Old Dominion University

ODU Digital Commons

Computer Science Faculty Publications Computer Science

2023

An Approach to Developing Benchmark Datasets for Protein Secondary Structure Segmentation from Cryo-EM Density Maps

Thu Nguyen
Old Dominion University

Yongcheng Mu
Old Dominion University

Jiangwen Sun
Old Dominion University

Jing He
Old Dominion University

Follow this and additional works at: https://digitalcommons.odu.edu/computerscience_fac_pubs

Part of the Amino Acids, Peptides, and Proteins Commons, Artificial Intelligence and Robotics Commons, and the Computational Biology Commons

Original Publication Citation

Nguyen, T., Mu, Y., Sun, J., & He, J. (2023). An approach to developing benchmark datasets for protein secondary structure segmentation from Cryo-EM density maps. In *Proceedings of the 14th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics* (Article 102). Association for Computing Machinery. https://doi.org/10.1145/3584371.3612947

This Conference Paper is brought to you for free and open access by the Computer Science at ODU Digital Commons. It has been accepted for inclusion in Computer Science Faculty Publications by an authorized administrator of ODU Digital Commons. For more information, please contact digitalcommons@odu.edu.

An Approach to Developing Benchmark Datasets for Protein Secondary Structure Segmentation from Cryo-EM Density Maps

Thu Nguyen
Department of Computer Science
Old Dominion University
Norfolk VA USA
tnguy028@odu.edu

Yongcheng Mu
Department of Computer Science
Old Dominion University
Norfolk VA USA
ymu004@odu.edu

Jiangwen Sun
Department of Computer Science
Old Dominion University
Norfolk VA USA
jsun@odu.edu

Jing He[†]
Department of Computer Science
Old Dominion University
Norfolk VA USA
jhe@cs.odu.edu

ABSTRACT

More and more deep learning approaches have been proposed to segment secondary structures from cryo-electron density maps at medium resolution range (5-10Å). Although the deep learning approaches show great potential, only a few small experimental data sets have been used to test the approaches. There is limited understanding about potential factors, in data, that affect the performance of segmentation. We propose an approach to generate data sets with desired specifications in three potential factors - the protein sequence identity, structural contents, and data quality. The approach was implemented and has generated a test set and various training sets to study the effect of secondary structure content and data quality on the performance of DeepSSETracer, a deep learning method that segments regions of protein secondary structures from cryo-EM map components. Results show that various content levels in the secondary structure and data quality influence the performance of segmentation for DeepSSETracer.

CCS Concepts

• Computing Methodologies → Machine Learning

• Applied Computing → Life and Medical Sciences → Computational Biology

Keywords

Deep learning; Secondary structure; Cryo-EM; Protein; Benchmark data

ACM Reference format:

Thu Nguyen, Yongcheng Mu, Jiangwen Sun and Jing He. 2023. An Approach to Developing Benchmark Datasets for Protein Secondary Structure Segmentation from Cryo-EM Density Maps. In *Proceedings of the 14th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics. ACM-BCB, Houston, TX, USA, 8 pages.* https://doi.org/10.1145/3584371.3612947

1 INTRODUCTION

Deep learning has been widely applied in biological problems. With the fast accumulation of cryo-electron microscopy (cryo-EM) image data and 3-dimensional molecular structure data, an increasing number of deep learning methods have been developed in many subdomains of cryo-EM or cryo-electron tomography (cryo-ET). For example, deep learning approaches have been developed for picking out molecular particles from 2D cryo-EM images [1], for segmentation of protein secondary structures from cryo-EM 3D density maps [2-6], for deriving initial backbones from cryo-EM density maps [7], and for segmentation of cellular objects from cryo-ET images [8, 9]. The problem of segmentation of protein secondary structures from a cryo-EM maps of medium resolution (5-10Å) is to detect the location of helices and β-sheets in the cryo-EM density map (Fig. 1).

Many image processing methods have been proposed for segmentation of protein secondary structures [10-14]. In addition, recent deep learning methods have shown great potential leading to increased accuracy [2-6, 15]. However, different approaches were tested using different experimental data, and there has been limited study suggesting a proper procedure to construct a test data set. Emap2sec used 4-fold cross-validation to train and test using 43 experimental maps with the medium resolution [3]. The data set was obtained after two steps of screening. The first step eliminates low-quality maps that have lower than 0.65 cross-correlation score

[†]Corresponding author email: jhe@cs.odu.edu

This work is licensed under a Creative Commons Attribution International 4.0 License.
BCB '23, September 3–6, 2023, Houston, TX, USA
© 2023 Copyright is held by the owner/author(s).
ACM ISBN 979-8-4007-0126-9/23/09.
https://doi.org/10.1145/3584371.3612947

between the map and the atomic structure. The second screening removes maps if any of their chains share more than 25% sequence identity with any chain of a map in the data set [3]. EMNUSS used the same 43 experimental maps that was developed in EMap2sec method [6]. Emap2sec+, which segments both secondary structures and nucleic acids, was tested four times using medium-resolution experimental maps, each with 4 to 5 density maps. The small number of experimental maps used in testing may be related to limited cryo-EM maps that contain nucleic acids. Haruspex was trained and tested using high-resolution density maps with resolution better than 4Å [4]. It was not tested using medium-resolution maps. Our previous method DeepSSETracer was tested using cryo-EM density map components centered around protein chains, rather than entire density maps [5]. Structure validation has been an important problem in the cryo-EM community. There have been coordinated efforts to develop test data for structure validation [16-18], Map and Model Challenge of 2016 [17], and Model Metrics Challenge in 2019 [18]. The data used in the challenges were designed for validation of atomic structures, particularly for cryo-EM maps with better than 5 Å resolution, making them inapplicable for secondary structure segmentation from medium-resolution maps. A benchmark data set will likely advance the development of approaches for the segmentation of secondary structures.

