Deep Learning Approach to Identify Protein's Secondary Structure Elements

Mohammad Bataineh, Kamal Al Nasr(✉), Richard Mu, and Mohammed Alamri

Tennessee State University, Nashville, TN 37209, USA
Kalnasr@tnstate.edu

Abstract. Cryo-electron microscopy (cryo-EM) has become a crucial method for structure determination. Despite the substantial growth in deposited cryo-EM maps driven by advances in microscopy and image processing, accurately constructing models from these maps remains challenging. Extracting secondary structure information from EM maps is valuable for cryo-EM modeling. In this context, we introduce a novel deep learning secondary structure annotation framework specifically designed for intermediate-resolution cryo-EM maps, employing a three-dimensional Inception architecture. Testing it on diverse datasets, including maps with authentic intermediate resolutions, demonstrates its accuracy and robustness in identifying secondary structures in cryo-EM maps. We conducted a comparative analysis of our results against frameworks that exist in the state-of-the-art, and our framework demonstrated superior performance across nearly all secondary structure elements. We employed the F1 accuracy metric, yielding an average F1 score of 0.657 for helix, 0.712 for coil, and 0.596 for sheet predictions. Notably, certain helix and sheet predictions achieved an impressive F1 score of 0.881.

Keywords: Protein Modeling · Protein Secondary Structure Elements · Deep Learning · Cryo-EM Map · Inception Architecture · Machine Learning

1 Introduction

Progress in both microscopy tools and image processing algorithms has resulted in a growing abundance of cryo-electron microscopy (cryo-EM) maps [1–3]. Increasing the resolution of cryo-EM has opened doors to elucidating the structures of biological systems that were once considered too challenging to tackle, now achieving remarkable levels of detail [4,5]. It is important to note that the ultimate objective of cryo-EM is not merely the acquisition of 3D maps but the precise determination of atomic structure [6–8].

Constructing precise structural models for cryo-EM maps poses a significant challenge [9]. The methods typically employed, such as rigid fitting and flexible fitting, rely on pre-existing template structures for the accurate placement of atomic structures into EM maps. When template structures are lacking, the

W. Peng et al. (Eds.): ISBRA 2024, LNBI 14954, pp. 461–472, 2024.
https://doi.org/10.1007/978-981-97-5128-0_38

necessity for de novo modeling tools arises to construct complete atomic models within EM density maps.

Jiang et al. [10] developed Helixhunter, software for helix identification, length, and orientation using cross-correlation search and feature extraction on density maps. They achieved over 88% helix detection accuracy on 8 Å resolution simulated maps, with some misclassifications and missed helices. Kong and Ma [11] introduced Sheetminer, successfully identifying Beta-sheets in protein structures with promising results on Cryo-EM and X-ray density maps at various resolutions. Kong et al. [12] developed two methods for Beta-sheet identification, one relying on Sheetminer's output and the other using deconvolution for improved results. Despite advancements, de novo modeling tools face accuracy limitations, leading to a gap between the quantity of cryo-EM maps and successfully reconstructed 3D structures [13]. Machine learning models offer potential to enhance predictions based on their capabilities.

Li et al. [14] achieved significant progress in secondary structure prediction using a CNN framework, testing it on 25 simulated 8 Å cryo-EM maps, achieving an average sensitivity and specificity of 71.52% and 97.86% for Alph-helix and Beta-sheet detection, surpassing SVM methods. Subramaniya et al. [15] introduced Emap2sec, predicting secondary structure in cryo-EM maps. Their validation involved two datasets, yielding impressive results at 6 Å with an overall F1 score of 0.798, Alph-helix at 0.848, Beta-sheet at 0.828, and other structural elements at 0.672. At 10 Å resolution, results remained substantial, with Alph-helix, Beta-sheet, and other elements achieving F1 accuracy scores of 0.82, 0.75, and 0.64, respectively.

Shifting the focus to another framework, Haruspex, developed by Mostosi et al. [16], employed U-net architecture to predict secondary structure elements. Notably, Haruspex was primarily designed for detecting and annotating RNA/DNA and protein secondary structure elements within high-resolution cryo-EM maps. This framework showed promising results but faced challenges in 'unassigned' regions, resulting in an unbalanced classification that impacted its efficiency.

