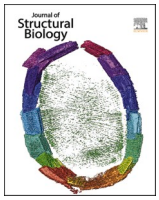
Journal of Structural Biology 214 (2022) 107915

Contents lists available at ScienceDirect

Journal of Structural Biology

journal homepage: www.elsevier.com/locate/yjsbi



Research Article

Automatic determination of the handedness of single-particle maps of macromolecules solved by CryoEM

J. Garcia Condado [a,b,c], A. Muñoz-Barrutia [b], C.O.S. Sorzano [c,*]

[a] Biocruces Bizkaia Instituto Investigación Sanitaria, Cruces Plaza, 48903 Barakaldo, Bizkaia, Spain
[b] Universidad Carlos III de Madrid, Avda. de la Universidad 30, 28911 Leganés, Madrid, Spain
[c] Centro Nacional de Biotecnologia (CNB-CSIC), Darwin, 3, Campus Universidad Autonoma, 28049 Cantoblanco, Madrid, Spain

ARTICLE INFO

Keywords:
Electron microscopy
Single Particle Analysis
Validation

ABSTRACT

Single-Particle Analysis by Cryo-Electron Microscopy is a well-established technique to elucidate the three-dimensional (3D) structure of biological macromolecules. The orientation of the acquired projection images must be initially estimated without any reference to the final structure. In this step, algorithms may find a mirrored version of all the orientations resulting in a mirrored 3D map. It is as compatible with the acquired images as its unmirrored version from the image processing point of view, only that it is not biologically plausible.

In this article, we introduce HaPi (Handedness Pipeline), the first method to automatically determine the hand of electron density maps of macromolecules solved by CryoEM. HaPi is built by training two 3D convolutional neural networks. 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 defines the overall map hand. The pipeline is trained on simulated and experimental data. The handedness can be detected only for maps whose resolution is better than 5 Å. HaPi can identify the hand in 89% of new simulated maps correctly. Moreover, we evaluated all the maps deposited at the Electron Microscopy Data Bank and 11 structures uploaded with the incorrect hand were identified.

1. Introduction

Single-Particle Analysis is an ill-posed problem because the reconstruction of 3D macromolecular structures from 2D images is not well determined. If all particle image orientations are mirrored, a map is reconstructed that is equally consistent with the measured data but non-superimposable over the map previously reconstructed, see Fig. 1. As proteins have a specific handedness (Efimov, 2018) only one of the two possible reconstructed maps is the correct reconstruction of the structure.

Currently, a trained biologist is required to look at the α-helices rotation to assess the handedness of the map. If incorrect, the reconstructed map is mirrored. The direction of rotation is easily determined at very high resolutions of 1 Å but can be difficult at lower resolutions even for experts, see Fig. 2. As the resolution decreases, the α-helix slowly transitions from a helix to a cylinder, which no longer has a hand (see Fig. 3). Hence, we propose HaPi (Handedness Pipeline) to automatically determine the hand of reconstructed maps using deep learning for resolutions of up to 5 Å.

To the best of our knowledge, there are no algorithms to detect the hand of reconstructed CryoEM maps automatically. The proposed model identifies Secondary Structure Elements (SSE) of interest in the volume and then uses these to detect the hand. There are several previous approaches to automatically determine SSE in electron density maps based on non-machine learning methods (Baker et al., 2012; Baker et al., 2007; Si and He, 2013; Zhou et al., 2017), machine learning methods (Rusu and Wriggers, 2012; Si et al., 2012) and more recently, deep learning techniques (Li et al., 2016; Maddhuri Venkata Subramaniya et al., 2019; Wang et al., 2021; He and Huang, 2021). As the latter have shown better performance, in this work, 3D Convolutional Neural Networks (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.

The number of reconstructed structures is quickly increasing as CryoEM is being widely adopted. HaPi is a valuable tool to guarantee the correctness of automatic image processing pipelines and as quality control in public databases like Electron Microscopy Data Bank (EMDB).