Developing benchmark set requires understanding the distribution of data over metrics that potentially influence the segmentation. A general hypothesis is to eliminate bad quality data in training and testing. However, there has been no study how quality affects the performance and how many data at what quality level should be included in a benchmark set to fairly represent the entire database. There has also been no study on other potential factors for segmentation, such as the complexity of structures, the size of secondary structures, the size of the protein, and repetitive sequences that are often in cryo-EM density maps.

In this study, we developed a data stratification approach to create data sets satisfying pre-defined specifications in sequence identify, secondary structure contents, and quality of data. The idea is to cluster the entire data set using potential factors and then to compose and select data sets satisfying specific requirements. This approach was implemented and has generated a test set and various training sets with different levels of data quality and secondary structures contents. Our results show that the distributions of data across different structure clusters and quality clusters in the training set can affect the performance of deep learning models of DeepSSETracer.

2 METHOD

Cryo-EM density maps with resolution range 5-10 Å were downloaded from Electron Microscopy Data Bank (EMDB) [19]. Their corresponding atomic structures were downloaded from Protein Data Bank (PDB) [20]. Since many cryo-EM density maps contain multiple copies of the same atomic structure chains, only one of the repeated chains in a map was used as an envelope to isolate the corresponding density region with Chimera [21]. The data set used in this paper contains 1292 atomic structure chains and their corresponding regions in the cryo-EM maps. The overall idea for stratification is to first create clusters from the entire data set based on each factor that potentially affects the performance of segmentation training and testing. Three factors are used in this study – the sequence similarity between protein chains, the secondary structure content (helix, β-sheets) in a chain, the structure-map fit that often reflects the quality of a map region (Fig. 1B).

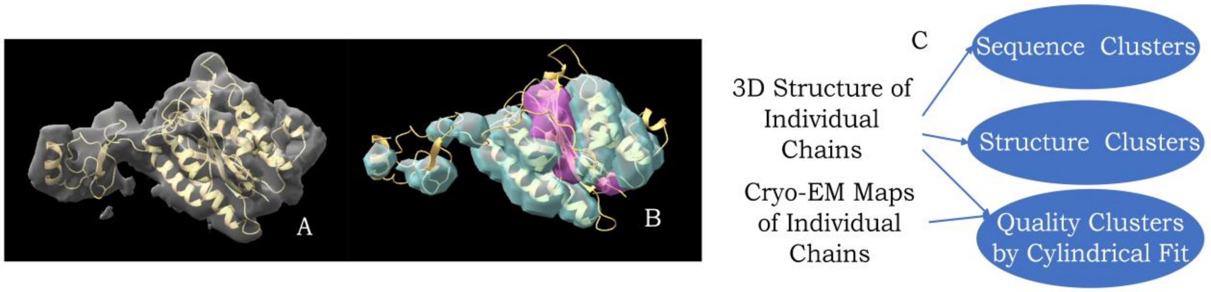
A
B
C
3D Structure of Individual Chains
Cryo-EM Maps of Individual Chains
Sequence Clusters
Structure Clusters
Quality Clusters by Cylindrical Fit

Figure 1: The segmentation of secondary structure problem and clusters of data in three metrics. (A): The atomic structure (ribbon) and its corresponding cryo-EM density map region (gray) for EMD-3850 PDB-5oqm chain 4; (B): Segmented helix (blue) and β sheet (pink) regions detected using DeepSSETracer [5]. The density map and the atomic structure are visualized in ChimeraX [22]. (C): Clusters created based on three metrics - sequence similarity, secondary structure content, and quality of structural fit in cryo-EM maps.

2.1 Creation of Sequence Clusters

A matrix of sequence identity scores for protein chains was created to represent the similarity between each pair of sequences across all obtained chains. The sequence identity score of any given pair was calculated by aligning two chain sequences using Needleman Wunsch algorithm [23] and defined as the percentage of identical amino acids in the alignment. The obtained similarity matrix was subsequently used to acquire sequence clusters of chains by the agglomerative hierarchical clustering algorithm. All clusters were created to ensure that any two chains with an identity score equal or above 40% were grouped into the same cluster. The agglomerative clustering algorithm is implemented in Python, using scikit-learn library [24]. The single-linkage with a distance threshold of 60 was used to maintain that any two chains with a sequence identity score equal to or above 40% were grouped into the same cluster, and any cluster shared no more than 40% sequence identity with the remaining clusters.

2.2 Creation of Structure Clusters

Seven variables describing varying aspects of secondary structure content in each chain were used to obtain structure clusters of chains. These variables include the chain length (i.e., the number of amino acid residues), the number of helix residues, the number of β-sheet residues, the number of helices, the number of β-strands, the average helix length, and the average of β-strand length. Each variable was normalized to between 0 and 1 before clustering. Specifically, the length of a chain was normalized by the minimum and maximum chain length in the dataset. The number of helix (β-sheet) residues was divided by the chain length, followed by min-max normalization. Similarly, the number of helices (β-strands) was also divided by the chain length followed by min-max normalization. The average length of helices (β-strands) in a chain was calculated as the number of total helix (β-strands) residues divided by the number of helices (β-strands) and was also normalized by min-max normalization. We plotted a dendrogram to visualize the hierarchical relationship of individual chains in our dataset and to determine the optimal number of structure clusters, then applied the agglomerative hierarchical clustering with Ward's linkage to obtain 4 clusters.

