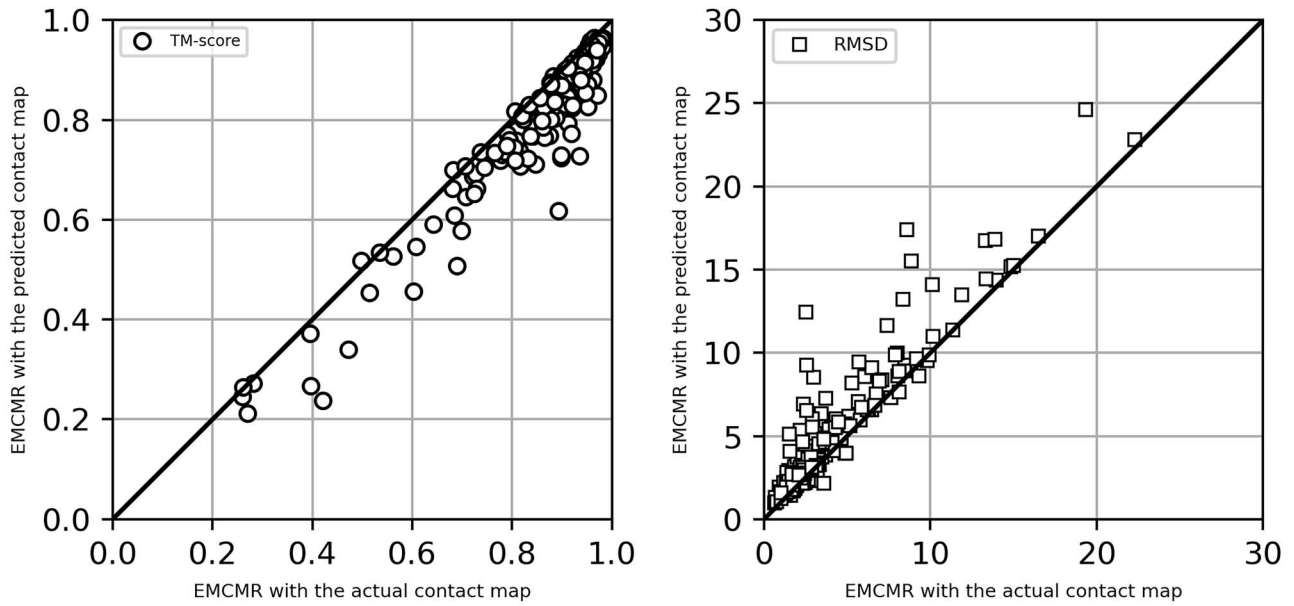


Figure 5. Comparison of results obtained by EMCMR with the actual contact/distance map and EMCMR with the predicted contact/distance map. Left and right figures are the TM-score and RMSD to the native structure, respectively.

