1 Abstract

Cellular processes are performed and regulated by assemblies of macromolecules. Full understanding of the function, associations and dynamics of such assemblies comes from their detailed atomic structural description. Traditionanlly structural analysis of a complex is made by integration of the results of variuos experinental techniques: X-ray crystallography, NMR, cryo electron microscopy (cryo-EM) and others. The output of a cryo-EM experiment is 3D density maps. In recent years, vast progress was achieved in obtaining high-resolution (3.5 Åand below) cryo-EM maps. However, modelling an atomic structure from a cryo EM map remains difficult. For high-resolution maps researchers mostly rely on methods developed for X-ray crystallography. Modelling protein assemblis into medium resolution maps usually based on image processing techniques and has still not yet achieved good performance.

Parallel to the advances in Cryo-EM in the last decade, deep neural networks achieved remarkable performance in various 2D and 3D image processing tasks. Those include Convolutional Neural Networks (CNNs) for image classification , Fully Convolutional Networks (FCN) for semantic image segmenation, Variational AutoEncoders (VAEs) or Generetive Adversarial Networks (GANs) for realistic image generation.

We propose a Deep Learnig Analysis Framework (DLAF) for integrative structural analysis of cryo-EM maps. The proposed framework integrates 3D imaging information from a cryo-EM map with existing sequence and structural data. Deep Convolutional Networks are used to locate structural motifs within a map,. CNNs are trained on a dataset adjusted to a specific problem. This adjustment is done using the sequence and structural data. Realistic cryo-EM map simulation plays a key role in the presented approach. While the simulation is crucial for DLAF, it is also a valuable tool for development and analysis of algorithms which work with cryo-EM and can contribute a lot to structural bioinformatcis community. Preliminary results of our study show that CNN are capable to successfully detect amino acids in high-resolution cryo-EM maps. We also developed a realistic simulation of a cryo-EM map using an adversarial deep learning approach.

2 Introduction

2.1 General

Cellular processes are performed and regulated by assemblies of macromolecules. Such assemblies, which are often referred to as molecular machines [1], vary widely in their lifespan, size, activity and dynamics. Complete understanding of the function and dynamics of an assembly is derived from its detailed atomic structure. Structural characterization of macromolecular assemblies represents a major challenge in structural biology. While there exist contemporary experimental techniques, each suffers from drawbacks. X-ray crystallography ([16])

is limited by the ability to grow suitable crystals and to build molecular models into large unit cells; NMR spectroscopy ([40]) is restricted by size; electron microscopy ([8]), affinity purification ([7]), yeast two hybrid ([37]) and FRET spectroscopy (Truong and Ikura 2001) suffer from low resolution of the corresponding structural information. Single particle cryo electron microscopy (s.p. cryo EM) plays an important role in biomolecular structure determination. For many years data obtained by s.p. cryo-EM was of low resolution due to technical limitations. However, recent technical improvements such as direct electron detectors combined with modern computational methods for data acquisition and image processing, coined as "resolution revolution" have led to an ability to obtain s.p. cryo-EM images of high resolution, i.e., 4 Å and better, [51], [18]. While the gold rush of high resolution s.p. cryo-EM data continues, the task of determining an atomic structure from s.p. cryo-EM image remains very labourous and can be thought of like a master artwork. Next, we provide a more detailed description of few computational methods that can be used in atomic structure determination from a s.p. cryo-EM image.

2.2 Protein Structure Modelling from cryo-EM maps

Intermediate resolution Most of the effort in modeling protein structures into intermediate resolution ($5 - 10$‘ Å) maps focuses on locating structural motives. PowerFit [12] and Multifit [47] compute candidate locations for predefined structural fragments within a map. EMatch [17], SSEHunter [4], StrandRoller [43], and EMBuilder [58] use geometry calculations and a template-based search to identify secondary structures. Machine Learning algorithms are widely used to identify secondary structure, e.g. SSELearner [44] (SVM) and [29] (Deep CNNs). EMatch [17] and MULTIFIT [47] exploit structural motives detection in order to model the entire protein complex. The modelling is based on rigid fitting of template protein structures into a cryo-EM map, while detected structural fragments serve as anchors.

At resolutions better than $4 - 5$ Å de novo modeling techniques are being exploited. In addition to adaptations of the standard X-ray crystallography modeling methods, which tend to be time consuming, several de-novo modeling techniques have been developed to deal specifically with cryoEM density maps [14]. Pathwalking [11] detects first pseudo-atom anchors and then applies the travelling salesperson (TSP) combinatorial optimization algorithm to detect the protein backbone. MAINMAST [46] detects a set of anchor points and calculates the backbone by applying a minimun spanning tree (MST) approach. A recently published method [59] 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.

2.3 Deep Learning

Deep neural networks are powerful learning models that achieve excellent performance in visual and speech recognition problems [9, 8]. Neural networks

achieve high performance because they can express an arbitrary computation that consists of a modest number of massively parallel nonlinear steps.

Convolutional neural networks (CNNs) were proposed by Yann LeCun in 1989 for zip code recognition [54]. 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 $\mathbf{y}$:

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

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:

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

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 [23].

The development of the techniques below was motivated by the observation that the current amount of experimental cryo-EM data is not sufficient for DL training and thus the data has to be augmented by data from other sources.

Domain shift. If a CNN is used on data with features distribution different from the data it was trained on, its performance degrades. The problem is known as **domain shift**. In this case a **domain adaptation** technique is required. In **transfer learning** [35] a pretrained net is fine tuned by training on a relatively small dataset, which has feature distribution resembling the query data. The main principle of **adversarial domain adaptation** techniques ([48], [21]) is to train a CNN such that in a prediction phase of the network only the features which are common to the train and the test domains are used.

2.4 Proposed Research

We propose to develop and apply deep learning algorithms for locating structural motifs in cryo EM maps of high and medium resolution. While all algorithms utilize deep CNN for 3D object detection and classification, the detection goal depends on the cryo EM map resolution:

1. CNN for detection of Amino Acids in High Resolution ($2 - 4$ Å) maps.

2. CNN for annotation of SSEs (helices, beta-strands) in Medium Resolution $(4-6$ Å) maps.

3. CNN for detection of functionaly significant regions, such as Binding Sites in Medium Resolution $(4-6$ Å) maps.

We anticipate a lack of comprehensive experimental training data set for all three CNNs mentioned above, we intend to address it as follows:

1. For High Resolution $(2-4$ Å) maps CNN is trained on existing crystallography data. Since the X-ray crystallography technique differs from single particle cryo-EM a domain shift problem is anticipated. We suggest to handle it using Deep Domain Confusion [49] technique.

2. We shell develop a realistic cryo-EM simulations using the VAE-GAN net ([28] , [53]).

In the final stage we shall attempt to develop a novel algorithm for de-novo modelling of protein structures from cryo-EM maps at appropriate resolution.