2.3 Creation of Quality Clusters

For a given density map, its quality was estimated by that of its helix regions. A previously developed metric, the cylindrical fit between atomic structures of helices and their respective density map regions [25] was used to assess the quality. Quality clusters of chains were subsequently obtained by applying thresholds on the quality scores, leading to four non-overlapping clusters indexed with integers from 1 to 4. Quality cluster 1 is considered as the highest quality cluster, in which the cylindrical fit scores (F1 measurement) of chains are between 0.7 and 1.0. The difference between precision and recall (Δ) are further used for the classification of clusters 2 and 3. Specifically, cluster 2 includes chains with cylindrical fit score ranging between 0.6 and 0.7, with Δ less than 0.15. Quality cluster 3 consists of chains under one of the two conditions: Either the cylindrical fit score between 0.6-0.7, and Δ>0.15, or the quality score between 0.55 and 0.6, and Δ<0.15. All remaining chains that do not enter into any of the clusters 1, 2 and 3 are placed in cluster 4.

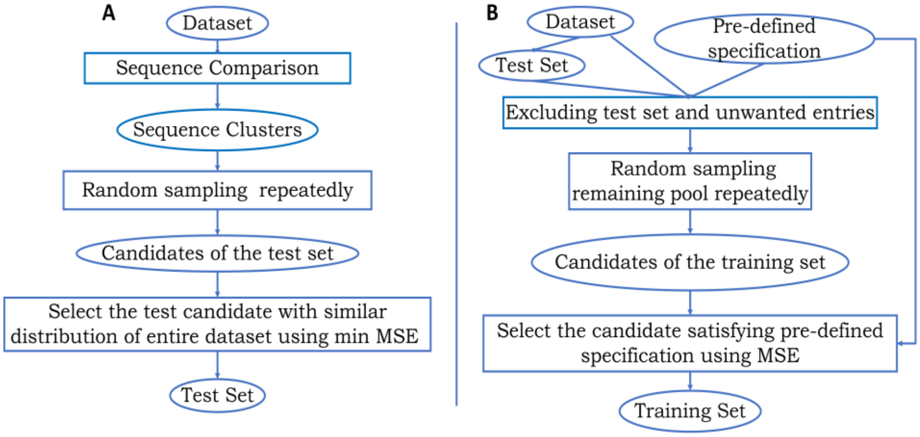
A
Dataset
Sequence Comparison
Sequence Clusters
Random sampling repeatedly
Candidates of the test set
Select the test candidate with similar distribution of entire dataset using min MSE
Test Set
B
Dataset
Pre-defined specification
Test Set
Excluding test set and unwanted entries
Random sampling remaining pool repeatedly
Candidates of the training set
Select the candidate satisfying pre-defined specification using MSE
Training Set

Figure 2: Constructing testing set (A) and training sets with varying desired content (B).

2.4 Test Set Construction

In general, the construction of a good test set involves the consideration of multiple factors, such as redundancy in density map regions, map quality, representativeness of targeted population, and structural complexity. Here, we implemented a method to construct a test set that has a distribution of chains in both structure content and map quality similar to those in the full dataset. Specifically, a total of 50 candidate test sets were first obtained by uniform random sampling from the full dataset (Fig. 2A). For each candidate set, the Mean Square Error (MSE) [26] defined in below was then calculated to measure the difference in the distribution of chains by structure content between the candidate set and the reference set, which, in this case, is the full set.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i^{ref} - y_i^{candidate})^2 \qquad (1)$$

where n is total number of structure clusters, y_i^{ref} is the proportion of chains in the full set that are from i-th structure cluster, and $y_i^{candidate}$ is the similar proportion in a candidate test set. A similar MSE was also calculated to measure the difference in the distribution over quality clusters. Finally, the candidate test set with the minimum average of the two MSEs was chosen as the test set.

2.5 Construct Training Sets with Pre-defined Specifications

The constructed sequence, structure, and quality clusters of chains provide a foundation for producing various data sets satisfying desired distributions over the three factors. For example, since the four structure clusters have distinct structural compositions (such as numbers and lengths of helices and β-strands), and therefore a data set with high content of β-sheets can be composed to include more chains from those structure clusters with high β-sheet composition. As illustrated in Figure 2B, once the test set was determined, a training set with a pre-defined specification was derived through sampling from a specially prepared pool of chains. The construction of this pool started with the inclusion of all chains in the entire data set except those included in the test set. The pool was further prepared to meet the pre-defined specification. As an example, a pre-defined specification (Specification 1 in Table 1) is intended to create a training set biased towards chains in structure clusters 1 through 4 with specification: <70%, 50%, 30%, 100%>. To achieve this, 30%, 50%, 70%, and 0% chains from structure clusters 1 through 4, respectively, were randomly chosen and removed from the pool. Once the pool was prepared, 50 times of random sampling was conducted to generate candidate training sets with each containing 400 chains. If maintaining the same distribution of map quality as in the full set is among the desired specification, a MSE was calculated as in Eq. (1) where the full set and the candidate training set were used for y_i^{ref} and $y_i^{candidate}$ respectively. The one with the minimum MSE was chosen to be the final training set satisfying the specification. The above procedure also allows construction of training sets that are biased towards any of the quality clusters while maintaining similar distribution of structural content as in the full set, such as in Specification 6 (Table 1).

Table 1: Structure and quality specifications used in generating training sets. [a]: percentages of chains from the four structure clusters (Specifications 1 to 3) and quality clusters (Specifications 4 to 6) entered into the sampling pool; [b]: numbers of chains from structure clusters s1 to s4 and quality clusters q1 to q4.

