

**Automated picking of amyloid fibrils from cryo-EM images for helical reconstruction with
RELION**

Kent R. Thurber^{1*}, Yi Yin², and Robert Tycko¹

¹Laboratory of Chemical Physics

National Institute of Diabetes and Digestive and Kidney Diseases

National Institutes of Health

Bethesda, MD 20892-0520

²Nuffield Department of Women's & Reproductive Health

University of Oxford

Oxford OX3 9DU, UK

*corresponding author: Dr. Kent Thurber, National Institutes of Health, Building 5, Room 403,
Bethesda, MD 20892-0520. phone: 301-451-7253, e-mail: thurberk@niddk.nih.gov

Abstract

Cryogenic electron microscopy (cryo-EM) is an important tool for determining the molecular structure of proteins and protein assemblies, including helical assemblies such as amyloid fibrils. In reconstruction of amyloid fibril structures from cryo-EM images, an important early step is the selection of fibril locations. This fibril picking step is typically done by hand, a tedious process when thousands of images need to be analyzed. Here we present a computer program called FibrilFinder that identifies the locations and directions of fibril segments in cryo-EM images, by using the properties that the fibrils should be linear objects and have widths within a specified range. The program outputs the fibril locations in text files compatible with the RELION density reconstruction program. After RELION is used to extract the particle image boxes contained in the fibril segments identified by FibrilFinder, a second program called FibrilFixer removes boxes that contain more than one fibril, for instance because two fibrils cross each other. As concrete and realistic examples, we describe the application of the two programs to cryo-EM images of two different amyloid fibrils, namely 40-residue amyloid- β fibrils derived from human brain tissue by seeded growth and fibrils formed by the C-terminal half of the low-complexity domain of the RNA-binding protein FUS. Both examples of amyloid fibrils can be picked from cryo-EM images using the same set of FibrilFinder and FibrilFixer parameters, showing that this software does not require re-optimization for each sample. A set of 1337 cryo-EM images was analyzed in 17 minutes with one multi-core computer. The new fibril picking software should enable the rapid analysis and comparison of more helical structures using cryo-EM, and perhaps serve as part of the greater automation of the entire structure determination process.

Keywords

electron microscopy, cryo-EM, amyloid fibril, automated particle picking

1. Introduction

Cryogenic electron microscopy (cryo-EM) is an increasingly powerful technique for determining the molecular structures of biological macromolecules, including proteins, protein complexes, and protein assemblies. Helical assemblies, such as amyloid fibrils, are one category of structure that can be determined by cryo-EM methods (Behrmann et al., 2012; Desfosses et al., 2014; He and Scheres, 2017; Rohou and Grigorieff, 2014). An important early step in the process of reconstructing a 3D helical density from cryo-EM images is identification of the locations and directions of the helical structures in the images. This step is typically done manually, often on thousands of cryo-EM images (Gallardo et al., 2020; Ghosh et al.; Hervas et al., 2020; Lee et al., 2020; Lu et al., 2020; Roder et al., 2020; Schweighauser et al., 2020; Zhang et al., 2020).

In order to automate this rather tedious process, we have written two MATLAB programs, called FibrilFinder and FibrilFixer, that can identify linear assemblies in cryo-EM images, output their locations as straight line segments, and remove particle boxes that contain contributions from more than one fibril. The output files are designed to be directly compatible with RELION 3.1 (He and Scheres, 2017; Scheres, 2012; Scheres, 2020). The RELION software package has been used in numerous recent structural studies of amyloid fibrils (Gallardo et al., 2020; Ghosh et al.; Hervas et al., 2020; Lee et al., 2020; Lu et al., 2020; Roder et al., 2020; Schweighauser et al., 2020; Zhang et al., 2020). FibrilFinder locates amyloid fibrils in cryo-EM images by using three main properties: (1) the fibrils are darker than the background of empty ice, (2) the fibrils have a width that is known approximately, and (3) the fibrils are elongated, nearly linear objects. The automated fibril picking algorithm can also handle cryo-EM images in which part of the support film of the sample grid is visible.

The automated fibril picking algorithm combines features from several previous works. First, ideas and code for filtering and width detection are adapted from the work of Yin and coworkers and Torrent and coworkers (Torrent et al., 2019; Yin et al., 2019), in which the location and characteristics of amyloid fibrils in negative-stain TEM images were identified. The detection of linear objects, such as protein fibrils, by convolution with a linear kernel has been described by Weber and coworkers (Weber et al., 2020) and Wagner and coworkers (Wagner et al., 2020). The combination of multiple detection features is necessary for accurate location of amyloid fibrils with low mass-per-unit-length in cryo-EM images which have low signal-to-noise.

In the initial fibril picking step by FibrilFinder, the fibril segments are allowed to cross one

another. However, particle boxes that include two crossing fibrils will probably not align well in the 3D density reconstruction (He and Scheres, 2017). To overcome this problem, FibrilFixer removes particle boxes where a second fibril is too close to the center of the particle box.

Section 2 below describes the algorithm and provides instructions for using FibrilFinder and FibrilFixer in conjunction with RELION 3.1. Section 3 describes applications to cryo-EM images of amyloid fibrils formed by 40-residue amyloid- β (A β 40) peptides and by a 115-residue C-terminal segment of the low-complexity domain of the FUS protein (FUS-LC-C, residues 110-214). Section 4 discusses other applications and potential limitations.

2. Automated fibril picking algorithm

2.1. Initial steps

FibrilFinder is written as a MATLAB script (FibrilFinder.m, available at github, <https://github.com/thurberk/fibrilfinder>). Overall, FibrilFinder integrates with RELION by automatically creating the fibril picking files for a RELION Manual Pick job. For automatic integration into a RELION project, before running FibrilFinder, it is necessary to start a Manual Pick job in RELION 3.1, open at least one image within the Manual Pick job, and save the results (without picking any fibrils). This generates many of the files and directories that RELION expects from a Manual Pick job, so that the FibrilFinder program described below only needs to generate the manualpick.star files that list the line segments representing the locations of the fibrils.

When FibrilFinder is run, the user is prompted to select a text input file containing the input parameters for the program. An example of an input file is included with the software at github (<https://github.com/thurberk/fibrilfinder>). Next, the user has the option to read the list of images in three different ways: (1) from a micrographs_ctf.star file created by a RELION CtfFind job, or (2) from a list hard-coded in the MATLAB code, or (3) to select the micrograph files interactively. The user also chooses between getting output that consists only of the manualpick.star files used by RELION or output that includes the MATLAB variables. The user can also choose to output the binned images and the binned images with fibril line segments shown, as MATLAB and Portable Network Graphics figure files.

