SEGEM: a Fast and Accurate Automated Protein Backbone Structure Modeling Method for Cryo-EM

1st Sheng Chen
School of Computer Science and Engineering
Sun Yat-sen University
Guangzhou, China
chenshengsemail@gmail.com

2nd Sen Zhang
School of Computer Science and Engineering
Sun Yat-sen University
Guangzhou, China
zhangs7@mail2.sysu.edu.cn

3rd Xiongjun Li
School of Computer Science and Engineering
Sun Yat-sen University
Guangzhou, China
709524591@qq.com

4th Yubao Liu
School of Computer Science and Engineering
Sun Yat-sen University
Guangzhou, China
liuyubao@mail.sysu.edu.cn

5th Yuedong Yang*
School of Computer Science and Engineering
Sun Yat-sen University
Guangzhou, China
yangyd25@mail.sysu.edu.cn

2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 978-1-6654-0126-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/BIBM52615.2021.9669647

***Abstract*—Cryo-electron microscopy (cryo-EM) technique has been widely used in protein structure determination, whereas it remains a challenge to automatically build accurate protein backbone structure from cryo-EM density map. A typical pipeline to automatically build a structure model from cryo-EM map is to first predict Cα sites and then assign them to protein sequence, which is a typical combinatorial optimization task of extremely high computational complexity. Here we propose SEGEM, a fast and accurate automated protein backbone structure modeling method for cryo-EM. We employed 3D Convolutional Neural Networks to predict Cα sites with their amino acid types from cryo-EM, and developed a highly parallel pipeline to assign Cα sites with their predicted amino acid types to protein sequence. We tested SEGEM on three benchmark datasets where it significantly outperformed several state-of-the-art prediction methods including MAINMAST, C-CNN and DeepTracer. In our method plus version SEGEM++, we combined SEGEM with the protein structure prediction algorithm AlphaFold2. SEGEM++ is capable to identify whether AlphaFold2 folds a good structure, and rectify the incorrectly folded region through protein threading on cryo-EM map. In our curated dataset of hard targets where AlphaFold2 predicted structures obtained an average RMSD of 7.87Å and GDT-TS score of 0.652 when superimposed to the native structure, SEGEM++ achieved a significantly better RMSD of 2.46Å and 0.676 GDT-TS score on average. Furthermore, with our highly parallel pipeline on 30 cores CPU, both SEGEM and SEGEM++ finished structure modeling within 10 minutes on average in our test datasets, indicating their potential in high throughout automated accurate backbone structure modeling for cryo-EM.**

***Index Terms*—cryo-EM, protein structure modeling, deep learning, 3D image semantic segmentation, protein structure prediction, AlphaFold2**

I. Introduction

Proteins perform a variety of functions, e.g., DNA replication, chemical reaction catalysis, and cell signals transmission, which are essential for organisms' life activities [1]. Protein function is determined by its structure [2]. The knowledge about protein structure is not only beneficial for drug design and vaccine engineering [3], [4], but also helps researchers to understand the complex mechanisms of life. As a consequence, it's crucial to deduce protein structure through wet-lab experiment and computational processing.

With the breakthrough of electronic instrument and image processing algorithm, cryo-electron microscopy (cryo-EM) has been developed rapidly in last decade and is widely used in protein structure determination [5]–[8]. cryo-EM is able to reconstruct 3D density maps in near-atomic resolution, which indicates the electronic density of the native protein structure. The number of 3D cryo-EM density maps deposited in EMDataResource (EMDB) is growing almost exponentially [9], which has given rise to the demand for automated structure modeling tools. Tools sprang up like bamboo shoots. The early work like EM-fold [10] and Gorgon [11] carried out pioneering explorations. Pathwalking [12]–[14] performed K-means clustering on the density map and using traveling salesman problem (TSP) solvers Concorde and LKH in protein threading. David Baker's lab developed a series of template fragment based method, e.g., Rosetta *de novo* [15] and RosettaES [16]. Phenix [17]–[20] is a software suite for cryo-EM protein structure modelling which ensemble a variety of image processing method, structure prediction model and backbone tracing tools. MAINMAST [21] identified the backbone structure by mean shift and employed tabu search algorithm in backbone tracing. Recently, method like A^2 Net [22], C-CNN [23], DeepTracer [24] and DeepMM [25] have explored the deep learning application in this field, whereas among them few works utilized protein sequence information. DeepMM grouped amino acid type into four-class to predict instead of predicting all 20 amino acid types. [25] DeepTracer predicted the total 20 amino acid types but was still lack in accuracy [24]

* corresponding author

Although advances have been made in the field of protein structure computational aided experimental determination, there is still an insurmountable gap between the number of determined protein sequences [26] and protein structures [27]. Can protein structure be computationally predicted for a given sequence? To give an answer to this question, CASP competition has been held for 14 sessions. The AlphaFold2 (AF2) algorithm [28], proposed by DeepMind, won CASP14 with a breaking average GDT score record of 0.92. After that, DeepMind cooperated with EMBL-EBI to release a large-scale predictive protein structure database, which covers 98.5% of human proteins [29]. It means that high-throughput accurate protein structure prediction is no longer far away. AF2 starts from multiple sequence alignments (MSAs) and structure template search, passes and transforms sequential and pairwise features interactively in attention-based Evoformer blocks and use SE(3)-equivariant Transformer network to finally end-to-end generate 3D coordinate of protein atoms [28].

We propose a novel cryo-EM protein backbone structure modelling method named SEGEM. We design a 3D image semantic segmentation framework to predict the Cα site and their amino acid type. After that the native protein sequence and predicted amino acid are used to build a score matrix, which could help us assign the predicted Cα to sequence efficiently. We then perform protein threading on each unassigned sequence segment and output the coordinate of each Cα site aligning to protein sequence. SEGEM significantly outperformed MAINMAST, C-CNN, DeepTracer in our test of experimental cryo-EM maps. Furthermore, in its plus version SEGEM++, we combined AF2 with SEGEM complementarily. SEGEM++ first tries to identify the correctly folded region in AF2 structure utilizing our predicted Cα probability density from cryo-EM maps. After that the high confidence AF2 structure fragments help us build a more accurate protein structure from cryo-EM map, which could rectify the incorrectly folded region in AF2 structure.