3 Research Goals And Significance

3.1 Detection of anchor amino acids in high resolution cryo-EM density maps

We propose a Deep Learning based method for the detection of high confidence anchor amino acid residues in high resolution cryo-EM maps. We focus on detection of **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. Reliable prior detection of amino acid anchors 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. The lack of sufficient experimental data required for the training stage is the main expected pitfall. We propose two different approaches to cope with the shortage of experimental data.

Integrating X-ray crystallography data with cryo-EM for structure determination .

X-ray crystallography is an invaluable tool for revealing the three-dimensional structure of molecules. Contemporary online databases: CSD (`http://www.ccdc.cam.ac.uk`) ,PDB (`http://www.rcsb.org/pdb`, and PDBe (`http://www.ebi.ac.uk/pdbe/node/1`) contain more than 100,000 biological macromolecules.

While having the same form of a 3D matrix of density values, Xray crystallography data differs from cryo EM data in the conditional distribution of the outputs given the inputs. The problem is known as *dataset shift*, [38]. This is due to different principles of cryo-EM and X-ray crystallography experiments, from

specimen preparation to data processing, (see Figure 1 and [56], [57], [52], [51] for details).

The proposed algorithm is to train a CNN on large amount of X-ray data (*source domain*) with the existing cryo-EM data (*target domain*) . A **domain adaptation** procedure will be applied to compensate for the difference between the source and the target domains.

The presented algorithm contains a novel approach to cryo-EM and X-ray processing in which both data types complement each other in the task of atomic structure determination. Running the algorithm on new experimental data will reveal structural properties at high precision, which is crucial for revealing the functionality and for drug design.

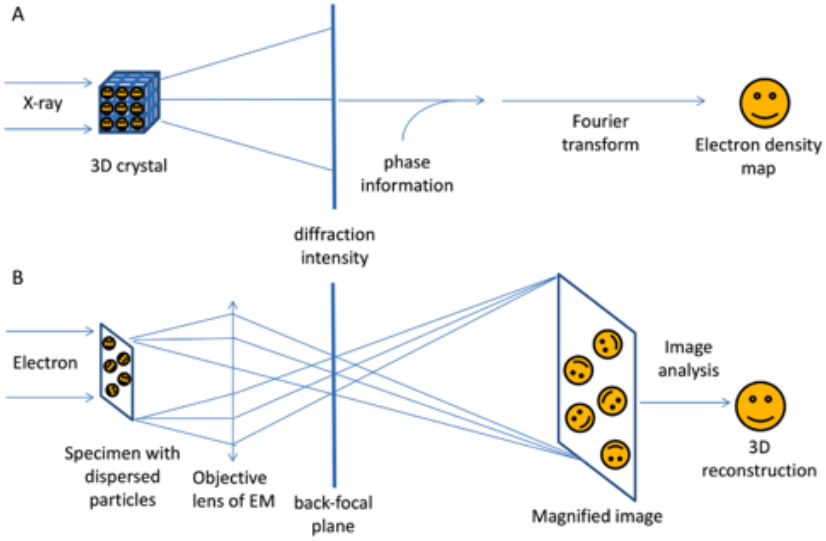


Figure 1: Technical difference between X-ray crystallography and single particle cryo-EM.

3.2 Simulating cryo-EM maps using Generative Adversarial Network

Our previous work on the AAnchor algorithm [41], showed that augmenting a training dataset by synthetic data improves a classification CNN performance. Contemporary models for generating synthetic cryo EM maps from atomic structure are incapable to generate realistic data. We propose to use Generative Adversarial Networks with Variation AutoEncoder (VAE-GAN) to create cryo-EM maps which are indistinguishable from experimental ones. GANs and VAEs are proven deep learning techinques for generating 2D and 3D images. Reliable cryo EM map simulation is of great significance for the protein structural modelling

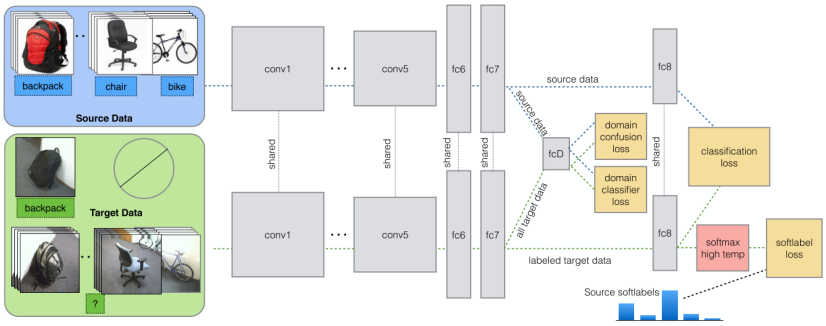


Figure 2: Deep Domain Confusion Architecture

task. In addition to augmenting the training dataset as in our case, such simulation is capable of improving performance of template matching based modelling methods: PowerFit [12], MultiFit [47], EMatch [17], and others.

3.3 Annotating Secondary Structure in Medium Resolution ($4-6$ Å) cryo-EM maps

We propose a Deep Learning based method for the detection of Secondary Structure Elements (SSE) in medium resolution cryo-EM map. Locating SSEs (helices, beta-strands) is of high importance to protein modelling and number of methods have been developed for the task. One group of the developed methods uses image processing tools, for locating cylinder-like (helices) and plane (beta-sheets) structures, [4], [27], [42], [55]. Another family of SSE detections methods uses ML approach [32] [45] including deep CNNs [50], [29], [31], [34].

While the above mentioned methods relied solely on structural features, we propose an **integrative** method which incorporates protein sequence information as well. The query protein sequences are used to search for homologs with known structure. A realistic cryo-EM simulation will be used to create a dataset from these homologous structures. The new dataset should have a similar features distribution as the query map. The created synthetic dataset will be used for fine tunnig of the detection CNN.

3.4 Detection of binding sites (BSs) in medium resolution ($4-6$Å) cryo-EM density maps

Whilst de-novo protein modeling from a median resolution map remains an unresolved problem, we focus on locating BSs. Our task is to mark voxels in the cryoEM map which belong to a protein-protein interface or a protein binding site. Since BSs dictate the molecular function, they are usually more evolutionary conserved and tolerate less flexibility than less functionally important parts of a protein structure. A deep learning approach will benefit from both of

the above mentioned properties. Evolutionary conservation leads to a greater amount of similar structures in existing databases. Due to the limited flexibility, structures under search should be similar or nearly similar to those in a database. Locating BSs is an important step towards full atomic structure determination, since it provides information about protein tertiary structure and separation to domains. From a biological point of view BSs are key to understanding protein functionality

Integrative algorithm for finding conserved structures in cryo-EM maps. The existing vast amount of data about conserved structures, is hard to utilize, since it belongs to different domains. Resolved protein structures are represented as lists of atom positions. Results of cryo-EM experiments are given in a form of a 3D matrix of electron density. Sequence data is represented as a list of chains, where each chain is an ordered sequence from an alphabet of twenty letters. To address the challenge of utilizing all three types of data, we suggest a deep learning algorithm architecture which combines a standard Convolutional Neural Network with Generative Adversarial Network, Transfer Learning and Multiple Sequence Alignment.