If the user opts to read the list of images from a RELION CtfFind job, the user is prompted to select the RELION project directory and the micrographs directory of the Manual Pick job started earlier in RELION. Finally, the user can request that FibrilFinder run on more than one

computer core. If more than one core is requested, a MATLAB parallel for loop (“parfor” statement) is started.

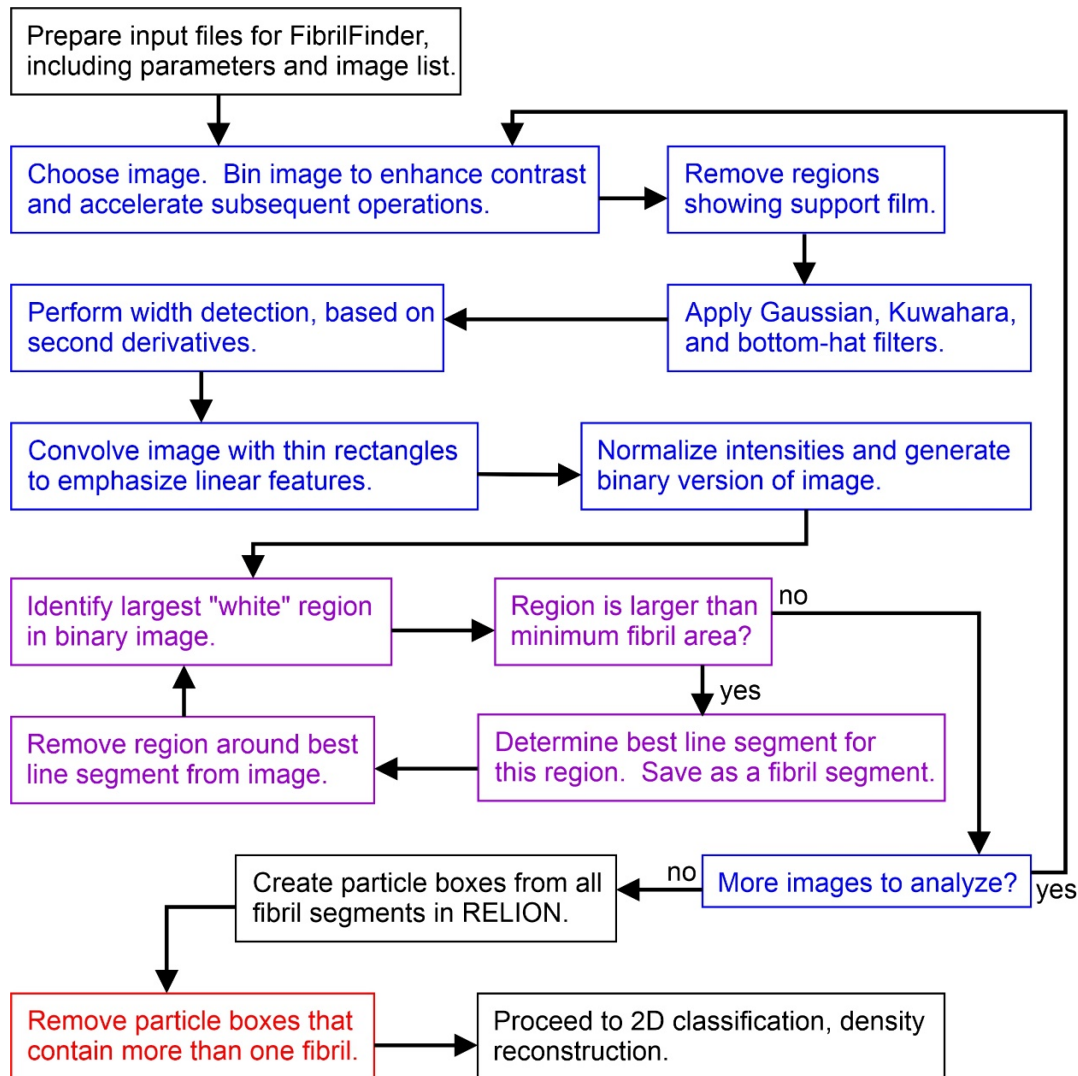


Figure 1: Flowchart for automated fibril picking. Operations performed in the image processing and fibril picking stages of the FibrilFinder algorithm are shown in blue and purple, respectively. The operation performed by FibrilFixer is shown in red.

2.2. Identification of fibrils in cryo-EM images with FibrilFinder

Fig. 1 contains a flowchart that summarizes the image processing steps within FibrilFinder. Fig. 2 shows examples of images after each processing step. The original cryo-EM images are in 32 bit MRC format with 3838 x 3710 pixels. The first step within FibrilFinder is to bin the image, as shown in Fig. 2A. This both increases the signal-to-noise of the image and speeds up the

processing because all subsequent calculations are performed on smaller images. In the examples in Section 3, we binned the images by a factor of 12 in each direction, so roughly 144 pixels in the original images are summed together for the later calculations. In the following discussion and in Table 1, pixel numbers refer to the original image, before binning. However, within the script, these numbers are adjusted by the binning factor.

Table 1: Parameters used for fibril detection for both A β 40 and FUS-LC-C fibrils. All parameters in pixel units are given for the pixels in the original images, which correspond to 1.08 Å/pixel for A β 40 and 1.07 Å/pixel for FUS-LC-C.

| Parameter | Value |
|--|-----------------------|
| Binning factor | 12 |
| RELION box size | 400 pixels |
| Box size used for fibril overlap detection | 300 pixels |
| Fibril width estimate | 90 pixels |
| Length for rectangle convolutions | 800 pixels |
| Width for rectangle convolutions | 24 pixels |
| Orientation increment for rectangle convolutions | 2° |
| Fibril line segment search angular range | $\pm 10^\circ$ |
| Fibril line segment search angular step | 1° |
| Fibril line segment search length step size | 24 pixels |
| Threshold for edge detection by Sobel algorithm | 0.05 |
| Threshold for binary image | 0.6 |
| Minimum fibril area | 28,800 square pixels |
| Gaussian filter standard deviation | 24 pixels |
| Kuwahara filter window size | 60 \times 60 pixels |
| Bottom hat filter disk radius | 300 pixels |

The next step is to remove image regions which show the support film of the cryo-EM grid, as shown in Fig. 2B. This is necessary because some of our images show the support film at the corner or edge of the image. If the support film region is left in the image, we find that spurious fibril segments can be picked on top of the film, because the film region is darker than the real fibrils in the ice. This is especially problematic because the edge between support film and ice can look like a linear object, and thus be picked incorrectly. To remove the support film region, the image is binned by another factor of 4. The Sobel algorithm (Duda and Hart, 1973) is then used to detect the edge between film and ice, with the threshold parameter set to 0.05. It is important that the edge detection threshold be low enough to detect the edge between the film and ice without

detecting the edges of the real fibrils. If more than one area is defined as a possible film region within the same image, the largest area is taken. The film region is also required to include one of the sides of the image. The side of the edge that is the film region is determined from the image intensity gradient across the edge, since the film region should be darker than the neighboring ice region. Once the support film region has been identified, pixels within this region are all set to a single intensity value. This value is the average of the pixel values of the closest non-film pixel for each pixel in the film region. This ensures that the support film region of the image is replaced with a constant pixel value that is typical of the closest ice-only pixels. Note that this approach to removal of support film regions prevents the picking of any fibril segments that lie within the support film region.

