uc3m | Universidad Carlos III de Madrid

Master Degree in Information Health Engineering
2020-2021

Master Thesis

"Automatic determination of the handedness of electron density maps of macromolecules solved by cryoEM"

Jorge García Condado

Carlos Óscar Sánchez Sorzano
Maria Arrate Muñoz Barrutia
Leganés, 2021

Automatic determination of the handedness of electron density maps of macromolecules solved by cryoEM

Abstract

HaPi (Handedness Pipeline) is the first method to automatically determine the hand of electron density maps of macromolecules solved by Cryo-Electron Microscopy. HaPi is built by training two 3D CNN models. The first determines α-helices in a map and the second determines whether the α-helix is left-handed or right-handed. A consensus strategy determines the overall map hand. The pipeline is trained on simulated and experimental data. HaPi is able to correctly identify the hand in 89% of new simulated maps. HaPi correctly identified the hand of all 10 randomly selected experimental maps whose hand was confirmed by visual inspection.

1 Introduction

During the last decades Single-Particle Analysis with cryo-electron microscopy (CryoEM) has revolutionised the field of Structural Biology helping us see macromolecules at atomic resolution [1]. Single-Particle Analysis applies a series of image processing methods to combine many projections of macromolecules to reconstruct their 3-dimensional (3D) structure, see Fig. 1. Particles are first identified in micro-graphs. These are used to reconstruct the 3D map of the Coulomb potential by assigning orientations to each particle. With these orientations an initial volume is reconstructed that is then refined. The determined experimental maps are then used to fit an atomic model of the macromolecule, see Fig. 2.

Single-Particle Analysis is an ill-posed problem because the reconstruction is not well determined. If all particle image orientations are mirrored a map is reconstructed that is equally consistent with the measured data, but nonsuperimposable over the map previously reconstructed, see Fig. 3. However, only one of the maps is the correct reconstruction.

The polypeptide chain of proteins is made up of L-amino acids. D-amino acids are never used. This confers a unique chirality to proteins and their secondary structure [2] resulting in the property of protein structures that is commonly referred to as handedness. As proteins have a specific handedness only one of the two possible reconstructed maps is the correct reconstruction of the structure. Hence, it is important to reconstruct with the correct handedness for a correct fitting of the atomic models.

Currently, a trained biologist is required to look at the α-helices rotation to asses the handedness of the map. If incorrect, the reconstructed map is mirrored. The direction of rotation is easily determined at very high resolutions of 1Å, but can be difficult at lower resolutions even for experts, see Fig. 4. As the resolution decreases the α-helix slowly transitions from a helix to a cylinder, which no longer has a hand, as seen in Fig. 5. Hence, we propose HaPi (Handedness Pipeline) to automatically determine the hand of reconstructed maps using deep learning for resolutions of up to 5Å.

To the best of our knowledge, there are no algorithms to automatically detect the hand of reconstructed CryoEM maps. The proposed model is based on first identifying Secondary Structure Elements (SSE) of interest in the volume and then, using these to detect the hand. There are several previous approaches to automatically determine SSE in electron density maps based on non-machine learning methods [3]–[6], machine learning methods [7], [8] and more recently deep learning techniques [9]–[12]. As the latter have shown better performance, in this work 3D CNNs are used to determine SSE of interest and detect the hand of a map from small boxes extracted from the map at the location of the SSE.

2 Datasets

Three different datasets were used, two datasets of non-redundant atomic models selected using PDB-Select [13] and downloaded from the protein databank (RCSB PDB [14]) and one dataset of experimental maps from the electron microscopy databank (EMDB [15]). The first dataset (Dataset I) of atomic models was used to simulate data to train, validate and test models on small boxes. The second dataset (Dataset II) of atomic models was used to validate and test models on whole simulated volumes. The EMDB dataset (Dataset III) consisted on experimentally determined structures.

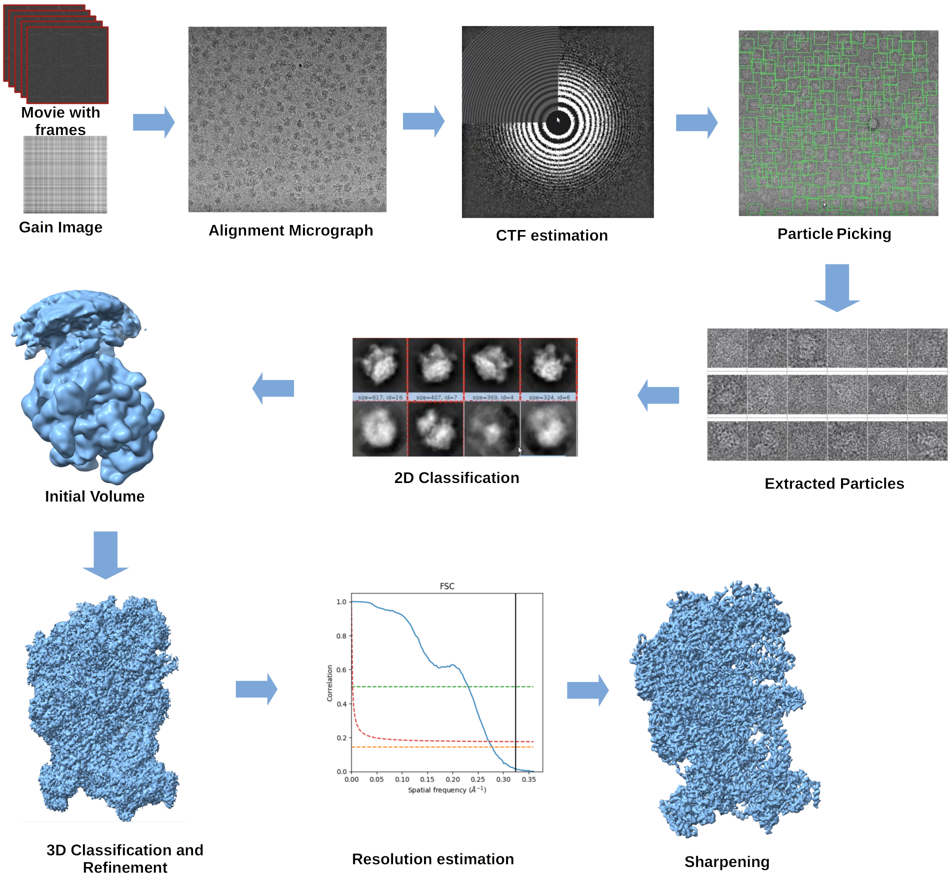


Fig. 1. Workflow of Single Particle Analysis. Cryo-Electron Microscopes capture 2D projections of the macromolecules embedded in amorphous ice. Movie frames taken by the microscope are aligned to obtain micrographs where the macromolecules can be seen. The contrast transfer function (CTF), which describes the aberrations of the microscope, is then estimated for each micrograph. From each micrograph particles are picked but these are very noisy 2D projections. Hence, particles which capture the same projection are combined to reduce noise by carrying out a 2D classification. These projections are used to reconstruct an initial volume. By carrying out refinement processes a cleaner experimental map is obtained whose resolution is then estimated. Finally, the map can be sharpened.

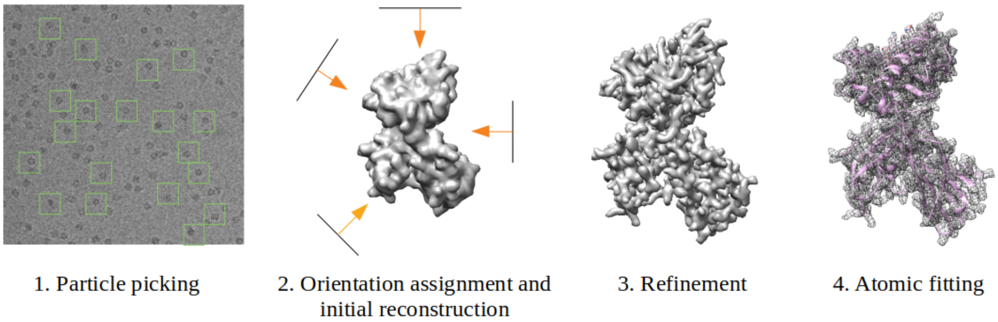


Fig. 2. Single Particle Analysis simplified pipeline from micrогaphs acquired containing macromolecules of interest to final atomic fitting of the structure. Cryo-Elctron Microscopes are used to take images of thousands of projections of a macro-molecular structure. These individual noisy images are then used to reconstruct the structure. Finally an atomic model of the macro-molecule is fitted to the reconstructed structure.

