Journal of Microscopy

Journal of Microscopy, Vol. 277, Issue 1 2020, pp. 12–22 doi: 10.1111/jmi.12858
Received 1 August 2019; accepted 17 December 2019

Automatic identification of crossovers in cryo-EM images of murine amyloid protein A fibrils with machine learning

MATTHIAS WEBER*, ALEX BÄUERLE†, MATTHIAS SCHMIDT‡, MATTHIAS NEUMANN*, MARCUS FÄNDRICH‡, TIMO ROPINSKI† & VOLKER SCHMIDT*

*Institute of Stochastics, Ulm University, Ulm, Germany

†Visual Computing Group, Institute of Media Informatics, Ulm University, Ulm, Germany

‡Institute of Protein Biochemistry, Ulm University, Ulm, Germany

Key words. Amyloid fibril, convolutional neural network, electron microscopy, semantic segmentation.

Summary

Detecting crossovers in cryo-electron microscopy images of protein fibrils is an important step towards determining the morphological composition of a sample. Currently, the crossover locations are picked by hand, which introduces errors and is a time-consuming procedure. With the rise of deep learning in computer vision tasks, the automation of such problems has become more and more applicable. However, because of insufficient quality of raw data and missing labels, neural networks alone cannot be applied successfully to target the given problem. Thus, we propose an approach combining conventional computer vision techniques and deep learning to automatically detect fibril crossovers in two-dimensional cryo-electron microscopy image data and apply it to murine amyloid protein A fibrils, where we first use direct image processing methods to simplify the image data such that a convolutional neural network can be applied to the remaining segmentation problem.

Introduction

The ability of protein to form fibrillary structures underlies important cellular functions but can also give rise to disease, such as in a group of disorders, termed amyloid diseases (Chiti & Dobson, 2017). These diseases are characterized by the formation of abnormal protein filaments, termed amyloid fibrils, which deposit inside the tissue (Chiti & Dobson, 2017). These fibrils, or intermediate structural states that occur in the course of fibril formation, are detrimental to their surrounding tissue and underlie the formation of disease. Examples hereof are Alzheimer's or Parkinson's diseases (Chiti & Dobson, 2017) or the various forms of systemic amyloidosis (Nienhuis *et al.*, 2016). Many amyloid fibrils are helically twisted (Annamalai *et al.*, 2016), which leads in cases of fibrils with an anisotropic cross-section to periodic variations in the apparent width of the fibril, when observing amyloid fibrils using microscopy techniques like cryogenic electron microscopy (cryo-EM) (Schmidt *et al.*, 2015, 2016; Close *et al.*, 2018; Radamaker *et al.*, 2019). Due to the two-dimensional (2D) projection, parts of the fibril orthogonal to the projection plane appear narrower than parts parallel to the plane. The parts of small width are called crossovers.

The distance between two adjacent crossovers is an important characteristic for the analysis of amyloid fibrils, because it is informative about the fibril morphology and because it can be determined from raw data by eye. A given protein can typically form different fibril morphologies. The morphology can vary depending on the chemical and physical conditions of fibril formation, but even when fibrils are formed under identical solution conditions, different morphologies may be present in a sample. As the crossovers allow to define fibril morphologies in a heterogeneous sample (Annamalai *et al.*, 2016; Liberta *et al.*, 2019), detecting crossovers is an important first step in the sample analysis.

EM-reconstruction software like Relion (He & Scheres, 2017), cryoSPARC (Punjani *et al.*, 2017) or EMAN2 (Tang *et al.*, 2007) allows for picking of fibrils using templates. But these techniques are especially designed for cryo-EM structure determination of single particles and not for a statistical analysis of an entire fibril sample. So far, the detection of fibrils in cryo-EM image data for statistical analysis to determine fibril morphologies has often been performed by labelling the crossovers locations by hand and measuring parameters such as fibril lengths, crossover distances, widths and curvatures manually. However, for large datasets, only a small number of fibrils can be analysed this way (Annamalai *et al.*, 2016), because this is a time-consuming and error-prone task. In the present paper, we propose an approach for the automatic detection of crossovers in 2D image data obtained by cryo-EM based on a combination of conventional image processing

Correspondence to: Matthias Weber, Institute of Stochastics, Ulm University, 89069 Ulm, Germany. Tel: +49 (0)731/50-23590; fax: +49 (0)731/50-23649; e-mail: matthias.weber@uni-ulm.de