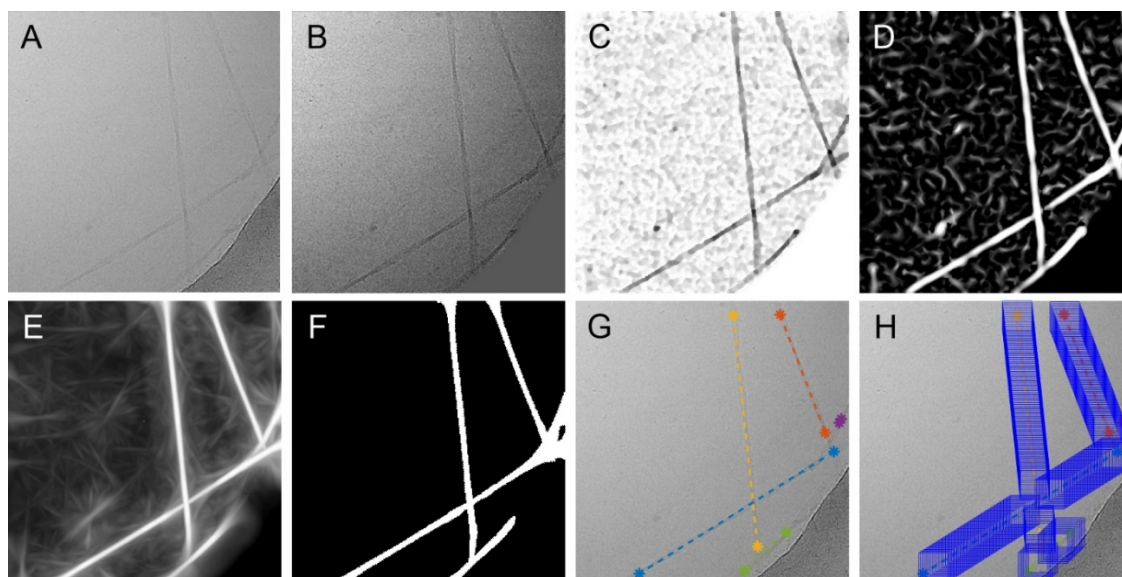


Figure 2: Illustration of steps in the automated fibril picking algorithm. (A) Example of a cryo-EM image of A β 40 fibrils after binning by a factor of 12 in each direction. A support film region appears in the lower right corner. (B) Image after removal of the support film region. (C) Image after application of Gaussian, Kuwahara, and bottom-hat filters. (D) Image after width detection based on second derivatives. (E) Image after convolution with thin rectangles to emphasizes linear objects. (F) Image after conversion to binary format by thresholding. (G) Fibril segments derived from the images in panels E and F, shown in color on the image from panel A. (H) Particle boxes (in blue) generated from the fibril segments in panel G by a RELION Extract job and FibrilFixer.

After removal of support film regions, three successive filters are applied to the image, producing the result shown in Fig. 2C. To increase signal-to-noise, a Gaussian filter is applied with a standard deviation of 24 pixels, followed by a Kuwahara filter with a 60×60 pixel window.

The Kuwahara filter is designed to reduce noise without overly smoothing edges (Balbi, 2007; Kuwahara et al., 1976). A bottom-hat filter (“imbothat” function in MATLAB) is then applied, using a disk with 300 pixel radius. This filter has the effect of removing slow variations in intensity across the image, while preserving objects that are darker than the background. Thus, the bottom-hat filter is intended to remove uneven illumination across the image.

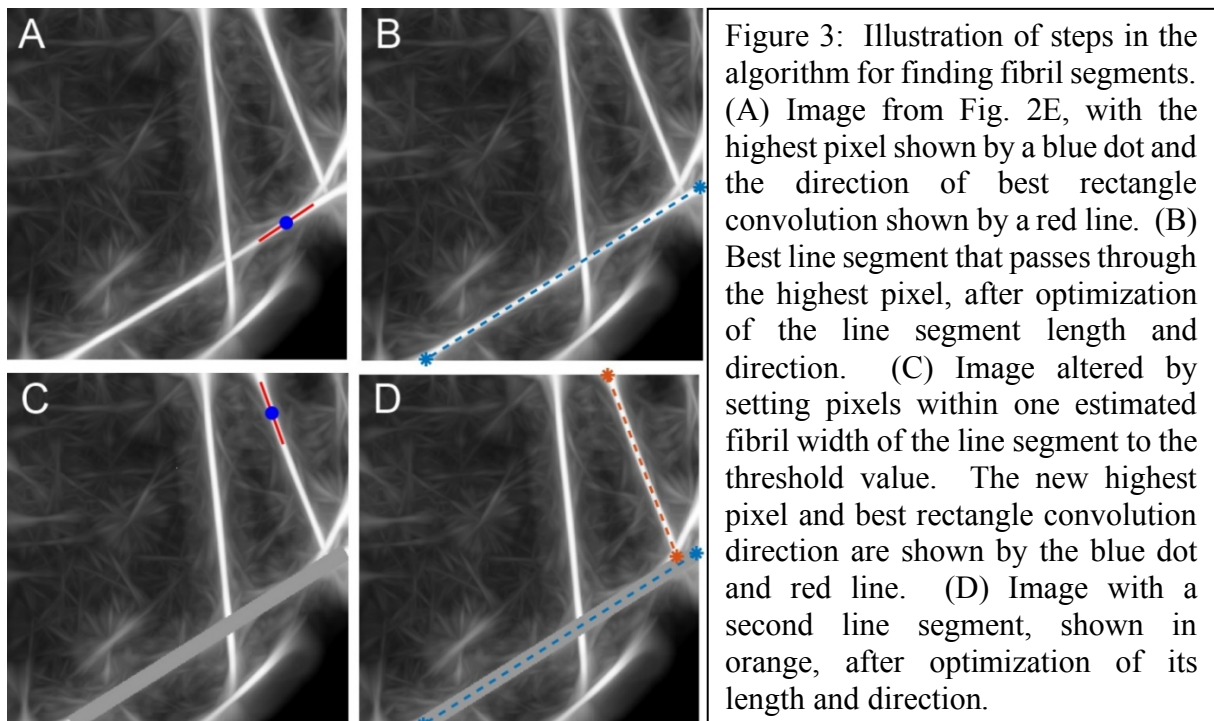
After filtering, objects with widths near the value expected for fibrils are emphasized using an algorithm originally developed for imaging of blood vessels (Jerman et al., 2016; Yin et al., 2019). The algorithm first filters the image with a Gaussian with a standard deviation of half of the estimated fibril width (see Table 1). Then the second derivatives are calculated to derive the 2D Hessian matrix at each pixel. Finally, the eigenvalues of the Hessian matrix are used to calculate the normalized output value, following the 2D version of equations 13-15 in the paper by Jerman and coworkers (Jerman et al., 2016). In this case, we apply the algorithm for a range of widths around the entered fibril width (here, 54 to 126 pixels) and keep the largest result. This processing step results in the image shown in Fig. 2D.