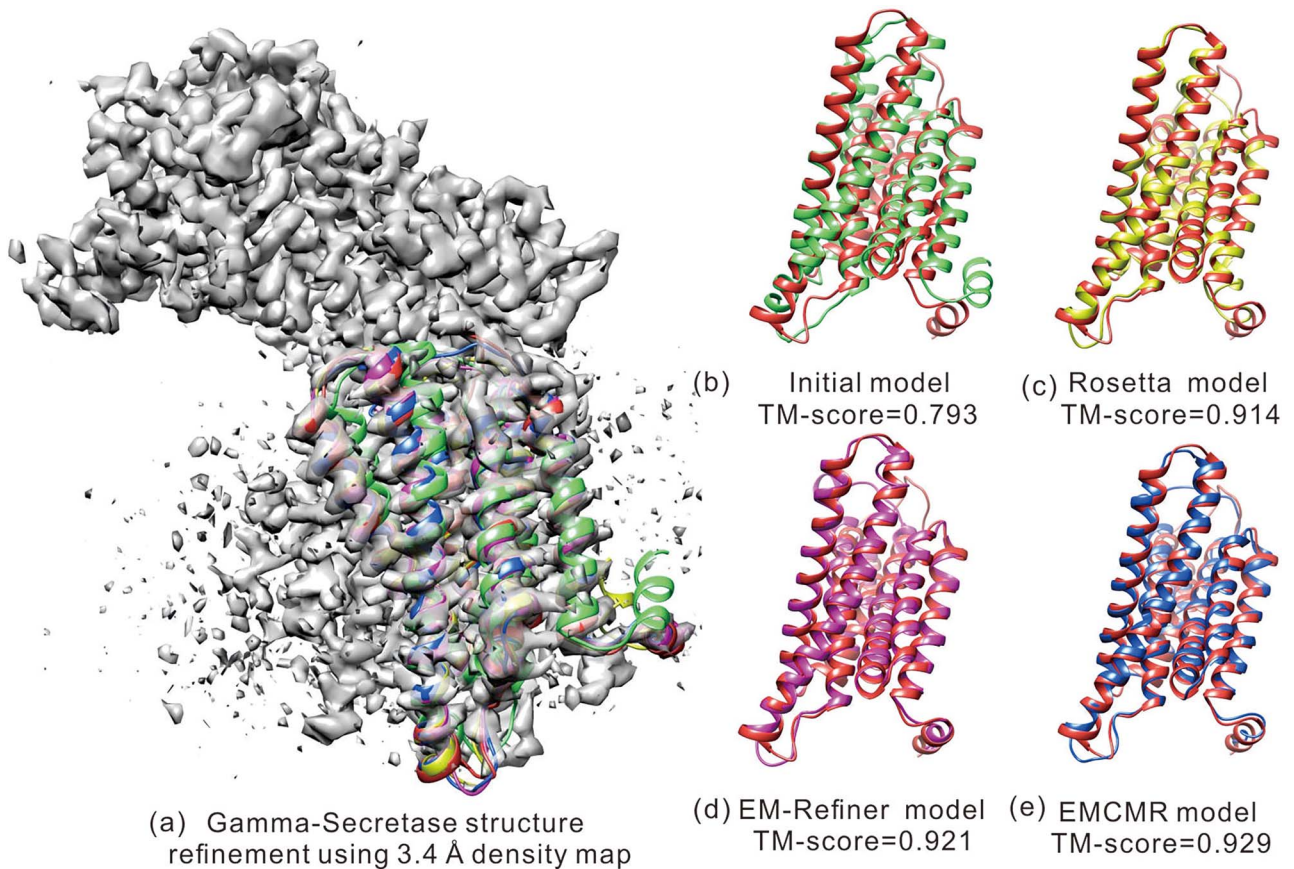


Figure 6. Refinement of a single chain of the γ -secretase protein based on a 3.4 Å experimental EM density map (A). The refined model by EMCMR (E) is overlaid onto the density map with a CC of 0.615 and a TM-score of 0.929 to the native structure (red). As a comparison, the initial C-I-TASSER model (B), the Rosetta refined model (C) and the EM-Refiner refined model (D) had CCs of 0.594, 0.601 and 0.615, respectively and TM-scores of 0.793, 0.914 and 0.921, respectively.

model in our refinement. For the refined model with TM-score < 0.5 after using the actual contact/distance map, the potential reason for this result was that

the final refined model was not only affected by the contact/distance map, but also by other factors, such as the resolution of density map, energy barrier and

Table 2. Summary of modeling results by C-I-TASSER structure prediction and the follow-up EM density map and contact/distance map refinement methods on 14 proteins with experimentally determined experiment maps

Protein Name	Methods	Resolution (Å)	Predicted structure		Refined structure	
			TM-score	RMSD	TM-score	RMSD
EMD10122_F	Rosetta	3.7	0.857	3.23 Å	0.878	3.11 Å
	EM-Refiner				0.903	2.77 Å
	EMCMR				0.911	2.56 Å
EMD10858_J	Rosetta	3.1	0.696	2.34 Å	0.846	1.61 Å
	EM-Refiner				0.920	1.01 Å
	EMCMR				0.923	0.97 Å
EMD3061_D	Rosetta	3.4	0.745	5.27 Å	0.826	3.78 Å
	EM-Refiner				0.853	3.20 Å
	EMCMR				0.859	3.12 Å
EMD8728_B	Rosetta	4.1	0.789	3.98 Å	0.916	2.07 Å
	EM-Refiner				0.927	1.97 Å
	EMCMR				0.932	1.84 Å
EMD8728_A	Rosetta	4.1	0.789	3.99 Å	0.926	1.96 Å
	EM-Refiner				0.932	1.83 Å
	EMCMR				0.948	1.50 Å
EMD3061_C	Rosetta	3.4	0.793	4.52 Å	0.915	2.68 Å
	EM-Refiner				0.921	2.63 Å
	EMCMR				0.929	2.53 Å
EMD30869_A	Rosetta	3.0	0.911	3.83 Å	0.967	2.49 Å
	EM-Refiner				0.947	3.15 Å
	EMCMR				0.956	2.77 Å
EMD6708_C	Rosetta	3.9	0.540	7.02 Å	0.713	5.91 Å
	EM-Refiner				0.716	5.10 Å
	EMCMR				0.722	5.47 Å
EMD6847_B	Rosetta	4.4	0.811	7.43 Å	0.828	7.88 Å
	EM-Refiner				0.841	7.28 Å
	EMCMR				0.896	2.18 Å
EMD20239_A	Rosetta	3.1	0.734	14.09 Å	0.815	13.94 Å
	EM-Refiner				0.818	13.92
	EMCMR				0.842	6.28 Å
EMD21923_C	Rosetta	3.3	0.756	9.60 Å	0.789	10.95 Å
	EM-Refiner				0.802	9.53 Å
	EMCMR				0.809	9.52 Å
EMD21040_A	Rosetta	3.8	0.719	5.35 Å	0.855	4.09 Å
	EM-Refiner				0.847	3.99 Å
	EMCMR				0.849	3.95 Å
EMD10090_A	Rosetta	1.8	0.557	4.27 Å	0.704	3.03 Å
	EM-Refiner				0.691	3.15 Å
	EMCMR				0.697	3.07 Å
EMD2221_D	Rosetta	8.4	0.709	7.09 Å	0.729	3.34 Å
	EM-Refiner				0.771	3.19 Å
	EMCMR				0.783	3.01 Å

