AAnchor: CNN guided detection of anchor amino acids in high resolution cryo-EM density maps

Mark Rozanov
Blavatnik School of Computer Science
Tel Aviv University, Israel
markroza@tau.ac.il

Haim J. Wolfson
Blavatnik School of Computer Science
Tel Aviv University, Israel
wolfson@tau.ac.il

Marik Privet

Abstract—**The recent cryo-EM resolution revolution enables the development of algorithms for direct de-novo modeling of protein structures into cryo-EM density maps. Here we present a machine learning based method for the detection of high confidence anchor amino acid residues in such a map. Such anchor residues can be exploited in several local de-novo modeling tasks, such as the reliable positioning of secondary structures, loop modeling and general fragment based modeling. In the experimental results we show the ability of the proposed procedure to locate and classify a significant number of amino acids in density maps of $3.1\mathring{A}$ (or better) resolution. Our performance analysis indicates that the main factor affecting the detection accuracy is the lack of sufficient experimental data for the training stage of the algorithm. Thus, our method is expected to improve significantly in the near future, due to the rapid increase in the release of novel high resolution cryo-EM maps.**

Index Terms—**cryo-EM maps, proteins, molecular modeling, machine learning, CNN.**

I. INTRODUCTION

Atomic accuracy models of protein structures are an invaluable tool for the elucidation of protein function. The recent "resolution revolution" [1] in cryo electron microscopy (cryoEM) has led to an ever increasing number of near atomic resolution density maps deposited in the EM databank EMDB [2]. While in 2002 the best structure deposited was at $9\mathring{A}$ resolution, recently ([3], [4]) key structures have been resolved at resolution better than $2.5\mathring{A}$.

Most of the techniques for modeling protein structures into intermediate resolution ($5 - 10$ Å) maps are based on rigid fitting of template protein structures into these maps. Many of these methods ([5]–[9]) use secondary structures as anchors for their fitting procedure.

At resolutions better than $4-5\mathring{A}$, the goal is more ambitious and de novo modeling techniques are being exploited. Some of them are adaptations of the standard X-ray crystallography modeling methods, however these tend to be time consuming. Recently, several de-novo modeling techniques have been developed to deal specifically with cryoEM density maps [10].

Pathwalking [11] detects first pseudo-atom anchors, which are based on high density local regions, and then applies the travelling salesperson (TSP) combinatorial optimization algorithm to detect a long enough path which should model the protein backbone. MAINMAST [12] applies quite a similar line of reasoning. It first detects a set of anchor points in the map, where the local density is above a given threshold (so, called, Local Dense Points). The threshold is set in such a way that the number of LDPs is much larger than the number of residues in the protein and roughly about half of its heavy atoms. A graph is built by connecting nearby points and a minimun spanning tree (MST) is calculated for this graph. Then, after a refinement procedure, the longest path in the MST is detected and the protein sequence is threaded along this path. A recently published method [13], which is based on the ideas of the Rosetta modeling suite, fits short sequence based structure fragment templates into the density map and applies a Monte Carlo simulated annealing procedure to detect a set of mutually compatible fragments. All of the above mentioned procedures require prior segmentation of the density map into its various protein chains.

Prior detection of reliable **amino acid anchors** in the density map, namely having knowledge of even a relatively small number of amino acids, whose identity and location has been established with high confidence can be used to guide the various de-novo modeling methods, as well as serve as a starting point for the development of novel methods. In particular, it could lead to the development of novel techniques, which do not require prior segmentation of the EM density map. While in high resolution maps (roughly,3.5 Å or better), sidechains become visible and individual rotamers may be distinguished ([10], [14]), no automated method has been suggested to detect specific amino acids in a cryo-EM density map.

In this work we present a machine learning (ML) algorithm nicknamed **AAnchor** (amino-acid anchor) for the detection and localization of amino acids in a high resolution cryo EM density map. The two primary goals of this study are (i) to develop an automatic tool for the detection and classification of amino acids in a near atomic resolution cryo EM map; and (ii) to provide a quantitative analysis of the ability to detect a specific amino acid in an experimental cryo EM map.

Detection of an amino acid in a density map can be decomposed into two tasks:

- **Classification** is defined as assigning one of 21 labels (20 amino acids plus none) to a voxel and its close neighborhood. A voxel is labeled as center of mass of an amino acid, if it is closer than 1.5 $\mathring{A}$ to this amino acid's actual center of mass.
- **Detection** is defined as the localization of amino acids of a specific type in the cryo EM map. The output of this task are the coordinates (of the center of mass) and type of the detected amino acids followed by an estimated detection confidence.

Both the classification and the detection tasks have been intensively studies in the fields of image processing and computer vision. Some of the best results for the classification task have been achieved using deep convolutional neural networks (CNNs) [15], [16]. The detection task is usually solved by combining a classification algorithm with some sampling technique (such as sliding window) and post processing filtering (such as SVM [17]).

Inspired by the similarity of our tasks to analogous image processing tasks, we built a CNN that classifies each voxel into 21 types. We trained and tested it on simulated and experimental cryo EM maps at various resolutions. Using the sliding window approach followed by post-processing we subsequently applied the trained classification CNN to the detection task.

The preliminary results of our algorithm are quite encouraging. For example, on the $3.1\mathring{A}$ cryo EM map of lysenin pore [18], the algorithm localizes above one hundred amino acids of different types with confidence above 80%. These preliminary results also indicate that the quality of detection is best for amino acids, which have a small number of rotamers and a large enough training set. This last observation also explains, why contrary to expectation, the results for $2.9\mathring{A}$ and 3.1 $\mathring{A}$ resolution maps are better then those for 2.2 $\mathring{A}$ maps. This is due to the limited training data set existing for the better resolution maps. Thus, with the rapid increase in high resolution cryoEM structures and the availability of larger trainng sets, the results of our machine learning based methodology are expected to improve significantly.

