How to cite: *Angew. Chem. Int. Ed.* **2020**, *59*, 14788–14795
International Edition: doi.org/10.1002/anie.202000421
German Edition: doi.org/10.1002/ange.202000421

Structural Biology **Hot Paper**

Haruspex: A Neural Network for the Automatic Identification of Oligonucleotides and Protein Secondary Structure in Cryo-Electron Microscopy Maps**

*Philipp Mostosi, Hermann Schindelin, Philip Kollmannsberger, and Andrea Thorn**

Abstract: *In recent years, three-dimensional density maps reconstructed from single particle images obtained by electron cryo-microscopy (cryo-EM) have reached unprecedented resolution. However, map interpretation can be challenging, in particular if the constituting structures require de-novo model building or are very mobile. Herein, we demonstrate the potential of convolutional neural networks for the annotation of cryo-EM maps: our network Haruspex has been trained on a carefully curated set of 293 experimentally derived reconstruction maps to automatically annotate RNA/DNA as well as protein secondary structure elements. It can be straightforwardly applied to newly reconstructed maps in order to support domain placement or as a starting point for main-chain placement. Due to its high recall and precision rates of 95.1 % and 80.3 %, respectively, on an independent test set of 122 maps, it can also be used for validation during model building. The trained network will be available as part of the CCP-EM suite.*

Introduction

The resolution revolution in single particle electron cryo-microscopy (cryo-EM) yields macromolecular structures of unprecedented resolution. These structures allow us to identify new drug targets, for example in the Zika virus,[1] to fight tuberculosis[2] or to understand the fundamental processes of life, such as the process of translation by ribosomes.[3] However, modelling an atomic structure to these maps remains difficult as researchers mostly rely on algorithms developed for the interpretation of crystallographic electron density maps. In X-ray crystallography, the measured diffraction corresponds to the amplitudes of the Fourier transform of the electron density, as the X-rays interact with the electrons in the molecular assemblies in a crystal and the phases are reconstructed only during refinement. In cryo-EM, on the other hand, the measured micrographs already contain phase information, but are very noisy, which is overcome by 3D-reconstruction and averaging. The individual micrographs show the interaction of the electron beam with the entire electrostatic potential of a single molecular assembly. Hence, cryo-EM reconstruction maps differ in both their nature and error distribution[4–6] from X-ray crystallographic electron density maps. Consequently, their modelling might be improved greatly by tools that consider these specific properties of the data at hand. Such modelling tools should not only provide good functionality, but also be easy to use and freely available to academic users worldwide.

Parallel to the advances in cryo-EM during the last decade, deep neural networks have achieved remarkable image segmentation capabilities,[7] making them the most powerful machine-learning approach currently available. Convolutional neural networks (CNN) combine traditional image analysis with machine learning by cascading layers of trainable convolution filters and are exceptionally well-suited for volume annotation. They have been successfully applied to biological problems, such as breast cancer mitosis recognition[8] and, in conjunction with encoder-decoder architectures, to volumetric data segmentation.[9,10] Given that a cryo-EM reconstruction map is essentially a three-dimensional image volume, CNNs seem a good choice for their annotation if good "ground truth" data to train the network could be provided.

In this work, we demonstrate that deep neural networks are not only capable of annotating protein secondary structure, but also oligonucleotides (RNA/DNA) in cryo-EM maps, and provide a pre-trained network, named *Haruspex*. Assigning a fold to regions in a cryo-EM map is the first step in modelling a structural region. This can be a major challenge, in particular for novice users, in low resolution regions, or when little is known about the composition of the macromolecular complex in question. Haruspex can be readily used to annotate cryo-EM maps, which will prove useful in model building and supporting the placement of known domain folds, thus accelerating the

[*] P. Mostosi, Prof. Dr. H. Schindelin, Dr. A. Thorn
Institute of Structural Biology,
Rudolf Virchow Center for Experimental Biomedicine,
University of Würzburg
Josef-Schneider-Str. 2, 97080 Würzburg (Germany)
E-mail: andrea.thorn@web.de
P. Mostosi, Prof. Dr. P. Kollmannsberger
Center for Computational and Theoretical Biology,
University of Würzburg
Campus Hubland Nord 32, 97074 Würzburg (Germany)

[**] A previous version of this manuscript has been deposited on a preprint server (https://www.biorxiv.org/content/10.1101/644476v3).

Supporting information and the ORCID identification number(s) for the author(s) of this article can be found under: https://doi.org/10.1002/anie.202000421.

modelling process and improving the accuracy of cryo-EM-derived molecular structures.