movement, suggested further efforts for how to combine these factors more effectively.

Figure S8 presents an ablation experiment with excluding density map from EMCMR to illustrate the effect of the density map for the protein structure refinement. The TM-score and RMSD of EMCMR would decrease to 0.718 and 6.68 Å from 0.818 and 5.04 Å in 218 test proteins, respectively. The main reason was that our method was a protein structure refinement program based on the density map and contact/distance map. The density map was the main limitation of our program and its missing would greatly affect our result. Second, the initial model was a high-accuracy structure after using the contact/distance map in C-I-TASSER. The density map and contact/distance map were complementary in the confidence score. The error information of the

contact/distance map was introduced into refinement owing to excluding the density map in EMCMR. All these reasons resulted in the lower TM-score and RMSD of refined models after excluding the density map in EMCMR.

Case studies on atomic structure refinement using experimental cryo-EM density maps

To illustrate the stability of the method, the method was tested on several experimental cryo-EM density maps with noise. The experimental cryo-EM density map was taken from the EMDR (electron microscopy data resource) dataset. The released fitted structures were used as a native structure to estimate the quality of refinement protein structure. The sequence was firstly extracted from the fitted structure to obtain the

initial structure. Next, the initial atomic structure was predicted from the sequences by using C-I-TASSER [27], excluding homologous templates with $\geq 30\%$ sequence identity to the query. The contact/distance map was predicted by trossetta [25]. The specific results are as follows.

Refining the Human Gamma-Secretase

Dysfunction of the intramembrane protease γ -secretase is thought to cause Alzheimer's disease, with most mutations derived from Alzheimer's disease mapping to the catalytic subunit presenilin 1 (PS1) [28]. The density map of γ -secretase with a 3.4 Å resolution was taken from the EMDR with ID name EMD-3061. The fitted PDB ID was 5A63, which was used as a native structure to estimate the quality of the refinement model with the sequence from 1 to 243 at chain C. Next, a single chain density map of γ -secretase was segmented from EMD-3061 by using UCSF Chimera [29].

From the query sequence, I-TASSER created a model with a TM-score of 0.793 to the native structure. After EMC MR refinement, the TM-score of the C-I-TASSER model increased to 0.929 from 0.793, where the RMSD and CC improved to 2.53 Å and 0.615 from 4.52 Å and 0.416, respectively (Figure 6). Rosetta and EM-Refiner were also run starting from the initial C-I-TASSER model, which produced a refined model with lower TM-scores of 0.914 and 0.921, respectively. In addition, the RMSD (2.63 Å) and CC (0.601) of the EM-Refiner model were slightly worse than those of EMC MR and this situation also occurred in Rosetta with the 0.594 CC and 2.68 Å RMSD. Given that the correlation between the model and density map for EMC MR and EM-Refiner were essentially equivalent at the resolution of the density map, the difference in model quality was most likely not a result of the CC score between the model and the density map. Rather, it indicated that the increased model quality by EMC MR was a result of the confidence score, which combined with the CC score and contact/distance map to perform the EM- and CM-based refinement. These data showed that EMC MR has also a better local structure than EM-Refiner in this example.