4 Preliminary Results

In our preliminary work we have developed AAnchor - a method for locating amino acids in a cryo EM maps of high resolution, and cryo-GAN - VAE-GAN based simulation of a cryo EM map.

4.1 cryo-GAN

4.1.1 Motivation

While there is a high demand for realistic cryo-EM simulation, the most popular existing tool (callled molmap) performs Gaussian Blurring on atoms.

While a synthetic map represents an "ideal" universe, in an experimental map the noise is not i.i.d gaussian and different regions have different volume density. Moreover, it is assumed that some of the physico-chemical properties which affect experimental cryo-EM maps such as charge and atomic bonds, are not captured by the existing simulation. Fig. 3 shows experimental and "molmap" map of the STING protein (pdb ID 6nt8, `https://www.rcsb.org/structure/6NT8`).

4.1.2 Methods

A three dimensional VAE-GAN network was used for map generation. The network architecture is shown in Fig. 4. The network input is an atomic structure fragment after voxelization, i.e., five (for each atom H, C, O, S, N) 3D matrices (called channels). The output is a cube of a cryo-EM map - 3D matrix of

density. Due to the normalization implied in the training phase, cryo-GAN generates cubes with mean value 0.5 and standard deviation of 0.16. The proposed cryo-GAN architecture differs from the well known image generating architectures ([28, 53]) in three aspects:

- Input: The input is five channels of 3D matrices, while in [28] and [53] the input is 2D images.
- Reconstruction Loss (RC Loss). Traditionally RC loss is the mean square of the difference between the obtained and the reference images. We added mean and standard deviation terms, i.e., there is a penalty if the mean value or the standard deviation of the created 3D cube differ from the required ones.
- Discriminator: The role of a discriminator is to distinguish between real (input) and fake (generated) map cube. Usually in GANs the output is generated from a random input, sampled from the same distribution. In our case the output is generated from a deterministic input - atomic model fragment, so we feed the atomic model fragment to the discriminator in addition to the generated map cube.

We implied the training strategy recommended in [53]:1) GAN loss was not backpropagated to the discriminator, 2) discriminator parameters were updated only if its accuracy fails below 80 %.

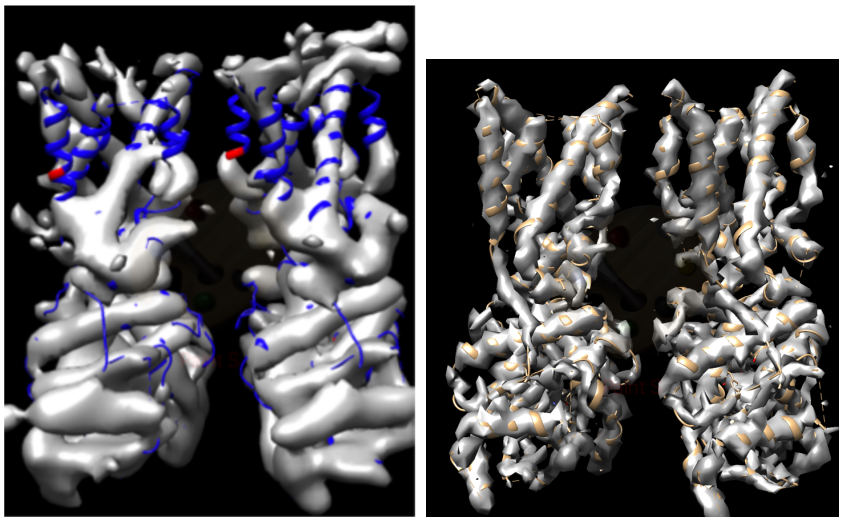
4.1.3 Results

Training phase results are shown in Fig. 5. Loss and discriminator convergence show that the network has converged and is not overfitted. The **generated synthethic map** of STING protein (pdb ID 6nt8) is shown in Fig. 6. While from visual inspection the map resembles the experimental map (see Fig. 3a), an additional evaluation criterion is required. To evaluate the quality of a generated map, an additional discriminator, named **evaluation discriminator** was trained independently from the cryo-GAN network. The evaluation discriminator was trained independently to distinguish real maps from those created by "molmap" and random data. Training results of the evaluation discriminator are shown in Fig 7. Evaluation of the map generated by cryo-GAN is performed by running the evaluation discriminator on each voxel of the map. 56 % of the cryo-GAN map voxels were labeled as "real".

4.2 AAnchor

AAnchor [41] succeeded to detect with confidence of above 80% a significant percentage of amino acids in cryo EM maps of resolutions of 3 Å and below. Analysis of the AAnchor detection performance brings to the following observations:

1. The CNN Classifier combined with an effective search algorithm is capable of detecting a sufficient number of amino acid "anchors".



(a) Experimental map of protein 6NT8, EMD-0505

(b) Synthetic Map of protein 6NT8, created using "molmal" command.

Figure 3: Cryo EM maps of protein 6NT8: experimental and simulated by molmap

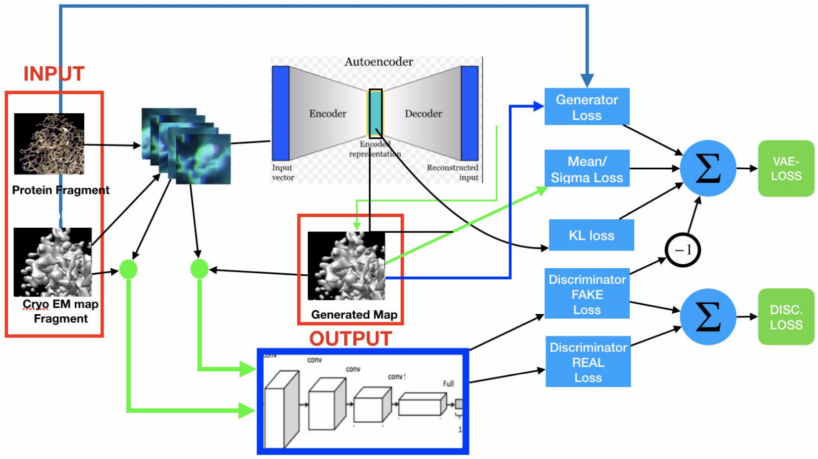


Figure 4: Cryo-GAN architecture.

2. The CNN classifier performance is critical for the amino acid detection performance. Designing and training a high precision classifier is key to a precise detection algorithm.