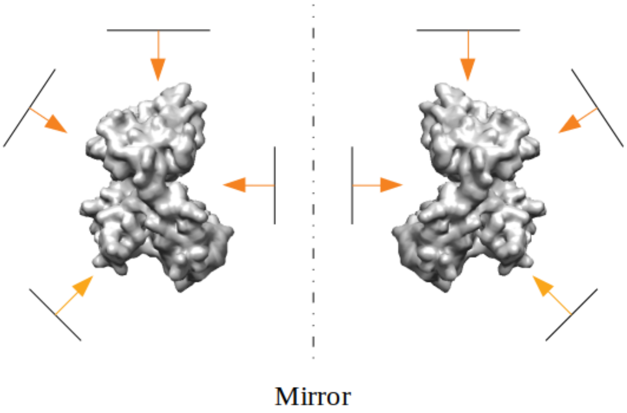


Fig. 3. Reconstruction of a structure from the same set of images but with mirror orientations assigned to each image which produces a mirrored version of the structure.

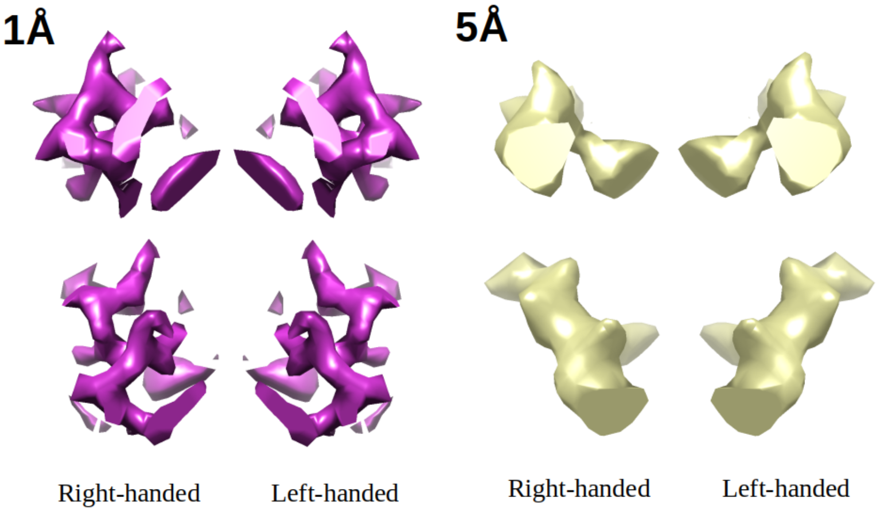


Fig. 4. A portion of the same α-helix at 1Å and 5Å with its true structure (right-handed) and mirrored version (left-handed) from different viewing angles.

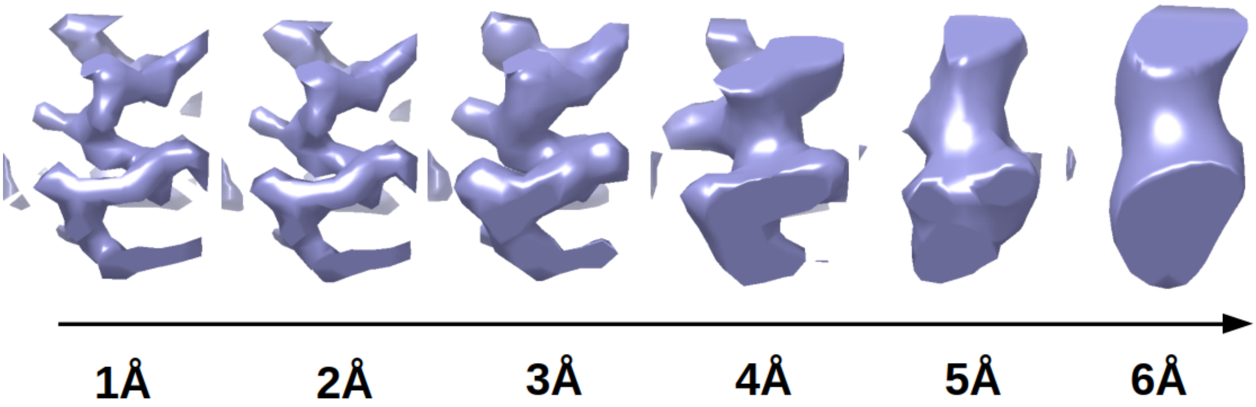


Fig. 5. Same α-helix with same viewing angle at different resolutions that shows transition from helical to cylindrical structure.

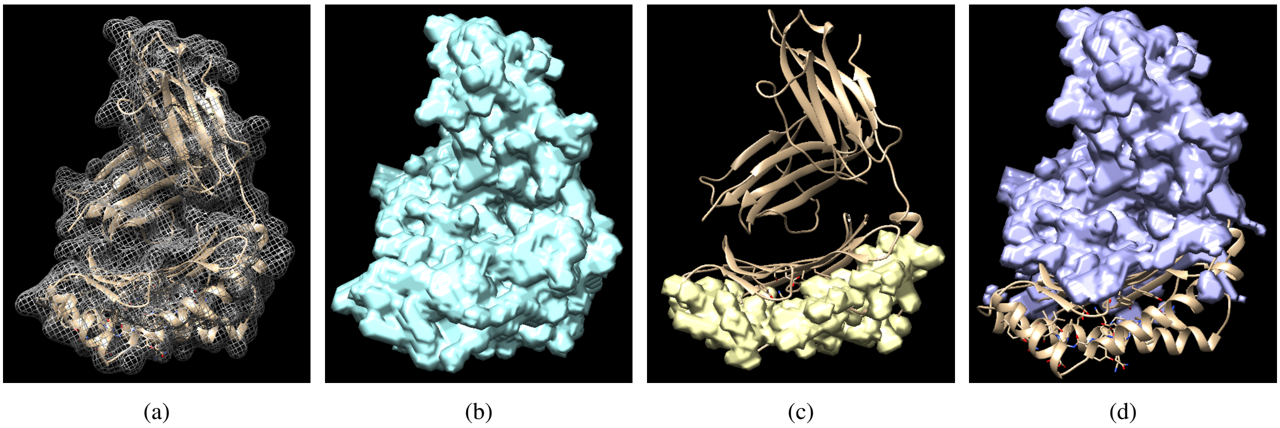


Fig. 6. Output of the simulation of the atomic structure 1AGC with α-helices as the Secondary Structure Element (SSE) of interest using Xmipp [16]: (a) V_f: Columb potential map of the structure; (b) V_{mask}: mask that covers the overall structure; (c) V_{maskSSE}: mask containing the SSE of interest; (d) $V_{\text{maskNoSSE}}$: mask containing the remaining part of the structure that does not have SSE of interest.

Dataset I was used to simulate electron density maps of macromolecules from atomic models. Boxes were then extracted from the simulated volume. Dataset I consisted of 12,343 atomic models. Boxes either contained a specific SSE such as an α-helix or a β-sheet, or a random part of the structure that did not contain the specific SSE. The dataset was simulated at resolutions of 1Å, 3Å, 5Å and 6Å focusing separately on α-helices and β-sheets at each resolution. 203,889 boxes were generated for the α-helices case, evenly split between containing α-helices and not containing α-helices. 25,776 boxes where generated for the β-sheets case, evenly split as well. The dataset was split into training, validation and test splits of 70%, 15% and 15% respectively.