* Corresponding author.
E-mail address: coss@cnb.csic.es (C.O.S. Sorzano).

https://doi.org/10.1016/j.jsb.2022.107915
Received 12 April 2022; Received in revised form 29 August 2022; Accepted 25 October 2022
Available online 29 October 2022

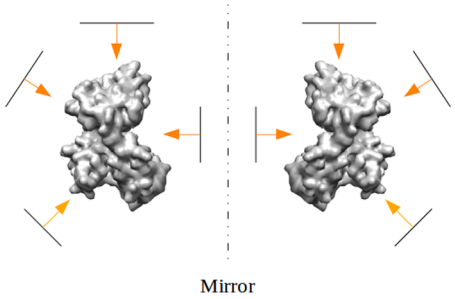


Fig. 1. Reconstruction of a structure from the same set of images but with mirror orientations assigned to each image which produces a mirrored version of the structure.

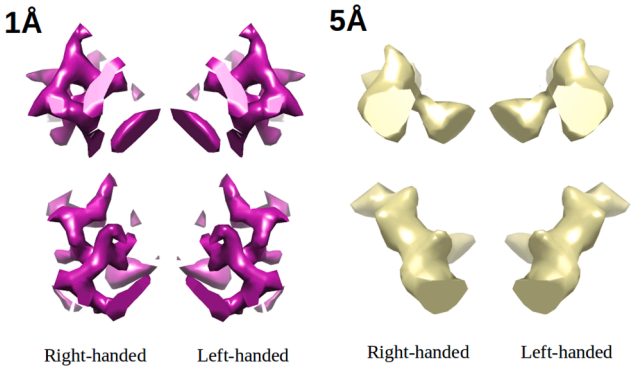


Fig. 2. A portion of the same α-helix at 1 Åand 5 Åwith its true structure (right-handed) and mirrored version (left-handed) from different viewing angles.

2. Methods

The HaPi package is freely available to use and documented via GitHub (https://github.com/JGarciaCondado/EMapHandedness). All the code for the methods described can be found in the same link. HaPi is also available in Scipion (Rosa-Trevín et al., 2016) and Xmipp (Sorzano et al., 2004).

2.1. Pipeline

HaPi determines the hand of electron density maps, Fig. 4 considering as inputs a Coulomb Potential map (V_f) and a mask of the non-background voxels (V_{mask}). Maps are first preprocessed by resampling to 1 Å/voxel and low-pass filtering to 5 Åto match the training data. This procedure reduces noise and homogenizes local resolutions.

AlphaVolNet determines the location of α-helices in the whole volume. It takes as input V_f and V_{mask} and outputs V_α, which is a mask containing the location of the α-helices found. It does so by taking at each non-background voxel location of V_f a box of dimensions $11 \times 11 \times 11$ voxels and passed through the trained 3D CNN α-SSE model. Then, if the label is above a threshold t_α, V_α is set to true at that location.

HandNet is used to determine the hand of a map from V_α and V_f. At each active voxel of V_α, a box of dimension $11 \times 11 \times 11$ voxels is extracted at that location from 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 boxes. The consensus value is the average of all hand predictions of each α-box.

2.2. 3D CNN model

The same 3D CNN model is used for the SSE and hand determination task. A 3D CNN is an extension of 2D CNNs that deals with volumes instead of images. The whole architecture of the 3D CNN design can be seen in Fig. 5. All 3D convolutional layers and the first connected layer are followed by ReLu activation functions. A sigmoid function follows the last fully connected layer. The 3D CNN has in total 104,080 parameters.

Input boxes are preprocessed by clipping all negative values to 0 and rescaling the box to be in the range 0 and 1.

The training data differs for each task. All proteins structures were

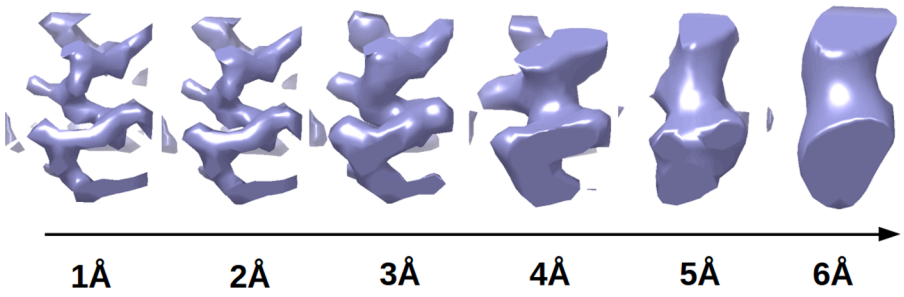


Fig. 3. Same α-helix with same viewing angle at different resolutions that shows transition from helical to cylindrical structure.