Results

Network Architecture and Implementation

In low-resolution cryo-EM maps, α-helices can often be discerned as long cylindrical elements. This has been exploited by the program helixhunter,[11] which searches for prototypical helices in reconstruction maps using a cross-correlation strategy. β-Strands are more difficult to identify as they are more variable in shape and therefore require morphological analysis.[12] A combination of these approaches led to the development of SSEHunter,[13] which uses a density skeleton to detect secondary structures. Deep learning offers an alternative approach: Fully convolutional networks[9,14] allow a swift generation of segmentation maps for objects of variable size. Here, we employ a state of the art U-Net-style architecture[9] to demonstrate that at an average map resolution of 4 Å or better, experimentally derived reconstruction maps allow the training of a well-performing network that can be used for a wide range of specimens—with no re-training necessary. The network was implemented with TensorFlow[15] and processes 40^3 voxel segments with a voxel size of 1.0–1.2 $Å^3$ (covering a secondary structure element and its immediate surroundings) to annotate 20^3 voxel cubes (corresponding to the center of the input volume). The output volume has four channels containing the probabilities that the voxel is part of an α-helical or β-strand protein secondary structure element, nucleotide, or unassigned. 40^3 voxel segments were chosen as a compromise between computational power and network complexity on one hand and covering the secondary structure including surrounding interaction partners on the other hand. A 40^3 voxel segment covers 40–48 $Å^3$; an average α-helix with 10 residues, for example, is 15 Å in length.[16]

The input is a single channel containing the reconstruction density. During prediction, this three-dimensional volume is passed through multiple convolutional layers (image filters) that extract learned image features relevant for structure detection, and through pooling layers, which select the most significant of the detected features. In the second ("upconvolutional") part of the network, these activations are combined with higher-level activations of the network to recover spatial detail. The output has four channels representing the probabilities for the four classes (helix, sheet, nucleotide, unassigned) and represents the annotation of the central 20^3 voxel cube of the input volume.

Training Data Selection

For network training, we pre-selected EMDB (Electron Microscopy Data Bank[17]) reconstruction maps with an average resolution of 4 Å or better as stated in the EMDB entry. From 576 entries (as of 15/2/2018), we picked 293 EMDB/PDB (Protein Data Bank[18]) pairs (Supporting Information, Table S1) by three criteria: 1) map and model represent the same structure and fit visually well to each other; 2) the presence of at least one annotated α-helix or β-sheet in the PDB model; 3) preference of higher resolution maps in case the same authors deposited several instances of the same macromolecular complex (as the model was most likely fitted to the highest resolution map). Maps with severe misfits, misalignments, or models without corresponding reconstruction density (and vice versa) were omitted. Visual evaluation was supplemented with a comparison between the entire map and the part which is occupied by the model using histograms, mean and median values; this provided an additional test of how well map and model fit each other. Furthermore, the training data were filtered by map root mean square deviations (r.m.s.d.) values (see below).

Cryo-EM maps are often post-processed, stitched or otherwise filtered, but it can be difficult to determine how exactly a given map has been altered. Hence, we did not apply any additional criteria pertaining to map modification and instead decided to train the network with all possible representations of the features in question. It is worth mentioning that some types of post-processing, such as map sharpening, are in principle equivalent to linear convolution filters. Convolutional neural networks (CNNs) can learn to apply or compensate for these during training (if they are relevant for predicting the annotated structure) and hence, can become insensitive to these procedures.

Training Data Annotation

To generate ground truth data for network training, a python script was implemented to automatically annotate the reconstruction map according to the deposited structural model as α-helical, β-strand, nucleotide or unassigned. The script extracts the original annotations from PDBML format[19] files using a custom parser. To obtain suitable training data, additional secondary structure information was necessary. We implemented a variant of the DSSP algorithm[20] omitting strand direction, and a torsion-angle-based secondary structure detection inspired by STRIDE:[21] annotated or DSSP-detected secondary structures were extended by neighbouring amino acids if they matched the same Ramachandran profile. Before usage, the voxel size of the reconstruction map was re-scaled to 1.1 Å if outside [1.0; 1.2] Å.

Following that, if a secondary structure was identified, and if the average main chain atom map r.m.s.d. (root mean square of the map density distribution) was above 2, all voxels within 3 Å of backbone atoms were annotated accordingly. Secondary structure residues below 2 but ≥ 1.0 r.m.s.d. were masked and excluded from error calculation during training. All voxels not within 5 Å of model atoms, but with density ≥ 1 r.m.s.d. were masked and excluded from training, as they had high density, but were not modelled. The remaining voxels were marked as "unassigned".

Network Training

The maps were split into a total of 2183 segments of 70^3 voxels, of which 110 segments (5%) were set aside for evaluation during network training. Each segment had to contain at least 100 atoms $\geq$1.0 r.m.s.d., a backbone mean density of $\geq$3 r.m.s.d., and at least 5% of the total segment volume annotated. The training data were augmented through on-GPU 90° rotations (24 possibilities), and by randomly selecting a 40^3 voxel sub-segment (translational augmentation).

The network was trained for 40000 steps with 100 segments employed per step. In training data generation, the average EMDB map had roughly 95% unassigned voxels after annotation with the PDB model. From this, we estimated that non-true negatives needed to be weighted approximately 16-fold stronger than true negatives. This was necessary as the majority of the space within a reconstruction map is not made up of secondary-structure/oligonucleotide-associated voxels and thus the network can reach approximately 70–90% accuracy by predicting "unassigned" (not α-helical, β-sheet or oligonucleotide) structure only.