Next, in order to emphasize linear objects and to determine their direction, the processed image is convolved at each pixel with thin rectangles (800 pixel length and 24 pixel width). A set of 90 rectangles is used, with orientations relative to the horizontal direction that vary from 0° to 178° in increments of 2° . We then create an image, shown in Fig. 2E, in which each pixel value is replaced by the maximum value from the set of 90 convolutions. We also save the orientation of the rectangle that yields the maximum convolution.

After the convolution step, the image is normalized to a maximum intensity value of 1.0. A binary image is then created, using an intensity threshold of 0.6. Intensity values above 0.6 are identified as possible fibril regions. Small regions (less than 28,800 square pixel area in the original image) are set to zero in the binary image. An example of the binary image is shown in Fig. 2F.

At this point, image processing is complete and fibril picking can begin. The image processing has produced three results for each pixel in the binned image: (1) the normalized value of the best rectangle convolution, (2) the direction of the best rectangle convolution, and (3) whether the pixel intensity is 0 or 1 in the binary image. Line segments representing fibrils are then picked serially, using the following steps: First, as shown in Fig. 3A, we find the largest connected region of the binary image with pixel values equal to 1 (using the MATLAB function

1 “bwlabel”) and select the pixel with the largest convolution value within this region. Pixels that
 2 are within half of the particle box size of the edge of the image are excluded from this selection.
 3 Second, we consider a set of line segments that run through the selected pixel with orientations
 4 that vary by $\pm 10^\circ$ from the orientation of the best convolution, with increments of 1° , and with
 5 lengths in increments of 2 pixels of the binned image (24 pixels in the original image). These line
 6 segments are then scored by summing over the difference between the convolution results and the
 7 threshold for all pixels that are closer to the line segment than 0.32 times the estimated fibril width.
 8 If the pixel value is exactly equal to the threshold (which should generally happen only if the pixel
 9 has already been used in a previous fibril segment), a small penalty (equal to -0.05) is applied to
 10 the score. The best line segment is kept as a fibril segment if it is longer than the particle box size
 11 and has a positive total score. Fig. 3B shows an example of one such line segment. Third, if the
 12 best line segment is kept as a fibril segment, pixels within the estimated fibril width are set to zero
 13 in the binary image and set to the threshold value in the normalized convolution image as shown
 14 in Fig. 3C. If the best line segment is not kept, only the pixel selected in the first step is set to zero
 15 in the binary image and to the threshold value in the normalized convolution image. As shown in
 16 Fig. 3D, these three steps are repeated until all white areas in the binary image are smaller than 1.5
 17 times the “minimum fibril area” parameter (see Table 1). The final fibril segments are shown in
 18 Fig. 2G.



2.3 Conversion to particle boxes in RELION

After all fibril segments in the binned image are found by FibrilFinder, the locations of the corresponding fibril segments in the full-size images are calculated. Because RELION creates a full particle box around the end points of each fibril segment, the endpoints of all segments are shifted inwards by half of the box length, so that particle boxes will not extend beyond the ends of fibrils. The final fibril segments are saved in a “manualpick.star” file for each image, following the format that RELION 3.1 would generate from a Manual Pick job. The result at this stage is a Manual Pick job that follows the RELION format, and thus can be viewed or modified from within RELION and then used for subsequent RELION processing steps. Typically, the next RELION processing step would be an Extract job to define the particle boxes from the fibril line segments.

2.4 Removal of fibril overlap

As illustrated in Fig. 4A, it is not uncommon for fibrils to cross one another in cryo-EM images. The FibrilFinder algorithm is deliberately designed not to exclude fibrils that cross one another. However, particle boxes that contain two crossing fibrils should be removed because the higher density of overlapped fibrils can affect the alignment of the 2D image to the 3D model (He and Scheres, 2017). Therefore, we have developed a second MATLAB script, called FibrilFixer.m (available at github, <https://github.com/thurberk/fibrilfinder>), to remove particle boxes that contain segments from more than one fibril. We have separated our MATLAB code into two separate scripts (FibrilFinder & FibrilFixer), so that we do not have to duplicate the features of the RELION Extract job within our own code. After FibrilFinder has picked the fibril line segments, the RELION Extract job is run to define the particle boxes created from each fibril. After the Extract job, FibrilFixer is run to remove particle boxes that contain more than one fibril. Figs. 2H and 4B show final results in which particle boxes that contain more than one fibril have been removed.

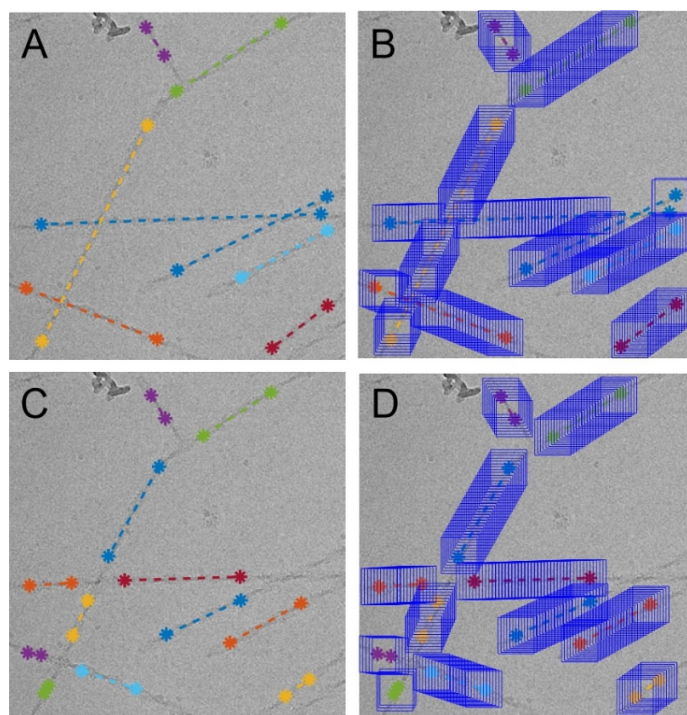


Figure 4: Comparison of results from automated and manual picking of fibril segments in one cryo-EM image of FUS-LC-C fibrils. (A) Fibril segments determined with FibrilFinder, RELION 3.1, and FibrilFixer. (B) Particle boxes generated from segments in panel A. (C) Fibril segments picked manually. (D) Particle boxes generated from segments in panel C.