To summarize, our contributions are as follows:

- We designed a novel 3D image semantic segmentation framework to predict Cα site and amino acid type.
- We developed a fast and accurate pipeline to parallelly assign predicted Cαs to protein sequence through aligning predicted amino acid type with native sequence.
- We created a paradigm that complementarily combines cryo-EM protein structure modeling with AlphaFold2 (AF2) structure prediction, where modeling on cryo-EM map helps to identify and rectify the incorrectly folded region in AF2 structure, and the high confidence AF2 structure fragments improve the structure modeling from cryo-EM.

II. Materials and Method

A. Method Overview

Identifying Cα sites in cryo-EM density map is the key point to build a protein backbone structure. We regard this as a 3D image semantic segmentation task. In our method, we designed a simple but efficient 3D image semantic segmentation framework (Fig.3 D) to identify Cα sites and predict their amino acid type and secondary structure from cryo-EM density maps. Fig.1 illustrates the Cα site predict pipeline, which outputs the predicted Cα coordinates from a cryo-EM density map. Assigning the predicted Cαs to protein sequence is a typical combinatorial optimization task, which can be extremely high in computational complexity. We designed a highly parallel pipeline to solve this problem, as is shown in Fig.2. In this pipeline we first trace each predicted Cα with its spatial neighbors to generate local Cα traces. For each Cα trace its sequence profile can be predicted using our amino acid type predictor. After that we align all Cα traces to native protein sequence segments, calculating a matching score matrix. All these steps are accelerated by multi-thread computation to about 23 times faster using 30 cores of Intel(R) Xeon(R) Gold 6248R CPU. We then utilize the score matrix to build a high confidence base structure model. At last a prune-search based protein threading is performed to finish the final model. Our method output a pdb file of Cα coordinates aligned to protein sequence.

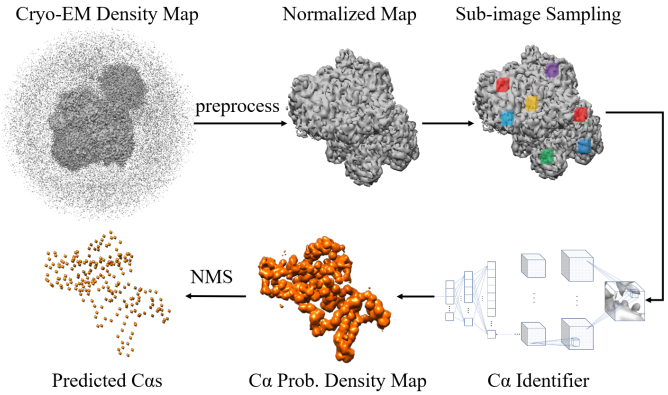

Fig. 1. The Cα predict pipeline, where the Cα probability density map represents the predicted probability for each voxel to be Cα site, and NMS is the abbreviation of Non-Maximum Suppression algorithm

B. Training Datasets

Training a deep neural network is in high demand for a large training dataset. Therefore, we curated two training datasets, one of which contains 41428 simulated cryo-EM maps, named as TR_SIM, and another consists of 2088 experimental cryo-EM maps, called TR_EXP.

The simulated maps in TR_SIM are generated by the EMAN2 script pdb2mrc.py [30] from PDB files of a protein dataset used in our previous study [31]. Every employed protein is R-factor $>$ 1.0, sequence length $\geq$ 30, and sequence identity $\leq$ 25% with each other. In this study, an additional filter was applied on TR_SIM to remove chains whose sequence identity $>$ 25% with any chains in our simulated test set, and thus only 10357 chains were left for simulated data generation. In order to avoid over-fit and obtain robust performance on various resolution density maps, the

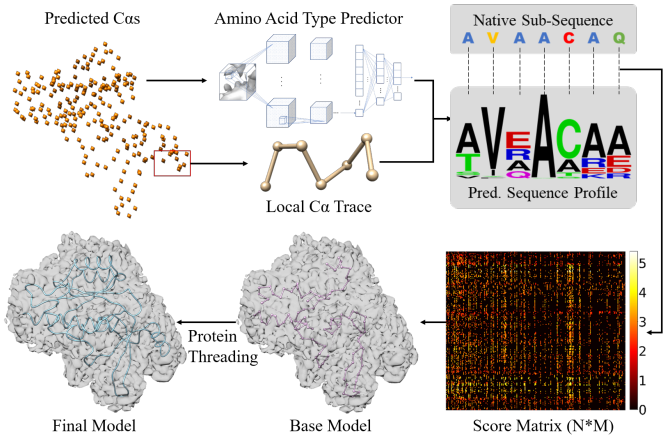

Fig. 2. Pipeline to assign predicted Cα to protein sequence, in which the sequence profile is the 20 amino acid type probability predicted by amino acid type predictor for all Cαs in a Cα trace, and the score matrix is of size $(N \times M)$ where N represents the number of protein sequence segments and M is the number of Cα traces.

neural network was trained with density maps with various resolution. Given a PDB file, we generate 4 simulated cryo-EM maps with various random resolution ranging from 3-4Å, 4-5 Å, 5-6Å and 6-7Å respectively. Finally, we obtained 41428 simulated cryo-EM density maps for TR_SIM. The experimental training set TR_EXP is composed of about two thousand cryo-EM maps from EMDB server. We collected maps deposited from January 2017 to April 2020. From them, maps without associated pdb were excluded. Furthermore, we removed maps whose protein sequence identity is more than 25% with any chains in our experimental test set. At last, we maintained 2088 maps, whose EM method is single particle, specimen type is particle, microscope model is FEI TITAN KRIOS, and resolution is from 2Å to 5Å.