Later Wang et al. [17] presented Emap2sec+, an updated version of Emap2sec. Where deep Residual convolutional neural network architecture was developed, ResNet. To predict the secondary structure elements, they tried to classify each voxel into one of three elements: Alph-Helix, Beta-sheet, or others. Emap2sec+ was trained and tested on the simulated and experimental datasets. In the simulated dataset, 108 non-redundant maps at 6 and 10 Å were used for training and testing. For the experimental dataset, 83 cryo-EM images were used for training and testing as well. In addition, Emap2sec+ outperformed Haruspex in predicting protein secondary structure in the F1 score term.

He and Huang [18] recently introduced EMNUSS, a robust framework utilizing advanced U-net architecture (nested U-net or U-net++) with skip connectors to enhance predictive power in secondary structure element (SSE) prediction. EMNUSS showcased its capabilities across diverse datasets, spanning simulated, mid-resolution, and high-resolution maps, demonstrating significant potential in

SSE prediction. It consistently outperformed other frameworks in various evaluation metrics, such as F1 scores and Q3 accuracy, making it a notable contribution to the field of SSE prediction.

Limitations, such as using improper grid intervals as seen in Haruspex and relying on small input chunks in Emap2sec and Emap2sec+, result in constraints that may not accommodate an average secondary structure element larger than those input chunks. In order to address the limitations of current methods, we've introduced an innovative deep learning framework for predicting secondary structures in authentic cryo-EM maps at intermediate resolutions. Our approach utilizes a three-dimensional (3D) Inception architecture, enabling rapid and precise prediction of protein secondary structures in cryo-EM maps of diverse dimensions. Our method has demonstrated a substantial enhancement in performance, particularly when applied to experimental maps at middle resolutions. Our approach has been tested exclusively on intermediate-resolution maps, making it specifically tailored for such data. This opens up opportunities for future research to test our approach on low- or high-resolution maps, or to develop a new model designed to handle these different resolutions.

2 Methods

2.1 Dataset

In order to develop and assess the performance of our framework, we compiled a diverse and non-redundant set of intermediate-resolution electron microscopy (EM) maps for experimentation. Our initial search was conducted within the EMDataResource database, targeting EM maps with resolutions falling within the 4 to 10 Angstrom range, while also ensuring the availability of associated PDB files. However, these criteria alone did not suffice to create a dependable dataset. To prevent the inclusion of identical or highly similar EM maps, we implemented additional selection constraints. Specifically, we excluded any chains displaying the following characteristics: 1. Presence of missing residues. 2. Absence of secondary structure information. 3. A sequence identity similarity exceeding 25% with any chain already present in the dataset.

Following the application of these aforementioned selection conditions, we successfully curated a collection of 487 unique chain maps. We divided these maps into two distinct sets for training and testing purposes: 455 maps were randomly selected for the training set, while the remaining 32 maps constituted the testing set.

2.2 Network Architecture

The Inception architecture represents a significant milestone in computer vision and deep learning. In our work, we have incorporated the Inception architecture, creating our own custom variant of this 3D deep convolutional neural network (CNN) architecture. Our proposed architecture, as depicted in Fig. 1, comprises

several key elements. It commences with a Stem layer, responsible for preprocessing the 3D data. This is followed by a Maxpool layer, subsequently leading to the core Inception blocks layer. After that, an Up-sampling layer is applied, and the network concludes with a Final layer, which reduces the neuron count to 4 to align with the number of classes we intend to predict. A more detailed breakdown of the Inception blocks and their sublayers can be found in Fig. 2.