For comparison with Figs. 4A and 4B, Figs. 4C and 4D show results from manual picking of fibril segments, in which regions where fibrils cross one another are avoided by dividing the fibrils into shorter non-crossing segments. Although the final set of particle boxes from the manual approach is similar to the final set from the automated approach, there is an important difference. When a fibril is divided into two segments to avoid a crossing, RELION treats the two fibril line segments as separate fibrils. Therefore, no correlations of psi or rot alignment angles are included between particle boxes within the two separate fibril segments. In contrast, when lines that define fibrils are allowed to cross one another, and particle boxes that contain multiple fibril segments are removed subsequently, correlations of particle alignment angles can be retained for all remaining boxes that come from the same fibril.

3. Results

3.1 Application to $A\beta 40$ fibrils

The FibrilFinder algorithm was first developed and tested on a set of cryo-EM images of $A\beta 40$ fibrils. As previously described (Ghosh et al., 2018; Ghosh et al.; Qiang et al., 2017), these fibrils were prepared by seeded growth, using an amyloid-containing extract from cerebral cortical tissue of an Alzheimer's disease patient as the source of seeds. Also as previously described (Ghosh et al.), images were obtained with a Titan Krios microscope, operating at 300 keV, and a

Gatan GIF Quantum K2 camera camera with an initial pixel size of 0.54 Å and defocus values from -0.5 µm to -3.0 µm. Images were binned by a factor of two in each direction before any analysis, resulting in an effective pixel size of 1.08 Å. Examples of the images, after further binning by a factor of 12 in each direction, are shown in Figs. 2A and S1.

Input parameters for FibrilFinder are listed in Table 1. Examples of fibril line segments identified by FibrilFinder are shown in Figs. 2G and S1. From 1337 images, a total of 330,643 particles were generated by FibrilFinder, RELION 3.1, and FibrilFixer. Fibrils were also picked manually in the same images, resulting in 383,717 particles. We then ran reference-free 2D classification in RELION 3.1, using both sets of particles. High-resolution 2D classes from particles produced by automated and manual analysis are compared in Figs. 5A and 5B. The full sets of 2D classes are shown in Figs. S2 and S3. With automated fibril picking, 26 2D classes with estimated resolution ≤ 4.5 Å were generated, which contain 99.974% of the particles. With manual fibril picking, 37 2D classes with estimated resolution ≤ 4.5 Å were generated, which contained all but one of the particles. These estimated resolutions are derived by RELION from the alignment statistics, but the high resolution of many of these 2D classes can be seen by the presence of the beta-strand spacing at ~ 4.8 Å. Although automated fibril picking resulted in a larger fraction of particles that would be discarded before 3D density reconstruction, the fraction of discarded particles would still be very small.

As another method to evaluate the quality of automated fibril picking, we chose 10 random images (see Fig. S1 and Table S1) and used those images to estimate the recall and precision, metrics that are often used for image recognition tasks. The recall is the percentage of available good fibrils that are picked, while the precision is the percentage of the picked fibrils that are good. If n_{tp} is the number of "true positives" (i.e., correctly-picked fibril segments), n_{fn} is the number of "false negatives" (i.e., fibril segments that are missed), and n_{fp} is the number of "false positives" (i.e., mistakenly-picked fibril segments), then the recall value is $n_{tp}/(n_{tp}+n_{fn})$ and the precision value is $n_{tp}/(n_{tp}+n_{fp})$. Values of n_{tp} , n_{fn} , and n_{fp} were determined by comparing the results from automated fibril picking with the results from manual fibril picking.

For the A β 40 fibrils, we obtain a recall of 91% and a precision of 95%. For these fibrils, the parameters used by FibrilFinder (see Table 1), especially the binary image threshold, were chosen to improve the precision at the cost of lower recall, based on the idea that including non-fibril particles is worse than missing a small fraction of the real fibrils. It should also be noted that

these metrics do not entirely represent the goals of the FibrilFinder algorithm, which is designed to have capabilities that are not readily available from manual picking. In particular, beyond simply duplicating manual picking at a faster speed, FibrilFinder allows the entire length of a fibril to be identified, even if the fibril crosses a second fibril.

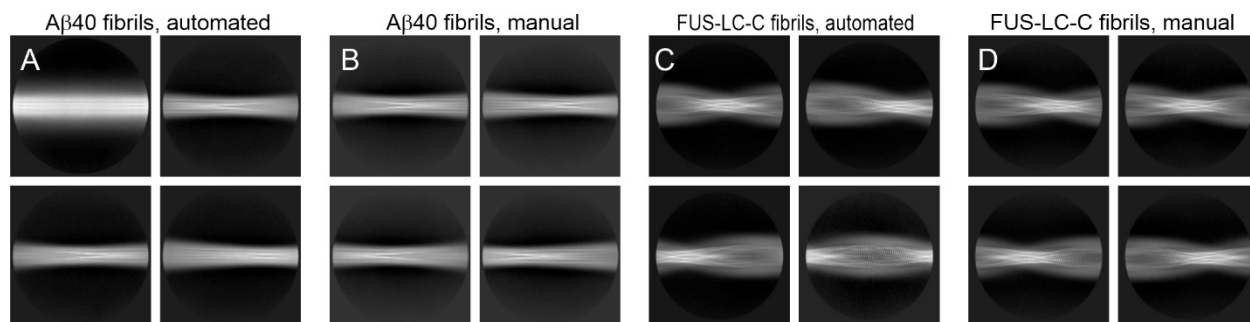


Figure 5: Comparison of 2D classes generated by RELION 3.1, based on automated fibril picking (A,C) and manual fibril picking (B,D) for A β 40 fibrils (A,B) and FUS-LC-C fibrils (C,D). The four classes with the highest resolution estimates are shown in each panel. The full sets of 2D classes are shown in Figs. S2, S3, S5, and S6 of the supplementary material.