C. Image preprocess

The cryo-EM maps were deposited by researchers that employed a verity of microscope equipment, electron doses, electron detectors, and experimental procedures. It results in the large variance in density distribution, local resolution and noise intensity. Therefore, its crucial to adopt an image pre-processing step to normalize the maps, making it more suitable for image semantic segmentation. An experimental EM map usually contains more than one chain. To focus on our target chain, a single subunit is first zoned out from a whole density map with a distance cutoff of 5Å [21], [23], [25]. After that, the subunit is transposed to the same coordinate system of its ground truth pdb file. Then we reshape the map by trilinear interpolation so that every voxel occupied exactly 1Å× 1Å× 1Å. Finally, we normalize the map by formula (1):

$$M_{ijk} = \begin{cases} 0, M_{ijk} < \theta_{med} \\ \dfrac{M_{ijk} - \theta_{med}}{\theta_{top1}}, \theta_{med} \le M_{ijk} < \theta_{top1} \\ 1, \theta_{top1} \le M_{ijk} \end{cases} \quad (1)$$

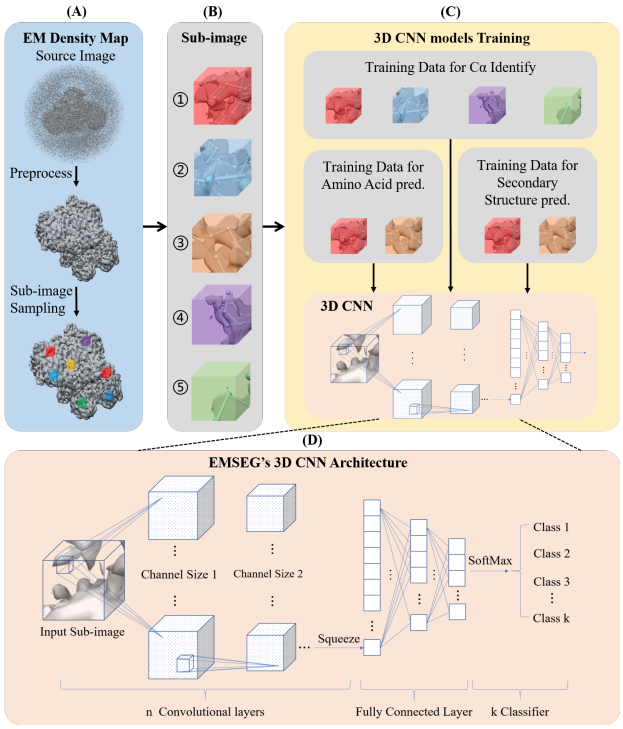

Fig. 3. The CNN model training procedure: (A) image preprocessing and group sub images into (B) 5 types of training data, which would be fed into (C) three CNN models for Cα identification, amino acid type and secondary structure prediction, whose network architecture illustrated in (D) where the output class number k varies in different tasks (3 for Cα and secondary structure identification, and 20 for acid amino type prediction).

, where M_{ijk} is the density value of [i, j, k] voxel, and θ_{med} represents the median density in the whole map, and θ_{top1} is defined as the top 1% density.

D. Image Semantic Segmentation

In our method, we designed a simple but efficient 3D image semantic segmentation framework based on 3D Convolutional Neural Networks (CNN), as is shown in Fig.3D. Considering the memory and time consumption, we set the size of input image to 11×11×11. A batch of input data is firstly fed into a bottleneck block which enlarged the channel size to 16. The bottleneck block is followed by 5 convolution blocks. In each block we employ convolutions with kernel size of $3 \times 3 \times 3$, stride of 1 and no padding. We don't use pooling since it introduces the translation invariance which is not conducive to image semantic segmentation tasks. Output of the last convolution block is squeezed and then fed into a fully-connected network with one hidden layer.

It is challenging to train a well-performed model to distinguish the Cα atoms with other atoms. Thus we grouped the sub-images into 5 types shown in Fig.3B: ① centering at native Cα site, ② centering at other atoms, ③ centering voxel nearby a Cα atom, ④ centering voxel nearby other atoms, and ⑤ otherwise. For Cα identification, we labeled sub-images ① as the *Cα* class, ②④ as *other-atom* class, and ⑤ as *non-atom* class. Considering the native sample number of these 3 classes

are extremely imbalance, we applied down-sampling in other-atom class and none-atom class to make sure the number of training data for three classes were 1:1:1. We did not sample sub-images ③ for Cα identification since it might confuse the model. However, for amino acid and secondary structure prediction, we sampled both ① and ③ as training data since training on sub-images ③ is beneficial for model robustness when the predicted Cα are not precise enough.

In Cα site prediction, it's unnecessary and time-wasted to identify Cα for all voxels since the density map is sparse and most of Cαs are in high density region. So we only sample sub-images whose central $3 \times 3 \times 3$ densities are over average. It could filter out about 99% of the sub-images, but retain more than 99% of the native Cα sites, which greatly reduce the time consumption with a tolerable accuracy lost. We feed the sampled sub-images into Cα identification CNN to obtain a predicted Cα probability density map. After that the Non-Maximum Suppression algorithm (NMS) picks out the local maximum Cα probability density voxels as predicted Cα sites whose amino acid type would also be predicted and be used for assigning the predicted Cαs to protein sequence.

E. Assigning Cα to Sequence

We designed a pipeline (Fig.2) to assign the predicted Cαs to protein sequence, which includes the following steps:

The first step is Cα local tracing. For each predicted Cα, we trace its neighbors (Cα distance within 2Å-6Å) iteratively to generate Cα traces. The number of generated Cα traces increases exponentially with their tracing depth, which results in higher computational complexity. On the other hand, aligning longer traces with sequence strengthen the sequential and structural constraints, which would feed back a more accurate matching score. In our method, tracing depth is set to 7 or 11, trying to find a balance between low computational complexity and high accuracy.