II. METHODS

A. The Task

Given an electron density map of a protein/macromolecular assembly, our task is to detect voxels in this map, which correspond to the location of the centers of mass of specific amino acids. The goal is to report only those amino acids, which have been detected with high confidence, which we call "anchors". The two main building blocks of our method are a **classification** CNN and a **detection** algorithm. The detection algorithm generates candidates for the classification CNN and filters out candidates of expected low accuracy.

B. A CNN for the classification task

Convolutional neural networks (CNN) were proposed by Yann LeCun in 1989 for zip code recognition [19]. A CNN consists of alternating convolutional and pooling layers optionally followed by fully connected layers. The first and last layers are the input and output layer respectively, while the other layers are referred to as hidden layers.

Formally, a CNN of depth D is a composition of D parametrized functions $\{f_1, \cdots, f_D\}$, which maps an input vector $\mathbf{x}$ to an output vector $\mathbf{y}$:

$$\mathbf{y} = f(\mathbf{x}) = f_D(\mathbf{z}, w_D, b_D) \circ, \ldots, \circ f_1(\mathbf{x}, w_1, b_1), \quad (1)$$

where w_k and b_k are the weights and biases vectors for the function f_k. The functions f_k are the previously mentioned layers.

Given a set of labeled data pairs $\{(\mathbf{x}^i, \mathbf{y}^i)\}_{i=1}^M$, the training process of a CNN defined by (1) is a process of a numerical solution of the optimization problem:

$$\text{Find } \{w_k, b_k\}_{k=1}^D \text{ which minimize:}$$
$$\frac{1}{M} \sum_{i=1}^M d\left(f(\mathbf{x}_i), \mathbf{y}_i\right), \quad (2)$$

where $d(\cdot, \cdot)$ is the loss function expressing a penalty for an incorrect classification. For a comprehensive discussion of CNNs the reader is referred to [20].

C. Softmax CNN for the Amino Acids Classification

We formulate the amino acids classification problem as a multiclass classification of volume cubes of a cryo-EM map. For a cryo EM map $E \in \mathbb{R}^{Length \times Height \times Width}$, denote by $\mathbf{x} \in \mathbb{R}^{L \times L \times L}$ (L in our experiments was $11\mathring{A}$) a local volume cube of the map, which is examined to contain the center of mass of an amino acid (of a specific type) in the proximity of the local cube's center. Let $c \in \{1, \cdots, C\}$ denote the corresponding label/type of the volume cube $\mathbf{x}$, in which C is the total number of classes. In our case C=21, which represent the 20 amino acids as well as a local volume, which does not contain a center of mass of an amino acid close to its center.

We train a CNN network of depth D, such that

$$c = \underset{i}{\operatorname{argmax}} \mathbf{p}\left(\mathbf{y}_D(\mathbf{x})\right), \quad (3)$$

where $\mathbf{p}\left(\mathbf{y}_D(\mathbf{x})\right) = f(\mathbf{x})$ is the output of the last layer of the CNN obtained by applying the softmax function to output $\mathbf{x}$ of the previous layer. The softmax function $\mathbf{p}(\mathbf{y}) = (p_1(\mathbf{y}), \cdots, p_C(\mathbf{y}))$ is used to transform the vector of C (arbitrary) real values to a vector, where each value is in the $(0, 1)$ interval and the sum of all the values is 1. It is defined as:

$$p_i(\mathbf{y}) = e^{y_i} / \sum_{k=1}^C e^{y_k}. \quad (4)$$

Since $\sum_i p_i(\mathbf{y}) = 1$ and $0 \leq p_i(\mathbf{y}) \leq 1$, the output of the softmax function is often (informally) treated as probabilities of a cube $\mathbf{x}$ to have label i.

D. Network Architecture and Training Details

The detailed architecture of the applied softmax CNN is presented in Table I. We use the rectified linear (ReLU) activation function which is known to provide the best learning rate in image classification tasks [15]. Keras [21] is used to implement and train the softmax CNN. The training time for one epoch was 3 minutes for one million samples on a server with 4 NVIDIA Titan black GPUs, each with 2880 cores.

E. Preparation of the Datasets for the Training Stage

We describe the datasets used for training of the CNN and their structure. The input to this stage are pairs of a cryoEM map with an atomic structure fitted to it. From each map we extract a set of volume cubes of size $L \times L \times L$ (L in our experiments is $11\mathring{A}$) centered at each amino acid of the map. Such a cube is labeled by the fitting amino acid type. In practice, we, usually, label 8 cubes with this label, namely, if the coordinates of the amino acid center of mass are not integer, we assign this label to all the nearby cubes with integer coordinates at their center. We normalize the density of a labeled cube so that the mean density is 0 and the standard deviation is 1. This is due to the fact that the average density of a cryoEM map varies between different regions. In addition to the amino-acid induced cubes, we sample a sufficient number of density cubes that do not represent centers of mass of amino acids and assign them the 21'st ("zero") label.

The labeled volume cubes from all the training maps of the dataset are fed into the CNN training procedure.

We have applied the above described procedure to several diverse datasets. Since there is not enough experimental data to perform proper training, we have used both available experimental datasets as well as datasets of simulated EM maps at the required resolutions. Simulated and experimental datasets were created for each of the resolution spans presented in Table II.

The **simulated** dataset consists of structures from the Dunbrack Rotamers Library [22] and cryo EM maps created using the UCSF Chimera [23] *molmap* command. Each map was created at a randomly selected resolution within a given resolution span.

The **experimental** dataset consists of publicly available cryo EM maps from the EMDataBank [2] within the required resolution span together with aligned/fitted PDB structures.