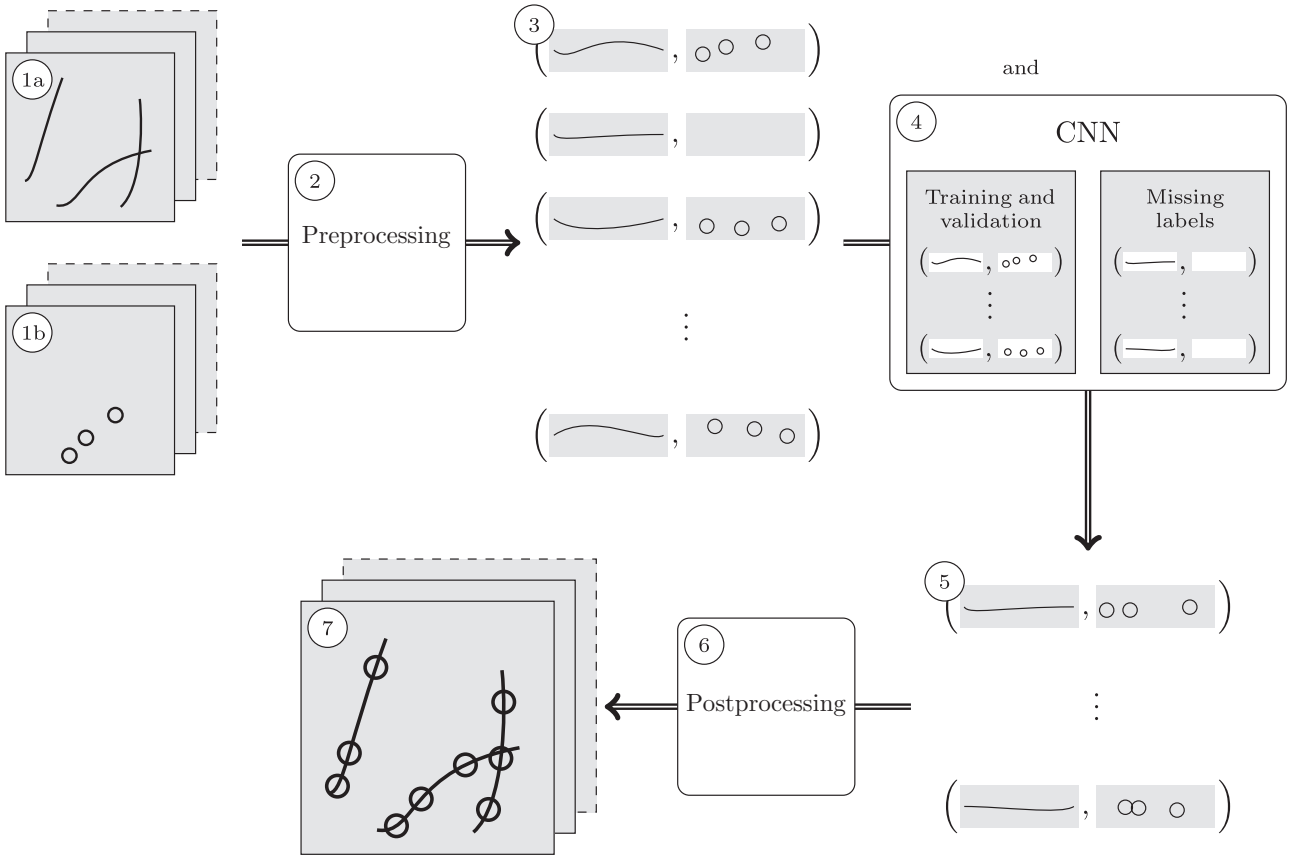

Fig. 1. Overview of the proposed methodology. The data used in our approach consist of cryo-EM images of fibrils (1A) and hand-labelled crossover locations (1B) for some of these fibrils. The presented preprocessing method (2) transforms this data into pairs (3) of extracted, realigned fibrils and crossovers, if available. Fibrils for which corresponding crossovers are known are used to train and validate a CNN (4). Applying the CNN to the remaining fibrils yields probable crossover locations for each fibril (5). Postprocessing (6) ensures that artefacts are removed and only valid crossovers remain in the final output (7).

methods with machine learning techniques. In contrast to existing tools, this method is specifically designed for the precise localization of crossover locations for the purpose of statistical analysis. Even for challenging data scenarios like overlapping fibrils and artefacts, entire fibrils (i.e. their crossovers) are correctly labelled.

The detection of specific locations in 2D images can be understood as an image segmentation task for which convolutional neural networks (CNNs) are often used. In particular, in recent years, encoder–decoder architectures (Ronneberger *et al.*, 2015; Liu *et al.*, 2018) were established for this kind of problems. However, CNNs, as all machine learning techniques, heavily depend on the presence of training data of sufficient quantity and quality. Although, in principle, it would be possible to obtain a suitable amount of hand-labelled image data of fibrils to perform successful training of a neural network, cryo-EM image quality poses challenges for this approach. Low contrast and image artefacts can make it infeasible to label some crossovers by hand. Furthermore, each image contains many possibly overlapping fibrils. Both problems lead to missing labels in the present data which make the direct training of a CNN impossible.

For further analysis of crossover locations, the quality of the detected crossovers may be even more important than the quantitative yield: Firstly, wrongly detected crossovers can obviously not be used further and would need to be removed labouriously. Secondly, the knowledge of crossover locations is especially useful when entire fibrils are labelled and no crossovers are missing for the labelled fibrils. A direct training of a CNN based on incomplete data where not all crossovers are labelled and regions with certainly no crossovers are not known would induce problems in the segmented data processed with this badly trained CNN.

Our approach therefore combines conventional tools of morphological image analysis (Soille, 2013) with machine learning techniques. The combination of these two methodologies has already proven valuable for different kinds of segmentation tasks (Petrich *et al.*, 2017; Furat *et al.*, 2019). Figure 1 shows an overview of the proposed approach which can also serve as an outline for this paper.

Thus, the rest of the paper is organized as follows. To begin with, we describe the data consisting of cryo-EM images of fibrils and hand-labelled crossovers for some fibrils and explain the basic problems when using this data. Then, we present a preprocessing method which is capable of extracting the rough shapes of fibrils from the given data. As this method does not rely on hand-labelled crossover locations, it can be applied to the whole dataset and serves to simplify the data and remove artefacts. In the next step, we introduce a CNN based on the U-Net architecture which is trained using the previously enhanced data. After applying this CNN to all of the preprocessed fibrils, we perform a final postprocessing step to remove wrongly detected crossovers. Finally, we present the results of applying the proposed technique to cryo-EM image data of fibrils. Furthermore, we assess its performance and compare it to a direct application of a CNN.

Data

Previously recorded cryo-EM 2D image data of murine AA amyloid fibrils (Liberta *et al.*, 2019) serve as the basis of our approach to the automated detection of crossovers. To obtain these images, fibrils were extracted from mice with systemic amyloidosis and applied in water onto a holey carbon coated grid, blotted and finally plunge-frozen into thin vitreous ice. A total of 1063 images (3838 px × 3710 px, pixel size: 1.36 Å) of the fibrils frozen in vitreous ice were collected at 300 kV in the transmission mode with a K2 summit (Gatan) direct electron detector. Each resulting image shows a different part of the sample and an unknown number of possibly overlapping fibrils of different lengths and shapes. For this sample, we previously showed that approximately 94% of all fibrils are of the same fibril morphology with a width at the widest point of 11.8 ± 0.5 nm and a crossover distance of 75.7 ± 1.3 nm (Liberta *et al.*, 2019). An example of the used image data is shown in Figure 2. Note that the fibrils are not evenly distributed over the images: Although there exist images showing barely any fibril, others show huge clumps of overlapping fibrils.

For a subset of 669 images, hand-labelling of crossovers has been performed in such a way that the positions of crossovers are entirely known for a total of 1069 fibrils. However, not all fibrils have been included in the hand-labelling and, more relevant, there exist no images in which all fibrils have been labelled. This is partly due to noise and overlapping fibrils which make it hard to hand-label some fibrils.

Thus, without further knowledge of the image structure, the hand-labelled data can only be used to determine crossover positions which are certain. For any nonlabelled region in the image data, we cannot directly conclude from the hand-labelled data if there might be a fibril or even a crossover. In the following, we present an approach to overcome these issues by first extracting the rough shapes of fibrils.

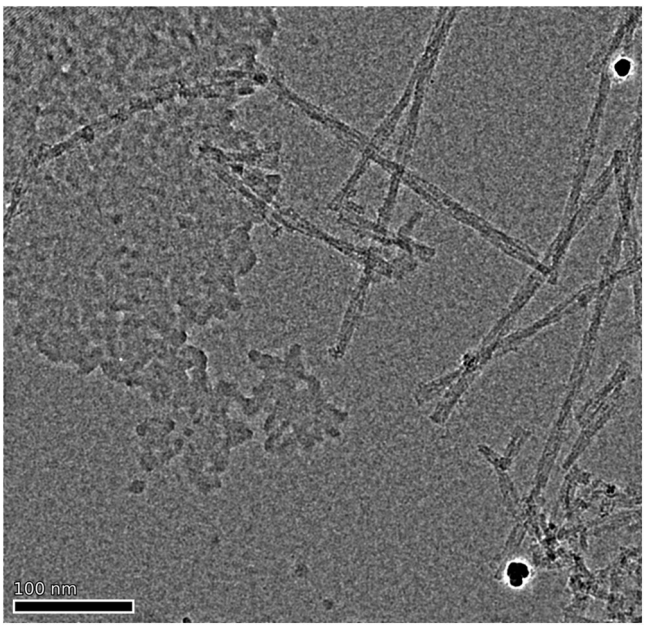

Fig. 2. Cryo-EM image of murine AA amyloid fibrils. Note that contrast and noise level have already been improved for visualization; fibrils would only be barely visible in the raw data.

Preprocessing

As already mentioned above, the raw data consist of 2D greyscale images of different (potentially overlapping in 2D projection) fibrils oriented mostly parallel to the projection plane, see Figure 3(A). Note that the raw image data are subject to heavy noise and low contrast. Furthermore, many fibrils overlap each other making a visual detection of crossovers practically impossible. Both problems lead to an incomplete labelling of crossovers in the images designated as training data, making a direct application of a CNN to the given data infeasible.