The second step is to calculate the matching score of Cα traces and sequence segments. We first predict amino acid type for every predicted Cα, after which we obtain an amino acid scoring matrix noted as A. A_{jk} represents the predicted probability that the j_{th} Cα belong to the k_{th} type amino acid. Then we can calculate the amino acid type matching score Sa_{ij} for the i_{th} position in protein sequence and j_{th} Cα:

$$\mathrm{Sa}_{i,j} = \mathrm{A}_{j,a_i} \tag{2}$$

, where a_i represents the i-th amino acid type in native protein sequence.

Sa is then used in the matching score function of Cα traces and sequence segments. We designed a scoring function defined as follow:

$$\mathrm{S}_{t,s} = \sum_{d=0}^{m-1} \mathrm{Sd}_{t_d,t_{d+1}} \times (\mathrm{Sa}_{s_d,t_d} + \mathrm{Sa}_{s_{d+1},t_{d+1}}) \tag{3}$$

, where t and s represent a Cα trace and a sequence segment respectively, d denotes the tracing depth and m is the max tracing depth (7 or 11), and we defined the distance score Sd as:

$$\mathrm{Sd}_{j_1,j_2} = \begin{cases} 1, |\mathrm{D}_{j_1,j_2} - 3.8| < 0.5 \\ 1 - \dfrac{|\mathrm{D}_{j_1,j_2} - 3.8| - 0.5}{2}, 0.5 \le |\mathrm{D}_{j_1,j_2} - 3.8| < 2.5 \\ 0, 2.5 \le |\mathrm{D}_{j_1,j_2} - 3.8| \end{cases} \tag{4}$$

, where D_{j_1,j_2}.is the distance between the j_{1th} and j_{2th} predicted Cα.

The third step is to build a high confidence base model. A naive way is to assign the best matched Cα trace to each sequence segment with a matching score cutoff, whereas it may cause structure clash or result in spatial distance gap between sequential neighbors. We adopted a carefully designed strategy to avoid structure clash and make sure that sequential neighbors are always contacted in spatial, as is described in Algorithm 1.

Algorithm 1 Build the Base Model

Require: Cα trace list, $traceL$
Require: protein sequence segments list, $segmentL$
Require: trace-segment matching score matrix, $scoreM$
Require: matching score cutoff, $scoreThrh$
  for $segment$ in $segmentL$ **do**
    find $bestTrace$ **in** $traceL$ **by** $scoreM$
  end for
  sort $segmentL$ **by** $bestTraceScore$ from max to min
  for $segment$ in $sortedSegmentL$ **do**
    if $bestTraceScore \le scoreThrh$ **then**
      break
    end if
    if $bestTrace$ clash with any $pickedTrace$ **then**
      continue
    end if
    if $segment$ gap with any $assignedSegment$ **then**
      continue
    end if
    pick $bestTrace$ and assign to $segment$
  end for

The final step is to perform protein threading on the remaining Cαs, aligning to the unassigned sequence segments. We trace Cα from the left or right endpoint, which is assigned with a certain Cα in base model, until it reach another end or failed. But it suffers from high computational complexity in long sequence segment. So we developed a strategy of breadth-first search with pruning to accelerate that, as is introduced in Algorithm 2. At last we obtain a final model, which consists of Cα coordinates aligned to protein sequence.

F. SEGEM++ with AlphaFold2

In our method plus version SEGEM++, we first modify the scoring function to calculate the matching score of Cα traces and sequence segments, noted as S^{++}:

$$\mathrm{S}^{++}_{t,s} = \mathrm{S}_{t,s} - \sigma_{t,s} \tag{5}$$

, where S is the original scoring function, t and s represent a Cα trace and a sequence segment respectively, and $\sigma_{t,s}$ is

Algorithm 2 Breadth-First Search with Pruning

Require: $C\alpha$ trace list, $traceL$
Require: remaining $C\alpha$ list, $C\alpha L$
Require: max number of $C\alpha$ traces, $maxNum$
Require: scoring function for $C\alpha$ traces, $scoringF$
Require: breadth first search function, BFS
  while not reach the end or failed **do**
    $traceL \leftarrow BFS(traceL, C\alpha L)$
    if length of $traceL \geq maxNum$ **then**
      $newTraceL \leftarrow$ empty list
      sort $traceL$ **by** $scoringF$ from max to min
      for t in $sortedTraceL$ **do**
        if t not similar with $traces$ in $newTraceL$ **then**
          add t **to** $newTraceL$
        end if
      end for
      $traceL \leftarrow newTraceL$
    end if
  end while

the RMSD calculated by superimposing the Cα trace t to AlphaFold2 (AF2) structure on s.

After building a base model, we superimpose AF2 structure to this base model, and calculate the matching score of each Cα in AF2 structure to our predicted Cα probability density map. We regard this as a confidence score to identify whether AF2 folded correctly in this area. The scoring function is described as below:

$$C_{xyz} = \sum_{i=x-2}^{x+2} \sum_{j=y-2}^{y+2} \sum_{k=z-2}^{z+2} P_{ijk} \times D^2_{xyz,ijk} \qquad (6)$$

, where C_{xyz} denote the confidence score of a Cα (coordinate [x,y,z]) in superimposed AF2 structure, [i,j,k] is one of its 5×5×5 neighbor coordinate, P represents our predicted Cα probability and D calculate the spatial distance.

Once we get the confidence score for all Cαs of AF2 structure, a confidence cutoff is set to get a list of high confidence correctly folded AF2 structure fragments. We then use these fragments as new base model, and perform protein threading on the predicted Cα probability density map to build the final model (similar to the final step in SEGEM's modeling pipeline).