A **data augmentation** procedure was employed to increase the dataset and reduce the overfitting effect. Each map was rotated at a random angle together with the corresponding fitted atomic-resolution model and the rotated map and model were added to the dataset. We performed 10 rotations for each protein map. Since a virus structure already consists of repetitions of small sub-structures at different poses, virus maps were excluded from the augmentation procedure.

F. Detection

The workflow of the proposed detection method is illustrated in Figure 2. In the preprocessing phase an input cryoEM map is resampled to a grid of $1\mathring{A} \times 1\mathring{A} \times 1\mathring{A}$ voxels. Using the

TABLE I: Architecture of Softmax CNN for Amino Acids Classification

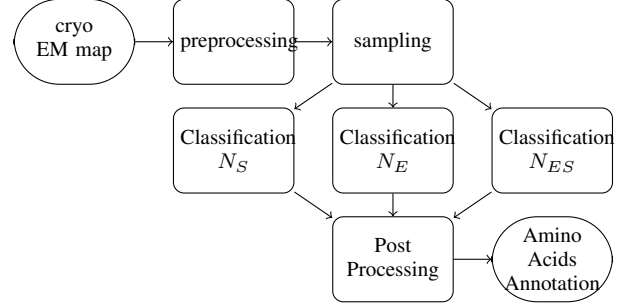| Layer	Type	Filter Dimensions	Input/Output Dimensions
1	Input		$11 \times 11 \times 11 \times 1$
2	3D Conv	$3 \times 3 \times 3 \times 1 \times 50$	$9 \times 9 \times 9 \times 50$
3	3D Conv	$2 \times 2 \times 2 \times 50 \times 50$	$8 \times 8 \times 8 \times 50$
4	Max Pool	$2 \times 2 \times 2$	$4 \times 4 \times 4 \times 50$
5	Fully Connected	$4 \times 4 \times 4 \times 50 \times 100$	1×100
6	SoftMax	1×100	1×21

Fig. 1: Workflow for the detection procedure

sliding window approach we sample the volume cubes from the resampled map (see Figure). Cubes with average density value less than the average density of the map are filtered out. The remaining cubes density is normalized to 0 mean and standard deviation 1. For each cube $\mathbf{x}$ we obtain the predicted value from the three pretrained classification CNNs: $N_S(\mathbf{x})$, $N_E(\mathbf{x})$, $N_{ES}(\mathbf{x})$.

The predicted label and confidence are calculate by combination of the results of three CNNs. The combination method depends on the amino acid type (Table V), and is one of the following: one of $N_S(\mathbf{x})$, $N_E(\mathbf{x})$, $N_{ES}(\mathbf{x})$, majority of $N_S(\mathbf{x}), N_E(\mathbf{x}), N_{ES}(\mathbf{x})$, mean value of $N_S(\mathbf{x}), N_E(\mathbf{x}), N_{ES}(\mathbf{x})$, or mean value of $N_S(\mathbf{x})$, $N_E(\mathbf{x})$. A cube centre is marked as an amino acid centre if the resulting confidence is above a threshold. The threshold value depends on the amino acid type and calculated in the training phase.

III. EXPERIMENTAL RESULTS

A. The Classification Task

We start by presenting the results of the classification task. The classification task, unlike the detection task, assumes that an approximate location of an amino acid is given. Though the detection task is the more interesting one, analysis of the simpler classification task is key to understanding the algorithm behavior and improving its performance.

Once trained, the CNN is used to perform prediction on a validation dataset, which is disjoint from the training set. For each input cube with known label j, the softmax CNN produces "probabilities" $p_k, k = 0, \cdots, C$. The predicted label i the one giving maximal "probability", and the confidence is p_i.

1) Confusion Matrix and Reliability Curve: For an amino acid labeled j, we are interested in the following questions:

1) What are our chances to detect it correctly, and what are the chances of missing it with an amino acid i?
2) Does our confidence level p_j reflect the probability of the true detection? Namely, does a p_j fraction of all the inputs predicted as j represent correct predictions.

Commonly accepted tools for quantitative answers to these questions are the **confusion matrix** [24] and the **reliability curve** [25].

An entry $a_{i,j}$ of a **confusion matrix** A is defined as the ratio $a_{i,j} = \frac{T_j^i}{N_j}$, where N_j is the number of input cubes labeled j and T_j^i the number of inputs labeled j with predicted label i. An ideal classification algorithm will result in $a_{i,i} = 1$ and $a_{i,j} = 0$ if $i \neq j$. We also refer to $a_{i,i}$ as the **total accuracy** of label i.

It is desired that softmax CNN be **well calibrated** [25], i.e., the predicted confidence should reflect the ground truth probability. A **reliability curve** is a plot of ground truth accuracy prediction vs reported confidence. For a perfectly calibrated network the probability curve is an identity function. For the experimental results the ground truth accuracy is estimated by grouping predictions in interval bins according to their reported confidence [25].

2) 2.2Å Resolution:

a) Simulation Results: We start by analyzing the results obtained on simulated data. Though it is commonly accepted that simulated data is not as reliable as real data, analysis of classification CNN performance on simulated data provides important insight on the real data behavior. Also, simulated data results provide an upper limit of what can be achieved with the selected CNN architecture.