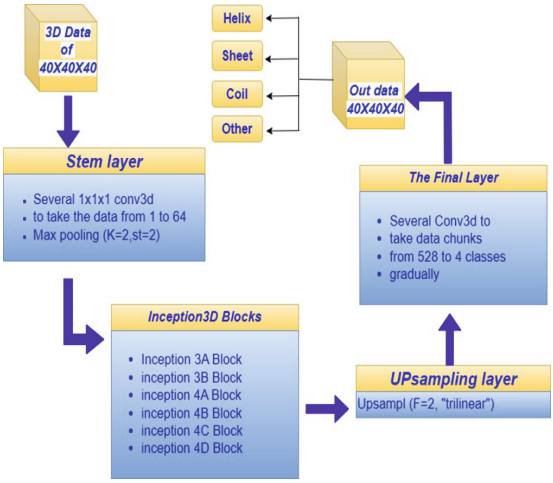


Fig. 1. Our proposed 3D deep convolutional neural network (CNN) architecture, it comprises several key elements. It commences with a Stem layer, which is responsible for preprocessing the 3D data. This is followed by Maxpool layer, Inception blocks layer, Up-sampling layer, and the Final layer, reducing neurons to 4 for class prediction alignment.

2.3 Processing Training Data

To prepare our training maps for analysis, we conducted the following steps. One of our goals was to annotate the dataset for both training and testing stages effectively. Uniform Grid and Resolution: We started by implementing a standardized interval grid for all maps using trilinear interpolation. This process ensured that each interval was precisely set to 1.0 Å, creating consistency across the dataset. Ground Truth Assignment: Subsequently, we assigned ground truth labels to each voxel to identify secondary structure elements. For this task, we associated each voxel with the nearest backbone atom (N, C, C-alph, or O atom) within a 3.0 Å radius.

In cases where no backbone atoms were found within this radius, we assigned a background label to the voxel. The input size for the framework was standardized to $40 \times 40 \times 40$, ensuring a consistent format for analysis. Density Value Normalization: Finally, we normalized the density values for each voxel to fall within the range of 0 to 1. This normalization process allowed us to exclude

any chunks with all values equal to zero, preventing their participation in the training process. This approach contributed to the effectiveness of the training procedure.

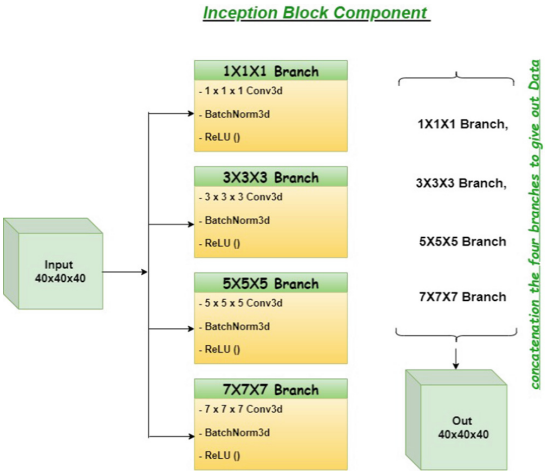


Fig. 2. Inception blocks components and other sublayers.

2.4 Network Training

We created three distinct Inception architectures, applying identical hyperparameters to each. The input data consisted of chunks with dimensions $40 \times 40 \times 40$. We partitioned 10% of the training maps for validation. Our framework was implemented using PyTorch, with 150 epochs and a batch size of 16. We utilized the Adam optimizer and employed the cross-entropy loss function, setting the learning rate at 1e−3.

2.5 Evaluation and Comparison

To assess the outcomes of our study, we employed the F1 score metric, which represents the balanced combination of precision and recall when assessing assignments. This metric was utilized to gauge our framework's performance at the voxel level. For a comprehensive evaluation against the current state of the art, we opted to compare our results with EMNUSS, a recent and widely recognized and resilient framework used for forecasting secondary structure elements in Cryo-EM maps. We computed the F1 score for both frameworks, revitalizing some of the results to provide a more insightful assessment of their performance.

3 Results and Discussion

3.1 Comparison with EMNUSS

We conducted a comparative analysis of our framework and EMNUSS using a test set comprised of 32 experimental EM maps of middle-resolution, with resolutions spanning from 4.0 to 10.0 Å. In Fig. 3, we present a visual representation of the voxel F1 score comparisons between the two methods. The figure clearly demonstrates that our framework outperformed EMNUSS significantly when applied to the middle-resolution experimental maps.