III. Result

TABLE I
Average Performance Comparison on TS_SIM

Method	*time*	*AA acc*[a]	*RMSD*[b]	*CLICK-SO*[c]
MAINMAST	-	-	1.79Å	81.88%
SEGEM	**5.1 min**	**97.18%**	**1.11Å**	**98.47%**

[a]amino acid type prediction accuracy on native structure
[b]root mean square deviation(RMSD) of Cα atoms between predicted structure and natives structure.
[c]percentage of predicted Cαs within 3.5Å of the aligned native Cαs

A. Modeling on Simulated EM Maps

To evaluate the performance of our method on simulated cryo-EM density maps, we test it in a simulated map dataset curated by MAINMAST [21], named as TS_SIM. Density maps in this dataset were generated at a resolution of 5.0 Å using the e2pdb2mrc.py script [32]. Table I provide the average performance of MAINMAST and SEGEM on TS_SIM, where SEGEM obtained a significant outperformance than MAINMAST, with its 1.11Å average RMSD and 98.47% average CLICK-SO compared to the 1.79Å and 81.88% of MAINMAST (CLICK-SO is computed by the CLICK package [33], and the result of MAINMAST is from MAINMAST paper [21]). We studied the detailed performance of SEGEM and MAINMAST on every test case of TS_SIM, and found that SEGEM outperformed MAINMAST in almost all cases for both RMSD and CLICK-SO metrics. These should be mainly attributed to our employed semantic segmentation framework. Compared to the mean shift algorithm in MAINMAST, our CNN-based semantic segmentation framework is able to predict more accurate Cα coordinate. It is worth mentioning that SEGEM correctly predict 97.18% amino acid type for native Cα sites on TS_SIM, which helps us assign more predicted Cα sites to protein sequence on our base model. Larger base model means less protein threading, while our highly parallel pipeline to build the base model would take up a larger proportion in the entire modeling process. As a consequence, SEGEM finished the end-to-end structure modeling with 5.1 minutes per case on average using 30 cores of Intel(R) Xeon(R) Gold 6248R CPU, significantly faster than 31.7 minutes running on single core environment.

B. Modeling on Experimental EM Maps

For experimental maps, we employed two benchmark set. The first benchmark set from C-CNN [23] consists of 50 experimental maps, named as TS_CC. The second one is composed of 18 test cases from MAINMAST [21], so we called it TS_MM.

TABLE II
Average Performance Comparison on TS_CC

Method	*False Positive*[a]	*CC-RMSD*[b]	*Cα in 3Å*[c]
Phenix	-	1.22Å	66.8%
C-CNN	3.74%	1.23Å	88.5%
SEGEM	**3.26%**	**1.13Å**	**95.5%**

[a]percentage of Cα atoms in the predicted model that are more than 3Å away from any native Cα atom.
[b]RMSD between native Cα and its closest predicted Cα.
[c]percentage of native Cα atoms within 3Å of its closest predicted Cα.

First, we compared SEGEM with C-CNN [23] and Phenix [19] on TS_CC. C-CNN adopted a different RMSD metric with MAINMAST, and we call it CC-RMSD, which represents the RMSD between native Cα and its closest predicted Cα. C-CNN also employed the false positive rate and Cα in 3Å rate to evaluate the method performance. We test SEGEM with the same evaluation metrics and the results are provided in Table

II. SEGEM achieved the lowest false positive rate, the lowest CC-RMSD and the highest Cα in 3Å rate when compared to C-CNN and Phenix (the results of C-CNN and Phenix are from C-CNN paper [23]). DeepTracer [24] is another method developed by C-CNN's authors, which evaluate its amino acid prediction and secondary structure prediction accuracy on a subset of TS_CC. Table III makes a performance comparison between SEGEM and DeepTracer on this subset. SEGEM achieved better performances in all five aspects: lower average false positive rate (3.2% than 5.8%), more Cα in 3Å (94.7% than 87.5%), smaller RMSD (1.12Å than 1.18Å), higher prediction accuracy of amino acid type (48.3% than 25.2%) and secondary structure (84.5% than 81.7%). We studied the detailed performance for amino acid type prediction of all test cases, and found that SEGEM obtained over 40% accuracy for 20 maps in total 30 test maps while DeepTracer got none.

TABLE III
AVERAGE PERFORMANCE COMPARISON ON A SUBSET OF TS_CC

Method	*False Positive*	*CC-RMSD*	*Cα in 3Å*	*AA acc*	*SS acc*[a]
DeepTracer	5.80%	1.18Å	87.5%	25.2%	81.7%
SEGEM	**3.20%**	**1.12Å**	**94.7%**	**48.3%**	**84.5%**

[a]secondary structure predict accuracy for native Cα sites

Test dataset TS_MM contains 18 test maps that have been modeled by MAINMAST. We test SEGEM on these maps, and the result is shown in Table IV. MAINMAST utilized a series of tools including MDFF [34] to refine their structure. Even so, SEGEM obtained the best average RMSD of 11.68Å, better than that of MAINMAST refined models (15.69Å). For the coverage evaluation, SEGEM achieved 91.76%, which is far better than Rosetta's 68.33%, and slightly outperforms MAINMAST refined models (91.39%). We noticed that the average AA accuracy on experimental test set (47.27% in TS_MM and 48.3% in TS_CC) is significantly inferior to that of the simulated test set (97.18%). This should be attributed to the variant local resolution and noise density in experimental maps, which make them hard to be normalized and thus difficult for CNN training and prediction. We then studied the correlation between AA accuracy and the model performance. As is shown in Fig.4, AA accuracy shows a significant negative correlation with RMSD and an obvious positive correlation with coverage. We also found that the RMSD and coverage were almost stable around 0-5Å and 95-100% when the AA accuracy rate exceeded 50%. In future, SEGEM is expected to achieve better performance with the improvement of AA accuracy leading by advancing 3D image semantic segmentation algorithm and cryo-EM image processing technology.