Figure 2a shows the confusion matrix obtained while the total accuracies are summarized in Figure 2b. The amino acids with the highest accuracy are: LEU, GLY, ALA, VAL, LYS, TYR, PRO. This indicates that the number of rotamers [22] of an amino acid has a major effect on the accuracy. Since an amino acid with a large number of rotamers can exhibit numerous conformations, it is harder for a classification algorithm to adjust its parameters for the given amino acid. Except LYS, all of the above amino acids have a relatively small number of rotamers. Moreover, all the amino acids with a small number of rotamers have high classification accuracy, except Glutamic acid and Glutamine which are mutually mixed in classification. The size of the dataset is another dominant factor for the classification accuracy. In a Machine Learning approach a large training dataset enables robust parameter estimation and reduces the effect of overfitting. Figures 3a and 3b show the correlation between the training dataset size and the resulting accuracy.

b) Experimental Data Results: We used a 2.2 cryo-EM map of β-galactosidase ([3]. EMD-2984) to study the CNN performance. At the time of the research, only three additional cryoEM maps with resolution better than 2.3Å were available: EMD-8762 [26],EMD-8194 [27], and EMD-
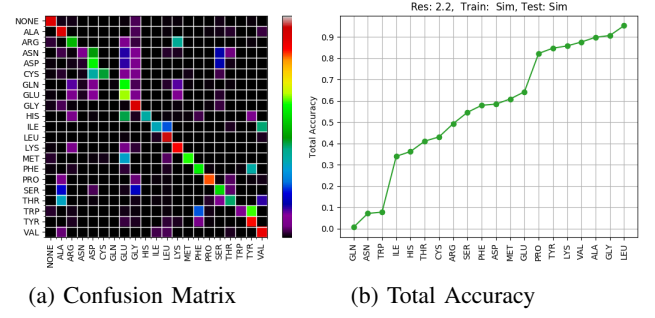

(a) Confusion Matrix   (b) Total Accuracy

Fig. 2: Classification Results: Resolution 2.2Å, Train: Simulated, Test : Simulated

(a) Accuracy vs Dataset size   (b) Accuracy vs Dataset size normalized to the number of Rotamers
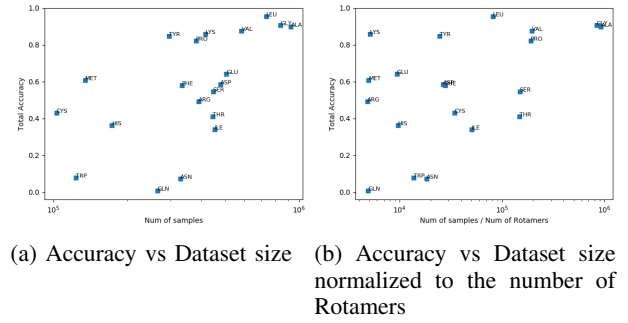
Fig. 3: Effect of the Training Dataset Size on the Classification Accuracy: Resolution 2.2Å, Train: Simulated, Test : Simulated

3295 [4]. Even augmented , this is definitely not enough for proper training of the CNN. The confusion matrix and the total accuracies are shown in Figures 4a and 4b. The accuracy dependence on the dataset size is shown in Figures 7a and 7b. Despite significantly lower accuracies, the tendency observed on simulated data is preserved.

In order to improve the classification accuracy we tried two methods to overcome the lack of data:

1) We trained up to 6 independent CNNs and averaged the obtained probabilities.
2) We combined the real maps with the simulated maps and created a new dataset.

The first method, averaging over multiple networks, did not result in any significant improvement of the classification accuracy. This indicates that the obtained low accuracy is not due to overfitting, but due to the small training set. Best classification results were achieved by combining in the training set simulated and experimental data in equal proportion. We denote the CNN trained on this combined dataset as N_{ES}^{22}. The detection accuracies for this case are shown in Figures 6a and 6b.

Figure 8 shows the estimated reliability curves for N_{ES}^{22}. Due to lack of data, reliability curves cannot be estimated for every input, but the overall tendency is that the output of the softmax layer of N_{ES}^{22} underestimated the classification accuracy.
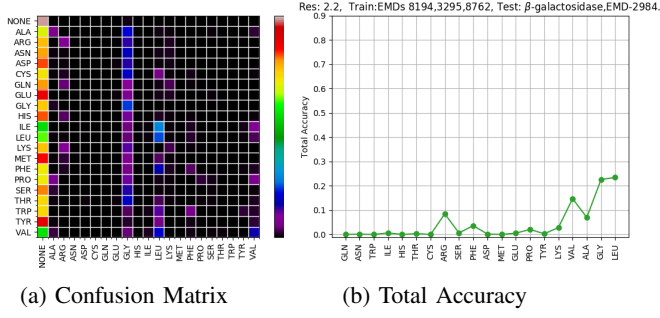
(a) Confusion Matrix   (b) Total Accuracy

Fig. 4: Classification Results: Resolution $2.2\mathring{A}$, Train: EMDs 8762,3295,8184, Test : β-galactosidase, [3].

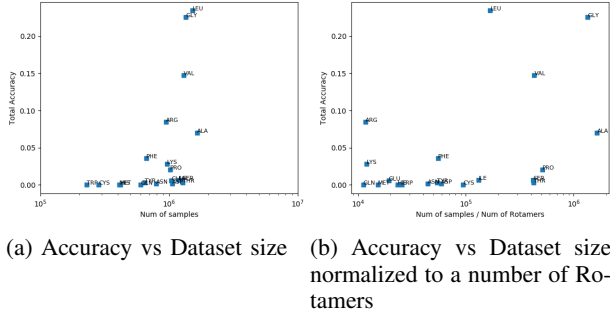(a) Accuracy vs Dataset size   (b) Accuracy vs Dataset size normalized to a number of Rotamers

Fig. 5: Effect of Training Dataset Size on the Classification Accuracy: Train: EMDs 8762,3295,8184, Test : β-galactosidase, [3].
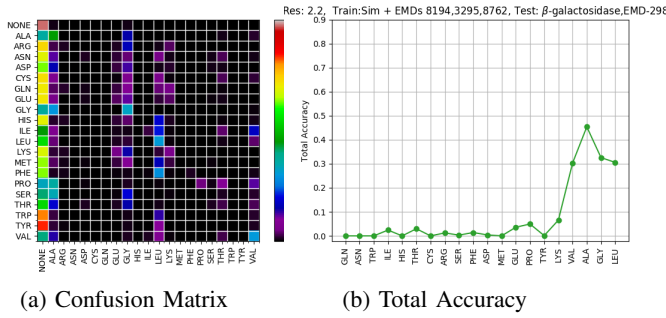

(a) Confusion Matrix   (b) Total Accuracy

Fig. 6: Classification Results: Resolution $2.2\mathring{A}$, Train: Simulation + EMDs 8762,3295,8184, Test : β-galactosidase, [3].