Dataset II was used to simulate whole electron density maps of macro-molecules from atomic models. The dataset consists of 3,119 atomic models simulated at 5Å. These were split into validation and test sets of 30% and 70%, respectively.

Dataset III consisted of all deposited experimental maps in EMDB with resolution below or at 5Å. In total 8,061 maps were downloaded. 19,971 boxes, evenly split between containing α-helices and not containing α-helices, were extracted from 262 of these structures. The dataset was split into training, validation and test splits of 70%, 15% and 15% respectively to train an α-helix determination model. 78 structures were then used to test the precision of the model on whole unseen experimental volumes. Also 10 experimental maps with high resolution of ≤ 3Å were randomly selected to test the accuracy of the pipeline on experimental data manually labelled.

3 Methods

The HaPi package is freely available to use and documented via github. All the code for the methods described can be found in the same link.

3.1 Simulation

Atomic models are used to simulate Coulomb potential maps to create a labelled data set for training. For this the Xmipp library [16] was used. Three sets of volumes were generated: a volume containing the Columb potential at each voxel (V_f), a mask of the overall structure (V_{mask}), and a mask of the location of the SSE of interest (V_{maskSSE}). A mask containing voxels with no SSE of interest is determined as $V_{\text{maskNoSSE}} = V_{\text{mask}} \cap \overline{V_{\text{maskSSE}}}$. Volumes have a voxel size of 1Å. An example of the four volumes generated for each structure can be seen in Fig. 6.

The network is trained from boxes extracted from V_f as seen in Fig. 7. V_{maskSSE} is first eroded to separate regions in the mask. Each region contains an α-helix. The centroid of each region is determined. The number of centroids found is n. At each of these centroid locations a box of dimen-

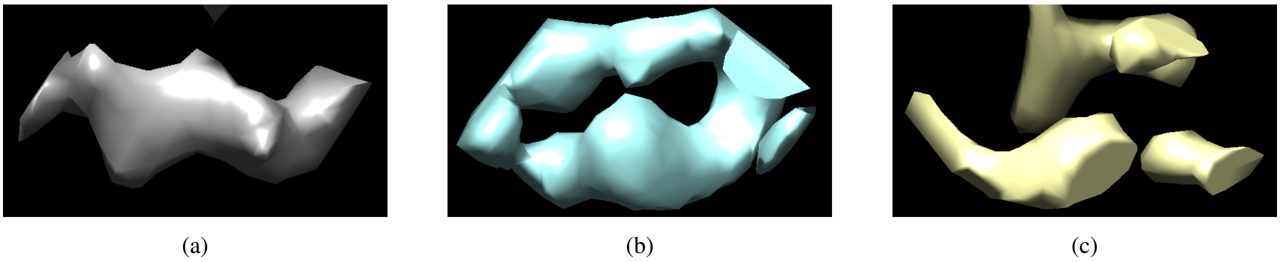


Fig. 8. The three types of Secondary Structure Elements that are found within simulated structures simulated at 5Å: (a) α-helix; (b) β-sheet; (c) No clear Secondary Structure Elements.

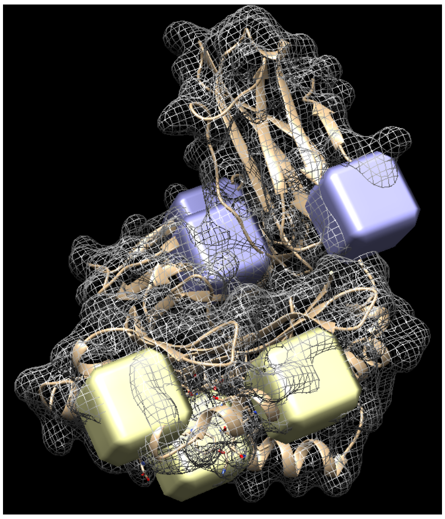


Fig. 7. Extracted boxes from V_f, the Columb potential map of the structure. In yellow boxes containing Secondary Structure Elements (SSE) of interest, in this case α-helices. In purple boxes containing no SSE of interest.

sions 11 pixels × 11 pixels × 11 pixels is extracted from V_f. $V_{\mathrm{maskNoSSE}}$ is first eroded so as to avoid selecting points that could contain parts of α-helices. Then, n points are selected at random from the eroded $V_{\mathrm{maskNoSSE}}$. Boxes are also extracted from V_f at those points.

There are three types of structures that can be found in a box: an α-helix, a β-sheet, or another part of the structure with no clearly defined SSE. The three types are visualised in Fig. 8.

The fitted atomic model of experimental structures is used to extract boxes containing α-helices from experimental maps. Xmipp is used to obtain a mask of the location of the alpha helices (V_{maskSSE}) and a mask of the overall structure (V_{mask}) from the atomic model. These masks are then aligned with the experimental map by applying the same transformation required to align the experimental map to a simulated map of the structure from the atomic model. The experimental map is resized to 1Å and filtered to 5Å . The same process as in simulated data is applied to obtain $V_{\mathrm{maskNoSSE}}$ and extract boxes containing α-helices and no α-helices from the experimental map.

3.2 Model

The same 3D CNN model is used for the SSE determination task and the hand determination task. A 3D CNN is an extension of 2D CNNs that deals with volumes instead of images. The full architecture of the 3DCNN designed can be seen in Fig. 9. All 3D convolutional layers and the first connected layer are followed by ReLu activation functions. The last fully connected layer is followed by a sigmoid function.

Boxes are preprocessed before being input into the model. First, any value below 0 is thresholded to 0. Then, each box is normalised to the range 0 to 1 by:

$$b_{norm} = \frac{b - \min(b)}{\max(b) - \min(b)}$$

3.3 Training

The training strategy is similar for both tasks. Dataset I boxes are assigned a label of 0 or 1. For the task of SSE determination a label of 1 is given if it contains the SSE of interest and 0 if it does not. For the hand determination task a label of 1 is given if the SSE is left-handed (this is the hand that is seldom found in nature) and a label of 0 if it is right-handed. A binary cross-entropy loss function is used for training with an Adam optimiser, a learning rate of 0.001 and batches of size 2,048. The model is trained

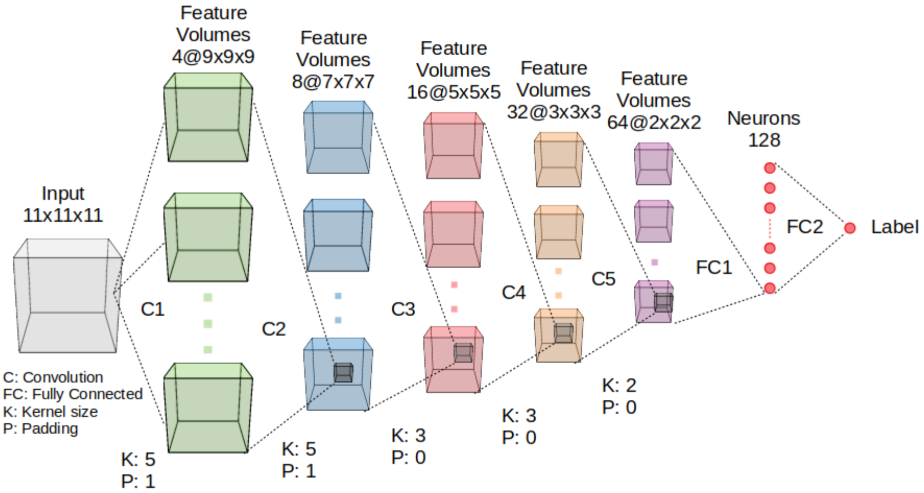


Fig. 9. 3D CNN model diagram used for both Secondary Structure Element (SSE) determination and hand determination. The input to the network is a box containing the electron density at each voxel normalised to the range 0 to 1. The output of the network is a value between 0 and 1. In the SSE determination task, a label of 1 represents that the SSE of interest is found within the box and 0 means it is not. In the hand determination task, a label of 1 means the box is left-handed and 0 means it is right-handed.

for 50 epochs. An early stopping strategy is adopted where the model saved at the epoch where the validation set loss stagnates is used as a the final model.