	Specification <%,%,%,%>[a]	Training sets <s1, s2, s3, s4, q1, q2, q3, q4>[b]
1	Structure: <70,50,30,100> Quality: same as full set	<148,143,56,53,31,64,63,242>
2	Structure: <30,70,50,100> Quality: same as full set	<58,188,105,49,24,67,67,242>
3	Structure: <50,30,70,100> Quality: same as full set	<104,71,180,45,31,71,60,238>
4	Structure: same as full set Quality: <0,0,100,100>	<124,113,145,18,64,180,156,0>
5	Structure: same as full set Quality: <0,0,0,100>	<91,136,152,21,0,0,0,400>
6	Structure: same as full set Quality: <0,100,100,25>	<99,121,157,23,0,145,127,128>

3 RESULTS

The full data set contains 1,292 protein chains, each having its corresponding region isolated from cryo-EM density maps. To understand the overall characteristics of the aggregated data, we examined the distribution of all chains in length (i.e., number of amino acid residues), percentage of helix and β-sheet content, and number of helices and β-strands. The most popular chain length is about 150 amino acid residuals; and chains are predominantly within 400 residuals (Figure 3A). Large number of chains have around 40% of their residues from helices and 20% from β-sheets, although substantial number of chains vary widely in length and are without β-sheets (Figure 3B). In majority of the chains, both the numbers of helices and β-strands are within 20. The distribution of these chains is somewhat uniform in the two numbers (Figure 3C), meaning no obvious correlation between the two.

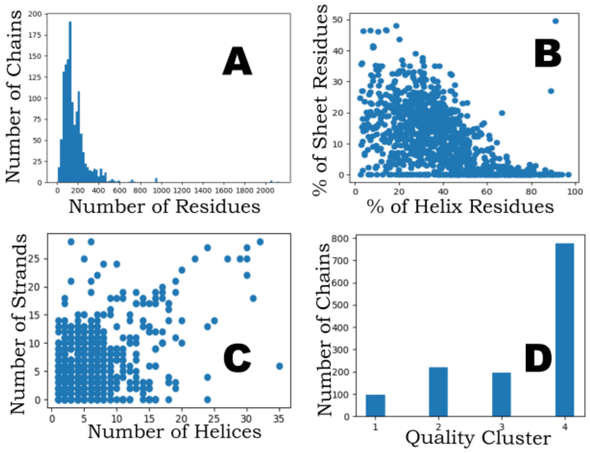

Figure 3: Distribution of chains in the aggregated data over various properties. (A) length, (B) percentage of helix and β-strand residues, (C) number of helices and β-strands, (D) number of chains in each quality cluster.

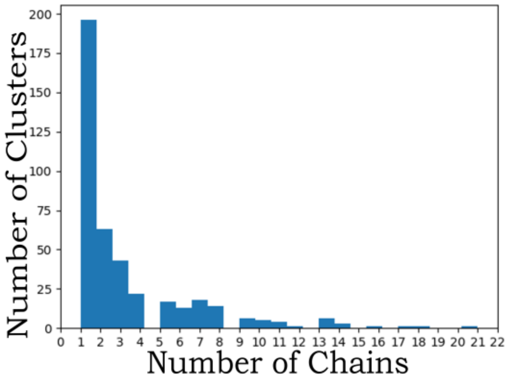

Figure 4: Distribution of cluster size (i.e., number of chains) among the 414 sequence clusters.

3.1 Sequence Clusters

Cluster analysis based on pairwise sequence identity led to a total of 414 distinct clusters, guaranteeing that the chains in the same cluster share at least 40% sequence identity with each other, and every cluster shares at most 39.9% sequence similarity with any other. The distribution of size among these clusters is provided in Figure 4. The largest cluster contains 21 chains, and 22 clusters have more than 10 chains. There are 196 clusters that contain only one member, indicating that the aggregated data set has 196 chains with less than 40% sequence identity shared with any other chain. To simplify the procedure, the test set was constructed via random sampling from these 196 unique clusters, and the set with the lowest average MSE was chosen.

3.2 Structure Clusters

With the agglomerative hierarchical clustering on the seven variables characterizing secondary structural content of chains, we obtained four structure clusters indexed using integers from 1 to 4. The numbers of chains in the four clusters are 318, 402, 500, and 72, respectively (Table 2). Structure cluster 3 comprises the majority of protein chains in the full data set. The average length among the chains in this cluster is 215 residues, echoing the peak in the histogram of chain lengths (Figure 3A). On average, each chain in this cluster has 40% residues in helices and 10% residues in β-sheets, suggesting that this type of secondary structural composition is the most common in our dataset. Structure cluster 1 contains chains with helix-rich structure, as chains in this cluster have 68% of residues in helices, on average, and almost no β-strands. Cluster 2 has the highest β-sheet content, with over 22% β-sheet residues, followed by cluster 3 with 10%. Chains in cluster 4 are characterized by 32% residues in helices and almost no β-sheet content. This cluster likely contains more loops, representing the minority chains in our dataset, as the number of chains in this cluster is significantly lower than the other three.

Table 2: Structural characteristics of chains in the four obtained structure clusters. Avg.: average; res.: residues; #: number.