To overcome these issues, we first use a method for preprocessing the raw cryo-EM image data. This method splits the images into pieces showing single fibrils which are then used for the training of a neural network.

First steps in preprocessing include the removal of grey value gradients, which do not carry information but instead need to be considered as artefacts using a Gaussian high-pass filter, and a Gaussian smoothing for noise reduction. Furthermore, the grey value range of all images is set to a predefined scale by adjusting the mean and standard deviation of grey values. By this, we account for deviations in exposure and measurement of different images. The scale is chosen to accommodate 99.9% of the original range in an 8-bit image. Next, the rough shapes of the fibrils are extracted using a local thresholding approach, see Figure 3(D). The image representing the local threshold values is obtained by applying a Gaussian smoothing (Russ, 2011) with a relatively large standard deviation such that the shapes of fibrils persist only faintly. Additionally,

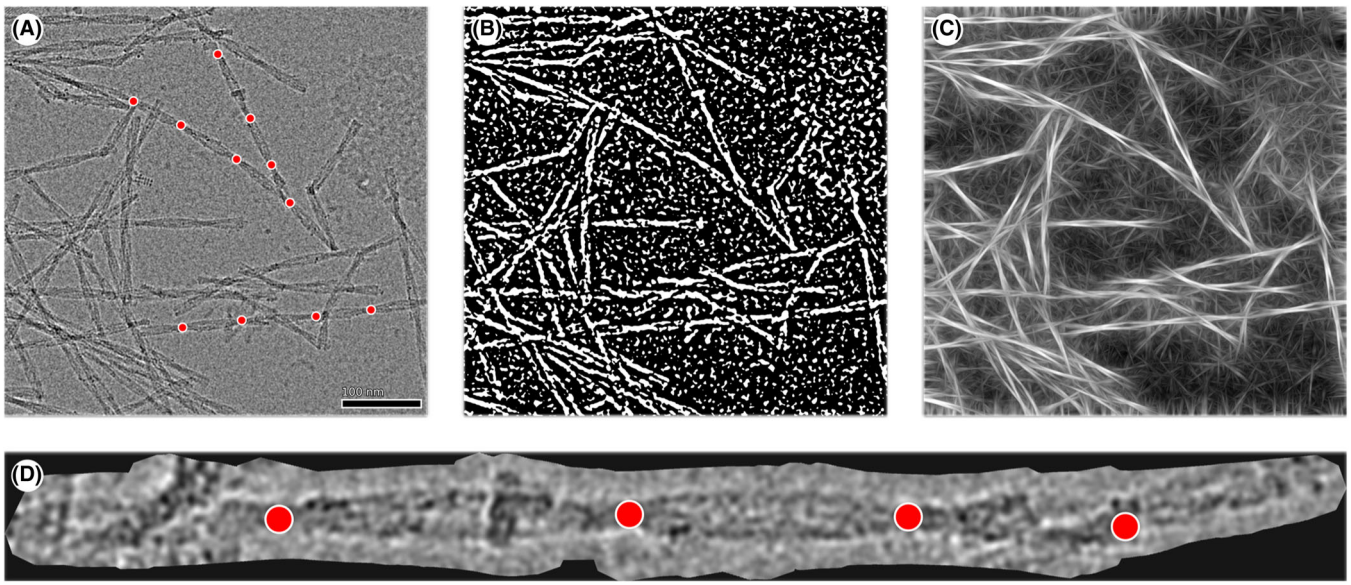

Fig. 3. Overview of the preprocessing steps. (A) Sample image from the histogram-normalized raw data and the hand-labelled positions of crossovers (red dots). (B) Rough fibril shapes. (C) Pointwise maximum of 60 convolutions of (B) with differently oriented line kernels. (D) Example of a detected fibril.

the local threshold values are multiplied by 0.97 and pixels are set to white in the resulting image if their grey value in the original image is below the corresponding threshold value, i.e. dark elements (fibrils) remain. The multiplicative correction of 0.97 is chosen empirically to improve the binarization. Yet, the result does contain a significant amount of artefacts and not yet a precise representation of the fibrils.

Thus, a method similar to the Hough transform (Duda & Hart, 1972) is applied to precisely extract the shapes of the fibrils.

This method is based on convolutions of the original image with suitably chosen kernels. Considering different kernels, convolutions can be used for various operations of image processing like sharpening or blurring (Russ, 2011). Moreover, convolutions can be used to detect objects of a known shape in an image, by choosing a kernel similar to the object of interest. This property is used in various situations of image analysis, e.g. for edge detection by employing a kernel depicting a strong gradient in the direction orthogonal to the desired edge. In a more general way, convolution with a kernel depicting a given object highlights the respective positions of all occurrences of the given object.

Note that this method of object detection is robust with respect to noise and minor deviations in the object's shape and still works when the object is partly covered by other objects. This is important for application to our problem as fibrils can overlap in the given image data.

We apply this method of object detection to our data using kernels which depict straight lines with 60 different orientations equally spaced in $\alpha \in [0, \pi]$. For any fixed α, the convolution highlights fibril segments which are oriented accordingly, see Figure 4(B). By applying a simple global thresholding to this greyscale image, the desired fibril segments can be extracted.

However, some postprocessing still needs to be performed to separate real fibril segments from artefacts, see Figure 4(C). Filtering the connected components of the obtained binary image using an area threshold, i.e. only keeping components which are larger than a given size, and a morphological closing/opening (see Soille, 2013) with a line segment oriented in the direction α as structuring element solves this problem reasonably well. Finally, the direction of the principal axis β of each region in the binary image is computed via principal component analysis (PCA, Lee *et al.*, 2006). While PCA and other statistical tools are often used in the cryo-EM context to obtain distinguishing features of a class of samples (Heel & Frank, 1981), we employ PCA solely to detect the geometrical orientations of single regions in the binary image. For these, the principal axis is given by the major axis of the ellipse which best fits the given region and corresponds to the perceived orientation of the extracted elongated regions. Only regions for which the predicted orientation α and the orientation β computed by PCA coincide to some extent are considered valid fibril segments.

Because the procedure described above may split fibrils into multiple fragments, the next step merges regions belonging to the same fibril, see Figure 5. Therefore, the previously described convolution-based extraction of fibril segments is performed for a certain (finite) number of directions equally spaced in the interval $[0, \pi]$. For each extracted fibril segment F_i, the calculated orientation β_i obtained from PCA is stored. Now, for each pair of fibril segments F_i, F_j, the angular deviation in orientation $\Delta_{i,j} = \min(|\beta_i - \beta_j|, |\beta_i - \beta_j - 2\pi|)$ and the