3. Synthetic Data can be used to improve classification performance if there is lack of experimental data.

5

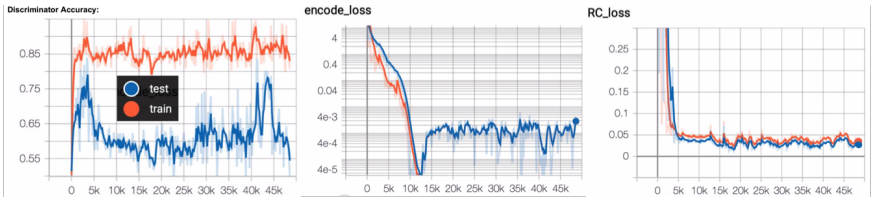


Figure 5: Training Phase Results for Cryo-GAN. Discriminator accuracy on train data of 80 % agrees with training strategy. Discrimator Accuracy of 50 % on test data means that the discriminator cannot distinguish between real and synthetic maps. RC loss and Encode loss converge on train and test data, meaning that both the Encoder and the Generator converged

4. Appropriate confidence estimation can be obtained from a classification CNN using a simple calibration process.

4.2.1 AAnchor overview

Given a cryo EM map at high resolution, AAnchor finds the position and type of amino acids within the map. The method is divided to two main procedures: classification and detection.

1. The **classification** problem is defined as the assignment of one out of 21 labels (20 amino acids plus "none") to a voxel and its close neighborhood. An amino acid is assigned to a voxel if its center of mass is within 1.5 Å from the voxel position.

2. The **detection** problem is defined as the localization of amino acids of a specific type in the cryo EM map. The output of the detection problem are the coordinates of the detected amino acid center of mass followed by estimated confidence.

Inspired by the similarity of our tasks to the known image processing problems we built a CNN that classifies each voxel to 21 types. We trained and tested this CNN on simulative and experimental cryo EM maps at resolutions 2.2 Å , 2.9 Å , and 3.1 Å . Using the sliding window approach and post-processing filtration, we applied the trained classification CNN to the detection problem.

4.2.2 The Classification CNN performance

Once trained, the CNN is used to perform prediction on a test 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 is the one giving maximal "probability", and the confidence is p_i. We show the classification results by plotting the obtained **confusion matrix** ([19]).

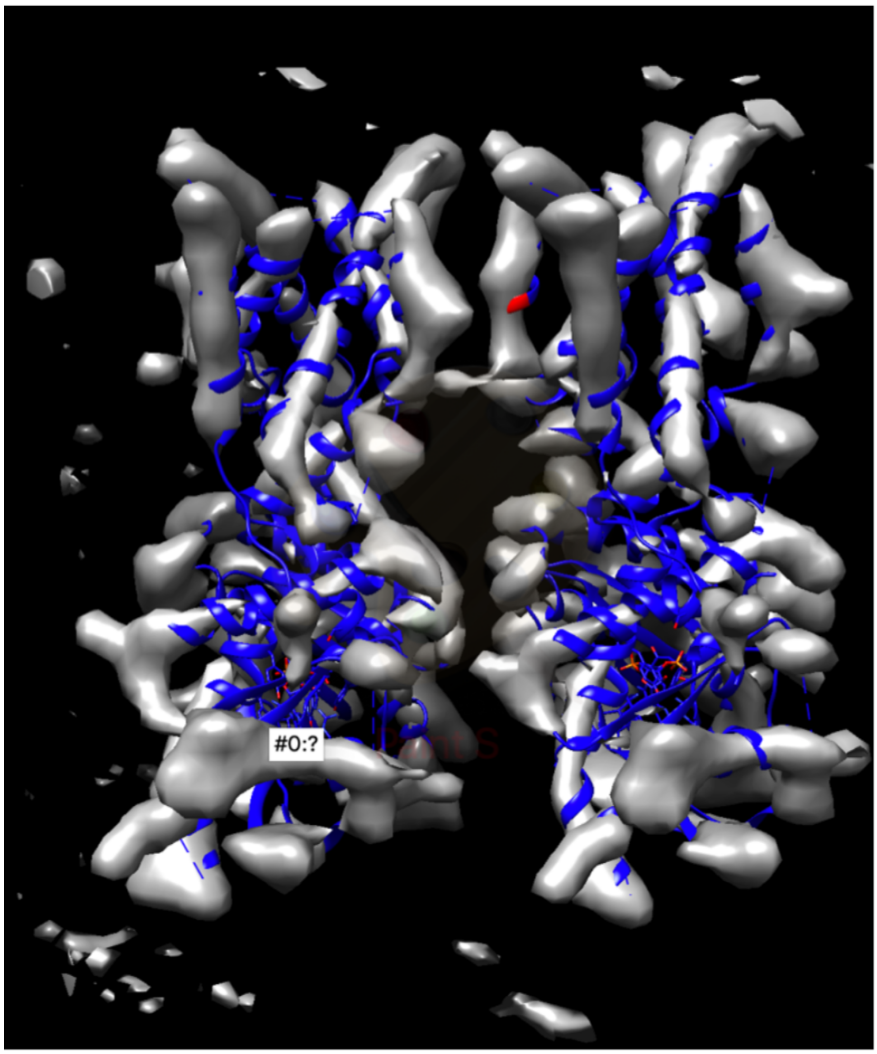


Figure 6: Genarated map of STING protein (pdb ID 6nt8)

4.2.3 Classification Accuracy

2.2 **Å Resolution.** We have the CNN performance on a 2.2 Åcryo-EM map of β-galactosidase [6] (EMD-2984). At the time of the research, only three additional cryoEM maps with resolution better than 2.3 Å were available: EMD-8762 [15],EMD-8194 [33], and EMD-3295 [5]. Even augmented, this is definitely not enough for proper training of the CNN. The confusion matrix obtained for the CNN trained on the experimental data is shown in Figure 8a. A much better accuracy is achieved if the experimental data is mixed with the synthetic (simulated) data in the training data set. Figure 8b shows the confusion matrix

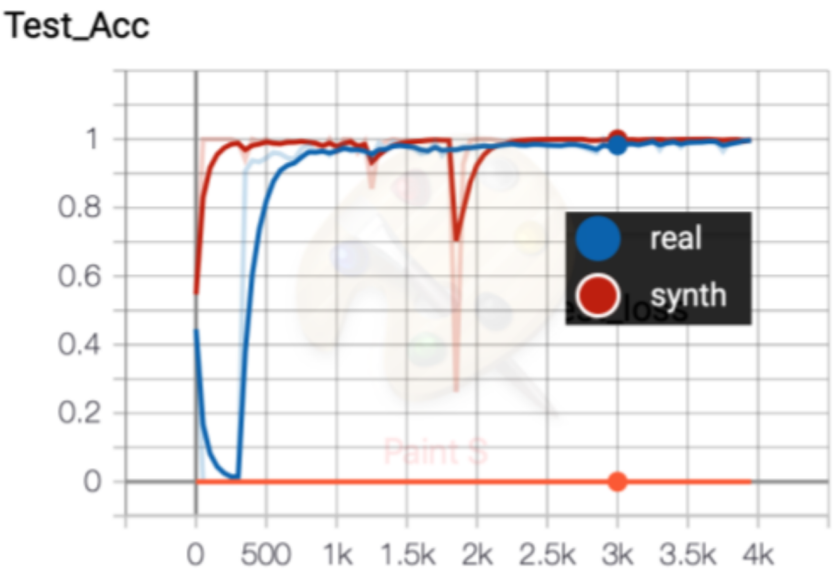
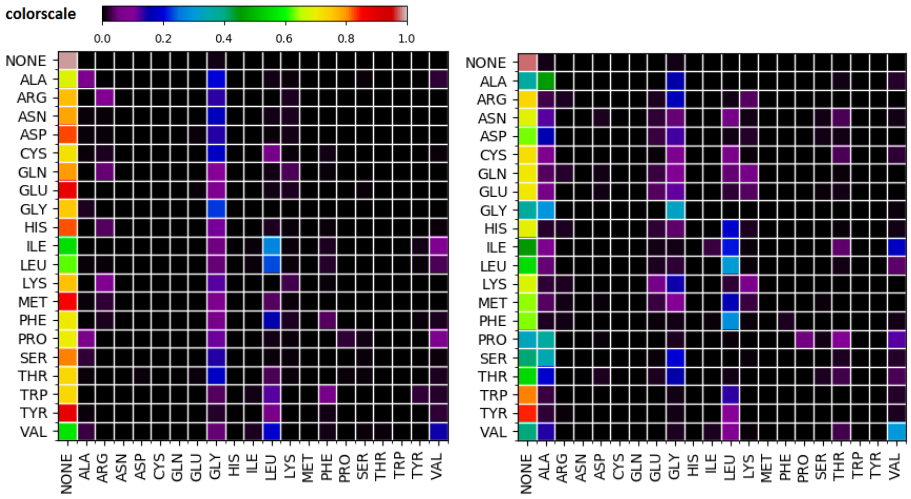