Network Performance Test

After training, the network was tested on an independent set of 122 EMDB maps (selected by the same criteria as training data and deposited after February 2018, for the complete list, see Supporting Information, Table S2). For evaluation, we investigated residues with mean backbone densities $\geq$1.0 r.m.s.d. and compared the predicted secondary structure on a per-residue basis with the one derived from the deposited PDB model. For this analysis, the r.m.s.d. value given in the header of each map file was used. Using this criterion, the network achieved similar performance on training, evaluation, and test data. Over all test maps, there were 75.4% true positives t_p (correctly predicted residues), 18.8% false positives $\underline{f_p}$ (wrongly predicted residues) and 4.0% false negatives f_n (non-predicted residues), resulting in a median recall rate $100*t_p(t_p+f_n)^{-1}$ of 95.1% and a precision $100*t_p(t_p+f_p)^{-1}$ of 80.3%. Precision and recall did not correlate significantly with average resolution (as given in the EMDB entry), Molprobity[22] score or deposition date.

The corresponding residue-level F_1 score (harmonic mean of precision and recall) on the test set for Haruspex (87.05%) is the highest reported so far on a per-residue-level when compared to other recent work.[23–25] Direct comparison of these values is, however, difficult since these other networks were tested on small test sets of lower-resolution simulated and experimental maps, whereas we used a large set of exclusively experimentally derived higher-resolution maps. Moreover, these networks did not annotate oligonucleotides, which affects the composition of the F_1 score. In a recent preprint,[26] the authors use deep learning for atom-level prediction and report 88.5% correctly predicted C_α atoms on 50 pre-cleaned experimental maps at 4.4 Å or better, which suggests similar performance for their intermediate secondary structure prediction.

As a typical example, human ribonuclease P holoenzyme (EMDB entry 9627) illustrates the power of our approach (see Figure 1). Haruspex is not only able to accurately predict the RNA vs. protein distribution in this complex, but also correctly assigns secondary structure elements in the protein areas with only a few exceptions. These notably include a stem-loop element in the RNA (upper left in the structure), regions that resemble β-sheets but do not follow the characteristic hydrogen bonding pattern, as well as secondary structure elements currently not covered by Haruspex, such as polyproline type II (P_{II}) helices (Figure 2C,D). Additional examples are shown in Figure 3.

Haruspex Usage

Haruspex can be used as a command line tool, which reads in an MRC format reconstruction map. No further parameters are needed and a prediction for a single map takes approximately 30 seconds to a few minutes on a normal workstation, depending on the available hardware (it can be used with or without GPU); on an older laptop, the annotation may take as long as 45 minutes for a very large structure. The output consists of four MRC format maps corresponding to the α-helical and β-strand protein, nucleotide, and "unassigned" portion of the input map. These maps can be displayed in Coot,[27] Pymol[28] or Chimera[29] and together represent the entire input map.

Discussion

Network Performance

Herein, we have described the development of the neural network Haruspex for the annotation of protein secondary structure and RNA/DNA in cryo-EM reconstruction maps in order to facilitate the modelling of such maps. We trained Haruspex on 293 experimentally derived reconstruction maps with a resolution of 4 Å or better and obtained recall and precision rates of 95.1% and 80.3%, respectively, on an independent test set of 122 maps. The pre-trained network can be readily applied to annotate newly reconstructed maps to support domain placement or to supply a starting point for main-chain placement.

When considering the 18.8% false positives and 4.0% false negatives, two fundamental limitations in the annotation of EMDB maps should be mentioned: firstly, the map can be wrongly modelled (see Figure 2C), which biases our annotation towards human modelling errors. Secondly, the deposited model may have been built employing additional information, such as structure-specific information from an external source, for example backbone folds established prior by crystallographic means,[30] NMR or structure prediction, or more than one map generated from different particle alignments.[31] This would in particular introduce higher rates of false negatives at the outer edges of the map, where the model covers secondary structure that was established by other