3) Resolutions $2.9\mathring{A}$ *and* $3.1\mathring{A}$ *:* Coarsening the resolution has two opposite effects on the classification accuracy. Intuitively, in higher resolution maps the amino acids shape is sharper, but lower resolution maps benefit from a larger training dataset. The size of the training dataset for three tested resolutions is shown on Table II. The effect of a map resolution on the classification accuracy is shown in Figures 9a and 9b. While having only a small effect on simulated data (Figure 9a), the effect of resolution on the real data is well expressed (Figure 9b). Accuracies for $2.9\mathring{A}$ and $3.1\mathring{A}$ are significantly better than those for $2.2\mathring{A}$. This is clearly due to

(a) Accuracy vs Dataset size   (b) Accuracy vs Dataset size normalized to the number of Rotamers
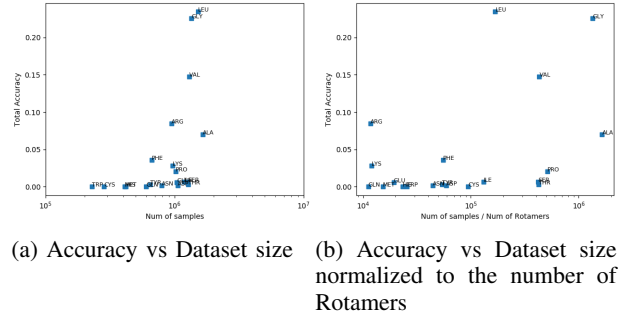
Fig. 7: Effect of Training Dataset Size on the Classification Accuracy: Train: EMDs 8762,3295,8184, Test : β-galactosidase, [3].

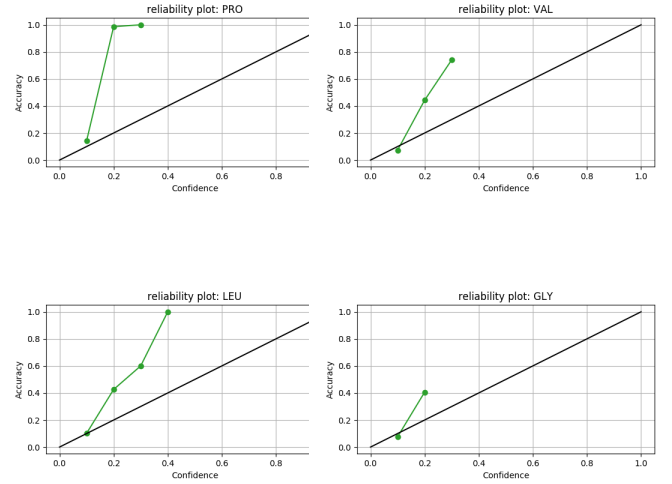Fig. 8: Estimated Reliability Curves for PRO, LEU, VAL, GLY : Resolution $2.2\mathring{A}$, Train: Simulation + EMDs 8762,3295,8184, Test : β-galactosidase, [3].

the increased training dataset. The only exception is Alanin, which is probably too small to be detected at resolutions above $2.2\mathring{A}$. However, moving from $2.9\mathring{A}$ to $3.1\mathring{A}$ we see that for the majority of the amino acids the total accuracy values are decreased. Thus, in this transition the effect of increasing the training dataset size did not compensate for the degradation in map precision.

The confusion matrices for $2.9\mathring{A}$ Anthrax toxin protective antigen pore and $3.1\mathring{A}$ Lysenin Pore are presented in Figures 10a and 10b, respectively. The Reliability Curves for the $2.9\mathring{A}$ Anthrax toxin protective antigen pore and $3.1\mathring{A}$ Lysenin Pore are presented in Figures 10a 10b, respectively. While at resolution of $2.9\mathring{A}$ the estimated confidence is still less than the observed one, at $3.1\mathring{A}$ resolution the estimated confidence is of good precision.

Resolution Span	Maps in Training Set	Map for Validation
$1.8 - 2.3\mathring{A}$	3	EMD-2984 , $2.2\mathring{A}$ *beta* - galactosidase [4]
$2.7 - 2.9\mathring{A}$	8	EMD-6224 ,$2.9\mathring{A}$ Anthrax toxin protective antigen pore [28]
$2.9 - 3.1\mathring{A}$	13	EMD -8015, $3.1\mathring{A}$ Lysenin Pore [18]

TABLE II: Experimental Train and Test data for Various Resolutions

(a) Simulated Data

(b) Real Data (Mixed with Simulation for Resolution $2.2\mathring{A}$
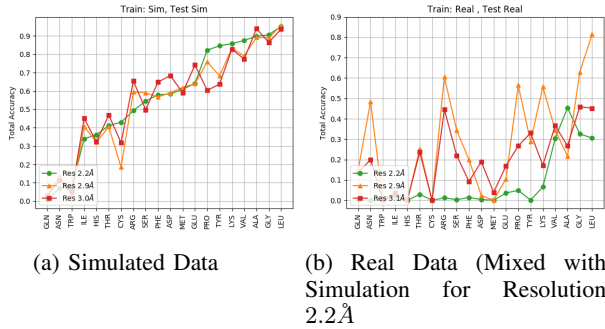
Fig. 9: Total classification accuracies for different resolutions.