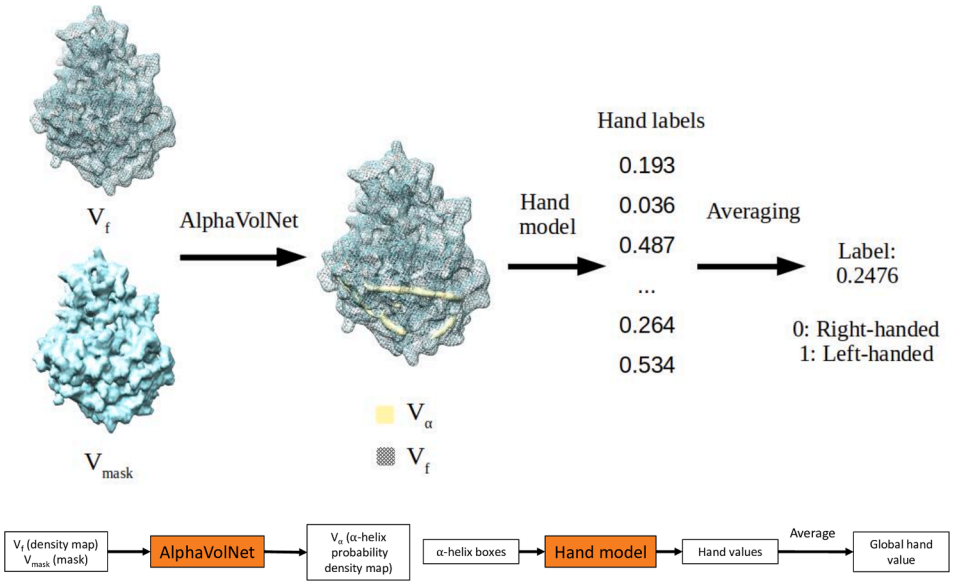


Fig. 4. Diagram of the HaPi pipeline with the inputs and outputs at the different stages and the models used to generate each. The input to HaPi is V_f (a Coulomb potential map) and V_{mask} (a mask of the overall structure). These are first passed through AlphaVolNet, which outputs a mask V_α of the proposed location of α-helices (yellow). Boxes are extracted from V_f at the location of voxels in V_α that contain an α-helix. They are 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.

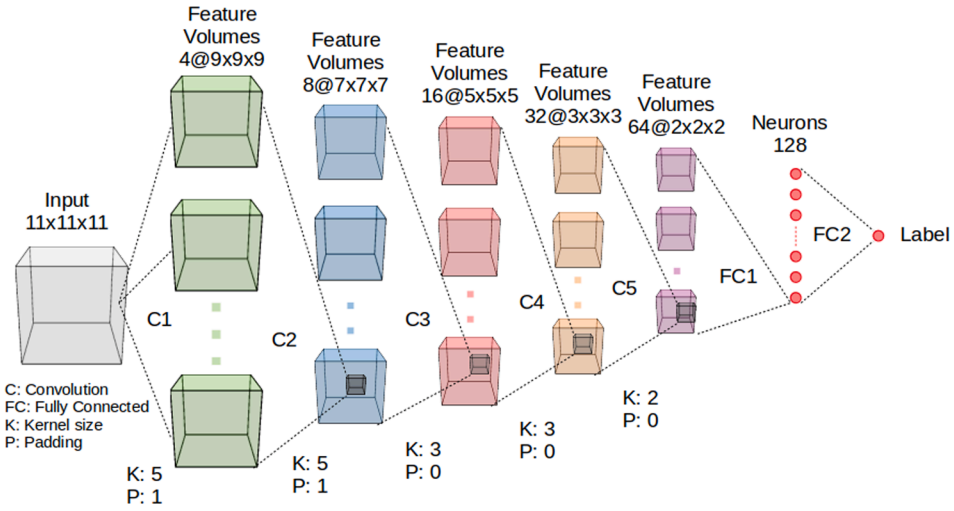


Fig. 5. 3D CNN model diagram used for both Secondary Structure Element (SSE) determination and hand determination. The input to the network is a box containing the electron density at each voxel normalised to the range 0 to 1. The output of the network is a value between 0 and 1. In the SSE determination 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.