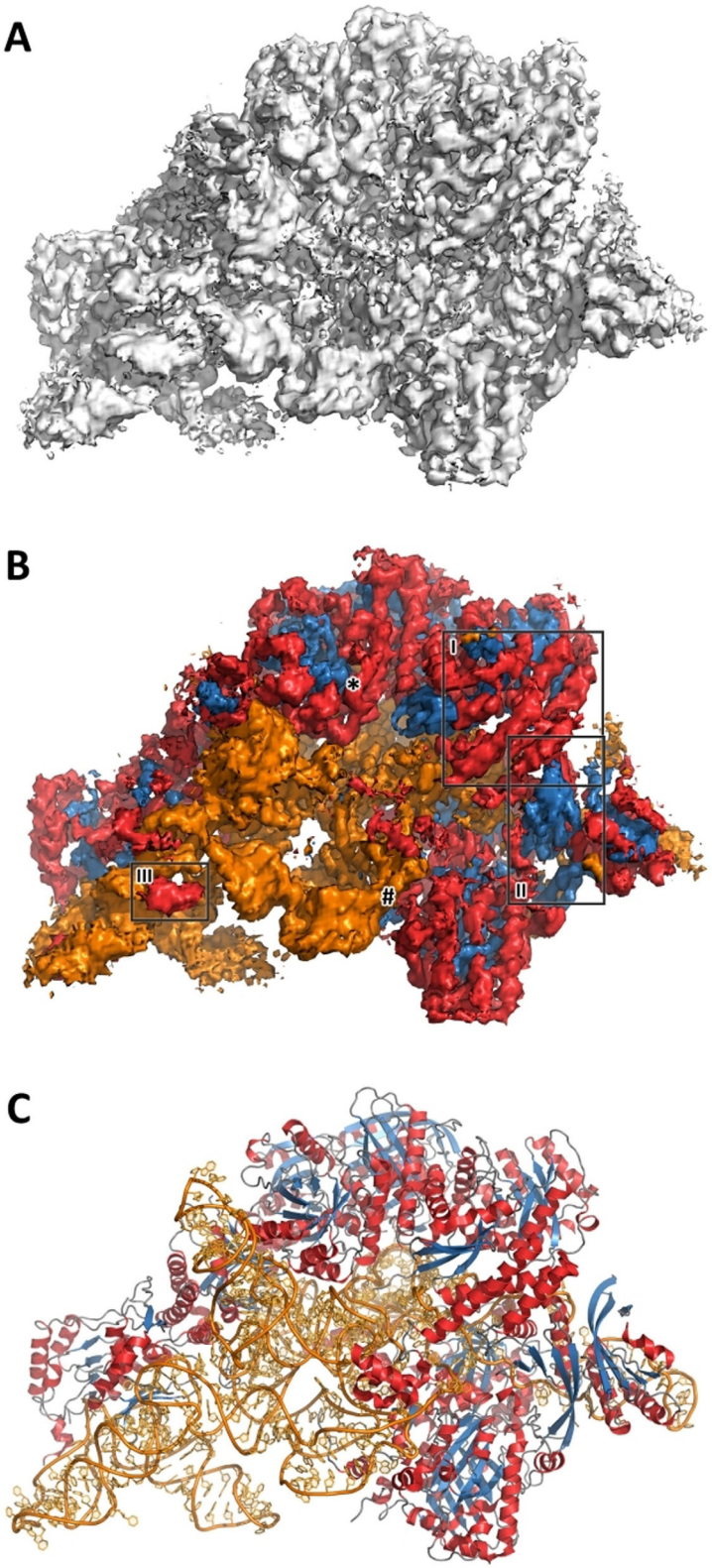

Figure 1. Typical example of Haruspex annotation. A) Reconstruction map for the human ribonuclease P holoenzyme (EMDB entry 9627). Manual assignment of secondary structure features can be difficult, in particular if the composition of a macromolecular complex is unknown. The surface shown corresponds to an r.m.s.d. of 0.04 with no carving. B) Secondary structure, as identified by our network in the map, is projected onto the surface. Orange corresponds to RNA/DNA; red to helices and blue to sheets. This was a fairly typical test case with 70.5% true positives, 18.8% false positives, and 10.7% false negatives. Recall was 86.8% and precision 79.0%. Region I) depicts a well-predicted α-helical structure, II) a β-sheet, and III) RNA misinterpreted as an α-helix. C) The deposited model PDB 6AHU for this map is shown in comparison. The regions depicted in Figure 2C and 2D are marked # and *, respectively.