(a) Accuracy vs Dataset size

(b) Accuracy vs Dataset size normalized to a number of Rotamers
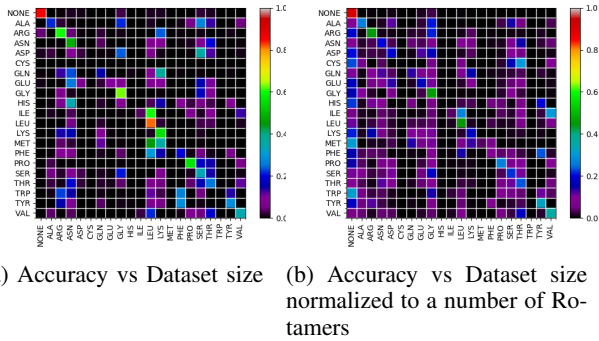
Fig. 10: Classification Accuracy Results for Resolution 2.9 and 3.1

Fig. 11: Estimated Reliability Curves for PRO, LEU, VAL, GLY : Resolution $2.9\mathring{A}$, Train: Experimental Data, Test : Anthrax toxin protective antigen pore [28].

Fig. 12: Estimated Reliability Curves for PRO, LEU, VAL, GLY : Resolution $3.1\mathring{A}$, Train: Experimental Data, Test : Lysenin Pore [18].

B. Detection Results

In the detection (also called localization) task, the exact location of an amino acid is unknown. Whilst the localization of all amino acids of a protein seems to be a hard problem at this time, detection of a subset of the amino acids obtained with high confidence is achievable. In this paper we focused on identifying **anchors**, i.e., amino acids, which have been located and labeled with confidence above 80%.

For each resolution case we tested a number of different classification methods, while the pre-processing and post-processing phases remain unchanged.

Table III summarizes the detection picks for the $2.2\mathring{A}$ map of *beta*-galactosidase [3]. There we succeed to detect amino acids of four types: ARG, LEU, PRO, VAL with confidence above 80%. The fraction of the detected high confidence amino acids is up to 10% from the total amount of the same type.
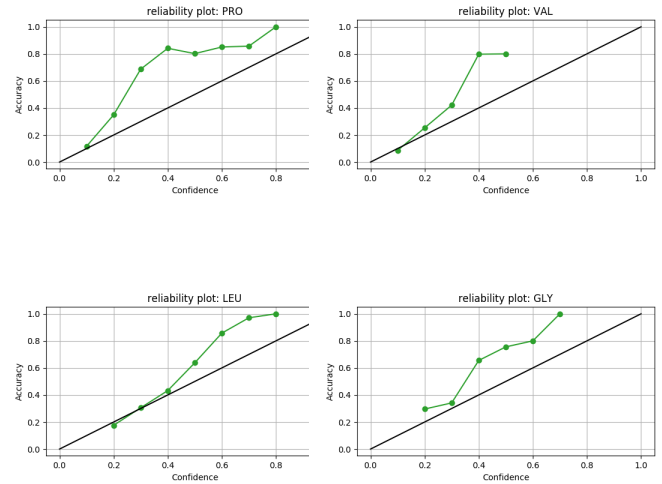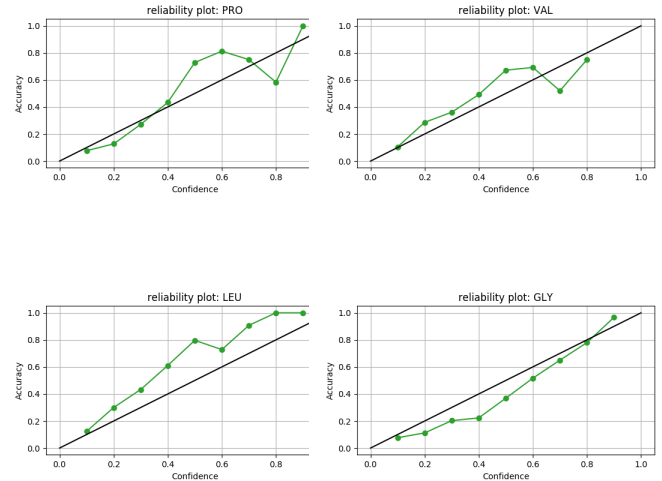
We expect this result to improve, as more high resolution cryo EM maps are being released.

Table IV summarizes the detection picks for the test map of the 2.9 $\mathring{A}$ resolution anthrax toxin protective antigen pore [28]. We succeded to detect 70% of the Prolines. Figure 13 illustrates the results of proline detection. The detection rate of ARG, LEU, TYR,VAL was about 20%.

Table V summarizes the detection picks for the test map of 3.1 $\mathring{A}$ resolution Lysenin Pore [18]. We detected more than 20% of the Leucine residues. We also succeeded to detect about

Amino Acid Type	Picks with Confidence of 80%	Total in Protein	Best Method
ARG	17	133	$maj(N_S, N_E, N_{ES})$
LEU	15	231	$mean(N_S, N_E, N_{ES})$
PRO	10	189	N_{ES}
VAL	10	119	$mean(N_S, N_E, N_{ES})$

TABLE III: Detection Results 2.2 Å cryo-EM single particle reconstruction of beta galactosidase (emd-2984)

Amino Acid Type	Picks with Confidence of 80%	Total in Protein	Best Method
ASN	5	287	N_E
ARG	20	119	N_E
LEU	40	231	N_E
LYS	30	189	$mean(N_S, N_E)$
PRO	80	133	$maj(N_S, N_E, N_{ES})$
TYR	18	91	N_{ES}
VAL	35	161	N_E

TABLE IV: Detection results for 2.9 Å CryoEM single particle reconstruction of an "anthrax toxin protective antigen pore (emd-6224)"

10% of LYS, TYR, ARG, GLY and PRO. Figure 14 illustrates the detection results for LEU, LYS, TYR and GLY at map resolution of 3.1Å.