Weight initialisation differs for the SSE and hand models. For SSE determination, the model is initialised with the default PyTorch settings for weight initialisation. For the hand model, the 1Å model is initialised with the default PyTorch settings. Then, a transfer learning approach is used on the 5Å and 6Å model for hand determination. The weights of the model trained on 1Å data are used as the initial weights for training the model at 5Å and 6Å.

3.4 Volumes

The HaPi pipeline is used to process whole volumes rather than individual boxes, see Fig. 10. HaPi takes as input a Coulomb Potential map (V_f) and a mask of the non-background voxels (V_{mask}). Models used in the pipeline are those trained at 5Å on experimental data for SSE determination and on simulated data for hand determination.

Experimental maps have to be preprocessed. First they are resampled to 1Å voxel size to match the training data voxel size. Each map has a resolution below 5Å, but models are trained at 5Å so they are low-pass filtered to 5Å. This reduces any noise and homogenises local resolutions. V_{mask} is obtained by thresholding at the specified contour level for visualisation set by the researchers that deposited the map.

The model trained to determine α-helices is wrapped into what we have called AlphaVolNet. AlphaVolNet determines the location of α-helices in the whole volume. AlphVolNet takes as input V_f and V_{mask} and outputs V_α, which is a mask containing the location of the α-helices found. At each voxel location where V_{mask} is true a box is extracted at that location from V_f and passed through the trained α-SSE model. Then if the label is above a threshold t_α, at that location V_α is set to true. Dataset II is used to determine t_α by choosing t_α that maximises accuracy on the validation set.

HandNet is used to determine the hand of a map from V_α and V_f. At each true voxel of V_α, a box is extracted at that location of V_f and passed through the trained hand model. Then, a hand value is given to the map by consensus of all the labels of the passed boxes. The consensus value is the average of all hand predictions of each α-box.

4 Results

4.1 Simulated boxes

Dataset I was simulated focusing on α-helices at 3 different resolutions: 1Å, 5Å, and 6Å. Models were then trained to determine the hand of the simulated data without transfer learning strategies. The models 5Å_TL and 6Å_TL were trained on 5Å and 6Å data respectively and starting from

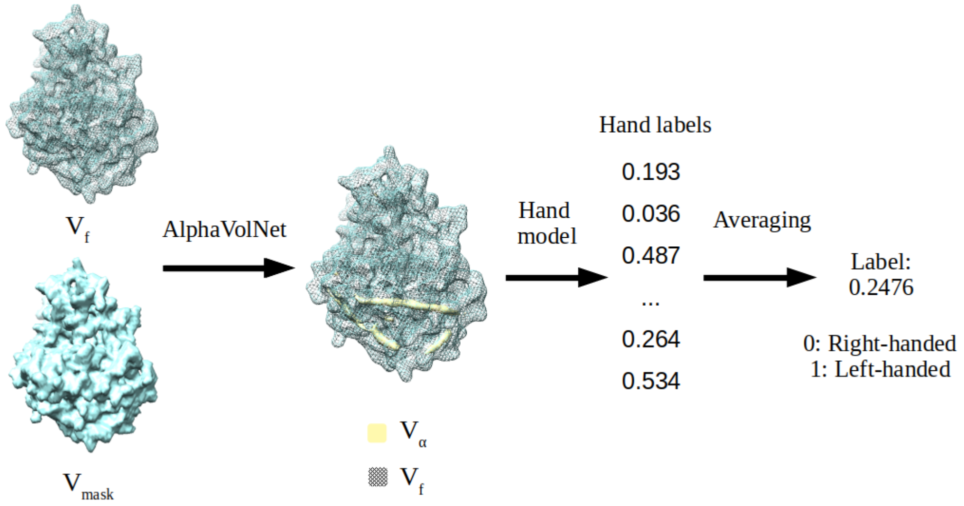


Fig. 10. Diagram of the HaPi pipeline with the inputs and outputs at the different stages and the models used to generate each. The input to HaPi is V_f: a Columb potential map and V_{mask}: a mask of the overall structure. These are first passed through AlphaVolNet which outputs a mask V_α of the proposed location of α-helices. Boxes are extracted from V_f at the location of voxels in V_α that contain an α-helix and passed through the hand model to be given a hand value. Values are averaged to assign a value to the overall map. The pipeline has 208,163 parameters.

the weights of the 1Å model. The results can be found in Table 1. The best model is 5Å_TL as it achieves >90% accuracy on both 1Å and 5Å datasets.

Table 1. Hand accuracy on the models evaluated with resolutions of 1Å, 5Å, and 6Å.

	Dataset		
Models	1Å	5Å	6Å
1Å	0.990	0.605	0.503
5Å	0.503	0.500	0.505
5Å_TL*	**0.949**	**0.955**	**0.520**
6Å	0.503	0.500	0.505
6Å_TL*	0.532	0.520	0.522

*Models trained with a transfer learning strategy using the weights of 1Å model for initialisation instead of random weights.

Dataset I was also used to simulate β-sheet boxes. Models were trained for both α-helices and β-sheets at 1Å, 3Å and 5Å. The results on the accuracy of determining each individual SSE at a time can be found in Table 2. The results on the accuracy of determining the hand from each type of SSE can be found in Table 3.

4.2 Simulated volumes

Dataset II's validation set was used to study the effect of varying the α-threshold (t_α) on the networks performance. In whole volumes, precision is used rather than accuracy as a metric when identifying α-helices, since volumes contain mostly non-alpha-helix voxels. Consequently, we are mostly interested in the ratio of true positive to false positives so as to pass on to the hand model as few non-α-helices as possible. Accuracy is still used as a performance metric fo the hand determination task.

Table 2. Determination accuracy for Alpha and Beta Secondary Structure Elements at different resolutions (1Å, 3Å and 5Å).

	1Å	3Å	5Å
α-helix	0.985	0.985	0.981
β-sheet	0.854	0.857	0.847

Table 3. Hand determination accuracy trained on Alpha and Beta Secondary Structures at different resolutions (1Å, 3Å and 5Å).

	1Å	3Å	5Å
α-helix	0.990	0.991	0.955
β-sheet	0.647	0.641	0.610

The effect of t_α on the precision of determined α-helices and the accuracy of the hand prediction can be seen in Fig. 11. Increasing t_α increases precision and slightly increases accuracy. The effect of varying t_α on the average prediction assigned to each map according to its true label

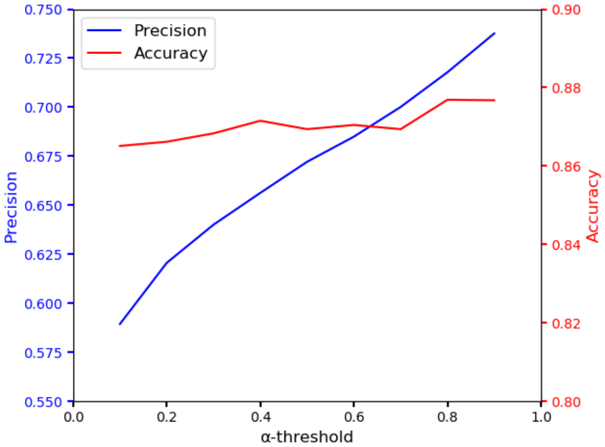


Fig. 11. Effect of varying α-threshold (t_α) on the precision and the accuracy of the pipeline. The α-threshold controls how stringent the model is in accepting a box as containing an α-helix. Precision measures how many boxes labelled as containing an α-helix actually contain an α-helix. The accuracy measured is the percentage of simulated maps whose hand was correctly identified.

can be seen in Fig. 12. Increasing t_α increases the average value given to true positives and decreases the average value given to true negatives. AlphaVolNet captures the centroids of the α-helices as seen in Fig. 13. Fig. 14 shows the histogram of hand predictions for each volume and the confusion matrix for different t_α. Increasing t_α decreases the bias as seen in the confusion matrix and causes the left and right peaks to get closer to 0 and 1 respectively.