3.2 Application to FUS-LC-C fibrils

As a test of the robustness of the FibrilFinder algorithm, we also analyzed a set of cryo-EM images of FUS-LC-C fibrils without changing the input parameters. Molecular weights of A β 40 and FUS-LC-C differ by a factor of 2.3. Molecular structures of our A β 40 and FUS-LC-C fibrils are also rather different, as indicated by reconstructed density maps with 2.77 Å and 2.62 Å resolution, respectively (Ghosh et al.; Lee et al., 2020). As previously described (Lee et al., 2020), images of FUS-LC-C fibrils were obtained with a Titan Krios microscope, operating at 300 keV, and a Gatan K2 Summit camera, with a pixel size of 1.07 Å and defocus values from -1.0 μ m to -2.5 μ m. Examples of the images, after further binning by a factor of 12 in each direction, are shown in Figs. 4A and S4.

Examples of fibril line segments identified by FibrilFinder are shown in Figs. 4A and S4. From 1461 images, a total of 422,263 particles were generated by FibrilFinder, RELION 3.1, and FibrilFixer. Fibrils were also picked manually in the same images, resulting in 276,535 particles. In this case, manual fibril picking was performed conservatively to avoid all fibril crossings and near-overlap of fibrils. In contrast, automated fibril picking was able to generate longer fibril

segments and a larger number of particles. High-resolution 2D classes from particles produced by automated and manual analysis are compared in Figs. 5C and 5D. The full sets of 2D classes are shown in Figs. S5 and S6. With automated fibril picking, 29 2D classes with estimated resolution ≤ 4.5 Å were generated, containing 99.79% of the particles. With manual fibril picking 44 2D classes with estimated resolution ≤ 4.5 Å were generated, containing 98.2% of the particle boxes. For FUS-LC-C fibrils, automated fibril picking resulted in a larger number of particles that could be used for 3D density reconstruction.

For the FUS-LC-C fibrils, we randomly chose 10 images (see Fig. S4 and Table S2) and evaluated the recall and precision of automated fibril picking as described above. The recall was 83% and the precision was 94%.

3.3 Speed of automated fibril picking

Automated fibril picking with FibrilFinder took roughly 18 s per image per computer core, using an Intel® Xeon® Processor E5-2695, MATLAB version 2020a, and the CentOS 7 Linux operating system. Analysis of the entire A β 40 data set took 7.3 h using only one core. Analysis of the entire FUS-LC-C data set took 6.8 h. The automated fibril picking algorithm is easily run in parallel across multiple computer processor cores, because the analysis of each image is completely independent from analyses of other images. We therefore implemented a “parfor” loop in the MATLAB code to run the image analyses in parallel. With this simple modification, analysis of the entire A β 40 dataset on one computer with two E5-2695 processors and 28 total cores was completed in 17 min. As expected, the speed-up is nearly proportional to the total number of cores used.

The fact that automated fibril picking is much faster than manual fibril picking potentially allows larger numbers of images and larger sets of fibril samples to be analyzed. If necessary, FibrilFinder can be run many times, with different sets of input parameters, to optimize the output. However, as described above, successful analyses of cryo-EM images of two different fibrils with a single set of input parameters indicates that parameter optimization may not be necessary, provided that the images have similar overall quality.

4. Discussion

Software described in Section 2 automates the process of fibril picking in large sets of cryo-

EM images of amyloid fibrils. Results for cryo-EM images of A β 40 fibrils and FUS-LC-C fibrils presented in Section 3 show that automated fibril picking with FibrilFinder, followed by creation of particle boxes with standard procedures in RELION 3.1 and removal of particle boxes that contain more than one fibril with FibrilFixer, produces a set of particle boxes that is comparable to the set of particle boxes that would be produced by manual fibril picking.

There has been previous work to automate the picking of amyloid fibrils (He and Scheres, 2017; Huber et al., 2018) or other linear objects such as tobacco mosaic virus and F-actin filaments (Wagner et al., 2020) from cryo-EM images. However, fibril picking was still done manually in recent successful structural studies of amyloid fibrils by several different research groups (Gallardo et al., 2020; Ghosh et al.; Hervás et al., 2020; Lee et al., 2020; Lu et al., 2020; Roder et al., 2020; Schweighauser et al., 2020; Zhang et al., 2020), suggesting that there is still a need for effective automated fibril picking software. In this work, we have taken an approach that allows for full automation of the picking process. Previous automated picking approaches have relied on manual picking of some images, in order to generate reference images or 2D classes (He and Scheres, 2017; Huber et al., 2018). Also, neural network algorithms have been used, which are trained from manually picked images (Wagner et al., 2020). In addition to their neural network algorithm, Wagner and coworkers have recently published an automated picking method for linear objects based on convolutions (Wagner et al., 2020). Even for this method, they use an optimization of their threshold parameters. Although methods that require training or optimization for different samples may be useful, we have designed our algorithm to run on multiple samples without requiring changes in the parameters used by the algorithm. The fact that FibrilFinder could be applied successfully to cryo-EM images of two different amyloid fibril samples without adjustment of input parameters suggests that this automated approach will have wide applicability without serious sample-dependent idiosyncrasies.

Another important aspect of the FibrilFinder algorithm is the removal of image regions that show the support film of the cryo-EM grid, before the rest of the image analysis. Since the fibrils are identified as linear objects with a certain range of widths, if support film regions are not removed, the edges of the film are easily mis-identified as possible fibrils. In addition, because image regions containing the support film are typically darker than just ice, the overall intensity scale is different in an image that includes a support film region. Recently, Sanchez-Garcia and coworkers published a method (called MicrographCleaner) that is specifically designed to remove

1 image regions undesirable for picking, such as support film regions or ice crystal contamination
 2 (Sanchez-Garcia et al., 2020). This pre-identification of undesirable regions is likely to be an
 3 important tool for any particle picking algorithm.

4 The FibrilFinder algorithm is not perfect. We find that good fibril segments that are missed
 5 by FibrilFinder tend to be concentrated in a small number of images with a high density of fibrils.
 6 This is because the algorithm relies on identifying the amyloid fibrils against a uniform
 7 background. If there are too many amyloid fibrils in the image, this can reduce the contrast
 8 between an amyloid fibril and the average “background” of the processed image. Fibrils that are
 9 close together or cross also inhibit their detection by the apparent width. Also, if there is a
 10 significant amount of dark non-fibril contamination, this can skew the intensity threshold for
 11 generating a binary image (*e.g.*, generating Fig. 2F from Fig. 2E), so that fibrils are missed. In the
 12 applications discussed above, the threshold value was chosen high enough to reduce the number
 13 of non-fibrils mistakenly picked, at the cost of missing some fibrils. In addition, we have the
 14 “Minimum Fibril Area” parameter, which ignores small areas over the threshold, rather than trying
 15 to add them to a fibril. We tested adjusting this parameter, and found it has a fairly small effect.
 16 Reducing the “Minimum Fibril Area” enables a few more fibrils to be picked (generally short
 17 ones), but at a cost in picking more non-fibrils. For example, by cutting the parameter in half (to
 18 14,400 square pixels), for the A β 40 sample, 11% more fibrils are added, but at the cost of adding
 19 5% more bad (non-fibril) picked segments. Dark non-fibril contamination can also sometimes
 20 be mistakenly picked as a fibril. If the contamination is dark enough, it may still be above the
 21 detection threshold, even after being de-emphasized by the width detection and linear convolution
 22 steps.