Our proposed method significantly outperforms EMNUSS across all classes in terms of voxel F1 scores, achieving an Overall F1 accuracy of 0.739 compared to EMNUSS's 0.277. For Helix prediction, the Proposed Method F1 scores 0.657 versus EMNUSS's 0.14, and for Sheet prediction, it achieves 0.596 F1 score against EMNUSS's 0.093. In Coil prediction, the Proposed Method also leads with a 0.712 F1 score compared to EMNUSS's 0.598. This demonstrates the superior reliability and effectiveness of our Proposed Method for protein structure prediction.

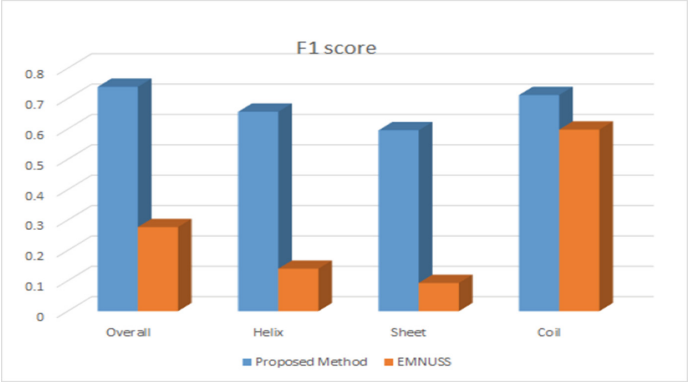
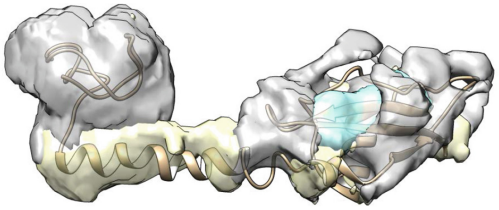
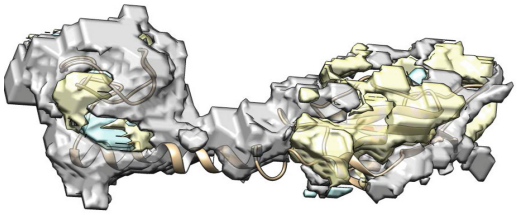


Fig. 3. Average voxel F1 scores, comparison of our framework, and EMNUSS for different secondary structure classes on the middle-resolution experimental map.

In Fig. 4, the comparison between EMD-1263H map visualizations using both the proposed method and EMNUSS reveals significant performance differences. Our framework achieves a notable overall F1 accuracy of 0.778, surpassing EMNUSS by a large margin (0.277), particularly excelling in identifying helices and strands. Additionally, our method outperforms EMNUSS in predicting secondary structure classes, including alpha helices, beta-sheets, and coils, with higher voxel F1 scores. EMNUSS struggles with accuracy across classes, mislabeling helical structures within sheet regions and misidentifying coil and sheet regions consistently. While our method exhibits slightly lower F1 scores



(a) Our proposed method



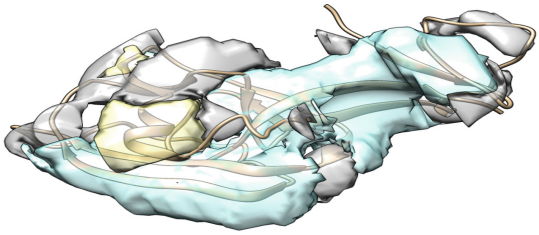
(b) EMNUSS

Fig. 4. EMD 1263-A prediction visualization.

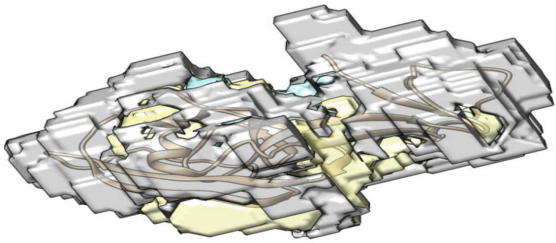
in sheet prediction (0.59), it provides smoother and more interpretable visualizations compared to EMNUSS, with highly accurate predictions for helix and coil regions, although minor misses exist. Overall, our approach demonstrates superior performance across all structural classes.