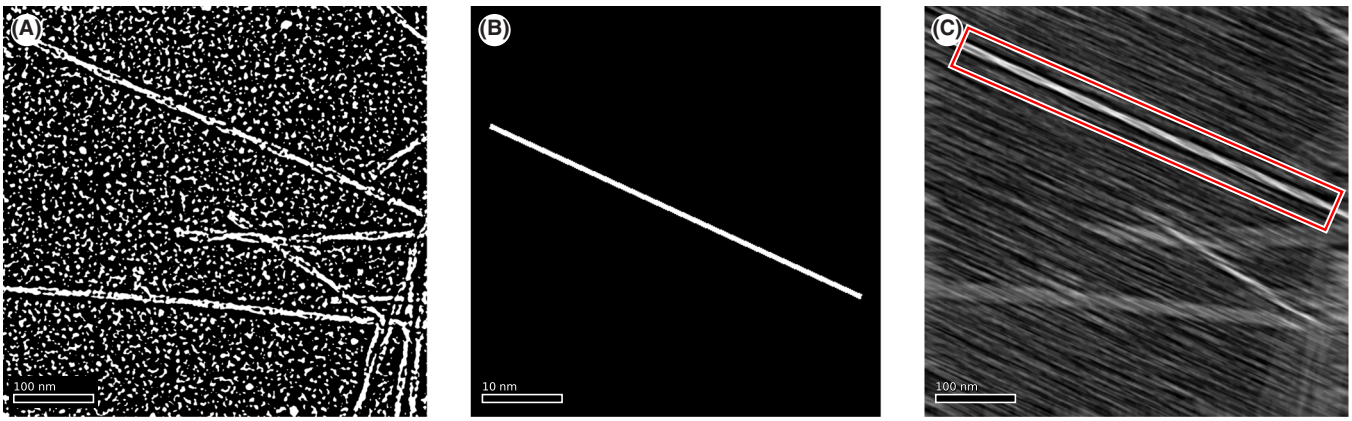

Fig. 4. Sample image visualizing the result (C) of the convolution of a binary image (A) with the shown kernel (B, not to scale). Fibril segments of any length which are oriented similar to the kernel are clearly visible in the result of the convolution (see the highlighted area) and can easily be separated from the background using a global threshold. Note that for curved fibrils, this does only extract segments of the fibril for each given orientation. These parts are then combined in a later step.

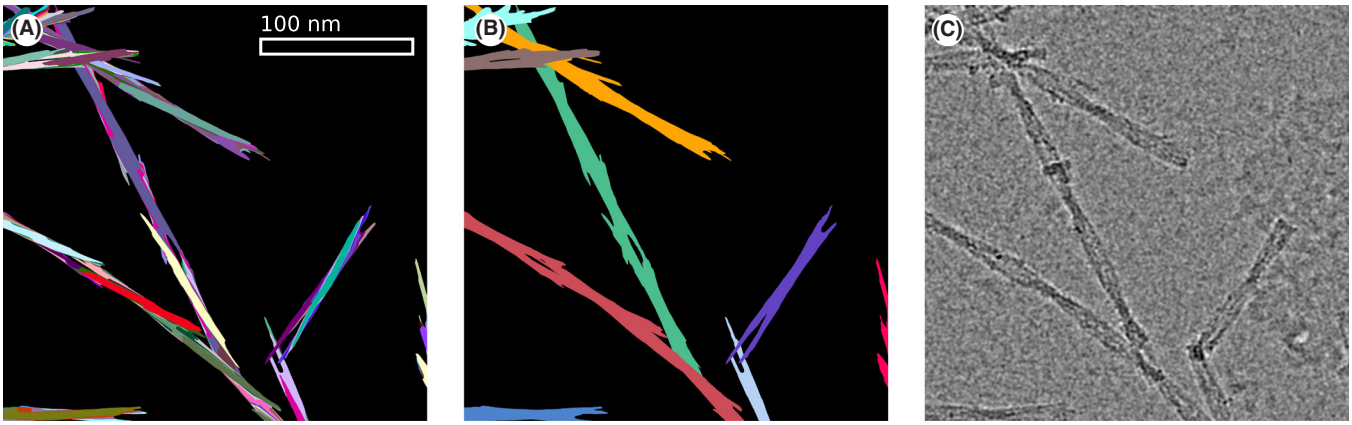

Fig. 5. (A) Fibril segments detected using kernels with different orientations. Each colour corresponds to a specific orientation. (B) Combination of many detected (overlapping) segments results in accurate detection, compare the original image (C). Note the curved red fibril in the lower left part of panel (B) which was combined from segments with considerably different orientations.

relative overlap $\cap_{i,j} = \lambda(F_i \cap F_j)/\min(\lambda(F_i), \lambda(F_j))$ are determined, where $\lambda(F)$ denotes the area of a fibril segment F. Then, fibril segments with relative overlap $\cap_{i,j} > 0.4$ and deviation in orientation $\Delta_{i,j} < 10^\circ$ are considered to belong to the same fibril. These numbers are based on the angular step size of 3° used for detection of fibril segments. As the measured orientations of the regions can be hugely affected by noise and inaccuracies, we allow for a fairly large deviation. Furthermore, based on visual inspection, 10° are sufficient to capture all nonbroken, curved fibrils encountered in the data.

Assessing the thereby detected regions by size and shape gives a reasonably good procedure for the final elimination of wrongly detected fibrils, as detailed in section 'Validation'. This leaves us with a set of fibrils for each input image. To simplify the process of further analysis, the cut-outs representing each extracted fibril (and hand-labelled crossovers) are rotated such that the fibrils are horizontally oriented, see Figure 3(D).

A subset of all fibrils for which hand-labelled crossover locations are known is then used to train a neural network as described in the next section.

CNN-empowered crossover detection

Roughly speaking, we now try to predict unknown crossover locations in 2D image data of fibrils employing the information we obtained from the hand-labelled data described in the previous section. A direct processing of the fibril shapes extracted in the previous steps using traditional morphological methods did not prove successful due to inaccuracies in the detected shapes, see Figure 5(B). However, the methods of statistical learning, in particular CNNs, provide a promising technique with regard to automated image analysis. If enough training data are available, they have shown to be successful in many domains, including microscopy (Dong *et al.*, 2015; Kraus *et al.*,

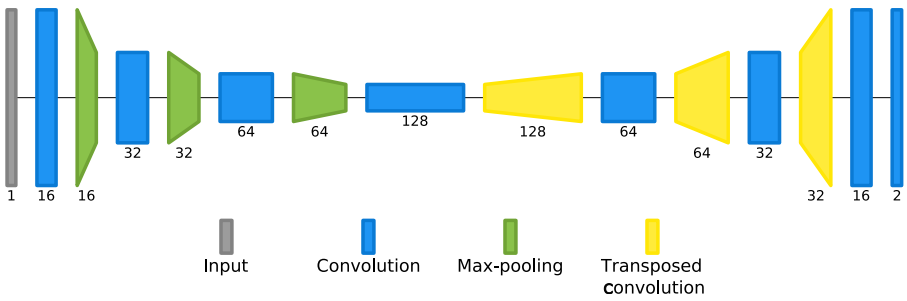

Fig. 6. The neural network architecture we use for segmenting fibril images. It transforms the input image into a 128-dimensional feature vector, for which the spatial resolution is reduced by a factor of eight. In the following, this vector is utilized to produce the two-dimensional segmentation map used as the crossover prediction. Due to its fully convolutional nature, the input and output sizes of the network are arbitrary.

2016; Petrich *et al.*, 2017; Rivenson *et al.*, 2017; Furat *et al.*, 2019). Integrating these techniques into the analysis process of fibril images is a promising approach that could drastically reduce the efforts needed to process these images using completely interactive approaches alone.