Table 2 summarizes the refinement results by the EM- and distance-based refinement methods on an additional set of 14 randomly selected proteins from the EMDR that had experimentally determined density maps. All the methods (Rosetta, EM-Refiner, and EMC MR) were tested on these 14 proteins, starting from the C-I-TASSER models. The average TM-score and RMSD of EMC MR were 0.861 and 3.48 Å, respectively, compared with the 0.849 and 4.48 Å for EM-Refiner and 0.836 and 4.77 Å for Rosetta. The corresponding *P*-values were $6.0E-3$ between the EMC MR and EM-Refiner and $4.0E-3$ between EMC MR and Rosetta, as calculated by two-tailed Student's *t*-tests between the TM-scores. These results showed that EMC MR has a better performance than EM-Refiner and Rosetta at structure refinement using experimental cryo-EM density maps.

Conclusion

With the application of deep machine learning in protein structural prediction, the accuracy of protein structure prediction has obtained great improvement and attracted attention from people. However, a large part of the structure still does not achieve the requirements of structural biologists and needs to be further refinement. At present, many of the approaches address this challenge by cryo-EM density map and contact/distance map. However, how to optimize the protein structure by using a combination of density map and contact/distance map is still a relatively new problem. In this work, a new protocol was developed to integrate cryo-EM density map and contact/distance map into the energy function to refine protein structure by introducing a confidence function. The test results on a set of 233 (218 + 15) membrane proteins showed the advantages of the protocol for protein structure refinement compared with many of the previous methods, especially at refining structural models with incorrect folds (typically with TM-scores < 0.5). EMC MR is freely available for download and could be applied to EM-based protein structure prediction and refinement.

Despite the encouraging results, some targets still need to be further refined, especially at structural models with incorrect folds. This constraint stems essentially from the absence of correlation between the density map/predicted model correlation score and the accuracy of contact/distance map when initial models have lower TM-score (TM-score < 0.5). This absence of correspondence between the amino acids of protein and the position of the density map considerably limits the accuracy of the final refinement structure as the force field is strongly weighted by the CC score and contact/distance map score in addition to physics- and knowledge-based energy terms. In this regard, the recent efforts utilizing deep-learning-based cryo-EM density maps are important to provide a $C\alpha$ atom locations model from the density map for protein refinement [30–32]. A combination of the predicted $C\alpha$ atoms model and contact/distance map to refine protein structure will be our next important work. Meanwhile, developments of advanced methods to improve the correlation between the amino acids of protein and the position of the density map should also be key to help improve the success rate of EM-based structure prediction and refinement.

Currently, the EMC MR was trained and tested on single-chain proteins. For modeling multi-chain proteins, the user needs to apply a segmentation program, such as 'Segment Map' in UCSF chimera [29], to obtain the density maps for each chain. Nevertheless, an optimal solution to cryo-EM-guided multi-chain structural construction may be to integrate the original density map data and contact/distance map with the flexible chain refinement and inter-chain assembly simulations; the work along this line is in progress.

Data and code availability

All data needed to evaluate the conclusions are present in the paper and the Supplementary Materials. The additional data and code related to this paper can be downloaded from <https://github.com/iobio-zjut/EMCMR>.

Key Points

- Advanced method for protein structure refinement using cryo-EM density maps and deep learning.
- Physics- and knowledge-based energy functions, integrated with Cryo-EM density map data and deep learning data, were used to optimize the protein structure in the simulation.
- Robust algorithms built on large-scale benchmark training and test.
- Significant ability to refine models on the density maps.

Authors' contributions

B.H.S., B.Z. designed the research; B.Z. performed the research; B.H.S., D.L. and G.J participated in discussions and analyzed the data; Y.Z. participated in discussions and proofread the manuscript; B.Z. and G.Z. wrote the manuscript.

Supplementary Data

Supplementary data are available online at BIB.

Funding

This work has been supported by the educational commission of Zhejiang Provincial of China (grant number GZ21461030004), the Fundamental Research Funds for Provincial Universities of Zhejiang (grant number LQ22F020028), the National Nature Science Foundation of China (grant numbers 62173304, 61773346, 61725302 and 62073219), the Key Project of Zhejiang Provincial Natural Science Foundation of China (grant number LZ20F030002).