chosen using PDB-Select (Griep and Hobohm, 2009), which has a non-redundant set of proteins with low mutual sequence identity to enable unbiased statistics. For the SSE determination task, 261 experimental maps and their respective fitted atomic models were downloaded from EMDB (Tagari et al., 2002). In total, 9,985 boxes of dimension $11 \times 11 \times 11$ voxels were extracted from the centroids of alpha helices in the structures and another 9,985 boxes of the same size from random parts of the structure. For the hand determination task, the Xmipp library (Sorzano et al., 2004) was used to simulate Coulomb potential maps of 12,343 atomic models using Electron Atomic Scattering Factors (Sorzano et al., 2015). From these, 101,944 boxes of dimension $11 \times 11 \times 11$ voxels from the center of α-helices were extracted to use for training. Half of the boxes were randomly flipped to obtain left-hand helices.

The training strategy is similar for both tasks. 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 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 the final model.

Weight initialisation differs for the SSE and hand models. For SSE determination, the model has been initialised with the default PyTorch settings for weight initialisation. The hand model was first trained on 1 Ådata and initialised with the default PyTorch settings. Then, a transfer learning approach is used to train on 5 Ådata for hand determination by using the weights of the model trained on 1 Åon initialization.

3. Results

The training and validation curves for each of the models trained can be seen in Fig. 6. The network is unable to learn how to identify the hand of boxes at 5 Åresolution without transfer learning as seen in the high loss value in Fig. 6. When the weights are initialized at 1 Åthe loss value significantly decreases during training. The validation set loss stagnates for the models trained after only 15 epochs.

A dataset consisting of 3,119 atomic models was used to simulate maps at 5 Å. These were split into validation and test sets of 30% and 70%, respectively. The validation set was used to set t_α, which controls the stringency of the model to accept voxels as containing an α-helix. It was set to $t_\alpha = 0.7$ to maximize hand accuracy in the validation set. HaPi