Characteristics	Cluster 1	Cluster 2	Cluster 3	Cluster 4
	(N=318)	(N=402)	(N=500)	(N=72)
Avg. # of residues	130	158	215	100
Avg. # of helices	5.3	3.34	6.92	3.23
Avg. # of β-strands	0.36	8.14	6.138	0.06
Avg. % of helix res.	68%	24%	40%	32%
Avg. % of β-strands res.	0.3%	22.3%	10%	0.1%
Avg. length of helices	18.6	11	12	10
Avg. length of β-strands	0.3	4.5	3.5	0

3.3 Quality Clusters

There are 96, 221, 196, and 779 chains included in the four quality clusters 1, 2, 3, and 4, respectively. Among these clusters, cluster 4 is the largest and contains maps that have the lowest quality, indicating that most of the chains in the aggregated set have poor structure-map fit [25]. Lower fit score indicates less cylindrical density at a helix region. Since a helix is expected to have a rough cylindrical shape at a density threshold, the structure-map fit score represents the best cylinder score among all density thresholds. A low score suggests either the density map has low quality at the helix region or the erroneous placement of the atomic structure of the helix.

3.4 Test Set

We created a test set consisting of 50 chains by following the procedures depicted in Figure 2A. Specially, all 50 chains came from the 196 unique clusters (containing one chain in each cluster). This means for every chain in the test set, there is no other that has considerable amount (over 40%) of sequence identity in the entire dataset (also among the rest 49 chains in the test set). This test set has 9,804 residues in total, with 43.16% of all residues from helices, and 12.48% from β-sheets.

Since we chose the candidate set with the lowest MSE calculated as in Eq. (1) to be the final test set, it is representative of the full dataset in terms of both secondary structure content and map quality. The numbers of test chains in structure clusters 1, 2, 3, and 4 are, respectively, 12, 16, 19 and 3. These are proportional to the sizes of the four clusters in the entire chain population as indicated in Table 2. The numbers of test chains from the four quality clusters 1 through 4 are 7, 7, 6, and 30 respectively. This distribution also reflects that of the entire dataset, as out of the total 50 test chains, 20 (40%) belong to high-medium quality clusters (i.e., clusters 1, 2, and 3), and 30 chains (60%) belong to low quality cluster (i.e., cluster 4). These numbers for the entire dataset are 39.94% and 60.06%, respectively. Note that the test set currently represents the quality distribution of the entire data set that includes significant portion of poor data. However, the same data stratification method can be applied to a subset of the entire data, after extremely poor data are excluded.

3.5 Training Sets with Various Pre-defined Specifications

We created six training sets using six distinct specifications (Table 1) to study the effect of various factors on the performance of trained models. For example, since structure cluster 1 contains mostly helices (Table 2) and Specification 1 uses a greater percentage of cluster 1 than does Specification 2, its training set contains more helices than that of Specification 2. This is observed in the number of chains from structure cluster 1 in Training sets 1 and 2 (148 and 58 chains respectively) (Table 1 rows 1 and 2). However, Specification 2 led to a training set with more β-sheet content, as it included a greater percentage of structure cluster 2, the richest of the four clusters in β-strands. These training sets enable the study of how changes in the ratio between helix and β-sheet content impact model performance. As another example, Specification 5 produced Training set 5, which contained only low-quality density maps, and hence can be used to study the performance of a model that is trained with only low quality data.

Table 3: Performance of four DeepSSETracer models in detection of helices and β-sheets from cryo-EM density map components. The performance was measured by F1 score (%). The four models were trained using four training sets generated using Specifications 1 and 2 in Table 1. Two training sets (A and B) were generated for each specification. N/A: no β-sheet in the chain; Chain ID: <EMDB/PDB/Chain>; H: F1 score of helix detection, β: F1 score of β-sheet detection.