Figure 5 showcases EMD 8169-C, a 6.56 Å map, with predictions from both our proposed method and EMNUSS. Our method notably outperforms EMNUSS, particularly excelling in accurately predicting sheet regions with a high F1 score of 0.872. Similarly, our method demonstrates strong performance in identifying coil regions, although occasional misses occur. However, helix prediction regions show lower accuracy, with some expected helical regions missed while others are incorrectly labeled as helices. In contrast, EMNUSS misclassifies segments, labeling them predominantly as coil or background and overlooking helix and sheet regions. Helix regions are entirely overlooked, with sheet regions misidentified as coil or background. The predicted sheet region by EMNUSS represents only a fraction of the actual area, located differently.

Figure 6 illustrates EMD-12221A, a 9.5 Å resolution map, lacking sheet regions in the protein chain. Our proposed framework showcases highly satisfactory results in predicting secondary structure elements, achieving an F1 score of 0.88 for helix prediction, with nearly flawless visualization despite occasional missing data. Coil prediction also demonstrates excellent visualization and an impressive F1 score. Conversely, EMNUSS performs poorly, largely failing to identify helix regions and misclassifying them as coil or background, while erroneously identifying coil regions as predominant elements. The most glaring error in EMNUSS predictions is its incorrect labeling of sheet regions despite their absence in this protein chain.

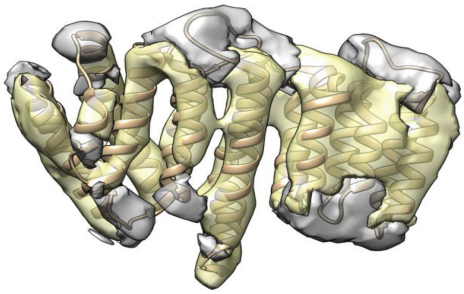


(a) Our proposed method

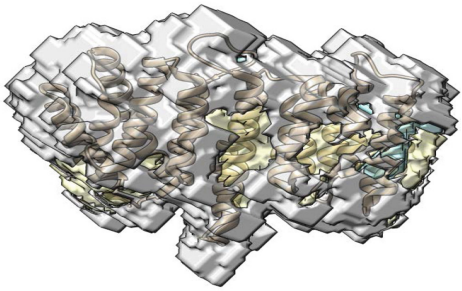


(b) EMNUSS

Fig. 5. EMD 8169-C prediction visualization.

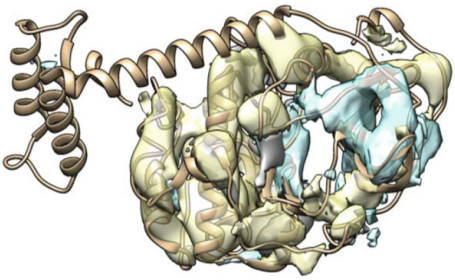


(a) Our proposed method

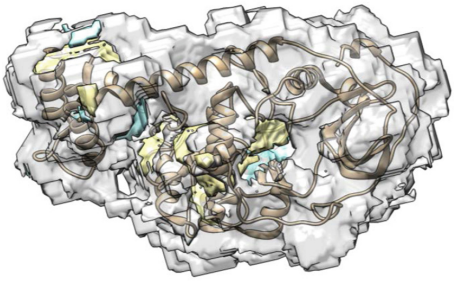


(b) EMNUSS

Fig. 6. EMD 12221-A prediction visualization.