Description of the CNN architecture

The task of automatically detecting crossovers in image data of fibrils can be seen as a classical image segmentation problem. For this type of problem, fully CNNs are well established, especially when dealing with complex data as given by the varying appearance of crossovers in microscopic image data. Fully CNNs use repeated layers of convolution of the input image with trainable kernels. To automatically detect crossovers in microscopy images of fibrils, we employ an encoder–decoder CNN. These types of networks, which have proven to be successful in related problem contexts (Ronneberger *et al.*, 2015; Liu *et al.*, 2018), transform the input image into a latent space, where, in our case, abstract representations of the fibril images are obtained at lower resolution, before reprojecting them onto the original resolution of the image. This kind of procedure is desirable as the aim is to predict pointwise probabilities for each pixel to be a crossover. The network we use divides fibril images into crossover and noncrossover pixels. The basic idea of our architecture is shown in Figure 6.

Training of the CNN

Fully convolutional networks like the architecture considered in the present paper are not designed for the detection of single points but instead for the segmentation of larger regions. This is partly due to the pooling and upsampling layers which cause high correlation between values of neighbouring pixels in the resulting image. Moreover, when individual pixels are labelled as crossover points, while the entire rest of the image is labelled as noncrossover, high accuracy can be reached by simply labelling the whole image as noncrossover. Thus, the network would optimize to classifying any input image entirely as noncrossover.

To circumvent these limitations of the learning process, we carefully choose the data used for training the neural network. As shown in Figure 7, we take small cut-outs of each image containing a horizontally aligned fibril. These cut-outs can either contain exactly one crossover or no crossover. Finding regions containing exactly one crossover is straightforward and is done by taking square regions of a given size around all hand-labelled crossovers. Note that the regions are chosen at random while still containing the crossover, resulting in patches featuring crossovers at different locations. This prevents the neural network from learning to predict crossovers simply based on their location.

Even though the hand-labelled data may be incomplete (as illustrated in Figure 7), by adjusting the size, we can guarantee that only one crossover is present in each cut-out if no other fibrils cross the present fibril in the given cut-out. Selecting regions which do not contain any crossovers requires some extra effort due to potentially missing labels. However, hand-labelling was performed such that no additional nonlabelled crossover lies between two hand-labelled crossovers. Thus, as shown in Figure 7, we can use any cut-out of the fibril which lies entirely between two labelled crossovers as further training data.

By this, we get two sets of square cut-outs of fibril images. The first set ('positive samples') consists of regions which contain exactly one crossover whose position is known from hand-labelled data. The second set ('negative samples') consists of regions which contain no crossover at all. Together, the two sets make up the input data for the training of the neural network. Furthermore, we require appropriate ground truth data for all cut-outs contained in the input data. For the 'negative samples', the ground truth is simply a black image of appropriate size representing the absence of crossovers. For the 'positive samples', we take a black image of appropriate size and place a white ellipse at the known position of the corresponding crossover. The ellipse's major axis is chosen parallel to the fibril whose orientation is known from preprocessing. By

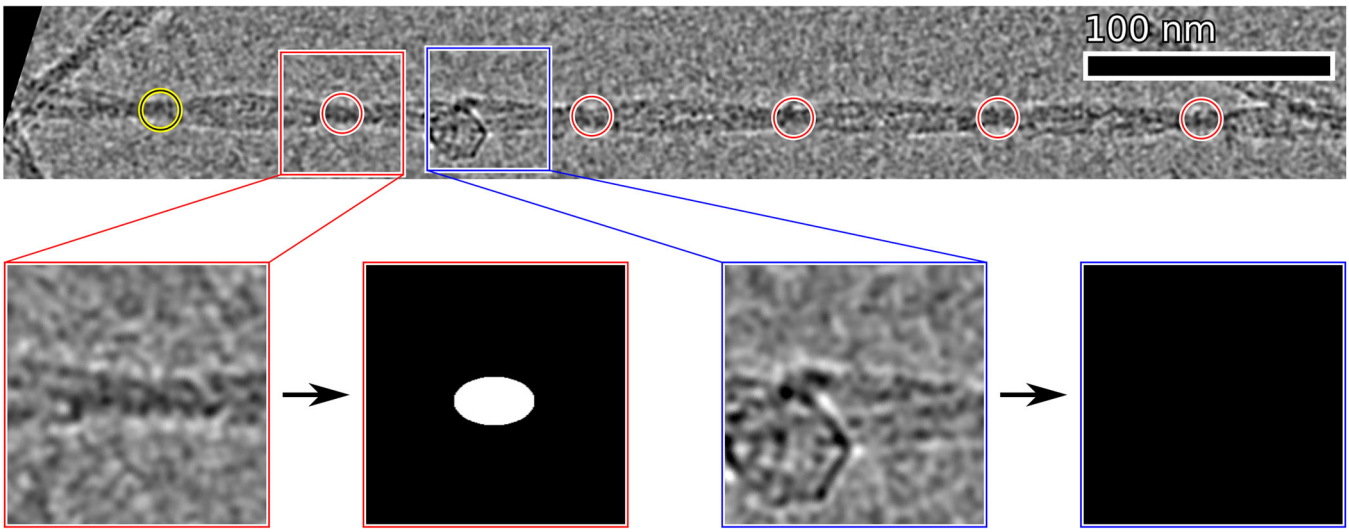

Fig. 7. Splitting up the available data into patches for training the neural network. The top row shows parts of a hand-labelled fibril after preprocessing. Red circles mark hand-labelled crossovers which can be used to cut out (square) patches for training. Between two hand-labelled crossovers, we can be sure that no further crossovers exists, so these regions can be used as negative samples. The black–yellow circle highlights a crossover which was not hand-labelled. The bottom row shows two pairs (input/ground truth) of training data for the neural network extracted from this fibril.

adjusting the size of the cut-outs and the size of the ellipses, we can account for class imbalance.

We use the data of input and ground truth images given by the method described above to train the neural network described in the previous section. For assessing the quality of a predicted output compared to the ground truth data, we need to define a so-called loss function which assigns a loss (i.e. a value specifying how 'good' the prediction is) to a pair of data (i.e. the CNN's output and the ground truth). We chose the mean cross-entropy loss (Goodfellow *et al.*, 2016) which operates on the output image $I_O : \{1, \dots, n_x\} \times \{1, \dots, n_y\} \times \{1, 2\} \to [0, 1]$ of the neural network prior to thresholding. Recall that this image represents the probabilities of each pixel belonging to a crossover or not. The corresponding ground truth data $I_T : \{1, \dots, n_x\} \times \{1, \dots, n_y\} \times \{1, 2\} \to \{0, 1\}$ take values 1 (the pixel belongs to a crossover) or 0 (the pixel does not belong to a crossover). The cross-entropy function is then defined for each pixel (x, y) and channel c by

$$L_{CE}(x, y, c) = -I_T(x, y, c) \log(I_O(x, y, c)) .$$

The total loss of an output image is just the mean $1/(n_x n_y) \sum_{x=1}^{n_x} \sum_{y=1}^{n_y} (L_{CE}(x, y, 1) + L_{CE}(x, y, 2))$ which equals 0 if $I_O = I_T$. The training of the neural network is performed using this loss function and the Adam optimiser (Kingma & Ba, 2015).

Due to the fully convolutional architecture, the (trained) network can operate on arbitrarily-sized patches to perform inference on unseen data. While training is performed on small image patches for computational reasons, the trained network can be used to detect crossovers on entire fibrils.