C. Modeling on Experimental EM Maps with SEGEM++

In our method plus version SEGEM++, we combined SEGEM with the protein structure prediction algorithm AlphaFold2. SEGEM++ is capable to identify whether AlphaFold2 folds a good structure, and rectify the incorrectly folded region through protein threading on cryo-EM map. For

TABLE IV
AVERAGE PERFORMANCE COMPARISON ON TS_MM

Method	*time*	*AA acc*	*RMSD*	*coverage*[a]
Rosetta	-	-	30.22Å	68.33%
MAINMAST org.	-	-	19.39Å	-
MAINMAST ref.	-	-	15.69Å	91.39%
SEGEM	**9.2 min**	**47.27%**	**11.68Å**	**91.76%**

[a]the fraction of Cα atoms in the native structure that are within 3.0 Å cutoff to any Cα atoms in the predicted structure.

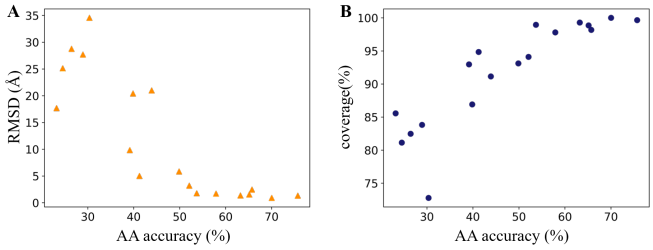

Fig. 4. Scatter plot of the model RMSD (A) and coverage (B) with amino acid type prediction accuracy on TS_MM.

SEGEM++ we build a test dataset TS_AF of 66 experimental cryo-EM maps deposited in EMDB after 2021-1-1 and reported resolution between 2Å-5Å. Then we run AF2 code to predict protein structure from their fasta files (randomly select one chain for each cryo-EM map). The max_template_date for AF2 to search template is set as 2020-05-14.

We first test the performance of SEGEM++ in identifying the correctly folded region of AF2 structures on TS_AF. We superimpose AF2 structure to native structure and calculate the aligned Cα distance. A large Cα distance indicates that AF2 did not correctly fold the protein structure in this area. We then study the correlation between Cα distance and the confidence score computed by SEGEM++. As illustrated in Fig.5A, with the increase of the confidence score, the average value and standard deviation of the Cα distance decline and trend towards 0 as our expectation. We also counted the number of Cαs in each confidence interval in Fig.5A. It can be found that our confidence cutoff of 5 could retain about 90% of Cαs while filtering out most of the incorrectly folded structure regions. We then study the correlation between average RMSD (RMSD between the filtered AF2 structure and native structure) and coverage (the percentage of remaining Cαs in AF2 structure) on TS_AF. As is shown in Fig.5B, the confidence score show negative correlation to both RMSD and coverage. The filtered AF2 structures retained more than 80% Cαs and obtained RMSD less than 1.5Å on average when the confidence score cutoff was set in 1-8, indicating the well-performance of SEGEM++ to identify the correctly folded region of AF2 structure.

We then evaluate the performance of SEGEM++ in building the final structure model. For every AF2 predicted structure, SEGEM++ base model and SEGEM++ final model in TS_AF, we calculate the GDT-TS score and RMSD when

TABLE V
AVERAGE PERFORMANCE COMPARISON ON TS_AF

Dataset	*TS_AF*		*TS_AF*$_{\mathbf{hard}}$	
Model	***RMSD***	***GDT-TS***	***RMSD***	***GDT-TS***
AF2 model	2.62Å	0.880	7.87Å	0.652
SEGEM++ base	1.05Å	0.852	1.59Å	0.585
SEGEM++ final	**1.30Å**	**0.881**	**2.46Å**	**0.676**

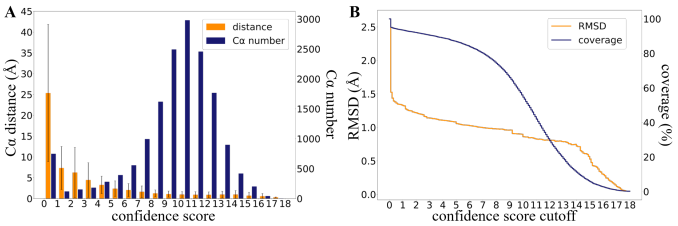

Fig. 5. Correlation between confidence score, Cα distance, and Cα number shown in histogram (A), and correlation between confidence score cutoff, RMSD and coverage of AF2 structure fragments ploted in (B)

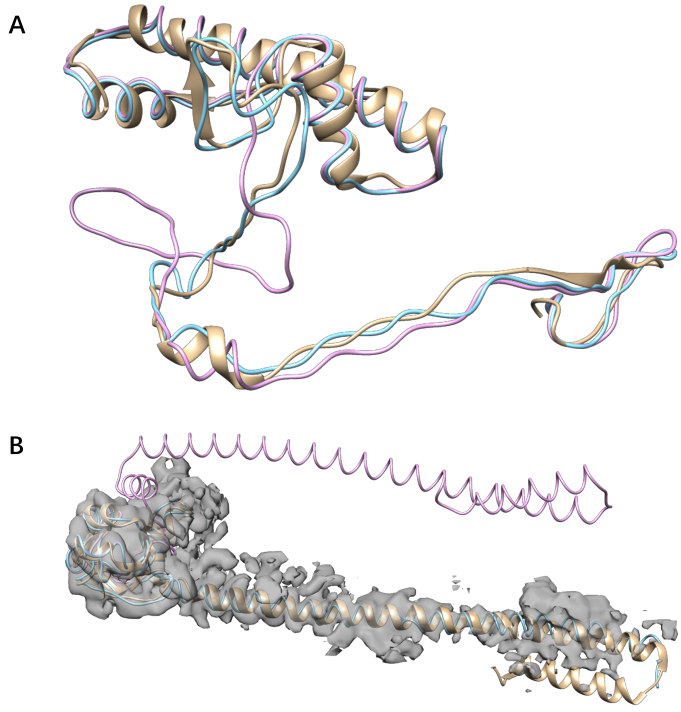

Fig. 6. Case study of (A) EM-31116, pdb-chain: 7egg-F, and (B) EM-31138, pdb-chain: 7egq-B, where golden ribbon is the native structure, pink line represents the AF2 structure superimposed to native structure, and blue line figure out SEGEM++ final structure; the transparent gray area in (B) represents the density value EM-31138 with a density cutoff.