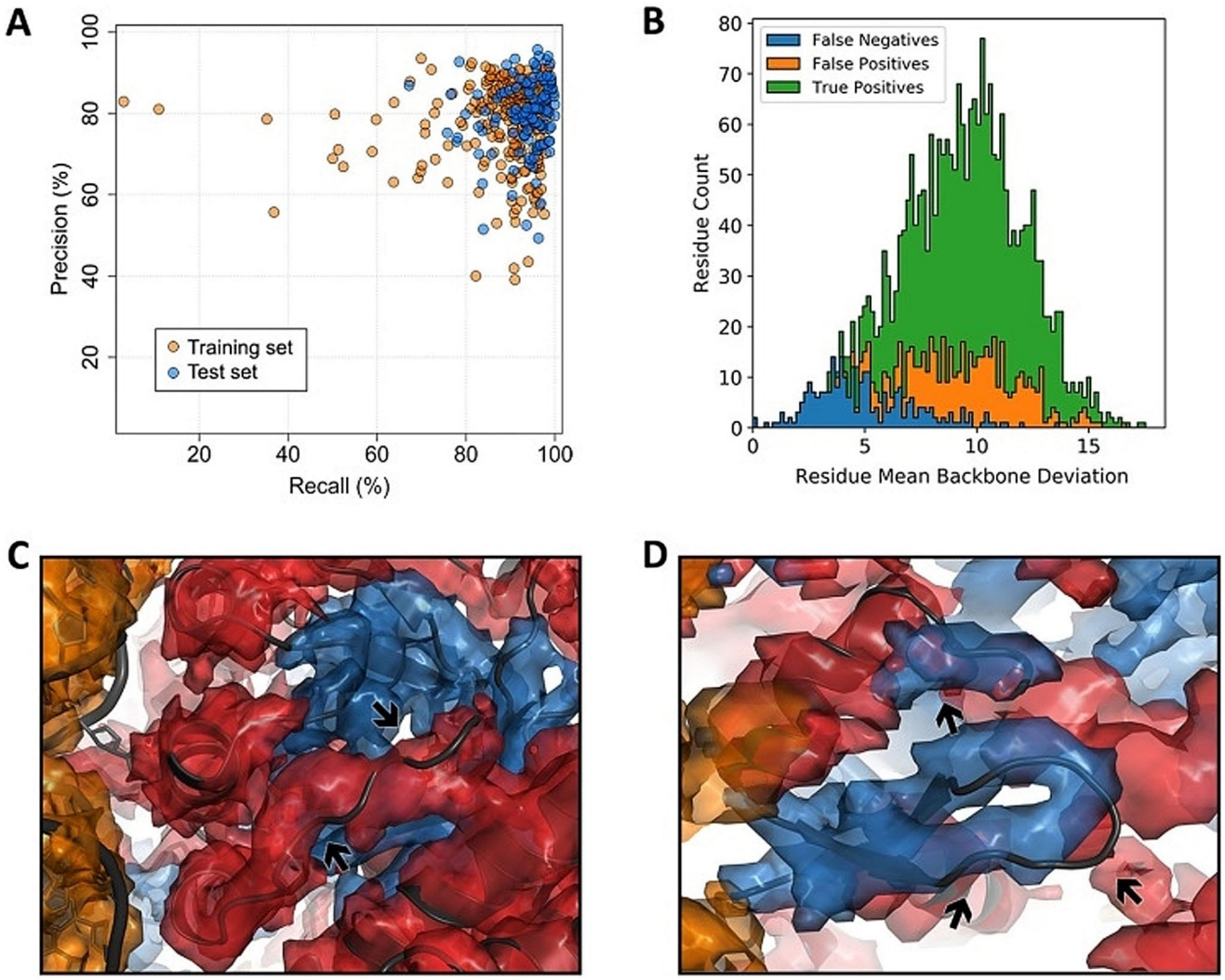

Figure 2. Network performance. A) Network precision vs. recall rates, with one marker per EMDB entry (training set entries are shown as orange, test set entries as blue markers). Both perform similarly well; with the training set producing a few more outliers. B) Frequency vs. map r.m.s.d. level for EMDB 9627 on a per-residue basis: True positives (green), false positives (orange), and false negatives (blue). This plot is typical: false negatives often occur in low-density map regions. C) α-Helical false positives (PDB 6AHU, residues 131–139 in chain J): The model partly occupies the conformational space of a polyproline type II helix (P_{II}), which is often misinterpreted as α-helical and may have been modelled incorrectly (given that the model does not completely fit the density). D) False positives in a β-sheet (6AHU, residues 215–221 in chain B). The deposited model does not maintain the hydrogen bonding that defines a regular β-sheet; to the network, however, the fold still "looks" like a β-sheet and a third segment (top) is also assumed to be part of it.

means, but the map does not provide enough information to make this assignment.

Closer inspection reveals that false positives are often elements closely resembling helices, sheets or RNA/DNA (see Figures 1, 2, and 3). In particular, semi-helical structures, β-hairpin turns, and residues belonging to polyproline type II (P_{II}) helices[32] are misclassified as α-helical, and loosely parallel structures without the typical hydrogen-bond pattern are frequently misclassified as β-strands. In the case of P_{II} helices, this is partly due to the STRIDE-like annotation. It would be very desirable to quantify the false positives in this respect, but this was not possible within the scope of this work, as no automatic annotation algorithms seem to exist for such cases. For the future development of Haruspex, predicting additional classes, such as β-turns, polyproline helices, and perhaps even membrane detergent regions would be very desirable, as this would potentially lower the number of incorrectly identified secondary structure elements, while at the same time supplying additional information to users.

Resolution Range and Comparison to Similar Algorithms

Haruspex was trained for average resolutions as low as 4 Å, and the median resolution of published cryo-EM maps is improving every year, and has been better than 4 Å since 2017 (see Figure S5 in the Supporting Information). Irrespective of this, we will extend our approach to lower resolution data in the future, where our automated annotations should be even more advantageous for users. Still, low resolution experimen-