Figure 7: Training Results of the evaluation discriminator

in this case.



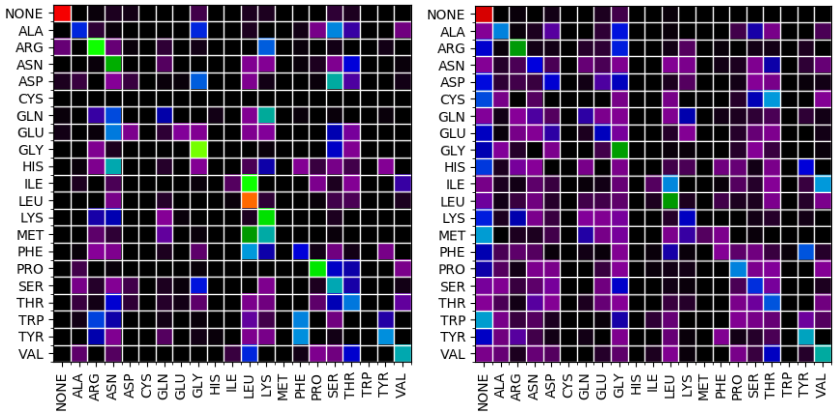
(a) Train: Experimental.
Test :Experimental

(b) Train: Sim+Exp.
Test :Experimental

Figure 8: Confusion matrix obtain for variuos training sets for Resolution 2.2 Å . An entry $a_{i,j}$ of a confusion matrix ([6]) A is defined as the ratio $a_{i,j} = \frac{T_j^j}{N_j}$, 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.

Resolutions 2.9 Å and 3.1 Å Coarsening the resolution has two opposite effects on the classification accuracy. Intuitively, in higher resolution maps the amino acids shape is sharper, however lower resolution maps benefit from a larger training dataset. The confusion matrices for 2.9 Å Anthrax toxin protective antigen pore and 3.1 Å Lysenin Pore are presented in Figures 9a and 9b, respectively.

Accuracies for 2.9 Å and 3.1 Å are significantly better than those for 2.2 Å . This is clearly due to the increased training dataset. The only exception is Alanine, which is probably too small to be detected at resolutions coarser than 2.2 Å . However, moving from 2.9 Å to 3.1 Å 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.



(a) Confusion matrix for 2.9 Å , test map EMD-6224

(b) Confusion matrix 3.1 Å , test map EMD-8015

Figure 9: Classification Results for Resolution 2.9 Å and 3.1 Å .

4.2.4 Detection Results

Whilst the detection (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. Our goal was to identify **anchors**, i.e., amino acids, which have been located and labeled with confidence above 80%.

Table 1 demonstrates the detection results for various map resolutions. The fraction of the detected amino acids is 10%-20%, depending on resolution. The best results were achieved for 2.9 Å resolution, where we detected 70% of prolines. Recall that all detections contain less than 20% of errors. We expect this result to improve, as more high resolution cryo EM maps are being released.

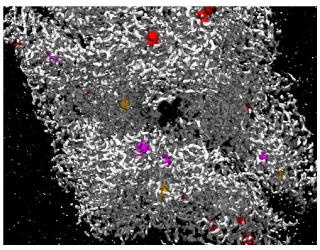
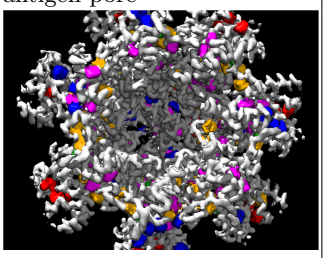
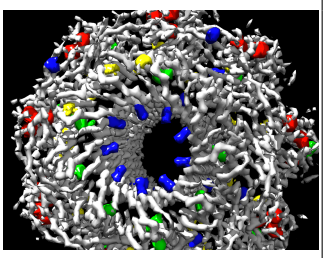
2.2 Å : EMD-2984 β -galactosidase	2.9 Å :EMD-6224 Anthrax toxin protective antigen pore	3.1 Å :EMD-8015 Lysenin Pore
		
ASN:— ARG: 17 out of 133 GLY: — LEU: 15 out of 231 LYS: — PRO: 10 out of 189 TYR: — VAL: 10 out of 119	ASN: 5 out of 287 ARG: 20 out of 119 GLY: — LEU: 40 out of 231 LYS: 30 out of 189 PRO: 80 out of 133 TYR: 18 out of 91 VAL: 35 out of 161	ASN: :— ARG: 10 out of 117 GLY: 25 out of 207 LEU: 25 out of 108 LYS: 20 out of 171 PRO: 9 out of 72 TYR: 18 out of 144 VAL: —

Table 1: AANchor detection results for resolutions 2.2 Å , 2.9 Å , and 3.1 Å

5 Future Research

Our main goal is to develop a tool for extracting structural information from a cryoEM map at various resolutions. Specifically, given a cryoEM map of an assembly, we intend to localize regions in the map of known and important atomic structure. Such regions are called anchors. The structure of the anchors will depend on the map resolution. For very high-resolution maps (3 Angstroms and better) amino acids and aromatic rings will be identified. For intermediate resolution maps (4- 6 Angstroms and better) binding sites and protein-protein interfaces will be identified.

The general strategy will be to train a classification Convolution Neural Network (CNN) for each anchor type. Given a small cube of cryoEM density a classification CNN will assign a label to the cube. Cubes where an atomic structure was identified with high confidence will be labelled as anchors. The whole map will be searched for anchors by transforming a classification CNN into a Full Convolutional Network (FCN).