Note that the use of simulated data was crucial for the 2.2Å resolution experiment. For resolutions 2.9Å and 3.1Å the role of simulated data is less important, since more experimental maps are available.

IV. WEB SERVER IMPLEMENTATION

We have developed a server of the detection/prediction stage of the AAnchor algorithm at http://bioinfo3d.cs.tau.ac.il/AAnchor/. The server outputs the detection results in PDB format, so that they could be displayed in commonly used programs such as UCSF Chimera [23]. Figure 15 illustrates the server output as shown in UCSF Chimera. Yellow spheres, which mark the predicted locations of prolines, are superimposed with the reference atomic model.

V. CONCLUSIONS

In this paper we present a CNN based method for localization and classification of amino acids in high resolution cryo EM maps. Whilst the *de-novo* detection of all protein

Amino Acid Type	Picks with Confidence of 80%	Total in Protein	Best Method
ARG	10	117	N_{ES}
GLY	25	207	N_E
LEU	25	108	N_E
LYS	20	171	N_E
PRO	9	72	$mean(N_R, N_E)$
TYR	18	144	$maj(N_R, N_E, N_{ES})$

TABLE V: Detection results for 3.1 Å CryoEM single particle reconstruction of an "Lysenin Pore (emd-8015)"

Fig. 13: Proline detection in 2.9 Å CryoEM single particle reconstruction of an "anthrax toxin protective antigen pore (emd-6224). Purple regions indicate the detected proline residues
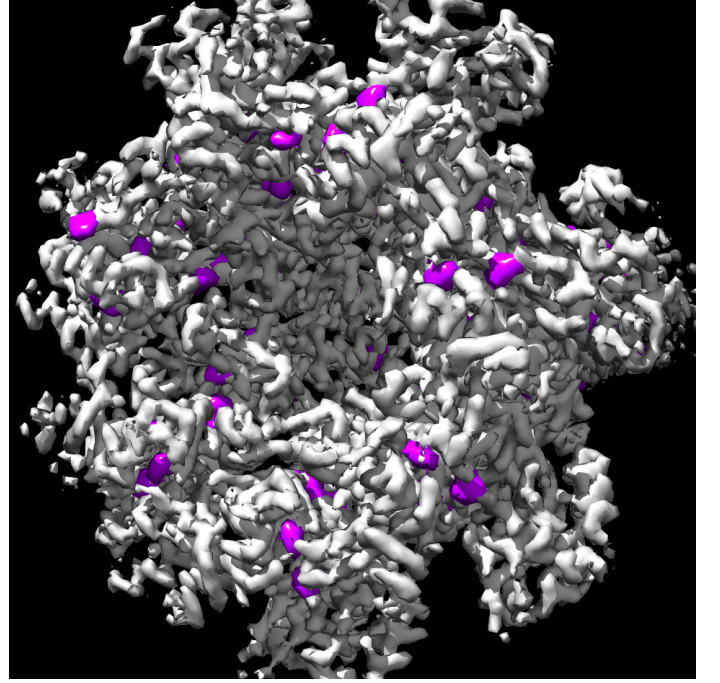

Fig. 14: Detection leucines, lysines, glycines and tyrosines in 3.1 Å CryoEM single particle reconstruction of single particle reconstruction of an "Lysenin Pore (emd-8015)". Red regions indicate the detected leucines, blue regions indicate the detected lysines, yellow regions indicate the detected glycines, and green regions indicate the detected tyrosines
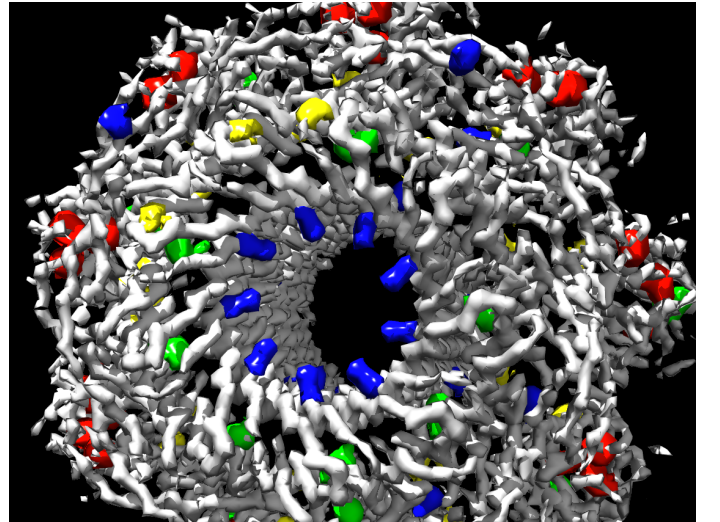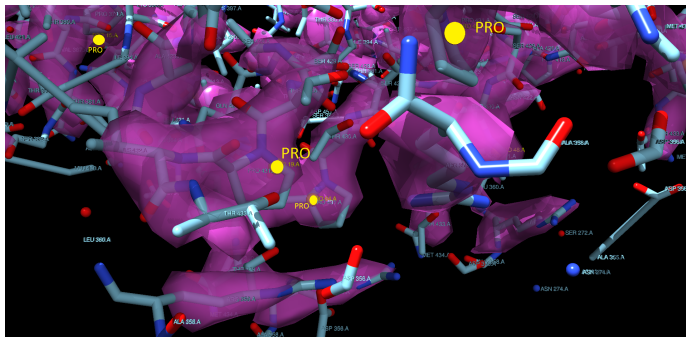
Fig. 15: Detection of proline residues in a 2.9 $\mathring{A}$ Cryo EM map. Yellow spheres show the predicted centers of mass for each of the detected amino acids.

residues is still a hard problem, we succeeded to detect with high confidence a significant percentage of some amino acids. Experimental results show that the proposed method is capable of detecting a sufficient number of amino acid "anchors" in a cryo Em map of resolution $3.1\mathring{A}$ or higher. The reported confidence of a detection is at least as the ground truth accuracy. These anchors can be further exploited in conjunction with several proposed modeling techniques as well as in the development of novel modeling methods.