superimposed to the native structure. As is shown in Table V, SEGEM++ final model significantly outperformed AF2 model by RMSD (1.30Å compared to 2.62Å) while their GDT-TS score are both approximately equal to 0.88. We then studied on a subset contains 11 test case of TS_AF , where the GDT-TS score of AF2 structure are all less than 0.80, named as TS_AF$_{hard}$. Table V provides the test result on TS_AF$_{hard}$, from which we found that SEGEM++ final model achieved both better RMSD (2.46Å) and GDT-TS (0.676) than AF2 model (7.87Å RMSD and 0.652 GDT-TS). These result indicates the ability of SEGEM++ to rectify the incorrectly folded AF2 structure. As a detailed example in Fig.6A, for EM_31116, SEGEM++ correctly rectified the incorrectly folded random coil in AF2 structure, achieving the 1.92 Å RMSD and 0.84 GDT-TS score, which significantly outperformed the 10.17Å and 0.71 of AF2. For another case EM_31138 in Fig.6B, SEGEM++ also successfully rectified incorrectly folded region in AF2 structure, while SEGEM++ didn't finish the complete protein threading in the tail. To study the possible infactor, we visualized the cryo-EM density map in Fig.6B and found local density variance in the uncompleted threading area, which makes it difficult for Cα site prediction and might gap the protein threading as a consequence. We also observed that SEGEM++ complete the structure modeling within an impressive speed: 3.9 minutes on average for TS_AF (excluding AF2 structure prediction time consumption), indicating the potential for high-throughput automated structural modeling from cryo-EM maps in future.

IV. CONCLUSION

In this study, we proposed a novel automated protein backbone structure modeling method SEGEM. Different from previous methods, SEGEM utilized the predicted amino acid type to develop a highly parallel pipeline to assign the predicted Cαs to protein sequence. As a consequence, SEGEM achieving state-of-the-art modeling accuracy and coverage with low time consumption. In its plus version SEGEM++ we employ protein structure prediction algorithm AlphaFold2 (AF2) to improve structure modeling from cryo-EM maps. SEGEM++ is proved to be discriminating at whether AF2 folded a good structure and is capable to rectify the incorrectly folded region, which make it obtain a near atomic modeling accuracy in our test dataset. Nevertheless, for some test case, SEGEM++ still failed to build a complete model. We study these case and deduce the reason to be the missing identification of some native Cα sites, which is a consequence of variant local resolution and noise density in cryo-EM map. But with the development of 3D image semantic segmentation algorithm and cryo-EM image processing technology, it's expected that our method would achieve higher modeling quality.

ACKNOWLEDGMENT

This study has been supported by the National Key R&D Program of China (2020YFB0204803), National Natural Science Foundation of China (61772566), Guangdong Key Field R&D Plan (2019B020228001 and 2018B010109006), Introducing Innovative and Entrepreneurial Teams (2016ZT06D211), Guangzhou S&T Research Plan (202007030010, 202002020047), Natural Science Foundation of Guangdong, China (2019A1515012207).

REFERENCES

[1] C. I. Branden and J. Tooze, *Introduction to protein structure*. Garland Science, 2012.

[2] J. S. Richardson, "The anatomy and taxonomy of protein structure," *Advances in protein chemistry*, vol. 34, pp. 167–339, 1981.

[3] X. Xu, P. Chen, J. Wang, J. Feng, H. Zhou, X. Li, W. Zhong, and P. Hao, "Evolution of the novel coronavirus from the ongoing wuhan outbreak and modeling of its spike protein for risk of human transmission," *Science China Life Sciences*, vol. 63, no. 3, pp. 457–460, 2020.

[4] L. E. Gralinski and V. D. Menachery, "Return of the coronavirus: 2019-ncov," *Viruses*, vol. 12, no. 2, p. 135, 2020.

[5] E. Nogales, "The development of cryo-em into a mainstream structural biology technique," *Nature methods*, vol. 13, no. 1, pp. 24–27, 2016.

[6] J. Frank, "Advances in the field of single-particle cryo-electron microscopy over the last decade," *Nature protocols*, vol. 12, no. 2, pp. 209–212, 2017.

[7] Y. Cheng, "Single-particle cryo-em—how did it get here and where will it go," *Science*, vol. 361, no. 6405, pp. 876–880, 2018.

[8] S. Raunser, "Cryo-em revolutionizes the structure determination of biomolecules," *Angewandte Chemie International Edition*, vol. 56, no. 52, pp. 16 450–16 452, 2017.

[9] A. Patwardhan, "Trends in the electron microscopy data bank (emdb)," *Acta Crystallographica Section D: Structural Biology*, vol. 73, no. 6, pp. 503–508, 2017.

[10] S. Lindert, R. Staritzbichler, N. Wötzel, M. Karakaş, P. L. Stewart, and J. Meiler, "Em-fold: De novo folding of α-helical proteins guided by intermediate-resolution electron microscopy density maps," *Structure*, vol. 17, no. 7, pp. 990–1003, 2009.