Application of the trained network

For the automatic detection of crossover locations, we apply the trained neural network to single fibrils which have been extracted from the original data using the preprocessing method described above, see also Figure 3(D).

While crossovers on most fibrils are correctly identified, the preprocessing method described above does propose some areas of the images as fibrils which are not suitable as input for the neural network, see Figure 8. On the one hand, some areas of the images are proposed as fibrils which, at visual inspection, are clearly noise and do not contain fibrils. On the other hand, fibrils which are crossed by many other fibrils are (correctly) detected by the preprocessing steps. While these do indeed contain crossovers, it is almost impossible to distinguish between crossovers and artefacts introduced by crossing fibrils. In both cases, the neural network cannot be expected to produce valid results.

Thus, we try to eliminate wrongly labelled crossovers in a final postprocessing step, see below.

Postprocessing

Recall that our aim is to obtain a set of fibrils whose crossovers are entirely labelled. For postprocessing, we thus consider the entire fibrils which have been extracted from the original image data using the preprocessing methods described above. A fibril should be classified as correctly labelled if all of its crossovers are correctly labelled and no additional points are labelled as crossovers. Missing or wrongly placed labels would lead to inaccurate further analyses. Thus, we develop a method to correct minor errors and detect wrongly labelled fibrils. This

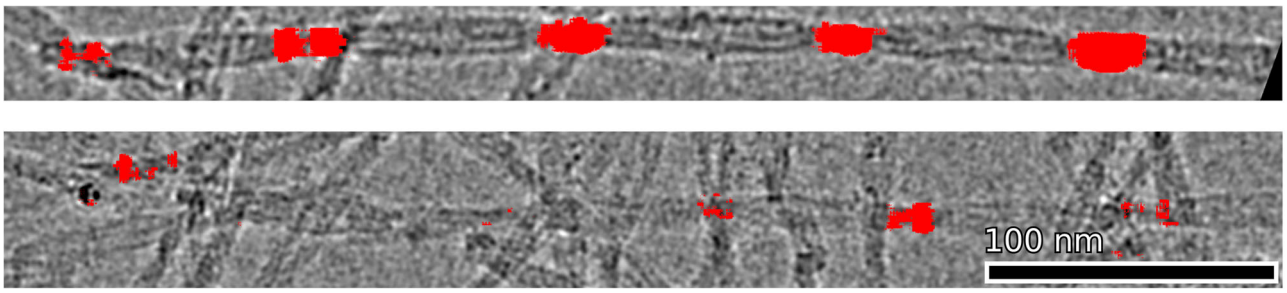

Fig. 8. Top: A correctly extracted fibril with precisely detected crossovers. Bottom: A fibril which is not clearly visible, resulting in a bad performance of the neural network.

includes the cases where the region proposed by the preprocessing method does not contain any (clearly visible) fibril as well as the cases where the neural network does not perform correctly. A sample of correctly and wrongly labelled fibrils is shown in Figure 8.

First, the detected crossovers are assessed by size of the detected region. Based on visual inspection of a small fraction of detected regions, it seems reasonable to assume that larger regions are associated with a more reliable detection. Thus, we apply a morphological closing (Russ, 2011) to eliminate minor noise and then remove small regions using another morphological opening. Due to the helical shape of a fibril, crossovers on a single fibril should form an approximately periodical pattern. The following method uses this information to classify correctly labelled fibrils. Using the known orientation of the fibrils—which, after preprocessing, is horizontal—we interpret the detected crossover locations as points $X = (x_1, \ldots, x_n)$, $x_i < x_{i+1}$ on a straight line. If the crossovers are detected correctly, they should form a semiperiodic pattern, i.e. $x_{i+1} - x_i \approx d$ for some constant crossover distance d. To find the value of d, we define the functions

$$f_c(t) = \sum_{i=1}^{n} \varphi(x_i, \sigma^2; t)$$

and

$$f_p(d_0, d; t) = \sum_{i=0}^{n-1} \varphi(d_0 + id, \sigma^2; t),$$

where $\varphi(\mu, \sigma^2; \cdot)$ is the probability density function of a Gaussian random variable with mean μ and variance σ^2. The function f_p corresponds to a proposed crossover distance d and offset d_0. For some assumed crossover pattern X and a proposed crossover distance d and offset d_0, these functions are visualized in Figure 9. By minimizing the difference Δ given by

$$\Delta = \int_{-\infty}^{\infty} \left| f_c(t) - f_p(d_0, d; t) \right| \mathrm{d}t,$$

employing a grid search, we obtain a prediction for the real crossover distance and offset. This prediction is robust to

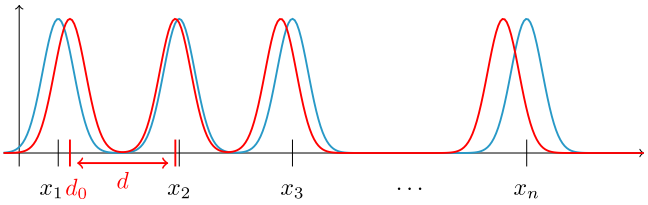

Fig. 9. Visualization of the superimposed Gaussian kernels used for estimating the crossover distance d and offset d_0 from the x-positions of the detected crossovers $x_1, \ldots, x_n$.

wrongly detected crossover locations which do not fit into the regular pattern: If three or more correctly detected consecutive crossovers are present, the integral will still be minimal for the correct distance and offset, even if another wrongly placed crossover is introduced.

Using the predicted offset d_0 and distance d, we can compute all possible locations of crossovers $d_0 + jd$. We generate a new set of crossovers by taking all detected crossovers x_i whose distance $\min_{j \in \mathbb{N}} |x_i - d_0 + jd|$ to the set of possible locations is below some heuristically chosen threshold between 10 nm and 25 nm. The precise value of this threshold proved irrelevant for the performance of our method. The new set of crossovers shows an approximately periodic pattern and can be used for further analysis.

Validation

For a validation, we applied the proposed approach to the data described in the previous sections. Before assessing the performance, let us recall the major goal of the approach, which is the extraction of fibrils and the detection of corresponding crossover locations from the image data. More precisely, each image patch returned by the postprocessing should show a horizontally aligned fibril and the corresponding predicted crossover locations which pass postprocessing should be correct. Furthermore, no crossovers should be missing. For simplicity, we will call image patches satisfying all three requirements *entirely labelled fibrils*. We will call image patches and the corresponding predicted crossover locations which are returned by the postprocessing steps

Table 1. Overview of the number of hand-labelled and not hand-labelled fibrils included in each processing step.

	Hand-labelled	Not hand-labelled	Total
Hand-labelled data (total)	1069	–	–
Proposed by preprocessing	743	1010	1753
Of which are actual fibrils	743	ca. 850	ca. 1592
Used for training	670	0	670
Used for validation	73	0	73
Accepted by postprocessing	393	179	572
Used for training	353	0	353
Postprocessed & not Used for training	40	179	219
Entirely labelled	39	ca. 140	ca. 179

Note: The total number of fibrils present in the data is unknown. Among the 1069 hand-labelled fibrils, 326 fibrils were not identified by the preprocessing. The number of included fibrils decreases with each processing step as possibly flawed data are removed from the set of considered fibrils. For not-hand-labelled fibrils, the approximate numbers are obtained extrapolating data from a visual inspection of a subset of fibrils.

postprocessed fibrils. Note that our main goal is met if all (or most) postprocessed fibrils are indeed entirely labelled fibrils, similar to avoiding false positives in a classification setting.