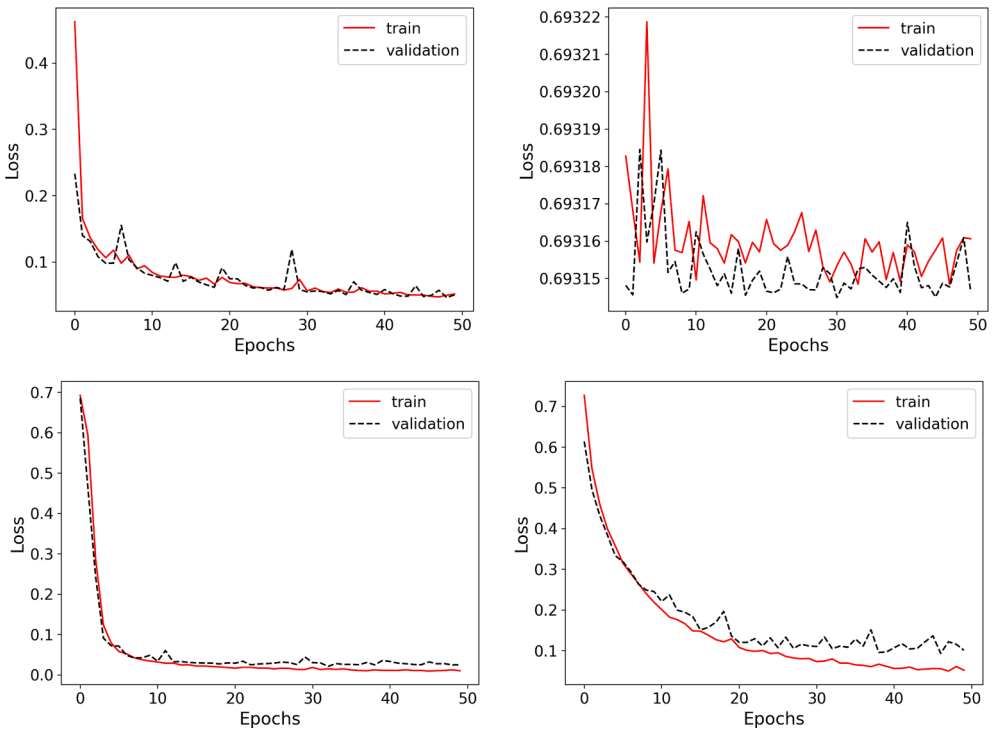


Fig. 6. Training and validation loss for each of the 3D CNN models trained. (Top left) Training and validation loss for α-helix determination at 5 Å. (Top right) Training and validation loss for hand determination at 5 Å. Clearly the network is not able to generalize as the loss stagnates at a high value and is unable to correctly determine the hand. (Bottom left) Training and validation loss for hand determination at 1 Å. (Bottom right) Training and validation loss for hand determination at 5 Åwith weights initialized at those of the model trained at 1 Å. With transfer learning strategies the model is able to learn how to correctly classify the hand at lower resolution.

is able to identify the hand in 89% of new simulated maps correctly. The resulting hand prediction values for the test set is shown in Fig. 7. In this figure we see three peaks. During testing maps with hand values < 0.5 are predicted to be right-handed and maps with hand values > 0.5 are predicted to be left-handed. This decision results in 89% of the maps correctly assigned. The peak around 0.15 contains right-handed simulated maps that are correctly labelled and the peak around 0.85 contains left-handed simulated maps that are correctly labelled. In the middle peak around 0.5 we have a mix of right-handed and left-handed simulated maps that are either correctly or incorrectly labelled depending on which side of the peak they land on and their hand. Our network relies very much on the detection of alpha helices. For those maps that do not have alpha helices or have short alpha helices, the network has difficulties to determine their hand and ends up assigning values around 0.5. Therefore, the errors come mostly from the central region. For experimental maps we chose then a threshold of 0.6 that reflects our position that, when in doubt, let us be conservative and assume that the map is in a correct position.

All deposited experimental maps in EMDB with resolution below or at 5 Åwere downloaded and passed through HaPi. In total, 8,061 maps were downloaded. Maps deposited in EMDB should all be right-handed. The resulting hand predictions for structures can be seen in Fig. 8. All structures with a high chance of being left-handed (those with hand value > 0.6, which are 285 structures) were manually checked to assess if they were left-handed. Eleven of these structures were indeed present in the database despite incorrect handedness (EMD-9890, EMD-10012, EMD-22052, EMD-22053, EMD-22056, EMD-22057, EMD-22058, EMD-11082, EMD-11083, EMD-20213 and EMD-23584). Assuming that all structures with hand value < 0.6 were uploaded to EMDB with correct handedness (right-hand), then the error rate is only $\frac{274}{8061} \times 100 = 3.4\%$. In Fig. 9, we show one of the examples of incorrectly deposited maps in which it can be seen that the atomic alpha helix turns in a different direction with respect to the turn of the alpha helix of the CryoEM map.

4. Discussion

HaPi is able to automatically determine the hand of reconstructed macromolecular structures for intermediate resolutions of 5 Åor below. Therefore, HaPi is a valuable tool for validation in databases to avoid incorrect structures to be uploaded. It will also reduce the time biologists dedicate to checking the handedness during image processing and facilitate the construction of automatic image processing pipelines. The method implemented in Scipion automatically return a flipped volume if the hand value is above a threshold set by the user (a value of 0.6 is recommended.) As the number of maps with a resolution of 5 Åor below is increasing at a fast pace, this tool can become very useful for many.

HaPi is robust as it has a high accuracy of 96.6% on previously unseen experimental data. Although one of the networks was trained on synthetic data, it can still generalize to experimental data. Some of the downloaded maps contained filaments, electron crystallography data and DNA. This type of data was not used for training. Hence, part of the error rate could be due to including these data during testing. Also, structures that do not contain a considerable number of α-helices are not well predicted as α-helices are the basis for determining the hand.