Chain ID	Specification 1				Specification 2			
	Set A		Set B		Set A		Set B	
	H	β	H	β	H	β	H	β
0139/6h58/x	49.8	33.5	52.2	47.7	49.8	54.5	48.6	47.9
1480/4v5z/B8	70.5	NA	73.5	NA	64.2	NA	73.5	NA
2221/2ynj/A	65.5	43.8	66.6	41.0	65.2	39.1	53.1	36.6
2678/4upc/A	60.2	41.2	65.1	47.0	69.1	60.5	66.9	43.3
2844/4ue5/E	24.1	52.0	47.6	40.9	30.1	57.6	37.8	40.8
2844/4ue5/S	9.5	NA	49.6	NA	40.0	NA	2.3	NA
2860/5afu/K	45.1	41.5	44.6	38.6	43.2	47.4	43.3	39.6
3049/3jaq/m	59.7	45.9	58.4	35.0	59.8	52.8	64.7	57.2
3101/5a9e/A	67.6	NA	68.3	NA	67.3	NA	65.7	NA
3440/5g5p/A	37.2	0.0	42.6	0.0	52.3	0.0	49.5	0.0
3491/5mdx/D	67.9	8.2	68.6	6.3	58.6	3.4	53.3	5.5
3491/5mdx/M	82.9	NA	82.1	NA	54.1	NA	63.7	NA
3544/5mq7/0A	72.1	61.0	71.3	51.1	70.2	65.8	69.4	62.3
3594/5n61/R	45.4	0.0	50.7	0.1	48.1	0.0	46.5	0.0
3683/5nrl/M	68.6	0.0	67.2	17.4	65.6	1.3	66.4	34.2
3850/5oqm/g	77.1	NA	76.9	NA	67.7	NA	74.7	NA
3896/6emk/E	22.7	NA	28.0	NA	29.5	NA	30.3	NA
3948/6esg/B	73.0	NA	76.7	NA	75.6	NA	77.3	NA
4041/5ldx/I	50.2	43.6	52.2	26.7	54.7	38.7	50.7	46.2
4041/5ldx/l	42.4	NA	38.4	NA	40.8	NA	41.4	NA
4041/5ldx/m	67.8	NA	70.2	NA	65.1	NA	65.2	NA
4041/5ldx/o	74.9	NA	77.1	NA	65.3	NA	76.8	NA
4089/5ln3/T	66.6	0.0	66.1	0.3	64.2	1.4	61.1	0.0
4089/5ln3/U	70.8	58.2	67.7	29.8	67.0	52.0	71.5	49.0
4098/5lqp/AB	66.4	68.1	56.3	64.8	56.9	65.1	67.5	65.2
4100/5lqx/H	59.9	55.6	53.1	54.4	69.2	59.7	68.7	61.0
4141/5m1s/B	50.0	53.7	45.3	49.7	50.0	57.5	49.8	56.9
4177/6f38/V	8.4	11.1	16.6	28.7	25.0	9.5	3.5	28.8
4182/6f42/P	35.1	0.0	26.2	21.8	32.0	21.2	36.1	22.9
5030/4v68/B1	26.1	16.4	25.9	9.2	26.2	37.0	26.6	16.5
5030/4v68/BF	66.2	42.4	65.9	57.4	67.0	53.9	69.1	61.0
5249/3izm/A	45.8	26.6	57.3	33.8	52.7	33.1	43.6	29.6
5592/4v6x/Ce	0.0	2.3	0.0	0.0	0.0	0.0	0.0	0.0
5943/3j6y/80	42.5	0.9	39.1	1.1	37.0	0.0	39.4	0.0
6149/3j8g/W	34.5	56.4	19.9	62.9	35.9	50.4	29.1	69.3
6456/3jbn/2	51.1	NA	53.4	NA	52.7	NA	53.9	NA
6695/5wyj/R1	49.9	44.3	50.3	35.6	55.0	51.6	51.2	47.3
6695/5wyj/U1	NA	53.1	NA	44.5	NA	54.7	NA	59.0
6810/5y5x/H	0.2	0.0	1.3	0.2	0.0	0.0	0.2	0.0
6889/5z56/6	16.8	1.6	18.7	8.4	19.9	0.0	14.8	21.6
8129/5j8k/D	56.4	30.2	58.9	25.5	60.9	29.2	59.0	24.4
8130/5j4z/BG	61.7	NA	56.9	NA	66.6	NA	56.5	NA
8143/5jpq/s	62.3	50.8	54.5	24.5	62.6	55.6	64.6	53.9
8148/5jb3/E	35.7	45.6	23.9	45.2	19.8	45.1	31.2	50.3
8400/5tcr/Q	76.4	58.6	73.9	61.8	74.7	66.5	64.2	62.1
8473/5tzs/d	48.5	50.9	39.7	50.4	0.9	34.7	43.3	49.0
8473/5tzs/l	57.8	51.8	43.8	38.1	24.1	3.4	58.1	41.6
8518/5u8s/2	59.9	46.6	61.8	54.8	63.3	56.3	61.9	47.6
8518/5u8s/B	61.1	30.4	64.1	26.3	63.4	20.5	67.3	27.0
8789/5w9n/B	24.2	53.5	14.7	46.2	7.0	54.8	1.7	49.2
Average	49.4	33.7	49.7	32.3	47.8	35.1	48.3	37.0

3.6 Performance of Models trained using Training Sets Generated with Two Specifications

To showcase the utility of the obtained test set and training sets, we trained models using training sets generated according to Specifications 1 and 2 (defined in Table 1). As discussed in section 3.5, specification 1 led to training set with higher helix content, while specification 2 led to training set with higher β-sheet content. Although Specification 1 and 2 have different desired distributions over structure clusters, they have the same distribution over quality clusters (Table 1). Both distributions over the quality clusters are same as that in the full data set. Since 779 of 1292 chains in the full data set belong to quality cluster 4 (Section 3.3), the worst of the four in quality, the training sets produced using Specification 1 and 2 have about 60% of the training data with poor quality. On the other hand, the test set was selected to resemble the distributions the same as those in the full data set, and therefore 30 of the 50 chains in the test set belong to quality cluster 4 (Section 3.4). Therefore, both the training sets and the test sets contain about 60% of the data with poor quality, and this may be reflected in the lower F1 scores of the performance. In fact, we observed a higher F1 score for both helix and β-sheet detection for models trained and tested using data without cluster 4 previously [5, 27].

To reduce observations due to random chance, two independent training sets (A and B) were generated for each specification (Table 3). With each training set, a model was obtained by training the U-Net deep neural network in DeepSSETracer [5]. The performance of all models was evaluated using the test set described in section 3.4 by calculating the F1 score. The average performance across all 50 test cases is consistent between the two replicates (A and B) for each specification (Table 3). Typically, machine learning models benefit from more training examples. So, as expected, models trained with data from Specification 1 performed better for helix than those trained with data from Specification 2, while the opposite is true for β-strands. For example, higher averaged F1 scores of 49.4% (trained using set A) and 49.7% (trained using set B) for helix detection were observed, when Specification 1 was used to generate the two training sets (Table 3). These F1 scores are higher than the corresponding F1 scores of 47.8% and 48.3%, when Specification 2 was used to generate its training sets A and B. In terms of β-sheet detection, the F1 scores (35.1% and 37.0%) for models trained using Specification 2 (more β-content) are higher than those (33.7% and 32.3%) for models trained using Specification 1.