5.1 Detecting Amino Acids in very high resolution maps.

We should continue to develop the techique of Amino Acid detection in High Resolution Maps. 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 centres of mass of specific amino acids. The goal is to report only those amino acids, which have been detected with high confidence, nicknamed "anchors". The detection procedure consists of extracting small cubes from a

map and applying a *selective classifier* to each cube.

Selective Classification . A selective classification is a classification with a reject option. A *selective classifier* is a pair of functions $f(x)$, called the *classifier*, and $g(x)$ is called *selection function*. The classification of a proposal x is as follows:

$$(f,g)(x) = \begin{cases} f(x) & if \quad g(x) = 1 \\ NONE & if \quad g(x) = 0 \end{cases} \tag{3}$$

We anticipate that integrating sythetic and X-ray crystallography data will result in improvement of the performance of our algorithm.

5.1.1 Training Phase

In a training process the parameters of a selective classification are adjusted by running backpropagation algorithm on a large amount of labeled data. The main limitation of the training phase is small amount of high resolution cryo EM maps. There are two additional data sources which we try to use in the training phase:

1. Crystallographic data. PDBe database [50] contains tens of thousands of electron density maps.
2. Synthetic Data i.e., cryo-EM maps obtained by simulation program from a given atomic structure.

Both sources mentioned above suffer from the **domain shift**, i.e., the difference in the data distribution between train and test sets. To address this issue, two **domain adaptation** approaches are proposed to bridge the gap between the source and target domains: utilizing crystallography data (Section 5.1.2) and semi-supervised approach (Section 5.2) .

5.1.2 Utilizing X-ray Crystallography Data

We suggest to use the Domain Confusion architecture presented in [25], see Fig. 2. Up to date for most of the molecular complexes in cryo-EM dataset X-ray data can be found for the same molecule or its close homolog. Meaning that a one-to-one correspondence is available for entries in the target and source domain. We will adjust the domain confusion network to take an advantage of the above fact by adding a new loss function member.

5.2 Semi-Supervised Deep Learning Approach for Anchors Detection

Semi-Supervised Deep Learning Approach presented at Fig 10 utilizes realistic cryo-EM map simulation for Anchors detection. The query protein sequence is

used to detect atomic models of structural homologs. This can be done by employing local alignment search tools (BLAST, PSI-PLAST, HHPred) or retrieving known structures from the protein family according to existing hierarchical classifications (SCOP [26], CATH [36]). Retrieved homolog structures are fed to the VAE-GAN simulation to create realistic cryo-EM maps (see Section 5.6 for details). Created synthetic cryo-EM maps together with atomic models are used for fine tuning of the pretrained neural network. Anchors locations are obtained by running a fine-tuned network on the query map. The presented approach has a number of advantages:

- Sequence information is utilized in the anchor location task.
- Using homolog structures ensures that the training dataset and the query map have similar feature distributions.
- Training datasets of arbitrary size can be created.

5.3 Annotating Secondary Structure in a Medium Resolution ($4-6$Å) cryo-EM maps

The task is to assign each voxel in the qiven cryo-EM map to one of the three secondary structures: helix, beta-strand, coil. This task is known as **semantic segmentation**. The voxel is classified according to the cube in the 3D map which surounds it. Classification CNN converted to Fully Convolutional Network (FCN) is used. The classification CNN is pretrained on a set of experimental cryo-EM maps of medium resolution. Fine tunnig achieved by training the last two layers on simulated data set as shown on Fig. 10 and Section 5.2

5.4 Locating Binding Sites in medium resolution cryo-EM maps

Proteins with similar function often have active sites which are structurally similar. Active site structures are usually conserved during evolution. Despite the above facts, locating an active site for an unknown functionality is a challenging task. This is due to the wide structual diversity of protein active sites. Their structural diversity can be reduced by existing sequence analysis algorithms. We propose a four -phased algorithm for integrative location of binding sites - Fig.11.

In the **sequence analysis** phase possible reqions of binding sites are located in the query protein sequence. This can be done using traditional tools for finding conserved regions (conSurf [3] and others [9], [20], [39], [30]) or new Deep Learning Tools (DeepBind [2], and others [13], [10]). Atomic models are created from the located sequence regions using existing modelling tools (Modeller , Rosetta). In the **simulation** phase generated atomic models are fed to VAE-GAN to create realistic synthetic cryo-EM maps of expected binding sites. In the **transfer learning phase** a pretrained CNN is fited to the new dataset,

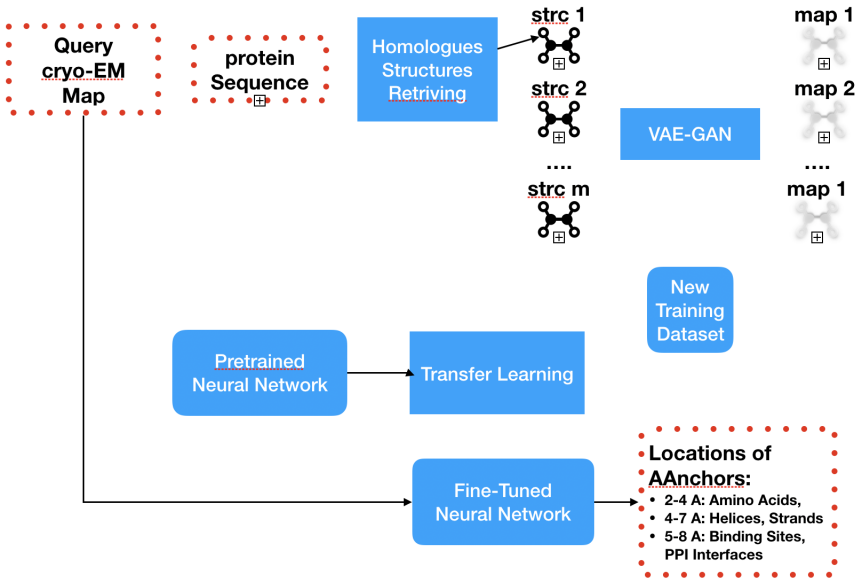


Figure 10: Semi Supervised ML Algorithm for Locating Anchors

which consists of generated atomic models and synthetic cryo-EM maps. Finally, in the **search** phase, the query map is fed to the fine-tuned neural network.

5.5 Calibration

In the anchor detection tasks(amino acids detection Section 5.1 and binding site detection Section 5.4) the confidence of reported detections matters. The goal is to filter out picks with confidence below a predefined threshold (say 80%). This can be done by **calibration** of a detection network, using post-training methods such as **temperature scaling** [24]. Another approach is to estimate the uncertainty of the neural network prediction by its training history [22].

5.6 Realistic cryo EM Simulation : Generative Adversarial Network

Realistic cryo EM map simulation is the core of the algorithm presented above. In the preliminary work we have developed cryo-GAN (Section 4.1). The simulation uses Variational AutoEncoder (VAE) architecture combined with Generative Adversarial Network (GAN) to create a cryo-EM map of an atomic structure. While the preliminary results show that cryo-GAN is able to generate realistic cryo-EM maps, additional performance evaluation is required. Moreover, we plan futher develop the simulation to make it a useful tool for algorithm developers working with cryo electron microscopy.