[11] M. L. Baker, S. S. Abeysinghe, S. Schuh, R. A. Coleman, A. Abrams, M. P. Marsh, C. F. Hryc, T. Ruths, W. Chiu, and T. Ju, "Modeling protein structure at near atomic resolutions with gorgon," *Journal of structural biology*, vol. 174, no. 2, pp. 360–373, 2011.

[12] M. R. Baker, I. Rees, S. J. Ludtke, W. Chiu, and M. L. Baker, "Constructing and validating initial cα models from subnanometer resolution density maps with pathwalking," *Structure*, vol. 20, no. 3, pp. 450–463, 2012.

[13] M. Chen, P. R. Baldwin, S. J. Ludtke, and M. L. Baker, "De novo modeling in cryo-em density maps with pathwalking," *Journal of structural biology*, vol. 196, no. 3, pp. 289–298, 2016.

[14] M. Chen and M. L. Baker, "Automation and assessment of de novo modeling with pathwalking in near atomic resolution cryoem density maps," *Journal of structural biology*, vol. 204, no. 3, pp. 555–563, 2018.

[15] R. Y.-R. Wang, M. Kudryashev, X. Li, E. H. Egelman, M. Basler, Y. Cheng, D. Baker, and F. DiMaio, "De novo protein structure determination from near-atomic-resolution cryo-em maps," *Nature methods*, vol. 12, no. 4, pp. 335–338, 2015.

[16] B. Frenz, A. C. Walls, E. H. Egelman, D. Veesler, and F. DiMaio, "Rosettaes: a sampling strategy enabling automated interpretation of difficult cryo-em maps," *Nature methods*, vol. 14, no. 8, pp. 797–800, 2017.

[17] P. D. Adams, P. V. Afonine, G. Bunkóczi, V. B. Chen, I. W. Davis, N. Echols, J. J. Headd, L.-W. Hung, G. J. Kapral, R. W. Grosse-Kunstleve *et al.*, "Phenix: a comprehensive python-based system for macromolecular structure solution," *Acta Crystallographica Section D: Biological Crystallography*, vol. 66, no. 2, pp. 213–221, 2010.

[18] P. V. Afonine, B. K. Poon, R. J. Read, O. V. Sobolev, T. C. Terwilliger, A. Urzhumtsev, and P. D. Adams, "Real-space refinement in phenix for cryo-em and crystallography," *Acta Crystallographica Section D: Structural Biology*, vol. 74, no. 6, pp. 531–544, 2018.

[19] T. C. Terwilliger, P. D. Adams, P. V. Afonine, and O. V. Sobolev, "A fully automatic method yielding initial models from high-resolution cryo-electron microscopy maps," *Nature methods*, vol. 15, no. 11, pp. 905–908, 2018.

[20] ——, "Cryo-em map interpretation and protein model-building using iterative map segmentation," *Protein science*, vol. 29, no. 1, pp. 87–99, 2020.

[21] G. Terashi and D. Kihara, "De novo main-chain modeling for em maps using mainmast," *Nature communications*, vol. 9, no. 1, pp. 1–11, 2018.

[22] K. Xu, Z. Wang, J. Shi, H. Li, and Q. C. Zhang, "A2-net: Molecular structure estimation from cryo-em density volumes," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 1230–1237.

[23] D. Si, S. A. Moritz, J. Pfab, J. Hou, R. Cao, L. Wang, T. Wu, and J. Cheng, "Deep learning to predict protein backbone structure from high-resolution cryo-em density maps," *Scientific reports*, vol. 10, no. 1, pp. 1–22, 2020.

[24] J. Pfab and D. Si, "Deeptracer: Predicting backbone atomic structure from high resolution cryo-em density maps of protein complexes," *bioRxiv*, 2020.

[25] J. He and S.-Y. Huang, "Full-length de novo protein structure determination from cryo-em maps using deep learning," *bioRxiv*, pp. 2020–08, 2021.

[26] R. Apweiler, A. Bairoch, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane *et al.*, "Uniprot: the universal protein knowledgebase," *Nucleic acids research*, vol. 32, no. suppl_1, pp. D115–D119, 2004.

[27] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The protein data bank," *Nucleic acids research*, vol. 28, no. 1, pp. 235–242, 2000.

[28] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko *et al.*, "Highly accurate protein structure prediction with alphafold," *Nature*, pp. 1–11, 2021.

[29] K. Tunyasuvunakool, J. Adler, Z. Wu, T. Green, M. Zielinski, A. Žídek, A. Bridgland, A. Cowie, C. Meyer, A. Laydon *et al.*, "Highly accurate protein structure prediction for the human proteome," *Nature*, pp. 1–9, 2021.

[30] F. DiMaio and W. Chiu, "Tools for model building and optimization into near-atomic resolution electron cryo-microscopy density maps," *Methods in enzymology*, vol. 579, pp. 255–276, 2016.

[31] J. Hanson, K. Paliwal, T. Litfin, Y. Yang, and Y. Zhou, "Accurate prediction of protein contact maps by coupling residual two-dimensional bidirectional long short-term memory with convolutional neural networks," *Bioinformatics*, vol. 34, no. 23, pp. 4039–4045, 2018.

[32] G. Tang, L. Peng, P. R. Baldwin, D. S. Mann, W. Jiang, I. Rees, and S. J. Ludtke, "Eman2: an extensible image processing suite for electron microscopy," *Journal of structural biology*, vol. 157, no. 1, pp. 38–46, 2007.

[33] M. N. Nguyen and M. Madhusudhan, "Biological insights from topology independent comparison of protein 3d structures," *Nucleic acids research*, vol. 39, no. 14, pp. e94–e94, 2011.

[34] R. McGreevy, I. Teo, A. Singharoy, and K. Schulten, "Advances in the molecular dynamics flexible fitting method for cryo-em modeling," *Methods*, vol. 100, pp. 50–60, 2016.