However, comparing these two sets does not yet measure the total yield of our approach. To assess this, we take into account the (total number of) *hand-labelled fibrils*, similar to the quantification of false negatives, as well as the image patches proposed by the preprocessing method which were used to train and apply the CNN. We will call these image patches *proposed fibrils*.

Note that hand-labelled crossovers corresponding to some of the entirely labelled fibrils, or postprocessed fibrils, respectively, were used to train the neural network. Thus, crossover data on these fibrils (image patches) cannot be used to assess the performance of our approach. Table 1 gives an overview of the data used for validation. We will discuss the detailed numbers of correctly detected fibrils and crossovers in the following, using different characteristics to assess the performance of the proposed approach, based on hand-labelled data.

From these considerations, the most important accuracy measure is the *precision* of the method consisting of, first, detection of fibrils; second, crossover detection on actual fibrils; and third, detection of entirely labelled fibrils. That is, which fraction of detected objects are correctly classified. Increasing the precision corresponds to decreasing the rate of false positives.

Performance of fibril detection

We applied the proposed methods to data from 1063 cryo-EM images. In 669 of these images, crossovers on a total of 1069 fibrils were labelled by hand. Based on Table 1, the preprocessing detected 743 out of 1069 hand-labelled fibrils (69.5%). However, we find additional fibrils, which have not been included in hand-labelling. Visual inspection of the proposed fibrils which were not contained in the hand-labelled data shows that approximately 850 out of 1010 (84.2%) additionally proposed image patches actually are fibrils. Thus, the total *precision* of fibril detection is approximately 90.9%. As missing hand-labels are mostly due to difficulties in visually assessing the structure of a fibril, this means that our approach is capable of processing poor-quality data. However, this does not imply that detection of crossovers is achievable for the majority of detected not hand-labelled fibrils. For many of the proposed fibrils which have not been hand-labelled, it can be hard to locate crossovers due to artefacts or overlapping fibrils. This is reflected in postprocessing accepting only 179 out of 1010 of these fibrils and corresponding detected crossovers (17.7%).

Performance of crossover detection

To validate the performance of the crossover detection performed by the considered CNN and subsequent postprocessing steps, we take into account the number of proposed fibrils and fibrils (and corresponding detected crossovers) which pass postprocessing. To validate the method in its entirety, we have to exclude fibrils which have been used for the training of the CNN. This leaves as set of postprocessed fibrils which have not been used for training. On 39 of these 40 fibrils, crossovers have been entirely labelled if we allow a deviation of 20 nm (compare to the crossover distance 75.7 ± 1.3 nm) between detected and hand-labelled crossovers, see Figure 10. For a deviation of 10 nm, we get 67.5% entirely labelled fibrils. This is related to the total number of correctly labelled crossovers on these 40 fibrils which is also shown in Figure 10. Thus, the labelling of fibrils with complete sets of crossovers has a precision of 97.5%. Without postprocessing, this number decreases slightly to around 66 of 73 fibrils or 90.4%.

However, these numbers are only true under the assumption that the objects detected by preprocessing actually are fibrils. Relaxing this assumption, we take into account objects proposed by preprocessing which are not hand-labelled fibrils, see Table 1. Note that not all fibrils are included in the hand-labelled data. This is the reason why some detected objects, despite not being hand-labelled, might actually be fibrils. Thus, we performed a visual inspection of the detected fibrils and crossovers to obtain the approximate numbers given in lines 3 and 10 of Table 1. On these data, approximately 84.2% of detected fibrils actually were fibrils. However, without postprocessing, only ca. 11% of actual fibrils (i.e. 13% of detected objects) were entirely labelled. Here, postprocessing

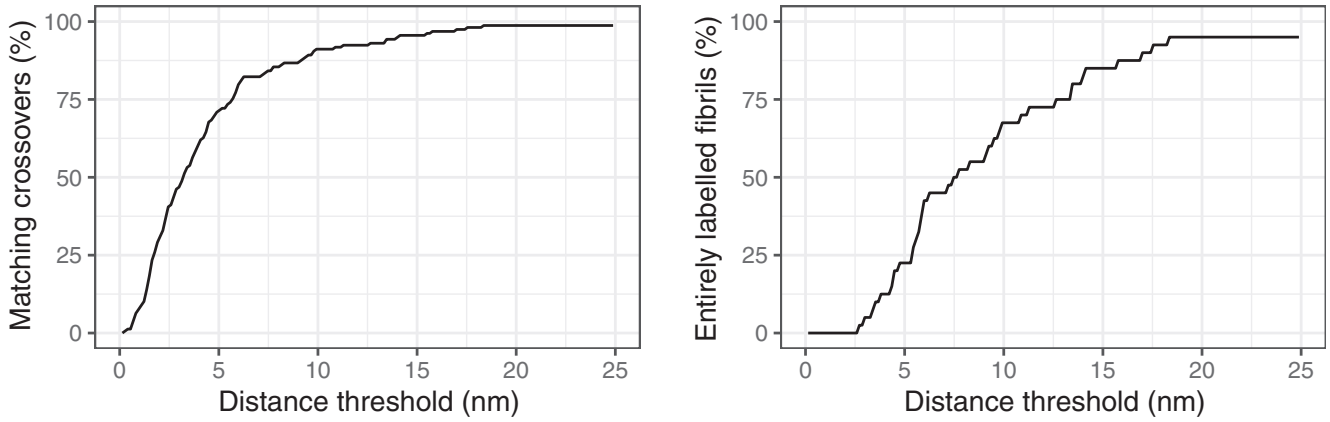

Fig. 10. The fraction of correctly located crossovers on the validation data (left) as well as the fraction of entirely labelled fibrils (right) depend on the permitted deviation from the hand-labelled data. In total, 158 hand-labelled crossovers exist on the given fibrils. Note that both fractions stay below 100% due to missing and additional, wrongly detected crossovers.

provides a huge advantage by increasing the precision to approximately 78%.

Combining both cases, we would expect an overall precision of approximately 91.5% with postprocessing and an overall precision of approximately 45.8% without postprocessing.

Comparison to direct application of a CNN

While the values presented above indicate an acceptable performance of the proposed approach, a comparison to the performance of directly applying an encoder–decoder shows a clear advantage. For this purpose, we performed no preprocessing on the original cryo-EM image data and trained a network of the same architecture as presented above using the given hand-labelled crossovers. We then applied the trained neural network to 50 raw images and postprocessed the extracted crossover locations by simply applying a morphological closing and a size threshold. This should remove some noise introduced by the neural network and keep only 'certain' crossovers. While the missing training data would suggest that not all crossovers are detected, we still expect that proposed crossovers reliably correspond to actual crossovers. However, even when considering only crossovers on actual fibrils for which hand-labelled crossovers exist, the *precision* of the crossover detection using solely a neural network comparable to the network considered in our method is only about 20%.

Moreover, our method does not only provide single crossover locations, but entirely labelled fibrils. Even when restricting the crossovers detected by the basic U-Net to crossovers lying on fibrils (for which hand-labelled crossovers exist), only 8 of 42 fibrils not used for the training of the network were entirely labelled, corresponding to a precision of 19%. Comparing this to 91.5% precision obtained by our method on *all* detected objects, this shows a clear benefit of the approach proposed in the present paper. A further advantage when applying the proposed technique is the additional information of approximate shape and location of fibrils at no further cost.