23 To compare the FibrilFinder algorithm with other existing automatic fibril picking
 24 methods, we have used RELION’s own automatic amyloid picking (He and Scheres, 2017), and
 25 the crYOLO software (Wagner et al., 2020) on the A β 40 fibrils (shown in Supplementary Figures
 26 S7 & S8). We should emphasize that in both of these cases, we are using their algorithm that relies
 27 on some human picking, to either define 2D classes for templates (in RELION), or for training of
 28 the neural network (crYOLO). This is in contrast to FibrilFinder which can be fully automated.
 29 RELION does have a fully automated picking procedure, but it is not necessarily intended for
 30 fibrils, and has not yielded useful results for our fibril samples. CrYOLO could be fully automated
 31 by training a model with a generic fibrillar sample, but we have chosen to test it by training with

1 our own sample.

2 For RELION, following the recommended procedure, we used 10 human-picked images
3 to generate 2D classes. Then from the 2D run, the 6 good 2D classes were chosen by hand, and
4 used as templates for the automatic amyloid fibril picking. The results for the 10 example images
5 of A β 40 were a recall of 95% and precision of 73%. The recall value is good, while the precision
6 is lowered mostly by the incorrect picking of carbon film edges and image edges as fibrils. Also,
7 the RELION template autopicking broke long straight fibrils into several line segments more often
8 than the FibrilFinder algorithm.

9 For the crYOLO comparison, we used 50 human-picked images as a training set for the
10 neural network algorithm. The resulting recall was 67%, and the precision 96%. As can be seen
11 in Figure S8, the trained cryYOLO algorithm does a good job avoiding picking non-fibrils, but has
12 a tendency to just choose short segments of longer fibrils. We set the crYOLO picking threshold
13 at 0.2 which did provide a slight improvement in recall over the default value of 0.3.

14 Overall, in comparison to these other methods on the A β 40 fibrils, FibrilFinder has both a
15 good recall of 91% and a good precision of 95%. A strength of the algorithm is the ability to pick
16 relatively long line segments on long straight fibrils. The A β 40 fibrils are a challenging sample to
17 automatically pick, because the wide projection has low signal-to-noise. The non-local nature of
18 the long rectangle convolution step helps to bridge the low signal-to-noise ratio region that occurs
19 between two narrow projection regions, as the fibril twists.

20 Although FibrilFinder is written to be compatible with RELION 3.1, the underlying
21 algorithms are obviously not specific to RELION. The automated fibril picking algorithm could
22 be adapted for use with other helical reconstruction software (Behrmann et al., 2012; Desfosses et
23 al., 2014; Rohou and Grigorieff, 2014). For use with RELION, FibrilFinder and FibrilFixer could
24 potentially be used in several ways beyond the self-contained automated picking described in this
25 article. First, FibrilFinder could be used as a method to rapidly generate 2D classes from various
26 samples or imaging sessions to evaluate quality or homogeneity of images. Rapidly generated 2D
27 classes could also be used in RELION's own automated picking option for fibrils, which relies on
28 2D classes as templates (He and Scheres, 2017). In addition, FibrilFixer could be used after
29 manual picking to remove particle boxes that contain overlapping or crossing fibrils. This would
30 allow one to ignore the crossing of fibrils during manual picking, thus reducing the number of
31 fibril segments to be picked and allowing RELION to consider the angular correlations among

1 particles along an entire fibril, even if it crosses another fibril.

2 As another example where FibrilFinder may be helpful in concert with other software,
3 Stabrin and coworkers recently published the TranSPHIRE software package for automated
4 analysis of cryo-EM images, including filamentous samples such as F-actin or an actomyosin
5 complex (Stabrin et al., 2020). Particle picking in TranSPHIRE is based on a neural network
6 model that can be trained in a feedback method from 2D classes identified as good. In their
7 published example of an actomyosin complex (Stabrin et al., 2020), user intervention is required
8 at two points. First, the neural network for particle picking should be pre-trained on a filamentous
9 sample, although this sample does not have to be closely similar to the filaments of interest. In
10 addition, the authors stop the automated algorithm after the first 2D classes are generated, in order
11 to choose good filament 2D classes by hand. Thus, picking of filaments with TranSPHIRE
12 currently requires manual intervention. Incorporation of a filament detection algorithm for
13 generation of the initial 2D classes, as described above, may enable full automation of filament
14 picking.

15 In conclusion, automated fibril picking from cryo-EM images can eliminate the tedious
16 task of manually picking amyloid fibril segments from the images. This can enable rapid screening
17 of multiple samples and could be included in a fully automated pipeline for image analysis and
18 reconstruction. Although we have only tested FibrilFinder on cryo-EM images of amyloid fibrils,
19 it is likely that that the same software can be used in studies of other types of protein filaments and
20 helical assemblies.

21
22

CRedit authorship contribution statement

Kent R. Thurber: Conceptualization, Methodology, Software, Writing-original draft;

Yi Yin: Software, Writing-review and editing;

Robert Tycko: Conceptualization, Writing-review and editing.

Declaration of competing interest

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

Acknowledgements

This work was supported by the Intramural Research Program of the National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health. This work used computational resources of the Biowulf cluster in the High Performance Computing facility of the National Institutes of Health. We thank Dr. Ujjayini Ghosh and Dr. Myungwoon Lee for providing the cryo-EM images of A β 40 and FUS-LC fibrils and for manual picking of fibrils in these images.