The performance of HaPi on volumes is then compared to the individual models performance on boxes in Table 4.

Table 4. Performance of models on boxes dataset vs. volumes dataset with t_α = 0.7

	SSE	Hand
Boxes	0.979	0.955
Volumes	0.692	0.892

The histogram containing the individual prediction labels of each map in the test set and the confusion matrix can be found in Fig. 15. It should be noted that out of the 236 incorrectly assigned hand labels, 133 have a precision of less than 0.2. The average prediction value for maps incorrectly labelled is 0.498 ± 0.080.

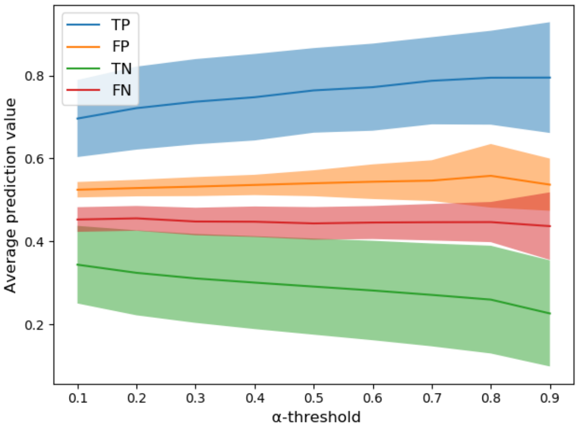


Fig. 12. Average hand prediction values for True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN) with standard deviations. If the predicted hand is above 0.5, it is assigned a label of 1 (left-handed), if not it is assigned a label of 0 (right-handed).

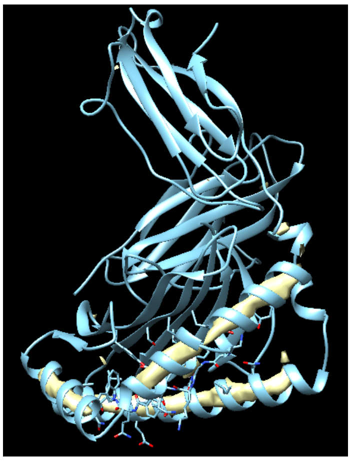
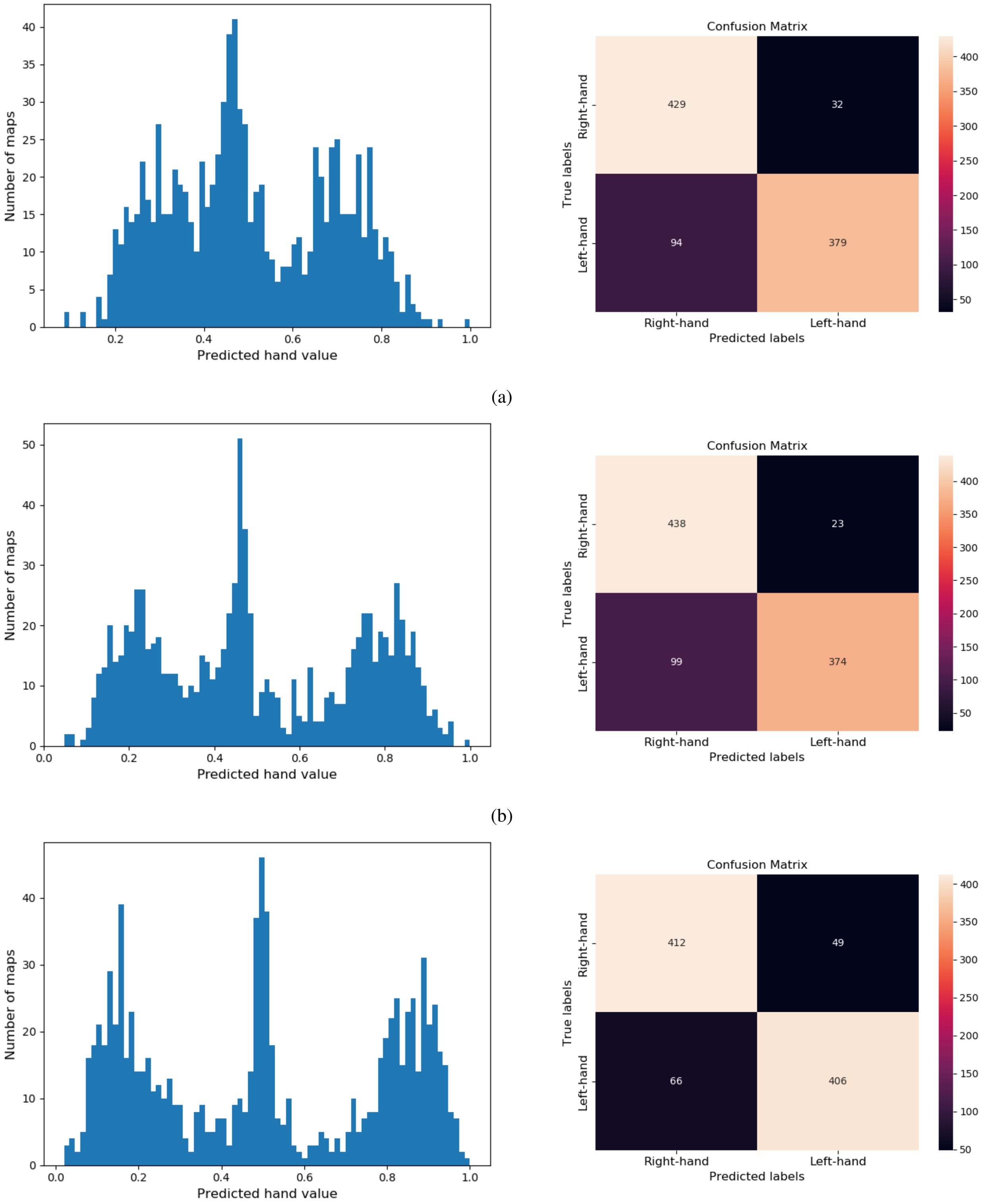


Fig. 13. Output of AlphaVolNet on 1AGC simulated structure with t_α = 0.85 in yellow and the atomic model of 1AGC. AlphaVolNet is a model that predicts the location of α-helices and t_α is the stringency of the model to accept voxels as containing α-helices. AlphaVolNet is clearly identifying the location of the center of α-helices.

4.3 Experimental data

AlphaVolNet trained on simulated data in some cases does not correctly identify α-helices in experimental maps. To illustrate this the experimental map of 7RH5 is passed through AlphaVolNet trained on simulated data (Fig. 16a).



(a)

(b)

(c)

Fig. 14. Effect of changing the threshold t_α on the individual predicted hand values for each map in the validation set of Dataset II (left) and the confusion matrix (right): (a) $t_\alpha = 0.1$, (b) $t_\alpha = 0.5$ and (c) $t_\alpha = 0.9$. t_α controls the stringency of the model to accept voxels as containing an α-helix. Simulated maps were left as is to be right-handed or mirrored to be left-handed. Each map was then passed through the HaPi pipeline to receive a hand value. Maps with values above 0.5 were given a left-hand label and maps with values below 0.5 were given a right-hand label.