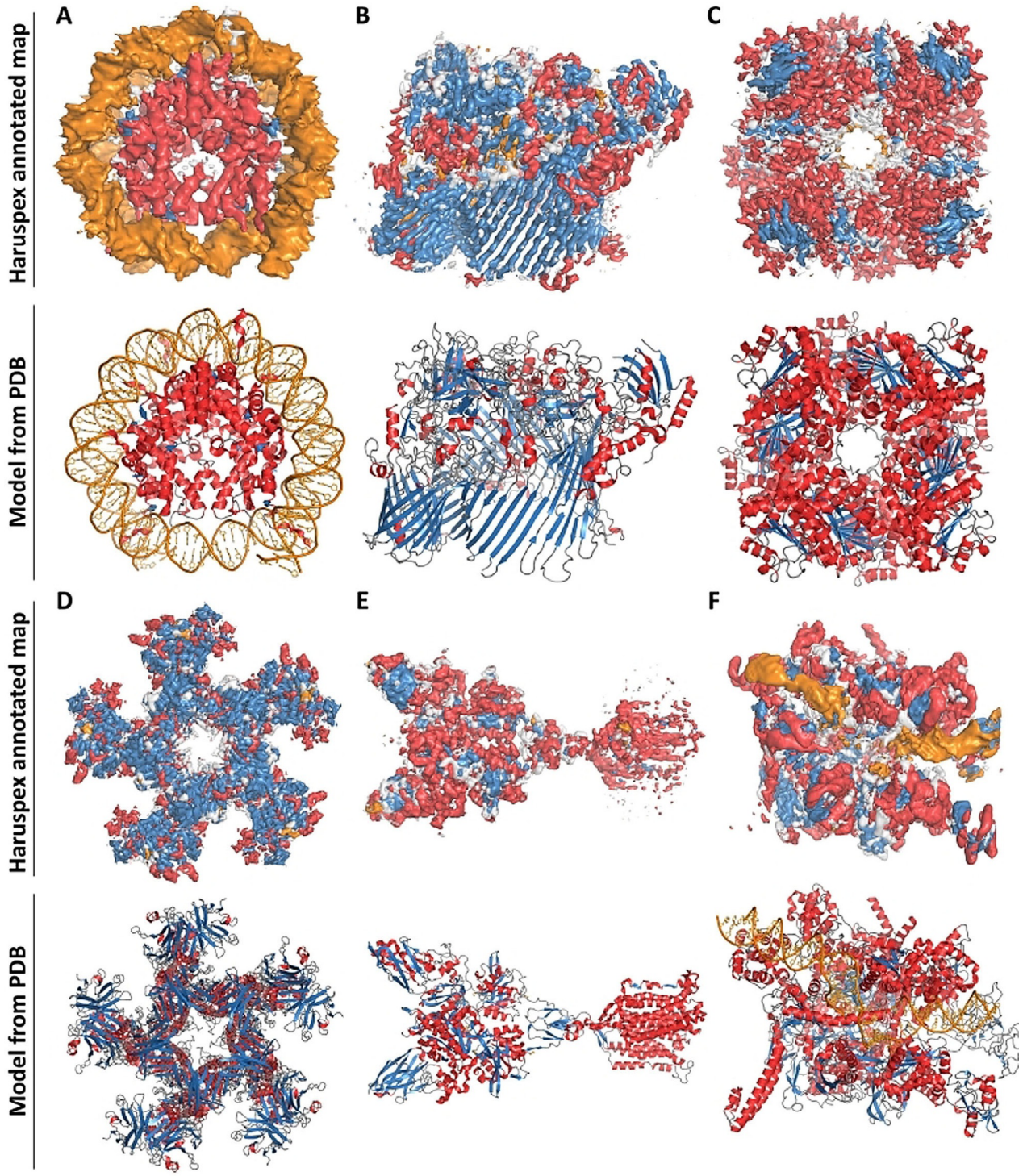

Figure 3. Additional examples from the test set. Top: Annotated map. Bottom**:** Deposited structure for comparison. Orange corresponds to RNA/DNA; red to helices; blue to sheets and grey regions were not assigned any secondary structure. A) Nucleosome from *Xenopus laevis*, average map resolution 3.8 Å (map: EMDB 4297, model: PDB 6FQ5): recall 98.5%, precision 94.0%. B) *Flavobacterium johnsoniae* Type 9 protein translocon, average map resolution 3.5 Å (map: EMDB 0133, model: PDB 6H3I): recall 96.3%, precision 49.3%. C) Leucine dehydrogenase from *Geobacillus stearothermophilus*, average map resolution 3.0 Å (map: EMDB 9590, model: PDB 6ACF): recall 89.8%, precision 85.7%. D) *Escherichia coli* Type VI secretion system, average map resolution 4.0 Å (map: EMDB 9747, model:PDB 6IXH): recall 95.9%, precision 70.9%. E) *Homo sapiens* metabotropic glutamate receptor 5, average map resolution 4.0 Å (map: EMDB 0345, model: PDB 6N51): recall 95.9%, precision 71.7%. F) Bacterial RNA polymerase-sigma54 holoenzyme transcription open complex, average map resolution 3.4 Å (map: EMDB 0001, model: PDB 6GH5): recall 94.2%, precision 67.5%.