HaPi is versatile because it does not output a discrete hand label but a value between 0 and 1. Closer to 0 means it is right-handed, and 1 means it is left-handed. Hence, HaPi estimates how sure it is about its decision. Changing the threshold of when to accept a structure as right-handed gives control to the user on stringency. The results on simulated data structures with values lower than 0.4 are highly likely to be right-handed, and those above 0.6 are likely to be left handed. HaPi is unsure of the hand for structures whose hand value is between 0.4 and 0.6, 90% of the structures whose hand was wrongly determined lied between these values.

The structures whose hand values lies between 0.4 and 0.6 and which HaPi has difficulty determining belong to two different groups. The first group would be structures that do not have clear α-helices or have only a few of them. As the algorithm searches for this, its predicting power is diminished if it does not find suitable candidates. The second group are structures whose resolution is very close to 5 Å. With lower resolution, there is less information about the hand encoded in the

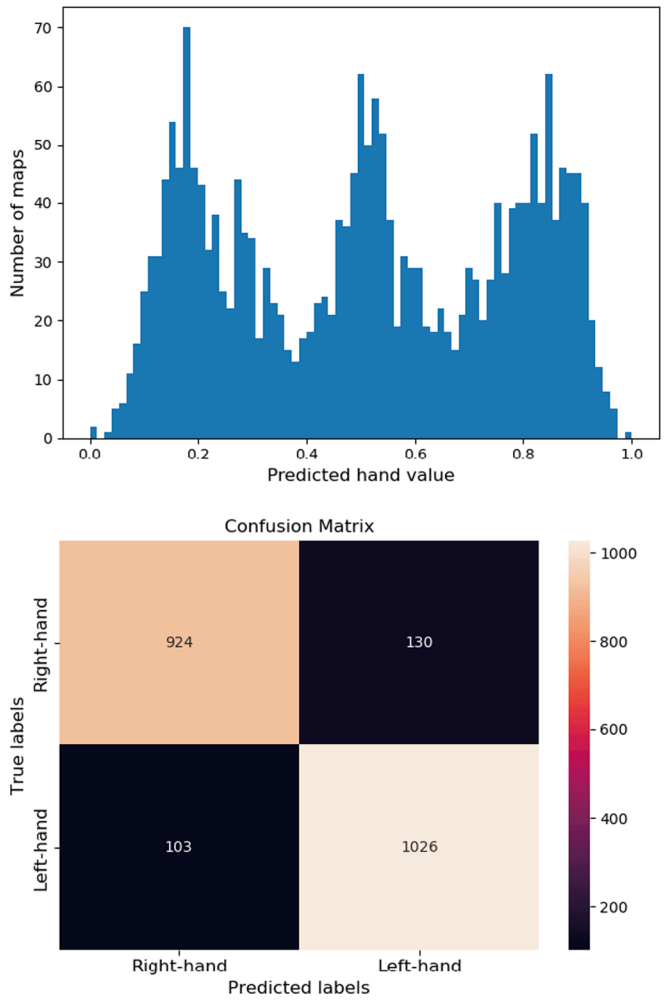


Fig. 7. Individual predicted hand values for each simulated map (5 Åresolution) in test set (above) and confusion matrix (below) for $t_\alpha = 0.7$. t_α controls the stringency of the model to accept voxels as containing an α-helix. Simulated maps were left as is to be right-handed or mirrored to be 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.

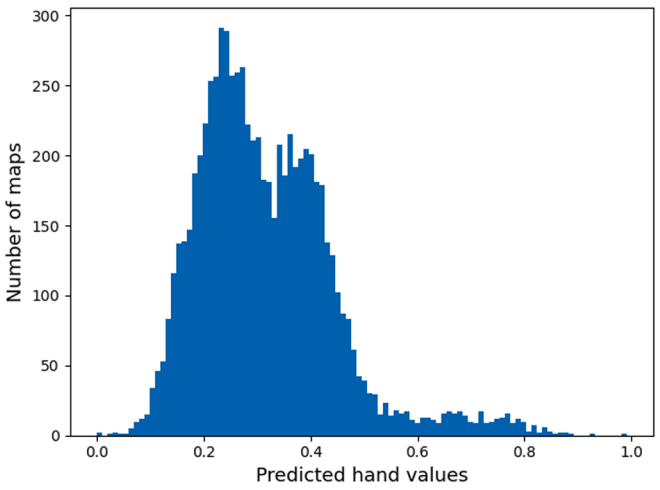


Fig. 8. Individual predicted hand values by HaPi for each map in the EMDB database with 5 Åor less for $t_\alpha = 0.7$. t_α controls the stringency of the model to accept voxels as containing an α-helix. Values close to 0 indicate the map is right-handed and values close to 1 indicate the structure is left-handed.