(a) Our proposed method

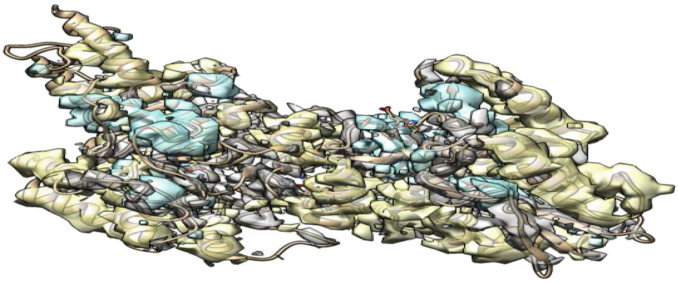


(b) EMNUSS

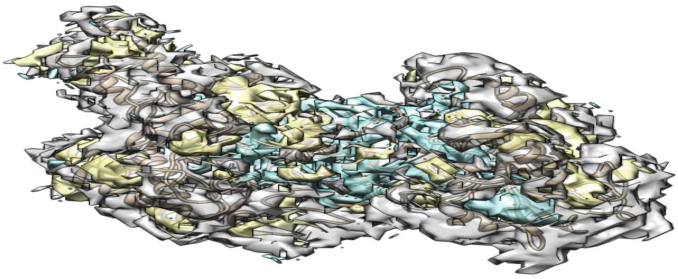
Fig. 7. EMD 21692-A prediction visualization.

Figure 7 depicts EMD 21316-A, presenting challenges for both our framework and EMNUSS in providing accurate visualizations. While our framework showed some success in predicting helix regions, several were missed, and the identification of sheet regions remained inadequate, resulting in a low F1 score. Although coil regions were relatively better predicted, many were still overlooked. Overall, our proposed method did not perform as well in predicting this protein chain compared to previous ones, exhibiting significant shortcomings. Despite EMNUSS achieving a higher overall F1 score, its visualization results in Fig. 7 reveal poor predictions, with most protein regions misclassified as coils and significant overlaps between different predicted classes. Specifically, EMNUSS completely missed identifying helix and sheet regions within the protein.

Figure 8 illustrates EMD 21156-A, portraying a complex and tangled protein structure. Our proposed method accurately predicts approximately 50% of the helices while struggling with sheet prediction, resulting in a low F1 score of 0.16, despite some correct identifications. However, it performs relatively well in predicting coil regions, with a decent F1 score of 0.6. In contrast, EMNUSS achieves higher F1 scores for both sheets and coils but produces a chaotic visualization lacking clarity, with numerous helices misidentified as coils and sheets appearing dislocated and misclassified. Despite the higher F1 scores, our proposed method's visualization offers better coherence and understandability



(a) Our proposed method



(b) EMNUSS

Fig. 8. EMD 21156-A prediction visualization.

compared to EMNUSS, indicating its superiority in presenting predictions despite similar struggles in accuracy.

3.2 Comparison with Different Architecture

To further scrutinize the effectiveness of our proposed framework, we've introduced an additional pair of Inception block architectures, we named them second and third designs. Our approach involves replacing the Inception blocks (as shown in Fig. 2) within our network to explore alternative designs and understand their impact on the results. In the following sections, we'll outline these two additional Inception block designs.

Table 1. Average Voxel F1 Scores for Various designs,

Class	Main design	Second design	Third design
Overall	0.739	0.705	0.739
Helix	0.657	0.597	0.654
Sheet	0.596	0.55	0.596
Coil	0.712	0.684	0.715

- Second design: The design includes four Conv3D branches also, each consisting of a sequence: a $1 \times 1 \times 1$ Conv3D layer, followed by a $3 \times 3 \times 3$ Conv3D layer, then a $5 \times 5 \times 5$ Conv3D layer. After this, there's a pool branch with a Conv3D layer. Additionally, before each of these branches, a $1 \times 1 \times 1$ kernel Conv3D sublayer this time to transform the data without big change in the kernel size.
- Third design: This architecture comprises four Conv3D branches: a $1 \times 1 \times 1$ Conv3D layer, followed by a $3 \times 3 \times 3$ Conv3D layer, then a $5 \times 5 \times 5$ Conv3D layer, and finally a pool branch with a Conv3D layer. Each branch is preceded by a $3 \times 3 \times 3$ kernel Conv3D sublayer, strategically included to prepare and enhance the 3D data, ultimately leading to improved performance.

We conducted a comparative analysis of three Inception block designs, all evaluated using the same test set, consisting of 32 experimental EM maps at middle resolution. Table 1 clearly illustrates that our primary design outperformed the second design significantly. However, the third design displayed highly competitive results, almost on par with the first design.

We attribute the excellent performance of the designs to the effective use of the $3 \times 3 \times 3$ kernel Conv3D sublayer, which proved to be well-suited for Cryo-EM maps, capable of extracting meaningful information. This kernel size is particularly conducive to this type of data.