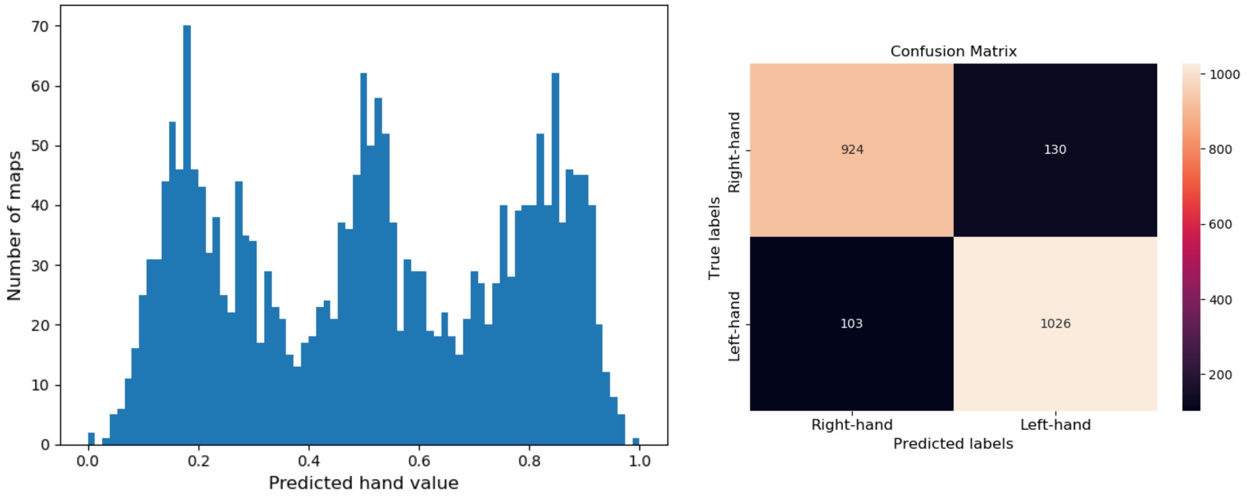


Fig. 15. Individual predicted hand values for each map in Dataset II test set (left) and confusion matrix (right) for $t_\alpha = 0.7$. t_α controls the stringency of the model to accept voxels as containing an α-helix. Simulated maps were left as is to be right-handed or mirrored to be left-handed. Each map was then passed through the HaPi pipeline to receive a hand value. Maps with values above 0.5 were given a left-hand label and maps with values below 0.5 were given a right-hand label

However, if the atomic structure of 7RH5 is used to simulate a map and this map is passed through AlphaVolNet the α-helices centres are clearly identified (Fig. 16b). This atomic structure was not used for training. Hence, it is a problem of generalisation to unseen experimental data rather than to new simulated data.

The SSE determination model was retrained on experimental data to correctly identify α-helices. When passing the experimental map of 7RH5 through AlphaVolNet trained on experimental data α-helices are correctly identified (Fig. 16c), although the precision of the algorithm is lower. Table 5 below compares the performance on experimental data of the model trained on simulated data to the model trained on experimental.

Table 5. Precision of α-determination models trained on simulated and experimental data and tested on experimental data with a threshold of $t_\alpha = 0.7$

	Dataset	
Model	Boxes	Volumes
Simulated	0.806	0.275
Experimental	0.936	0.467

The precision of the experimental model on experimental volumes is still lower than the precision of the simulated model on simulated models. However 7 out of the 78 tested models had zero precision because the models contained only α-helices with less than 7 residues which are too small to detect. If these structures are not included the precision increases to 0.501.

10 experimental maps with resolution lower than 3Å were randomly chosen to asses the accuracy of the hand determination capabilities of HaPi. All experimental structures were visually inspected and were clearly right-handed. The structures were passed through HaPi to obtain values below 0.5, indicating all were correctly identified as right-handed, see Table 6. All structures were then mirrored and passed again through HaPi. Values above 0.5 were obtained for all structures showing HaPi correctly identified them as left-handed. Fig. 17 shows a kernel density estimator fitted to each group of values, clearly separating them into two distinct groups with a threshold of 0.5.

The time taken on average to run an experimental map through HaPi including preprocessing is 30s on an NVIDIA Tesla T4 16GB.

5 Discussion

5.1 Training

Determination of the map hand from any part of the volume is difficult. Early experiments were unsuccessful when extracting boxes from random parts of the structure and

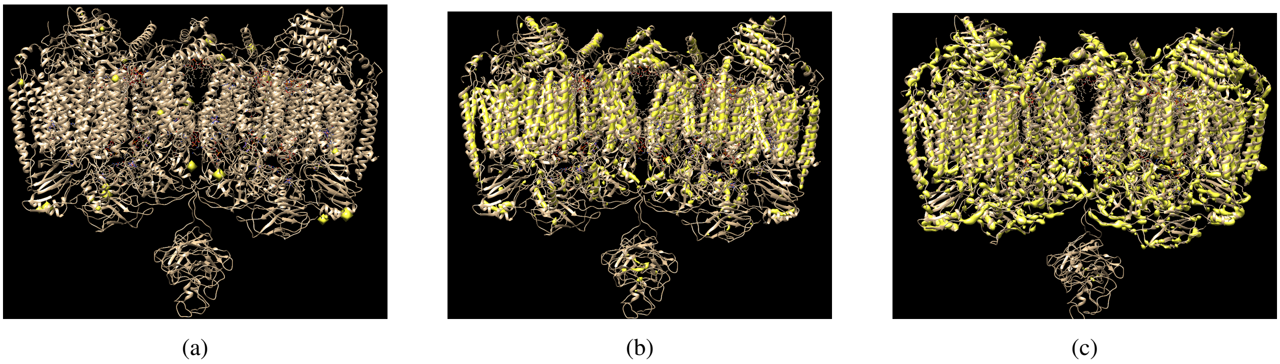


Fig. 16. α-helices determined by AlphaVolNet (a model that decides the location of α-helices within a map) when feeding an experimental map of 7RH5 to a model trained on simulated data (a) versus a simulated map of 7RH5 from its atomic model to a model trained on simulated data (b) versus an experimental map of 7RH5 to a model trained on experimental data (c) with the atomic model for reference.

Table 6. HaPi predicted hand values of 10 randomly selected experimental maps originally right-handed and left-handed when mirrored

EMD ID	Resolution (Å)	Original*	Mirrored*
12339	2.3	0.212	0.745
12343	2.8	0.247	0.734
12886	3.0	0.283	0.650
13008	2.7	0.148	0.788
20789	2.7	0.100	0.872
22245	2.7	0.258	0.721
23743	3.0	0.125	0.850
24455	3.0	0.227	0.738
30421	3.0	0.266	0.677
31135	2.7	0.132	0.814

*Prediction values are the probability that the structure is left-handed. Hence, values closer to 0 mean the map is right-handed and values closer to 1 mean the map is left-handed.

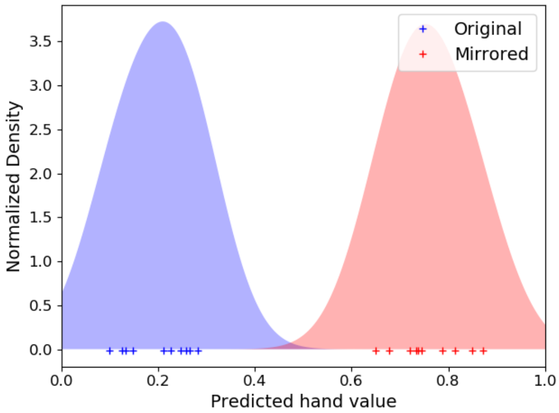


Fig. 17. Gaussian kernel density estimators of the hand value assigned by HaPi to right-handed experimental maps as is and when mirrored. Values closer to 1 mean the map is left-handed and closer to 0 mean the map is right-handed.

training on flipped and non-flipped versions of the boxes. This could be because the data is too heterogeneous and the model is unable to capture all the correlations. It could also be because a majority of the regions of a volume might have no actual information about the hand. Therefore, the algorithm is trying to look for information that is not there. Focusing on SSE that contain information about the hand was found to be a smarter approach to obtain better performance.