We noticed that changes in distribution of structure clusters in the training sets has a greater impact on β-sheet detection than helix detection, since the difference in the average β-sheet F1 scores between the two specifications are much higher than that of helix. This can be explained as our dataset contains more helix residues than β-sheet (Figure 3B). This imbalanced problem is also common in other datasets where coil is the majority component, and β-sheet is the minority class. Our method can be used to adjust the distribution over secondary structure content in a training set to target the minority group in the dataset. The ideas of combining clustering algorithm and under-sampling have been implemented in many studies and in various fields. Some of them include density-based majority under-sampling technique (DBMUTE) [28] and cluster-based hybrid sampling for imbalanced-data (CBHSID) [29]. In addition, our method can be used to design multiple custom training sets with different distribution of structure and quality clusters to study the effect of these features on model performance.

4 CONCLUSION

Segmentation of secondary structures from cryo-EM density maps with medium resolution is still a challenging problem due to quality of the maps at such a resolution range. Although deep learning methods have shown potential to enhance the segmentation, different methods were tested using different data. First of all, there are limited experimental data sets available to test the approaches. Currently there are only two test data sets using cryo-EM maps. One is the set of 43 medium-resolution cryo-EM maps that was used in a 4-fold cross-validation test [3]. This suggests that each of the four tests only uses about 11 maps in the test. The other is the set of 28 unique chain regions of cryo-EM map components [5]. Moreover, there has been no previous study supporting a method to establish a proper test set for the segmentation problem. Various studies are needed to understand potential factors that influence the segmentation. We developed a method to stratify data with three potential factors and created the sequence identity clusters, structure clusters for secondary structure characteristics, and quality clusters. We proposed an approach, in this paper, to compose data sets with desired specifications related to the three potential factors. Even though this approach has not been applied to study the actual effect of the potential factors, the methodology shown here can be used to attack the challenging problem of benchmark data establishment. Although our current study only focused on three potential factors, the stratification method could potentially be generalized to other factors that are considered important. The proposed approach was implemented to create a test set and various training sets to study the effect of the data on the performance of models trained using these data sets with distinct properties. The two training sets with higher content of helix perform better detection of helix overall, as expected, and the training set with lower content of β-sheet performs worse detecting β-sheet. The expected results show the potential of the stratification method for producing more balanced training sets to enhance overall performance of deep learning methods.

The work presented here is the first investigation for the problem of establishing benchmark data for protein secondary structure segmentation from cryo-EM maps at the medium resolution. The focus of this work is the development of a framework to attack the problem. Many more studies are needed, such as constructing larger data sets, understanding variables currently existing in the framework, making the specifications more flexible, and optimizing the selection of candidate data sets.

ACKNOWLEDGEMENT

We thank the support from NIH R01-GM062968

REFERENCES

[1] Al-Azzawi, A., Ouadou, A., Max, H., Duan, Y., Tanner, J. J. and Cheng, J. DeepCryoPicker: fully automated deep neural network for single protein particle picking in cryo-EM. *BMC Bioinformatics*, 21, 1 (2020/11/09 2020), 509.

[2] Li, R., Si, D., Zeng, T., Ji, S. and He, J. Deep convolutional neural networks for detecting secondary structures in protein density maps from cryo-electron microscopy. *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (15-18 Dec. 2016 2016), 41-46.

[3] Maddhuri Venkata Subramaniya, S. R., Terashi, G. and Kihara, D. Protein secondary structure detection in intermediate-resolution cryo-EM maps using deep learning. *Nature methods*, 16, 9 (Sep 2019), 911-917.

[4] Mostosi, P., Schindelin, H., Kollmannsberger, P. and Thorn, A. Haruspex: A Neural Network for the Automatic Identification of Oligonucleotides and Protein Secondary Structure in Cryo-Electron Microscopy Maps. *Angewandte Chemie International Edition*, 59, 35 (2020), 14788-14795.

[5] Mu, Y., Sazzed, S., Alshammari, M., Sun, J. and He, J. A Tool for Segmentation of Secondary Structures in 3D Cryo-EM Density Map Components Using Deep Convolutional Neural Networks. *Frontiers in Bioinformatics*, 1 (2021), :710119.

[6] He, J. and Huang, S.-Y. EMNUSS: a deep learning framework for secondary structure annotation in cryo-EM maps. *Briefings in bioinformatics*, 22, 6 (2021).

[7] Pfab, J., Phan, N. M. and Si, D. DeepTracer for fast de novo cryo-EM protein structure modeling and special studies on CoV-related complexes. *Proceedings of the National Academy of Sciences*, 118, 2 (2021/01/12 2021), e2017525118.

[8] Chen, M., Dai, W., Sun, S. Y., Jonasch, D., He, C. Y., Schmid, M. F., Chiu, W. and Ludtke, S. J. Convolutional neural networks for automated annotation of cellular cryo-electron tomograms. *Nature methods*, 14, 10 (2017/10/01 2017), 983-985.

[9] Xu, M., Chai, X., Muthakana, H., Liang, X., Yang, G., Zeev-Ben-Mordehai, T. and Xing, E. P. Deep learning-based

subdivision approach for large scale macromolecules structure recovery from electron cryo tomograms. *Bioinformatics*, 33, 14 (2017), i13-i22.

[10] Jiang, W., Baker, M. L., Ludtke, S. J. and Chiu, W. Bridging the information gap: computational tools for intermediate resolution structure interpretation. *Journal of molecular biology*, 308, 5 (2001), 1033-1044.

[11] Dal Palu, A., He, J., Pontelli, E. and Lu, Y. Identification of Alpha-Helices from Low Resolution Protein Density Maps. *Proceeding of Computational Systems Bioinformatics Conference(CSB)* (2006), 89-98.

[12] Baker, M. L., Ju, T. and Chiu, W. Identification of secondary structure elements in intermediate-resolution density maps. *Structure*, 15, 1 (Jan 2007), 7-19.