tal maps with a well-matching model for training and testing such a network are difficult to obtain. This obstacle has previously been faced by Si et al.[33] (SSELearner), Li et al.[23] and Subramaniya et al.[25] (Emap2Sec) who developed machine learning approaches for protein secondary structure prediction in cryo-EM maps, but not oligonucleotides,[23] and consequently resorted partly to simulated maps generated with pdb2mrc.[34] These simulated maps lack the error structure and processing artefacts found in experimentally derived reconstruction densities,[4–6] as they assume a perfectly processed data set of a homogenous sample where all atoms interact with the electron beam as if they were uncharged and unbound. Si et al. tested their support vector machine on 10 simulated maps of relatively small structures (less than 40 kDa) and, as available data were still very limited in 2012, only 13 experimental maps paired with individually selected training maps. Haslam et al.[24] used a 3D U-Net, which was trained on 25 simulated and 42 experimental maps between 3–9 Å resolution to predict helices and sheets obtaining an F_1 score $2(\text{recall}^{-1}+\text{precision}^{-1})^{-1}$ between 0.79 and 0.88. However, the network was only tested on six simulated maps and one experimentally derived map. We, on the other hand, used a total of 293 experimentally derived maps in a semi-automated workflow to provide a more realistic training environment. Furthermore, the amount of newly released high-resolution structures in conjunction with our processing infrastructure permitted us to test our network performance on a representative set of 122 unique depositions. The semi-automated workflow for the selection and annotation of training data (see the Methods section of the Supporting Information) allows for an easy expansion of ground truth data and re-training. However, given that Haruspex has already been trained on a diverse range of macromolecular structures, the network can be used to interpret any map at 4 Å or better without any additional (re-)training necessary.

Augmentation of Automatic Model Building

Haruspex ideally complements tools for automatic map-based structure building, such as MAINMAST,[36] RosettaES,[37] ARP/wARP,[38] phenix.map_to_model[39] or Buccaneer[40] by providing an independent method to locate secondary structure elements of proteins to assist the validation of an automatically built protein main-chain. Haruspex may even be employed in the future to serve as starting point for such methods. The ability of Haruspex to automatically recognize RNA/DNA is of particular interest for the analysis of ribosomes, spliceosomes, and polymerases, which all contain substantial amounts of oligonucleotides. As these and similar structures are among the most common specimens studied by single-particle cryo-EM, Haruspex, which, to our knowledge, is the first to use machine learning for the identification of nucleotides in cryo-EM reconstruction maps, offers a unique advantage for the analyses of these structures.

Conclusion

We demonstrate that a neural network can be used to automatically distinguish between nucleic acids and protein and to assign the two main protein secondary structure elements in experimentally derived cryo-EM maps. This technique will render the process of protein structure determination faster and easier. Haruspex was trained on a carefully curated ground truth dataset based entirely on experimental data from the EMDB. The pre-trained network can be straightforwardly applied to annotate newly reconstructed cryo-EM density maps. Besides guidance for domain placements, the network also proves useful for model validation during building due to its high median recall and precision rates of 95.1 % and 80.3 %, respectively, as has been demonstrated by early users at our institute, for example in the modelling of the mycobacterial type VII secretion system.[2] The newest version of Haruspex is online available at https://github.com/thorn-lab/haruspex and will be distributed as part of CCP-EM.[35] We plan to refine and adapt the network as new data become available, and extend the approach to lower resolution and more structural classes in the future.

Acknowledgements

We would like to thank Bettina Böttcher, Niko Grigorieff, Jane Richardson, Christopher Williams, Paul Emsley, Tom Burnley, and Jola Mirecka for fruitful discussions; Virginie Uhlmann and the image analysis journal club for helpful comments on the preprint; and Bernhard Fröhlich for great computational support. This work was supported by the Deutsche Forschungsgemeinschaft [DFG project TH2135/2-1]; the High Performance Computing Cloud of Würzburg University [DFG project 327497565] and the Rudolf Virchow Center for Experimental Biomedicine. Funding for publication charges: German Federal Ministry of Education and Research [05K19WWA].

Conflict of interest

The authors declare no conflict of interest.

Keywords: DNA structures · electron microscopy · neural networks · protein structures · RNA structures

[1] M. Sevvana, F. Long, A. S. Miller, T. Klose, G. Buda, L. Sun, R. J. Kuhn, M. G. Rossmann, *Structure* **2018**, *26*, 1169 – 1177.e3.
[2] N. Famelis, A. Rivera-Calzada, G. Degliesposti, M. Wingender, N. Mietrach, J. M. Skehel, R. Fernandez-Leiro, B. Böttcher, A. Schlosser, O. Llorca, et al., *Nature* **2019**, 1 – 21.
[3] J. Frank, R. K. Agrawal, *Nature* **2000**, *406*, 318 – 322.
[4] P. B. Rosenthal, *IUCrJ* **2019**, *6*, 3.
[5] R. A. Nicholls, M. Tykac, O. Kovalevskiy, G. N. Murshudov, *Acta Crystallogr. Sect. D* **2018**, *74*, 492.