The only fine-tuned hyperparameter of the 3D CNN was the number of epochs to train for. Hyperparameter fine-tunning was not required as high accuracy was obtained due to the shear size of data used for training. When using α-helices, 142,722 boxes were used for training. The algorithm was forced to generalise for it to be able to correctly identify such a wide range of boxes.

Transfer learning strategies are required to correctly distinguish the hand at intermediate resolutions of 5Å. As previously discussed, determining the hand of an α-helix at 5Å is non-trivial even for trained experts. When starting training with random weights the model is unable to converge to a good optimum set of parameters because the optimisation landscape is vast and has many local minima that the model might converge to. By starting from the 1Å model, the optimisation starts closer in the landscape to a lower minima as the model has already learnt to distinguish features that are related to hand determination.

5.2 Secondary Structure Elements

α-helices are better for determining the hand at box level compared to β-sheets (Table 3). It is easy to distinguish the hand of an α-helix at high resolutions. It is difficult to determine the hand of a β-sheets as it requires looking at the side-chains and their orientation. Hence, the α-helix model was used to build the whole pipeline.

The difference in accuracy in SSE determination for α-helices and β-sheets as seen in Table 2 is notable. This could be because the dataset of β-sheets is an order of magnitude smaller than α-helices. It could also be because the box dimensions are too small to capture all the features of a β-sheet. Boxes of dimension 11Å by 11Å by 11Å were chosen. This allows for two full turns of the α-helix to be present no matter the orientation as the pitch of an α-helix is 5.4Å [17]. The distance between adjacent amino acids along a β-strand is approximately 3.5 Å, in contrast with a distance of 1.5 Å along an α-helix [18]. The sideways distance between adjacent C-α-atoms in hydrogen-bonded beta-strands is roughly 5 Å [18]. Hence for the chosen box dimensions barley two chains of the β-sheets are captured and of each of these chains only 3 residues are captured compared to 7 for α-helices. As seen in Fig. 8, the α-helix at this box size can be clearly identified while the β-sheet looks like two chains rather than a sheet.

The minimum resolution at which the hand of an α-helix can be determined is that of its pitch. α-helices have an average pitch of 5.4Å. At resolutions above 5.4Å the structure is no longer a helix but a cylinder. Blurred helices become cylinders as information between amino acids below and above a turn merge to form a continuous chain resembling a solid structure rather than a spring coil. If a cylinder is reflected, its mirror version is superimposable over the non-mirrored cylinder. Therefore, a cylinder has no hand. The inability of the network to identify the hand at 6Å is because the information of the hand is not available at such resolution.

Being able to determine the hand of an experimental map at or below 5Å is still useful. Structures with resolutions below 5Å already represent more than 50% of all depositions at the EMDB, see Fig. 18. However, at resolutions below 4Å, the hand can be easily identified by manually inspecting the map as seen in Fig. 5. Still structures between 4Å and 5Å of resolution represent 13.1% of all deposited structures and this is likely to increase as the resolution of cyroEM maps improves. Hence, this will be a useful tool to easily and automatically determine the hand at this resolution range.

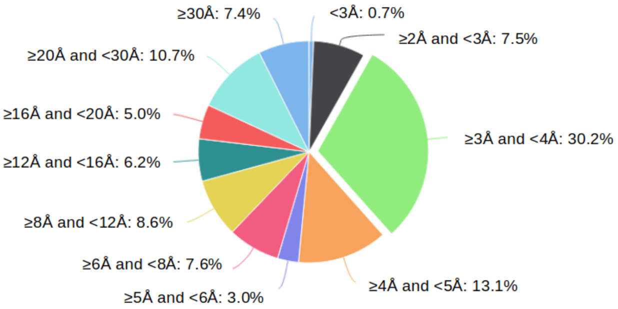


Fig. 18. The Electron Microscopy Data Bank (EMDB) entry resolution in shells distributions as of 2021. Statistics downloaded from EMDB site [15].

5.3 Volumes

In Fig. 15, three distinct peaks of predicted hand values are seen in the histogram. The left peak corresponds to right-handed maps that are correctly identified and are given a label close to 0. Conversely, the right peak corresponds to left-handed maps correctly identified and given a label close to 1. In the center there is a narrow peak that corresponds to those maps whose hand is not easily determined and are therefore given labels close to 0.5.

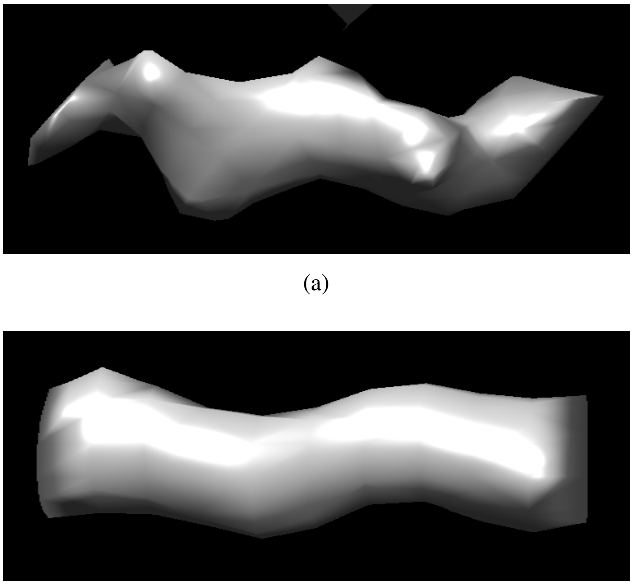
In Fig. 14, the effect of varying the threshold t_α on the predicted hand values is seen. t_α is a threshold at which a voxel is accepted to contain an α-helix. Increasing t_α makes the condition more stringent to accept a voxel as containing an α-helix. As shown in Fig. 11, increasing t_α improves the precision of the pipeline because the conditions to accept that a voxel contains an α-helix are more stringent, resulting in less false positives. Therefore, the left and right peaks in Fig. 14 move closer to 0 and 1, respectively as t_α increases. The hand model was not trained on non-α-boxes so, on average, non-α-boxes are given values around 0.5. When including false positives and averaging, the predicted hand value is pulled closer to 0.5. By reducing the number of false positives, the predicted value is less biased towards 0.5. Increasing t_α also reduces the bias as the set of false positives to false negatives becomes more balanced. However, increasing t_α has a minimal effect in improving the accuracy (Fig. 11). This is most likely because most of the structures that are mistaken are difficult to determine their hand or the α-helices were not properly identified.

From Fig. 12, the optimal t_α is deduced. Increasing t_α increases and reduces respectively the average predicted value of true positives and true negatives. However if t_α is too big the standard deviation increases because less boxes are available for consensus. Therefore, a threshold of $t_\alpha = 0.7$ was chosen because the true positives, false positives, true negatives and false negatives are clearly separated into four groups.

Fig. 12 contains valuable information about biases and the certainty of predictions. The false negative average value is further away from 0.5 than the false positive values, hinting that the model is biased towards assigning right-handed values. The pipeline gives values closer to 1 for left-handed structures than it gives right handed values closer to 0. Maps given values below 0.4 or above 0.6 are unlikely to be mislabelled but maps with values between 0.4 and 0.6 should be further tested.

There is a problem when dealing with experimental data for SSE determination but not for hand determination. The model works for previously unseen simulated data but not for experimental data. This suggests that the network has correctly learned to identify α-helices but has over-fitted the simulated data. It can clearly distinguish what is different between an experimental and simulated α-helix. Although visually experimental and simulate helices are similar more detail is found in simulated ones (Fig. 19), those details are what the network is focusing on. The network might be over parametrized allowing it to focus on small details in the box. Previous works [9]–[12] are all trained on simulated data and generalise to experimental data. However, they are trained on the more complex task of distinguishing several SSE at once, therefore forcing the model to focus on general features that distinguish each SSE. When trained on the more difficult task of hand determination, it generalises to properly identify the relevant features that contain information about the hand. A possible solution is to use a smaller network with less parameters to determine the SSE, so as to avoid it fixating on the small features that distinguish simulated and experimental data.