[13] Rusu, M. and Wriggers, W. Evolutionary bidirectional expansion for the tracing of alpha helices in cryo-electron microscopy reconstructions. *Journal of structural biology*, 177, 2 (2012), 410-419.

[14] Si, D. and He, J. Beta-sheet Detection and Representation from Medium Resolution Cryo-EM Density Maps. *BCB'13: Proceedings of ACM Conference on Bioinformatics, Computational Biology and Biomedical Informatics* (September 22-25 2013), 764-770.

[15] Wang, X., Alnabati, E., Aderinwale, T. W., Subramaniya, S. R. M. V., Terashi, G. and Kihara, D. Detecting protein and DNA/RNA structures in cryo-EM maps of intermediate resolution using deep learning. *Nature Communications*, 12, 1 (2021), 1-9.

[16] Henderson, R., Sali, A., Baker, M. L., Carragher, B., Devkota, B., Downing, K. H., Egelman, E. H., Feng, Z., Frank, J., Grigorieff, N., Jiang, W., Ludtke, S. J., Medalia, O., Penczek, P. A., Rosenthal, P. B., Rossmann, M. G., Schmid, M. F., Schröder, G. F., Steven, A. C., Stokes, D. L., Westbrook, J. D., Wriggers, W., Yang, H., Young, J., Berman, H. M., Chiu, W., Kleywegt, G. J. and Lawson, C. L. Outcome of the first electron microscopy validation task force meeting. *Structure*, 20, 2 (Feb 8 2012), 205-214.

[17] Kryshtafovych, A., Lawson, C. and Chiu, W. Evaluation of models in the 2016 cryo-EM model challenge. *Acta Crystallographica Section A Foundations and Advances*, 74 (07/20 2018), a123-a123.

[18] Lawson, C. L., Kryshtafovych, A., Adams, P. D., Afonine, P. V., Baker, M. L., Barad, B. A., Bond, P., Burnley, T., Cao, R., Cheng, J., Chojnowski, G., Cowtan, K., Dill, K. A., DiMaio, F., Farrell, D. P., Fraser, J. S., Herzik, M. A., Hoh, S. W., Hou, J., Hung, L.-W., Igaev, M., Joseph, A. P., Kihara, D., Kumar, D., Mittal, S., Monastyrskyy, B., Olek, M., Palmer, C. M., Patwardhan, A., Perez, A., Pfab, J., Pintilie, G. D., Richardson, J. S., Rosenthal, P. B., Sarkar, D., Schäfer, L. U., Schmid, M. F., Schröder, G. F., Shekhar, M., Si, D., Singharoy, A., Terashi, G., Terwilliger, T. C., Vaiana, A., Wang, L., Wang, Z., Wankowicz, S. A., Williams, C. J., Winn, M., Wu, T., Yu, X., Zhang, K., Berman, H. M. and Chiu, W. Cryo-EM model validation recommendations based on outcomes of the 2019 EMDataResource challenge. *Nature methods*, 18, 2 (2021/02/01 2021), 156-164.

[19] Lawson, C. L., Patwardhan, A., Baker, M. L., Hryc, C., Garcia, E. S., Hudson, B. P., Lagerstedt, I., Ludtke, S. J., Pintilie, G., Sala, R., Westbrook, J. D., Berman, H. M., Kleywegt, G. J. and Chiu, W. EMDataBank unified data resource for 3DEM. *Nucleic Acids Res*, 44, D1 (Jan 4 2016), D396-403.

[20] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. and Bourne, P. E. The Protein Data Bank. *Nucleic acids research*, 28, 1 (2000), 235-242.

[21] Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C. and Ferrin, T. E. UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem*, 25, 13 (Oct 2004), 1605-1612.

[22] Pettersen, E. F., Goddard, T. D., Huang, C. C., Meng, E. C., Couch, G. S., Croll, T. I., Morris, J. H. and Ferrin, T. E. UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Science*, 30, 1 (2021), 70-82.

[23] Needleman, S. B. and Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48, 3 (1970), 443-453.

[24] Pedregosa, F. a. V., G. and Gramfort, A. and Michel, V., and Thirion, B. a. G., O. and Blondel, M. and Prettenhofer, P., and Weiss, R. a. D., V. and Vanderplas, J. and Passos, A. and and Cournapeau, D. a. B., M. and Perrot, M. and Duchesnay, E. *Scikit-learn: Machine Learning in {P}ython*. City, 2011.

[25] Sazzed, S., Scheible, P., Alshammari, M., Wriggers, W. and He, J. Cylindrical Similarity Measurement for Helices in Medium-Resolution Cryo-Electron Microscopy Density Maps. *J Chem Inf Model*, 60, 5 (May 26 2020), 2644-2650.

[26] Allen, D. M. Mean Square Error of Prediction as a Criterion for Selecting Variables. *Technometrics*, 13, 3 (1971/08/01 1971), 469-475.

[27] Deng, Y., Mu, Y., Sazzed, S., Sun, J. and He, J. Using Curriculum Learning in Pattern Recognition of 3-dimensional Cryo-electron Microscopy Density Maps. *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics* (2020), 1-7.

[28] Bunkhumpornpat, C. and Sinapiromsaran, K. DBMUTE: density-based majority under-sampling technique. *Knowledge and Information Systems*, 50, 3 (2017/03/01 2017), 827-850.

[29] Palli, A. S., Jaafar, J., Hashmani, M. A., Gomes, H. M. and Gilal, A. R. A Hybrid Sampling Approach for Imbalanced Binary and Multi-Class Data Using Clustering Analysis. *IEEE Access*, 10 (2022), 118639-118653.