Conclusion

We proposed an approach for the automated detection of crossovers on 2D cryo-EM image data of AA amyloid fibrils. It was built around a CNN similar to the U-Net which was trained using hand-labelled data. A major improvement compared to the direct application of the CNN was achieved by a multistep preprocessing of the raw image data, using methods from classical image analysis, which extracted patches of horizontally aligned fibrils from the images. The neural network was trained on and applied to the thereby enhanced data. A final postprocessing using appropriately chosen parameters ensured that the results meet the required quality. This means, we were able to reduce wrongly detected crossovers to less than 5% of all crossovers, which is important for further analysis of the crossovers. While hand-labelling gives better results on most fibrils, the proposed approach was able to detect fibrils for which hand-labelling would have been too tedious and outperformed hand-labelling on many other fibrils. In comparison to a direct application of a CNN to the raw image data, the proposed approach shows outstanding accuracy with respect to false positives. Even though the total number of detected crossovers is affected by the focus on avoiding wrong detections, the overall performance of the proposed approach is satisfying.

While the results obtained for the specific type of fibrils considered in this paper are promising, further work may include incorporation of a more sophisticated type of postprocessing to obtain a higher yield in the total number of labelled fibrils. Moreover, as the presented method does not make any assumptions about the specific type of fibrils aside from minor

geometrical constraints, it may be used in further work to process and analyse different types of fibrils. The method described in this paper could thus be the basis for future applications, in which the morphological composition of a fibril sample is automatically assessed. The ability to analyse the morphological constitution of a sample is of general importance as it allows a more objective analysis of the fibril spectrum present in fibril extracts from patient tissue, and thus, of the pathogenic agents underlying amyloid diseases and their inherent clinical variability. In addition, quantification of the morphological composition of a fibril sample is an obvious first and indispensable step to control the fabrication of reliable and standardized amyloid fibril compositions in any form of biotechnological application of these fibrils.

Acknowledgement

This research was funded by the German Research Foundation (DFG) under grant numbers FA 456/23-1 and SCHM 997/30-1.

References

Annamalai, K., Gührs, K.-H., Koehler, R. *et al.* (2016) Polymorphism of amyloid fibrils in vivo. *Angew. Chem. Int. Ed.* **55**(15), 4822–4825.

Chiti, F. & Dobson, C.M. (2017) Protein misfolding, amyloid formation, and human disease: a summary of progress over the last decade. *Annu. Rev. Biochem* **86**(1), 27–68.

Close, W., Neumann, M., Schmidt, A. *et al.* (2018) Physical basis of amyloid fibril polymorphism. *Nat. Commun.* **9**(1), 699.

Dong, B., Shao, L., Da Costa, M., Bandmann, O. & Frangi, A.F. (2015) Deep learning for automatic cell detection in wide-field microscopy zebrafish images. In *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, pp. 772–776. IEEE.

Duda, R.O. & Hart, P.E. (1972) Use of the Hough transformation to detect lines and curves in pictures. *Commun. ACM* **15**(1), 11–15.

Furat, O., Wang, M., Neumann, M., Petrich, L., Weber, M., Krill III, C.E. & Schmidt, V. (2019) Machine learning techniques for the segmentation of tomographic image data of functional materials. *Front. Mater.* **6**, 145.

Goodfellow, I., Bengio, Y. & Courville, A. (2016) *Deep Learning*. MIT Press, Cambridge, MA.

He, S. & Scheres, S.H. (2017) Helical reconstruction in relion. *J. Struct. Biol* **198**(3), 163–176.

Heel, M.V. & Frank, J. (1981) Use of multivariate statistics in analysing the images of biological macromolecules. *Ultramicroscopy* **6**(2), 187–194.

Kingma, D.P. & Ba, J. (2015) Adam: a method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Kraus, O.Z., Ba, J.L. & Frey, B.J. (2016) Classifying and segmenting microscopy images with deep multiple instance learning. *Bioinformatics* **32**(12), i52–i59.

Lee, Y.-S., Koo, H.-S. & Jeong, C.-S. (2006) A straight line detection using principal component analysis. *Pattern Recognit. Lett* **27**(14), 1744–1754.

Liberta, F., Loerch, S., Rennegarbe, M. *et al.* (2019) Cryo-EM fibril structures from systemic AA amyloidosis reveal the species complementarity of pathological amyloids. *Nat. Commun.* **10**(1), 1104.

Liu, F., Zhou, Z., Jang, H., Samsonov, A., Zhao, G. & Kijowski, R. (2018) Deep convolutional neural network and 3d deformable approach for tissue segmentation in musculoskeletal magnetic resonance imaging. *Magn. Reson. Med* **79**(4), 2379–2391.

Nienhuis, H.L., Bijzet, J. & Hazenberg, B.P. (2016) The prevalence and management of systemic amyloidosis in western countries. *Kidney Dis* **2**(1), 10–19.

Petrich, L., Westhoff, D., Feinauer, J., Finegan, D., Daemi, S., Shearing, P.R. & Schmidt, V. (2017) Crack detection in lithium-ion cells using machine learning. *Comput. Mater. Sci.* **136**, 297–305.

Punjani, A., Rubinstein, J.L., Fleet, D.J. & Brubaker, M.A. (2017) cryoSPARC: algorithms for rapid unsupervised cryo-em structure determination. *Nat. Methods* **14**(3), 290.

Radamaker, L., Lin, Y.-H., Annamalai, K. *et al.* (2019) Cryo-EM structure of a light chain-derived amyloid fibril from a patient with systemic al amyloidosis. *Nat. Commun.* **10**(1), 1103.

Rivenson, Y., Göröcs, Z., Günaydin, H., Zhang, Y., Wang, H. & Ozcan, A. (2017) Deep learning microscopy. *Optica* **4**(11), 1437–1443.

Ronneberger, O., Fischer, P. & Brox, T. (2015) U-Net: convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (ed. by N. Navab, J. Hornegger, W.M. Wells & A.F. Frangi), pp. 234–241. Springer, Cham.

Russ, J. (2011) *The Image Processing Handbook*. 6th edn. CRC Press, Boca Raton, FL.

Schmidt, A., Annamalai, K., Schmidt, M., Grigorieff, N. & Fändrich, M. (2016) Cryo-EM reveals the steric zipper structure of a light chain-derived amyloid fibril. *Proc. Natl. Acad. Sci.* **113**(22), 6200–6205.

Schmidt, M., Rohou, A., Lasker, K., Yadav, J.K., Schiene-Fischer, C., Fändrich, M. & Grigorieff, N. (2015) Peptide dimer structure in an Aβ(1–42) fibril visualized with cryo-EM. *Proc. Natl. Acad. Sci* **112**(38), 11858–11863.

Soille, P. (2013) *Morphological Image Analysis: Principles and Applications*. Springer, Berlin.

Tang, G., Peng, L., Baldwin, P.R., Mann, D.S., Jiang, W., Rees, I. & Ludtke, S.J. (2007) EMAN2: an extensible image processing suite for electron microscopy. *J. Struct. Biol* **157**(1), 38–46.