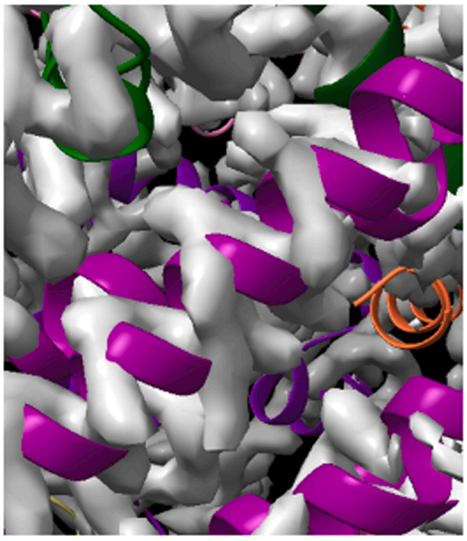


Fig. 9. EMD-22056 structure (grey surface) with superimposed apoferritin PDB (purple). The α-helix from the PDB (purple) turns right if you use your right hand and the experimental map alpha helix (grey surface) rotates left if you use your left hand. This showcases that the structure was uploaded with incorrect hand (left-hand) to the EMDB database.

structure, and therefore it is more difficult for HaPi to be sure about its decision.

An experiment was also run to determine which type of SSE was better suited for hand determination between α-helices and β-sheets. At 5 Åthe 3D CNN trained on α-helices had an accuracy of 95.5% when detecting the hand. On 5 Åβ-sheets it had an accuracy of 61.0% when detecting the hand; α-helices are better for determining the hand at box level than β-sheets. A biologist can easily distinguish the hand of an α-helix at high resolutions. Even at high resolutions, it is difficult to determine the hand of a β-sheet as it requires looking at the side-chains and their orientation. Hence, the α-helix model was used to build the whole pipeline.

The 3D CNN trained on data at 6 Ågave chance level results. The inability of the network to identify the hand at 6 Åis the result of the hand information not being available at such resolution. 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 5.4 Å(Schulz and Schirmer, 1979). At resolutions above 5.4 Å, 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.

Being able to determine the hand of an experimental map at or below 5 Åis still useful. Structures with resolutions below 5 Åalready represent more than 50% of all depositions at the EMDB. However, at resolutions below 4 Å, the hand can be easily identified by manual inspection of the map as seen in Fig. 3. Still, structures between 4 Åand 5 Åof 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 valuable tool to easily and automatically determine the hand at this resolution range.

5. Conclusion

HaPi has shown that it is possible to automatically determine the hand of a map without the necessity of inspection by a trained expert. This is even the case for intermediate resolutions where the hand is not

clear from visual inspection. HaPi offers a valuable tool for validation in databases and increases biologists' efficiency by reducing the need for hand inspection during the processing.

The automatic determination of the map hand is a very useful task in the construction of automatic image processing pipelines in CryoEM, and reduces the probability of depositing incorrect maps in public databases such as the EMDB.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work is supported in part by Ministerio de Ciencia, Innovación y Universidades, Agencia Estatal de Investigación, under grant PID2019-109820RB-I00, MCIN/AEI/10.13039/501100011033/, cofinanced by European Regional Development Fund (ERDF), "A way of making Europe."