[6] P. V. Afonine, B. P. Klaholz, N. W. Moriarty, B. K. Poon, O. V. Sobolev, T. C. Terwilliger, P. D. Adams, A. Urzhumtsev, *Acta Crystallogr. Sect. D* **2018**, *74*, 814 – 840.
[7] D. C. Ciresan, U. Meier, J. Masci, J. Schmidhuber, *The 2011 International Joint Conference on Neural Networks* **2011**, 1918 – 1921.
[8] D. C. Cireşan, A. Giusti, L. M. Gambardella, J. Schmidhuber, *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2013*, Springer, Berlin, **2013**, pp. 411 – 418.
[9] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, O. Ronneberger, *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, **2016**, pp. 424 – 432.
[10] T. Falk, D. Mai, R. Bensch, Ö. Çiçek, A. Abdulkadir, Y. Marrakchi, A. Böhm, J. Deubner, Z. Jäckel, K. Seiwald, et al., *Nat. Methods* **2019**, *16*, 67—70.
[11] W. Jiang, M. L. Baker, S. J. Ludtke, W. Chiu, *J. Mol. Biol.* **2001**, *308*, 1033 – 1044.
[12] Y. Kong, X. Zhang, T. S. Baker, J. Ma, *J. Mol. Biol.* **2004**, *339*, 117.
[13] M. L. Baker, T. Ju, W. Chiu, *Structure* **2007**, *15*, 7.
[14] E. Shelhamer, J. Long, T. Darrell, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2017**, *39*, 640 – 651.
[15] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. Corrado, A. Davis, J. Dean, M. Devin, et al., *TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems*, **2015**.
[16] C. I. Branden, J. Tooze, *Introduction to Protein Structure*, Garland Science, New York, **1999**.
[17] M. Tagari, R. Newman, M. Chagoyen, J.-M. Carazo, K. Henrick, *Trends Biochem. Sci.* **2002**, *27*, 589.
[18] H. Berman, K. Henrick, H. Nakamura, *Nat. Struct. Mol. Biol.* **2003**, *10*, 980.
[19] J. Westbrook, N. Ito, H. Nakamura, K. Henrick, H. M. Berman, *Bioinformatics* **2005**, *21*, 988 – 992.
[20] W. Kabsch, C. Sander, *Biopolymers: Original Research on Biomolecules* **1983**, *22*, 2577 – 2637.
[21] D. E. Tronrud, D. S. Berkholz, P. A. Karplus, *Acta Crystallogr. Sect. D* **2010**, *66*, 834 – 842.
[22] V. B. Chen, W. B. Arendall, J. J. Headd, D. A. Keedy, R. M. Immormino, G. J. Kapral, L. W. Murray, J. S. Richardson, D. C. Richardson, *Acta Crystallogr. Sect. D* **2010**, *66*, 12 – 21.
[23] R. Li, D. Si, T. Zeng, S. Ji, J. He, in 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), **2016**, pp. 41 – 46.
[24] D. Haslam, T. Zeng, R. Li, J. He, in Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, ACM, New York, NY, USA, **2018**, pp. 628 – 632.
[25] S. R. Maddhuri Venkata Subramaniya, G. Terashi, D. Kihara, *Nat. Methods* **2019**, *16*, 911 – 917.
[26] D. Si, S. A. Moritz, J. Pfab, J. Hou, R. Cao, L. Wang, T. Wu, J. Cheng, *Sci. Rep.* **2020**, *10*, 4282.
[27] P. Emsley, B. Lohkamp, W. G. Scott, K. Cowtan, *Acta Crystallogr. Sect. D* **2010**, *66*, 486 – 501.
[28] DeLano, Warren L., *PyMOL.*, Schrödinger, LLC, **2019**.
[29] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, T. E. Ferrin, *J. Comput. Chem.* **2004**, *25*, 1605 – 1612.
[30] M. G. Rossmann, M. C. Morais, P. G. Leiman, W. Zhang, *Structure* **2005**, *13*, 355 – 362.
[31] J. Zivanov, T. Nakane, B. O. Forsberg, D. Kimanius, W. J. Hagen, E. Lindahl, S. H. Scheres, *eLife* **2018**, *7*, e42166.
[32] S. A. Hollingsworth, P. A. Karplus, *Biomol. Concepts* **2010**, *1*, 271 – 283.
[33] D. Si, S. Ji, K. A. Nasr, J. He, *Biopolymers* **2012**, *97*, 698 – 708.
[34] G. Tang, L. Peng, P. R. Baldwin, D. S. Mann, W. Jiang, I. Rees, S. J. Ludtke, *J. Struct. Biol.* **2007**, *157*, 38 – 46.
[35] T. Burnley, C. M. Palmer, M. Winn, *Acta Crystallogr. Sect. D* **2017**, *73*, 469 – 477.
[36] G. Terashi, D. Kihara, *Nat. Commun.* **2018**, *9*, 1618.
[37] B. Frenz, A. C. Walls, E. H. Egelman, D. Veesler, F. DiMaio, *Nat. Methods* **2017**, *14*, 797 – 800.
[38] G. G. Langer, S. Hazledine, T. Wiegels, C. Carolan, V. S. Lamzin, *Acta Crystallogr. Sect. D* **2013**, *69*, 635 – 641.
[39] T. C. Terwilliger, P. D. Adams, P. V. Afonine, O. V. Sobolev, *Nat. Methods* **2018**, *15*, 905 – 908.
[40] K. Cowtan, *Acta Crystallogr. Sect. D* **2006**, *62*, 1002 – 1011.

Manuscript received: January 9, 2020
Revised manuscript received: March 11, 2020
Accepted manuscript online: March 18, 2020
Version of record online: May 11, 2020