4 Conclusion

We created an advanced deep learning framework designed for the prediction and annotation of protein secondary structures within EM density maps. This framework underwent comprehensive testing and evaluation using a dataset of middle-resolution experimental maps. The results clearly demonstrated that our framework substantially enhanced the accuracy of secondary structure detection, surpassing existing methods. Furthermore, we introduced two additional Inception block designs to investigate their influence on the results. A promising direction for future research involves developing an improved network architecture aimed at enhancing prediction accuracy.

References

1. Nogales, E.: The development of cryo-EM into a mainstream structural biology technique. Nat. Meth. **13**(1), 24–27 (2016)
2. Frank, J.: Advances in the field of single-particle cryo-electron microscopy over the last decade. Nat. Protoc. **12**(2), 209–212 (2017)
3. Cheng, Y.: Single-particle cryo-EM-how did it get here and where will it go. Science **361**(6405), 876–80 (2018)
4. Luque, D., Castón, J.R.: Cryo-electron microscopy for the study of virus assembly. Nat. Chem. Biol. **16**(3), 231–239 (2020)
5. 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. **432**(19), 5365–5377 (2020)

6. Yin, S., Zhang, B., Yang, Y., et al.: Clustering enhancement of noisy cryo-electron microscopy single-particle images with a network structural similarity metric. J. Chem. Inf. Model. **59**(4), 1658–1667 (2019)
7. Chen, M., Baldwin, P.R., Ludtke, S.J., et al.: De Novo modeling in cryo-EM density maps with Pathwalking. J. Struct. Biol. **196**(3), 89–298 (2016)
8. Chen, M., Baker, M.L.: Automation and assessment of De Novo modeling with Pathwalking in near atomic resolution cryoEM density maps. J. Struct. Biol. **204**(3), 555–563 (2018)
9. Terwilliger, T.C., Adams, P.D., Afonine, P.V., et al.: Cryo-EM map interpretation and protein model-building using iterative map segmentation. Protein Sci. **29**(1), 87–99 (2020)
10. Jiang, W., Baker, M.L., Ludtke, S.J., Chiu, W.: Bridging the information gap: computational tools for intermediate resolution structure interpretation. J. Mol. Biol. **308**(5), 1033–1044 (2001)
11. Kong, Y., Ma, J.: A structural-informatics approach for mining Beta-sheets: locating sheets in intermediate-resolution density maps. J. Mol. Biol. **332**(2), 399–413 (2003)
12. Kong, Y., Zhang, X., Baker, T.S., Ma, J.: A structural-informatics approach for tracing Beta-sheets: building pseudo-C(alph) traces for Beta-strands in intermediate-resolution density maps. J. Mol. Biol. **339**(1), 117–130 (2004)
13. Coskuner-Weber, O., Caglayan, S.I.: Secondary structure dependence on simulation techniques and force field parameters: from disordered to ordered proteins. Biophys. Rev. **13**(6), 1173–1178 (2021)
14. Li, R., Si, D., Zeng, T., Ji, S., He, J.: Deep convolutional neural networks for detecting secondary structures in protein density maps from cryo-electron microscopy. In: 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 41–46. IEEE (2016)
15. Subramaniya, S.R.M.V., Terashi, G., Kihara, D.: Protein secondary structure detection in intermediate-resolution cryo-EM maps using deep learning. Nat. Meth. **16**(9), 911–917 (2019)
16. Mostosi, P., Schindelin, H., Kollmannsberger, P., Thorn, A.: Haruspex: a neural network for the automatic identification of oligonucleotides and protein secondary structure in cryo-electron microscopy maps. Angew. Chem. Int. Ed. **59**(35), 14788–14795 (2020)
17. Wang, X., Alnabati, E., Aderinwale, T.W., Subramaniya, S.R., Terashi, G., Kihara, D.: Detecting protein and DNA/RNA structures in cryo-EM maps of intermediate resolution using deep learning. Biophys. J . **120**(3), 81a (2021)
18. He, J., Huang, S.Y.: EMNUSS: a deep learning framework for secondary structure annotation in cryo-EM maps. Brief. Bioinform. **22**(6), bbab156 (2021)