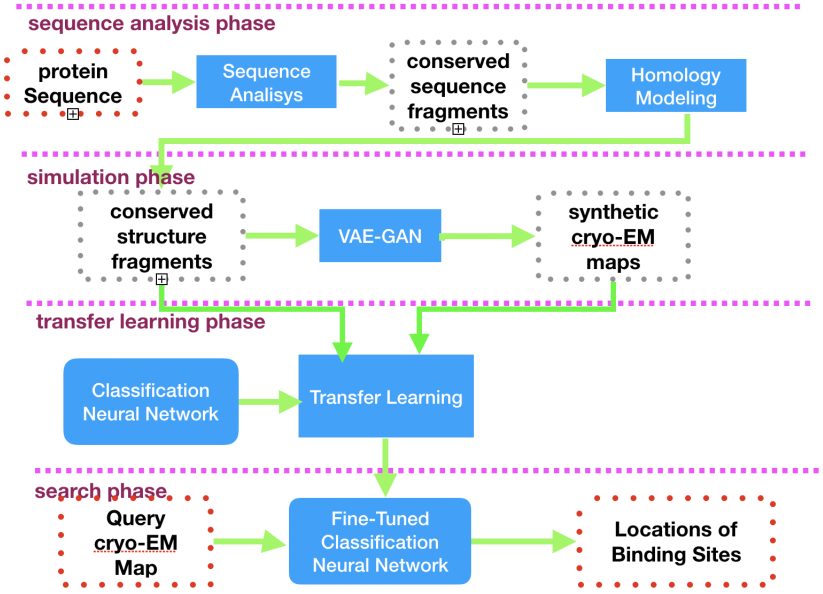


Figure 11: Finding Conserved Structures

6 Summary

As the time and cost of cryo electron microscopy experiments reduce , the ability of automatic structural analysis of experimental results becomes invaluable. We aim to show the potential of Deep Learning methods for analysis of cryo-EM maps. We proposed a number of algorithms that using DL extract structural information from a cryo EM maps of high and intermediate resolution. The proposed methods can be used for obtaining an atomic model of a molecule and for revealing its functionality. In the final stage of our work we will make an effort to integrate the proposed tools into an ab-initio modelling algorithm.

References

[1] Bruce Alberts. The Cell as a Collection of Protein Machines. *Cell*, 92:291–294, 1998.

[2] Babak Alipanahi, Andrew Delong, Matthew T. Weirauch, and Brendan J. Frey. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology*, 33(8):831–838, 2015.

[3] Haim Ashkenazy, Elana Erez, Eric Martz, Tal Pupko, and Nir Ben-Tal. ConSurf 2010: Calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Research*, 2010.

[4] Matthew L. Baker, Tao Ju, and Wah Chiu. Identification of Secondary Structure Elements in Intermediate Resolution Density Maps. *Structure*, 15(1):7–19, 2007.

[5] Soojay Banerjee, Alberto Bartesaghi, Alan Merk, Prashant Rao, Stacie L. Bulfer, Yongzhao Yan, Neal Green, Barbara Mroczkowski, R. Jeffrey Neitz, Peter Wipf, Veronica Falconieri, Raymond J. Deshaies, Jacqueline L.S. Milne, Donna Huryn, Michelle Arkin, and Sriram Subramaniam. 2.3 Å resolution cryo-EM structure of human p97 and mechanism of allosteric inhibition. *Science*, 2016.

[6] Alberto Bartesaghi, Alan Merk, Soojay Banerjee, Doreen Matthies, Xiongwu Wu, Jacqueline L S Milne, and Sriram Subramaniam. 2.2 A resolution cryo-EM structure of β-galactosidase in complex with a cell-permeant inhibitor. *Science*, 348(6239):1147–1151, 2015.

[7] Andreas Bauer and Bernhard Kuster. Affinity purification-mass spectrometry: Powerful tools for the characterization of protein complexes. *European Journal of Biochemistry*, 270(4):570–578, 2003.

[8] Wolfgang Baumeister and Alasdair C Steven. Macromolecular electron microscopy in the era of structural genomics. *Trends in Biochemical Sciences*, 25(12):624–631, 2000.

[9] John A. Capra and Mona Singh. Predicting functionally important residues from sequence conservation. *Bioinformatics*, 23(15):1875–1882, 2007.

[10] Ke Chen, Marcin J Mizianty, and Lukasz Kurgan. ATPsite : sequence-based prediction of ATP- binding residues. 9(Suppl 1):1–8, 2011.

[11] Muyuan Chen, Philip R. Baldwin, Steven J. Ludtke, and Matthew L. Baker. De Novo modeling in cryo-EM density maps with Pathwalking. *Journal of Structural Biology*, 2016.

[12] Gydo C.P.van Zundert and Alexandre M.J.J. Bonvin. Fast and sensitive rigid-body fitting into cryo-EM density maps with PowerFit. *AIMS Biophysics*, 2015.

[13] Yifeng Cui, Qiwen Dong, Daocheng Hong, and Xikun Wang. Predicting protein-ligand binding residues with deep convolutional neural networks. *BMC Bioinformatics*, 20(1):1–12, 2019.

[14] F. DiMaio and W. Chiu. Tools for Model Building and Optimization into Near-Atomic Resolution Electron Cryo-Microscopy Density Maps. In *Methods in Enzymology*. 2016.

[15] Yangchao Dong, Yue Liu, Wen Jiang, Thomas J. Smith, Zhikai Xu, and Michael G. Rossmann. Antibody-induced uncoating of human rhinovirus B14. *Proceedings of the National Academy of Sciences*, 2017.

[16] Jan Drenth. *Principles of Protein Separation*. Springer-Verlag, New York, 1999.

[17] Oranit Dror, Keren Lasker, Ruth Nussinov, and Haim Wolfson. EMatch: An efficient method for aligning atomic resolution subunits into intermediate-resolution cryo-EM maps of large macromolecular assemblies. In *Acta Crystallographica Section D: Biological Crystallography*, 2006.

[18] Jacques Dubochet, Joachim Frank, and Richard Henderson. Cryo-EM in drug discovery: achievements, limitations and prospects. *Nat. Rev. Drug Disc.*, 2018.

[19] Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 6 2006.

[20] J. D. Fischer, C. E. Mayer, and J. Söding. Prediction of protein functional residues from sequence by probability density estimation. *Bioinformatics*, 24(5):613–620, 2008.

[21] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, FranÃ§ois Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. In *Advances in Computer Vision and Pattern Recognition*. 2017.

[22] Yonatan Geifman, Guy Uziel, and Ran El-Yaniv. Bias-Reduced Uncertainty Estimation for Deep Neural Classifiers. 2018.

[23] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[24] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks. 2017.

[25] Judy Hoffman, Eric Tzeng, Trevor Darrell, and Kate Saenko. Simultaneous deep transfer across domains and tasks. *Advances in Computer Vision and Pattern Recognition*, (9783319583464):173–187, 2017.