We analyzed the detection process and factors affecting the classification task and concluded that the number of rotamers for a given residue and the size of the training data set have a dominant effect on the detection rate, while the amino acid size plays a secondary role. Contrary to expectation, the results for $2.9 - 3.1$ $\mathring{A}$ resolution maps have a better detection potential than the more accurate 2.2 $\mathring{A}$ maps. This is due to the limited training data set existing for the better resolution samples. Also the detection accuracy can be significantly improved by combining simulated and experimental cryo EM maps to compensate for the lack of experimental data.

As the number of released high resolution maps grows [2], the detection accuracy will rise. Applying post processing techniques such as Linear Regression [29] and SVM [17] may potentially decrease the number of false detections. Hopefully, with the accumulation of a sufficiently large number of experimental maps the classification of all voxels in a map to the different amino acids will be solvable using the Fully Convolutional CNN [30] technique.

REFERENCES

[1] W. Kühlbrandt. The resolution revolution. *Science*, 343(6178):1443–1444, 2014.
[2] C. L. Lawson, et al. Emdatabank unified data resource for 3dem. *Nucleic acids research*, 44(D1):D396–D403, 2015.
[3] A. Bartesaghi, et al. 2.2 A resolution cryo-EM structure of β-galactosidase in complex with a cell-permeant inhibitor. *Science*, 348(6239):1147–1151, 2015.
[4] S. Banerjee, et al. 2.3 å resolution cryo-em structure of human p97 and mechanism of allosteric inhibition. *Science*, pages 871–875, 2016.
[5] W. Jiang, et al. Bridging the information gap: Computational tools for intermediate resolution structure interpretation. *J. Mol. Biol.*, 308(5):1033–1044, 2001.
[6] K. Lasker, et al. Ematch: discovery of high resolution structural homologues of protein domains in intermediate resolution cryo-em maps. *IEEE/ACM TCBB*, 4(1):28–39, 2007.
[7] Z. Yu and C. Bajaj. Computational approaches for automatic structural analysis of large biomolecular complexes. *IEEE/ACM TCBB*, 5(4):568–582, 2008.
[8] M. Rusu and W. Wriggers. Evolutionary bidirectional expansion for the tracing of alpha helices in cryo-electron microscopy reconstructions. *J. Struct. Biol.*, 177(2):410–419, 2012.
[9] D. Si, et al. A machine learning approach for the identification of protein secondary structure elements from electron cryo-microscopy density maps. volume 97, pages 698–708. Wiley Online Library, 2012.
[10] F. DiMaio and W. Chiu. Tools for model building and optimization into near-atomic resolution electron cryo-microscopy density maps. In *Methods in enzymology*, volume 579, pages 255–276. Elsevier, 2016.
[11] M. Chen, et al. De novo modeling in cryo-em density maps with pathwalking. *Journal of structural biology*, 196(3):289–298, 2016.
[12] G. Terashi and D. Kihara. De novo main-chain modeling with mainmast in 2015/2016 em model challenge, 2018.
[13] R. Y.-R. Wang, et al. De novo protein structure determination from near-atomic-resolution cryo-em maps. *Nature methods*, 12(4):335–338, 2015.
[14] C. K. Cassidy, et al. CryoEM-based hybrid modeling approaches for structure determination. *Curr. Opin. MicroBio.*, 43:14–23, 2018.
[15] A. Krizhevsky, et al. Imagenet classification with deep convolutional neural networks. In *Adv. Neural Inf. Process Syst.*, pages 1097–1105, 2012.
[16] Y. Taigman, et al. Deepface: Closing the gap to human-level performance in face verification. In *Proc. IEEE CVPR*, pages 1701–1708, 2014.
[17] R. Girshick, et al. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proc. IEEE CVPR*, pages 580–587, 2014.
[18] M. Bokori-Brown, et al. Cryo-em structure of lysenin pore elucidates membrane insertion by an aerolysin family protein. *Nature Commun.*, 7:11293, 2016.
[19] Y. LeCun, et al. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.
[20] I. Goodfellow, et al. *Deep Learning*. MIT Press, 2016.
[21] F. Chollet et al. Keras, 2015.
[22] M. V. Shapovalov and R. L. Dunbrack Jr. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure*, 19(6):844–858, 2011.
[23] E. F. Pettersen, et al. Ucsf chimera - a visualization system for exploratory research and analysis. *J.Comp. Chem.*, 25(13):1605–1612, 2004.
[24] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.
[25] C. Guo, et al. On calibration of modern neural networks. In *Proceedings of the 34th Int. Conf. on Machine Learning*, 2017.
[26] Y. Dong, et al. Antibody-induced uncoating of human rhinovirus b14. *Proc. Natl. Acad. Sci.*, 114(30):8017–8022, 2017.
[27] A. Merk, et al. Breaking cryo-em resolution barriers to facilitate drug discovery. *Cell*, 165(7):1698–1707, 2016.
[28] J. Jiang, et al. Atomic structure of anthrax protective antigen pore elucidates toxin translocation. *Nature*, 521(7553):545, 2015.
[29] I. Naseem, et al. Linear regression for face recognition. *IEEE Trans. Pattern Anal. Mach. Intell*, 32(11):2106–2112, 2010.
[30] J. Long, et al. Fully convolutional networks for semantic segmentation. In *Proc. IEEE CVPR*, pages 3431–3440, 2015.