References

- Xu J, Mcpartlon M, Li J. Improved protein structure prediction by deep learning irrespective of co-evolution information. *Nat Mach Intell* 2021;**3**:601–9.
- Zhang C, Zheng W, Mortuza S, et al. DeepMSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. *Bioinformatics* 2020;**36**:2105–12.
- Li J, Xu J. Study of real-valued distance prediction for protein structure prediction with deep learning. *Bioinformatics* 2021;**37**:3197–203.
- Glaeser RM. How good can cryo-EM become? *Nat Methods* 2016;**13**:28–32.
- Yip KM, Fischer N, Paknia E, et al. Atomic-resolution protein structure determination by cryo-EM. *Nature* 2020;**587**:157–61.
- Patwardhan A. Trends in the Electron Microscopy Data Bank (EMDB). *Acta Crystallogr D Struct Biol* 2017;**73**:503–8. <https://doi.org/10.1107/S2059798317004181>.
- Bai X-C, McMullan G, Scheres SH. How cryo-EM is revolutionizing structural biology. *Trends Biochem Sci* 2015;**40**:49–57.
- Schröder GF, Brunger AT, Levitt M. Combining efficient conformational sampling with a deformable elastic network model facilitates structure refinement at low resolution. *Structure* 2007;**15**:1630–41.
- Gorba C, Miyashita O, Tama F. Normal-mode flexible fitting of high-resolution structure of biological molecules toward one-dimensional low-resolution data. *Biophys J* 2008;**94**:1589–99.
- Wang RY-R, Song Y, Barad BA, et al. Automated structure refinement of macromolecular assemblies from cryo-EM maps using Rosetta. *Elife* 2016;**5**:e17219.
- McGreevy R, Teo I, Singharoy A, et al. Advances in the molecular dynamics flexible fitting method for cryo-EM modeling. *Methods* 2016;**100**:50–60.
- Blau C, Lindahl E. All-atom ensemble refinement to cryo-EM densities with a bayesian measure of goodness-of-fit. *Biophys J* 2017;**112**:575a.
- Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;**596**:583–9.
- Li Y, Zhang C, Bell EW, et al. Ensembling multiple raw coevolutionary features with deep residual neural networks for contact-map prediction in CASP13. *Proteins* 2019;**87**:1082–91.
- Wang S, Li W, Liu S, et al. RaptorX-Property: a web server for protein structure property prediction. *Nucleic Acids Res* 2016;**44**:W430–5.
- Liu Y, Palmedo P, Ye Q, et al. Enhancing evolutionary couplings with deep convolutional neural networks. *Cell Syst* 2018;**6**:65, e63–74.
- DiMaio F, Tyka MD, Baker ML, et al. Refinement of protein structures into low-resolution density maps using rosetta. *J Mol Biol* 2009;**392**:181–90.
- Liu J, Zhou X-G, Zhang Y, et al. CGLFold: a contact-assisted de novo protein structure prediction using global exploration and loop perturbation sampling algorithm. *Bioinformatics* 2020;**36**:2443–50.
- Lomize MA, Pogozheva ID, Joo H, et al. OPM database and PPM web server: resources for positioning of proteins in membranes. *Nucleic Acids Res* 2012;**40**:D370–6.
- Zhang B, Zhang X, Pearce R, et al. A New Protocol for Atomic-Level Protein Structure Modeling and Refinement Using Low-to-Medium Resolution Cryo-EM Density Maps. *J Mol Biol* 2020;**432**:5365–77.
- Yang J, Yan R, Roy A, et al. The I-TASSER Suite: protein structure and function prediction. *Nat Methods* 2015;**12**:7–8.
- Godzik LA. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 2006;**22**:1658.
- Jinrui X, Yang Z. How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics* 2010;**26**:889–95.
- Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. *Proteins* 2004;**57**:702–10.
- Yang J, Anishchenko I, Park H, et al. Improved protein structure prediction using predicted interresidue orientations. *Proc Natl Acad Sci* 2020;**117**:1496–503.
- Chang Y, Bruni R, Kloss B, et al. Structural basis for a pH-sensitive calcium leak across membranes. *Science* 2014;**344**:1131–5.

27. Mortuza S, Zheng W, Zhang C, et al. Improving fragment-based ab initio protein structure assembly using low-accuracy contact-map predictions. *Nat Commun* 2021;**12**: 1–12.
28. Bai XC, Yan C, Yang G, et al. An atomic structure of human γ -secretase. *Nature* 2015;**525**:212–7.
29. Pettersen EF, Goddard TD, Huang CC, et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J Comput Chem* 2004;**25**:1605–12.
30. Si D, Moritz SA, Pfab J, et al. Deep learning to predict protein backbone structure from high-resolution cryo-EM density maps. *Sci Rep* 2020;**10**:1–22.
31. Terashi G, De Kihara D. novo main-chain modeling for EM maps using MAINMAST. *Nat Commun* 2018;**9**:1–11.
32. Kui X, Wang Z, Shi J, Li H, and Zhang QC. A2-net: Molecular structure estimation from cryo-em density volumes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019;**33**: 1230–7.