Previous works have had generalisation problems when moving from simulated data to experimental data acquired by cryoEM [19]. This could be because the simulated data does not resemble well enough the experimental data. For example α-helices are not structurally rigid [20] and the simulated data does not take this into account. Automatic SSE labels can be assigned to experimental data from fitted atomic models. Therefore, the SSE model was finally trained on experimental data. Training the hand model on experimental data would require manual labelling by an expert which is time consuming and often yields less datapoints and was not required by the high accuracy already achieved.



(a)

(b)

Fig. 19. Simulated α-helix (a) correctly identified as an α-helix by the trained model and experimental α-helix (b) incorrectly identified despite their visual similarity.

The precision of the experimental model on experimental data is not as good as the precision of the simulated model on simulated data. This is most likely because the experimental dataset is more heterogeneous. The experimental data is inherently noisy. The automatic assignment of labels could have also mislabelled some boxes as the alignment of the experimental map with the α-helix mask is not perfect. To improve precision post processing of the proposed α-helices could be considered.

The hand model trained on simulated data works on experimental data. HaPi was not only able to identify correctly that all maps were right-handed but also when the maps were mirrored it identified all as left-handed. This makes HaPi the first method to determine automatically the hand of experimental maps to our knowledge.

6 Conclusions and future work

HaPi has shown that it is possible to determine the hand of a map automatically without the necessity of inspection by a trained expert. This is even the case for intermediate resolutions were the hand is not clear from visual inspection. α-helices are the best SSE for identifying the hand over β-sheets. Deep learning techniques have been proven useful to complete such a task but using experimental data for training is a necessity for generalisation.

This was an initial exploration to determine the hand of CryoEM maps which to our knowledge has never been successfully attempted and will require future endeavours to improve the methods by:

- Testing on all downloaded EMDB maps (Dataset III) to identify any left-handed maps. This has been estimated to take 1 full week to run on an NVIDIA Tesla T4 16GB. Left-handed maps can be mirrored to obtain appropriate right-handed maps and carry out a refitting of the atomic model to obtain better fits.
- Making the method available to structural biology researchers in Scipion [21]. Scipion is a CryoEM image processing framework used by thousands of researchers world wide to process experimental data to reconstruct experimental maps.
- Training the model on a range of resolutions between 1Å and 5Å for better generalisation and to avoid having to filter maps to 5Å which could reduce information about the hand.
- Post-processing of the determined α-voxels by keeping those shaped like long cylinders to eliminate false positives and improve precision.

References

[1] S. Subramaniam, "The cryo-em revolution: Fueling the next phase," *IUCrJ*, vol. 6, no. Pt 1, pp. 1–2, 2019. DOI: 10.1107/S2052252519000277.

[2] A. Efimov, "Chirality and handedness of protein structures," *Biochemistry (Moscow)*, vol. 83, no. 1, pp. 103–110, Jan. 2018. DOI: 10.1134/S0006297918140092.

[3] M. L. Baker, M. R. Baker, C. F. Hryc, T. Ju, and W. Chiu, "Gorgon and pathwalking: Macromolecular modeling tools for subnanometer resolution density maps," *Biopolymers*, vol. 97, no. 9, pp. 655–668, 2012. DOI: 10.1002/bip.22065.

[4] M. L. Baker, T. Ju, and W. Chiu, "Identification of secondary structure elements in intermediate-resolution density maps," *Structure*, vol. 15, no. 1, pp. 7–19, 2007. DOI: 10.1016/j.str.2006.11.008.

[5] D. Si and J. He, "Beta-sheet detection and representation from medium resolution cryo-em density maps," in *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*, ser. BCB'13, Washington DC, USA: Association for Computing Machinery, 2013, pp. 764–770. DOI: 10.1145/2506583.2506707. [Online]. Available: 10.1145/2506583.2506707.

[6] N. Zhou, H. Wang, and J. Wang, "Embuilder: A template matching-based automatic model-building program for high-resolution cryo-electron microscopy maps," *Scientific Reports*, vol. 7, no. 1, pp. 1–9, 2017. DOI: 10.1038/s41598-017-02725-w.

[7] M. Rusu and W. Wriggers, "Evolutionary bidirectional expansion for the tracing of alpha helices in cryo-electron microscopy reconstructions," *Journal of Structural Biology*, vol. 177, no. 2, pp. 410–419, 2012. DOI: 10.1016/j.jsb.2011.11.029.

[8] D. Si, S. Ji, K. Al Nasr, and J. He, "A machine learning approach for the identification of protein secondary structure elements from electron cryo-microscopy density maps," *Biopolymers*, vol. 97, pp. 698–708, Sep. 2012. DOI: 10.1002/bip.22063.

[9] R. li, D. Si, T. Zeng, S. Ji, and J. He, "Deep convolutional neural networks for detecting secondary structures in protein density maps from cryo-electron microscopy," vol. 2016, Dec. 2016, pp. 41–46. DOI: 10.1109/BIBM.2016.7822490.

[10] S. R. Maddhuri Venkata Subramaniya, G. Terashi, and D. Kihara, "Protein secondary structure detection in intermediate-resolution cryo-em maps using deep learning," *Nature Methods*, vol. 16, Sep. 2019. DOI: 10.1038/s41592-019-0500-1.

[11] X. Wang *et al.*, “Detecting protein and dna/rna structures in cryo-em maps of intermediate resolution using deep learning,” *Nature Communications*, vol. 12, p. 2302, Apr. 2021. DOI: 10.1038/s41467-021-22577-3.

[12] J. He and S.-Y. Huang, “EMNUSS: a deep learning framework for secondary structure annotation in cryo-EM maps,” *Briefings in Bioinformatics*, May 2021. DOI: 10.1093/bib/bbab156.

[13] S. Griep and U. Hobohm, “Pdbselect 1992-2009 and pdbfilter-select,” *Nucleic acids research*, vol. 38, pp. D318–9, Sep. 2009. DOI: 10.1093/nar/gkp786.

[14] H. M. Berman *et al.*, “The Protein Data Bank,” *Nucleic Acids Research*, vol. 28, no. 1, pp. 235–242, Jan. 2000. DOI: 10.1093/nar/28.1.235.

[15] M. Tagari, R. Newman, M. Chagoyen, J.-M. Carazo, and K. Henrick, “New electron microscopy database and deposition system,” *Trends in biochemical sciences*, vol. 27, no. 11, p. 589, 2002. DOI: 10.1016/s0968-0004(02)02176-x.

[16] C. Sorzano *et al.*, “Xmipp: A new generation of an open-source image processing package for electron microscopy,” *Journal of structural biology*, vol. 148, no. 2, pp. 194–204, 2004. DOI: 10.1016/j.jsb.2004.06.006.

[17] G. E. Schulz and R. H. Schirmer, *Principles of protein structure*. Springer Science & Business Media, 1979, ISBN: 0387903348.

[18] J. Berg, J. Tymoczko, and L. Stryer, *Biochemistry. Vol.* 2002, ISBN: 0716730510.

[19] R. Sanchez-Garcia *et al.*, “Deepemhancer: A deep learning solution for cryo-em volume post-processing,” *Communications Biology*, vol. 4, p. 874, Jul. 2021. DOI: 10.1038/s42003-021-02399-1.

[20] S. Perticaroli *et al.*, “Secondary structure and rigidity in model proteins,” *Soft Matter*, vol. 9, pp. 9548–9556, Oct. 2013. DOI: 10.1039/C3SM50807B.

[21] J. Rosa-Trevín *et al.*, “Scipion: A software framework toward integration, reproducibility and validation in 3d electron microscopy,” *Journal of structural biology*, vol. 195, pp. 93–99, Apr. 2016. DOI: 10.1016/j.jsb.2016.04.010.