References

- Balbi, L., 2007. Faster Kuwahara filter,. MATLAB Central File Exchange, <https://www.mathworks.com/matlabcentral/fileexchange/15027-faster-kuwahara-filter>.
- Behrmann, E., Tao, G., Stokes, D.L., Egelman, E.H., Raunser, S., Penczek, P.A., 2012. Real-space processing of helical filaments in SPARX. *J. Struct. Biol.* 177, 302-313.
- Desfosses, A., Ciuffa, R., Gutsche, I., Sachse, C., 2014. SPRING: An image processing package for single-particle based helical reconstruction from electron cryomicrographs. *J. Struct. Biol.* 185, 15-26.
- Duda, R., Hart, P., 1973. Pattern classification and scene analysis. Wiley-Interscience, New York.
- Gallardo, R., Iadanza, M.G., Xu, Y., Heath, G.R., Foster, R., Radford, S.E., Ranson, N.A., 2020. Fibril structures of diabetes-related amylin variants reveal a basis for surface-templated assembly. *Nat. Struct. Mol. Biol.* 27, 1048-1056.
- Ghosh, U., Yau, W.M., Tycko, R., 2018. Coexisting order and disorder within a common 40-residue amyloid- β fibril structure in Alzheimer's disease brain tissue. *Chem. Commun.* 54, 5070-5073.
- Ghosh, U., Thurber, K.R., Yau, W.-M., Tycko, R., Molecular structure of a prevalent amyloid- β fibril polymorph from Alzheimer's disease brain tissue. *Proc. Natl. Acad. Sci. U. S. A.*, in press.
- He, S.D., Scheres, S.H.W., 2017. Helical reconstruction in Relion. *J. Struct. Biol.* 198, 163-176.
- Hervas, R., Rau, M.J., Park, Y., Zhang, W.J., Murzin, A.G., Fitzpatrick, J.A.J., Scheres, S.H.W., Si, K., 2020. Cryo-EM structure of a neuronal functional amyloid implicated in memory persistence in drosophila. *Science* 367, 1230-1234.
- Huber, S.T., Kuhm, T., Sachse, C., 2018. Automated tracing of helical assemblies from electron cryo-micrographs. *J. Struct. Biol.* 202, 1-12.
- Jerman, T., Pernus, F., Likar, B., Spiclin, Z., 2016. Enhancement of vascular structures in 3D and 2D angiographic images. *IEEE Trans. Med. Imaging* 35, 2107-2118.
- Kuwahara, M., Hachimura, K., Eiho, S., Kinoshita, M., 1976. Processing of RI-angiocardigraphic images, in: K. Preston and M. Onoe, (Eds.), *Digital processing of biomedical images*, Springer, Boston, MA.
- Lee, M., Ghosh, U., Thurber, K.R., Kato, M., Tycko, R., 2020. Molecular structure and interactions

- 1 within amyloid-like fibrils formed by a low-complexity protein sequence from FUS. *Nat.*
2 *Commun.* 11, 5735.
- 3 Lu, J.H., Cao, Q., Hughes, M.P., Sawaya, M.R., Boyer, D.R., Cascio, D., Eisenberg, D.S., 2020.
4 CryoEM structure of the low-complexity domain of hnRNPA2 and its conversion to
5 pathogenic amyloid. *Nat. Commun.* 11, 4090.
- 6 Qiang, W., Yau, W.M., Lu, J.X., Collinge, J., Tycko, R., 2017. Structural variation in amyloid- β
7 fibrils from Alzheimer's disease clinical subtypes. *Nature* 541, 217-221.
- 8 Roder, C., Kupreichyk, T., Gremer, L., Schafer, L.U., Pothula, K.R., Ravelli, R.B.G., Willbold,
9 D., Hoyer, W., Schroder, G.F., 2020. Cryo-EM structure of islet amyloid polypeptide
10 fibrils reveals similarities with amyloid- β fibrils. *Nat. Struct. Mol. Biol.* 27, 660-667.
- 11 Rohou, A., Grigorieff, N., 2014. FREALIX: Model-based refinement of helical filament structures
12 from electron micrographs. *J. Struct. Biol.* 186, 234-244.
- 13 Sanchez-Garcia, R., Segura, J., Maluenda, D., Sorzano, C.O.S., Carazo, J.M., 2020.
14 Micrographcleaner: A Python package for cryo-EM micrograph cleaning using deep
15 learning. *J. Struct. Biol.* 210, 107498.
- 16 Scheres, S.H.W., 2012. Relion: Implementation of a Bayesian approach to cryo-EM structure
17 determination. *J. Struct. Biol.* 180, 519-530.
- 18 Scheres, S.H.W., 2020. Amyloid structure determination in Relion 3.1. *Acta Crystallogr. Sect. D-*
19 *Struct. Biol.* 76, 94-101.
- 20 Schweighauser, M., Shi, Y., Tarutani, A., Kametani, F., Murzin, A.G., Ghetti, B., Matsubara, T.,
21 Tomita, T., Ando, T., Hasegawa, K., Murayama, S., Yoshida, M., Hasegawa, M., Scheres,
22 S.H.W., Goedert, M., 2020. Structures of α -synuclein filaments from multiple system
23 atrophy. *Nature* 585, 464-469.
- 24 Stabrin, M., Schoenfeld, F., Wagner, T., Pospich, S., Gatsogiannis, C., Raunser, S., 2020.
25 Transphire: Automated and feedback-optimized on-the-fly processing for cryo-EM. *Nat.*
26 *Commun.* 11, 5716.
- 27 Torrent, J., Martin, D., Noinville, S., Yin, Y., Doumic, M., Moudjou, M., Beringue, V., Rezaei,
28 H., 2019. Pressure reveals unique conformational features in prion protein fibril diversity.
29 *Sci Rep* 9, 2802.
- 30 Wagner, T., Lusnig, L., Pospich, S., Stabrin, M., Schonfeld, F., Raunser, S., 2020. Two particle-
31 picking procedures for filamentous proteins: SPHIRE-crYOLO filament mode and

1 SPHIRE-STRIPER. *Acta Crystallogr. Sect. D-Struct. Biol.* 76, 613-620.

2 Weber, M., Buerle, A., Schmidt, M., Neumann, M., Fandrich, M., Ropinski, T., Schmidt, V., 2020.
3 Automatic identification of crossovers in cryo-EM images of murine amyloid protein A
4 fibrils with machine learning. *J. Microsc.* 277, 12-22.

5 Yin, Y., Prigent, S., Torrent, J., Rezaei, H., Drasdo, D., Doumic, M., 2019. Automated
6 quantification of amyloid fibrils morphological features by image processing techniques,
7 p. 534-537, 2019 IEEE 16th International Symposium on Biomedical Imaging.

8 Zhang, W.J., Tarutani, A., Newell, K.L., Murzin, A.G., Matsubara, T., Falcon, B., Vidal, R.,
9 Garringer, H.J., Shi, Y., Ikeuchi, T., Murayama, S., Ghetti, B., Hasegawa, M., Goedert, M.,
10 Scheres, S.H.W., 2020. Novel tau filament fold in corticobasal degeneration. *Nature* 580,
11 283-287.

12