[26] Tim J P Hubbard, Bart Ailey, Steven E. Brenner, Alexey G. Murzin, and Cyrus Chothia. SCOP: A structural classification of proteins database. *Nucleic Acids Research*, 27(1):254–256, 1999.

[27] Wen Jiang, Matthew L. Baker, Steven J. Ludtke, and Wah Chiu. Bridging the information gap: Computational tools for intermediate resolution structure interpretation. *J. Mol. Biol.*, 308(5):1033–1044, 2001.

[28] Larsen. Autoencoding beyond pixels using a learned similarity metric. 2016.

[29] Rongjian Li, Dong Si, Tao Zeng, Shuiwang Ji, and Jing He. 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)*, 2016.

[30] Gonzalo Lopez, Paolo Maietta, Jose Manuel Rodriguez, Alfonso Valencia, and Michael L. Tress. Firestar - Advances in the prediction of functionally important residues. *Nucleic Acids Research*, 39(SUPPL. 2):235–241, 2011.

[31] B. Ma, T. Elkayam, H. Wolfson, and R. Nussinov. Protein-protein interactions: Structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proceedings of the National Academy of Sciences*, 2003.

[32] Lingyu Ma, Marco Reisert, and Hans Burkhardt. RENNSH: A novel α-helix identification approach for intermediate resolution electron density maps. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9(1):228–239, 2012.

[33] Alan Merk, Alberto Bartesaghi, Soojay Banerjee, Veronica Falconieri, Prashant Rao, Mindy I. Davis, Rajan Pragani, Matthew B. Boxer, Lesley A. Earl, Jacqueline L.S. Milne, and Sriram Subramaniam. Breaking Cryo-EM Resolution Barriers to Facilitate Drug Discovery. *Cell*, 2016.

[34] Philipp Mostosi, Hermann Schindelin, Philip Kollmannsberger, and Andrea Thorn. Automated interpretation of Cryo-EM density maps with convolutional neural networks. pages 1–11, 2019.

[35] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Learning and Transferring Mid-Level Image Representations using Convolutional Neural Networks. Technical report.

[36] C A Orengo, A D Michie, S Jones, D T Jones, M B Swindells, and J M Thornton. CATH â a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1109, 1997.

[37] Jodi R Parrish, Keith D Gulyas, and Russell L Finley. Yeast two-hybrid contributions to interactome mapping. *Current Opinion in Biotechnology*, 17(4):387–393, 2006.

[38] Quiñonero-Candela, Joaquin Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence. *Dataset Shift in Machine Learning*, volume 1291. 2010.

[39] A. Rausell, D. Juan, F. Pazos, and A. Valencia. Protein interactions and ligand binding: From protein subfamilies to functional specificity. *Proceedings of the National Academy of Sciences*, 107(5):1995–2000, 2010.

[40] Roland Riek, Jocelyne Fiaux, Eric B Bertelsen, Arthur L Horwich, and Kurt Wüthrich. Solution NMR Techniques for Large Molecular and Supramolecular Structures. *Journal of the American Chemical Society*, 124(41):12144–12153, 10 2002.

[41] M Rozanov and H J Wolfson. AAnchor: CNN guided detection of anchor amino acids in high resolution cryo-EM density maps. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 88–91, 2018.

[42] Mirabela Rusu and Willy Wriggers. Evolutionary bidirectional expansion for the tracing of alpha helices in cryo-electron microscopy reconstructions. *Journal of Structural Biology*, 2012.

[43] Dong Si and Jing He. Modeling Beta-traces for Beta-barrels from Cryo-EM Density Maps. (Figure 1):23–25.

[44] Dong Si, Shuiwang Ji, Kamal Al Nasr, and Jing He. A machine learning approach for the identification of protein secondary structure elements from electron cryo-microscopy density maps. In *Biopolymers*, 2012.

[45] Dong Si, Shuiwang Ji, Kamal Al Nasr, and Jing He. A machine learning approach for the identification of protein secondary structure elements from electron cryo-microscopy density maps. In *Biopolymers*, 2012.

[46] Genki Terashi and Daisuke Kihara. De novo main-chain modeling with MAINMAST in 2015/2016 EM Model Challenge, 2018.

[47] Elina Tjioe, Keren Lasker, Ben Webb, Haim J. Wolfson, and Andrej Sali. MultiFit: A web server for fitting multiple protein structures into their electron microscopy density map. *Nucleic Acids Research*, 2011.

[48] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017.

[49] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep Domain Confusion: Maximizing for Domain Invariance. 2014.

[50] S. Velankar, Y. Alhroub, C. Best, S. Caboche, M. J. Conroy, J. M. Dana, M. A. Fernandez Montecelo, G. Van Ginkel, A. Golovin, S. P. Gore, A. Gutmanas, P. Haslam, P. M.S. Hendrickx, E. Heuson, M. Hirshberg, M. John, I. Lagerstedt, S. Mir, L. E. Newman, T. J. Oldfield, A. Patwardhan, L. Rinaldi, G. Sahni, E. Sanz-García, S. Sen, R. Slowley, A. Suarez-Uruena, G. J. Swaminathan, M. F. Symmons, W. F. Vranken, M. Wainwright, and G. J. Kleywegt. PDBe: Protein Data Bank in Europe. *Nucleic Acids Research*, 40(D1):402–410, 2012.

[51] Catherine Vénien-Bryan, Zhuolun Li, Laurent Vuillard, and Jean Albert Boutin. Cryo-electron microscopy and X-ray crystallography: Complementary approaches to structural biology and drug discovery, 2017.

[52] Hong Wei Wang and Jia Wei Wang. How cryo-electron microscopy and X-ray crystallography complement each other, 2017.

[53] Jiajun Wu, Chengkai Zhang, Tianfan Xue, William T Freeman, and Joshua B Tenenbaum. Learning a Probabilistic Latent Space of Object Shapes via 3D Generative-Adversarial Modeling. Technical report.

[54] J. S. Denker D. Henderson R. E. Howard W. Hubbard Y. LeCun, B. Boser and L. D. Jackel. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, 1(4):541–551, 1989.

[55] Zeyun Yu and Chandrajit Bajaj. Computational approaches for automatic structural analysis of large biomolecular complexes. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2008.

[56] Giuseppe Zanotti. Cryo-EM and X-Ray Crystallography: Complementary or Alternative Techniques? *NanoWorld Journal*, 2016.

[57] Lingxiao Zeng, Wei Ding, and Quan Hao. Using cryo-electron microscopy maps for X-ray structure determination. *IUCrJ*, 2018.

[58] Niyun Zhou, Hongwei Wang, and Jiawei Wang. EMBuilder: A Template Matching-based Automatic Model-building Program for High-resolution Cryo-Electron Microscopy Maps. *Scientific Reports*, 2017.

[59] Niyun Zhou, Hongwei Wang, and Jiawei Wang. EMBuilder: A Template Matching-based Automatic Model-building Program for High-resolution Cryo-Electron Microscopy Maps. *Scientific Reports*, 2017.