The authors acknowledge the economical support from MICIN to the Instruct Image Processing Center (I2PC) as part of the Spanish participation in Instruct-ERIC, the European Strategic Infrastructure Project (ESFRI) in the area of Structural Biology and Grant PID2019-104757RB-I00 funded by MCIN/AEI/ 10.13039/501100011033/ and "ERDF A way of making Europe", by the "European Union". Comunidad Autónoma de Madrid through Grant: S2017/BMD-3817. Instituto de Salud Carlos III (project IMPaCT-Data, exp. IMP/00019), co-funded by the European Union, European Regional Development Fund (ERDF, "A way to make Europe"). CSIC and JAE Intro Program: JAEINT_20_01330. European Union (EU) and Horizon 2020 through grant: HighResCells (ERC - 2018 - SyG, Proposal: 810057)

References

Baker, M.L., Baker, M.R., Hryc, C.F., Ju, T., Chiu, W., 2012. Gorgon and pathwalking: Macromolecular modeling tools for subnanometer resolution density maps. Biopolymers 97 (9), 655–668.

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

Efimov, A., 2018. Chirality and handedness of protein structures. Biochemistry (Moscow) 83 (1), 103–110, 01.

Griep, S., Hobohm, U., 2009. PDBselect 1992–2009 and PDBfilter-select. Nucleic acids research 38, D318–9, 09.

He, J., Huang, S.-Y., 2021. EMNUSS: a deep learning framework for secondary structure annotation in cryo-EM maps. Briefings in Bioinformatics 22, 1–13, 05.

Li, R., Si, D., Zeng, T., Ji, S., He, J., 12 2016. Deep convolutional neural networks for detecting secondary structures in protein density maps from cryo-electron microscopy. In: Proceedings. IEEE International Conference on Bioinformatics and Biomedicine. Vol. 2016. pp. 41–46.

Maddhuri Venkata Subramaniya, S.R., Terashi, G., Kihara, D., 2019. Protein secondary structure detection in intermediate-resolution cryo-EM maps using deep learning. Nat. Methods 16, 911–917, 09.

Rosa-Trevín, J., Quintana, A., Cano, L., Zaldívar, A., Foche, I., Gutiérrez, J., Gomez-Blanco, J., Burguet-Castell, J., Cuenca-Alba, J., Abrishami, V., Vargas, J., Oton, J., Sharov, G., Vilas, J., Navas, J., Conesa, P., Kazemi, M., Marabini, R., Sorzano, C., Carazo, J., 2016. Scipion: a software framework toward integration, reproducibility and validation in 3d electron microscopy. Journal of structural biology 195, 93–99, 04.

Rusu, M., Wriggers, W., 2012. Evolutionary bidirectional expansion for the tracing of alpha helices in cryo-electron microscopy reconstructions. J. Struct. Biol. 177 (2), 410–419.

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

Si, D., He, J., 2013. 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. BCB'13. Association for Computing Machinery, New York, NY, USA, p. 764–770.

Si, D., Ji, S., Al Nasr, K., He, J., 2012. A machine learning approach for the identification of protein secondary structure elements from electron cryo-microscopy density maps. Biopolymers 97, 698–708, 09.

Sorzano, C., Marabini, R., Velázquez-Muriel, J., Bilbao-Castro, J.R., Scheres, S.H., Carazo, J.M., Pascual-Montano, A., 2004. XMIPP: a new generation of an open-source image processing package for electron microscopy. Journal of structural biology 148 (2), 194–204.

Sorzano, C.O.S., Vargas, J., Otón, J., Abrishami, V., de la Rosa Trevín, J.M., del Riego, S., Fernández-Alderete, A., Martínez-Rey, C., Marabini, R., Carazo, J.M., 2015. Fast and accurate conversion of atomic models into electron density maps. AIMS Biophysics 2, 8–20.

Tagari, M., Newman, R., Chagoyen, M., Carazo, J.-M., Henrick, K., 2002. New electron microscopy database and deposition system. Trends in biochemical sciences 27 (11), 589.

Wang, X., Alnabati, E., Aderinwale, T., Maddhuri Venkata Subramaniya, S.R., Terashi, G., Kihara, D., 2021. Detecting protein and DNA/RNA structures in cryo-EM maps of intermediate resolution using deep learning. Nature Communications 12, 2302, 04.

Zhou, N., Wang, H., Wang, J., 2017. EMBuilder: A template matching-based automatic model-building program for high-resolution cryo-electron microscopy maps. Scient. Rep. 7 (1), 1–9.