Universiteit Leiden
The Netherlands

Data Science and Artificial Intelligence

Automated Segmentation of Amyloid Fibres from Electron Micrographs

Adam Bujna

Supervisors:
Dr. Daan M. Pelt & Drs. Djim de Ridder

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

July 1, 2025

Abstract

Manual measurement of amyloid fibres from electron micrographs is a time intensive bottleneck in research on neurodegenerative diseases, and is complicated by high particle density, clustering, and image noise. This thesis investigates the efficacy of deep learning models for automating the segmentation and measurement of these fibres. We developed a framework for generating synthetic electron microscopy images with precise control over fibre count, clustering, signal-to-noise ratio, and image contrast. Using this, we generated synthetic data and trained and benchmarked state-of-the-art segmentation models, including two YOLOv11 variants and the Segment Anything Model (SAM) v2.1. We found that model performance is very robust to image noise, but degrades with increased fibre clustering and density. The YOLOv11-large model demonstrated the best resilience in highly clustered scenes, whereas SAMv2.1-tiny excels on well-separated fibres and generalises more effectively to real-world data after a brief fine-tuning period on a small set of 25 real images. Furthermore, we show that fibre length can be extracted from predicted masks with a mean median error across several images as low as 14.5%. This work shows that deep learning models trained on synthetic data can be successfully adapted for the analysis of real electron micrographs, supporting the development of automated tools for research on neurodegenerative diseases.

Contents

Introduction

Amyloid fibres are fibrous aggregates of misfolded protein and are associated with a number of neurodegenerative diseases such as Parkinson's disease [35]. They form when proteins, normally soluble and functional, fold incorrectly and aggregate into highly ordered structures [33]. Over time, these can develop into tangles, such as neurofibrillary lesions usually observed near degenerative nerve cells in Alzheimer's disease.

Suppressing the aggregation of these proteins into fibres and dissolving the fibres once they formed appears to be crucial in mitigating disease progression [32, 52]. Normally, cells prevent and dissolve amyloid protein aggregates from forming, but the failure of these mechanisms in disease has motivated the development of therapeutic interventions designed to artificially dissolve amyloid fibres [5]. However, the precise dynamics of fibre disaggregation remain unclear. This is of particular relevance, as fragmented fibres act as nucleation sites that seed further aggregation and can exhibit higher cytotoxicity than mature fibres [6, 53]. Moreover, the length and structural properties of the amyloid fibres themselves have an impact on their toxicity [11, 55].

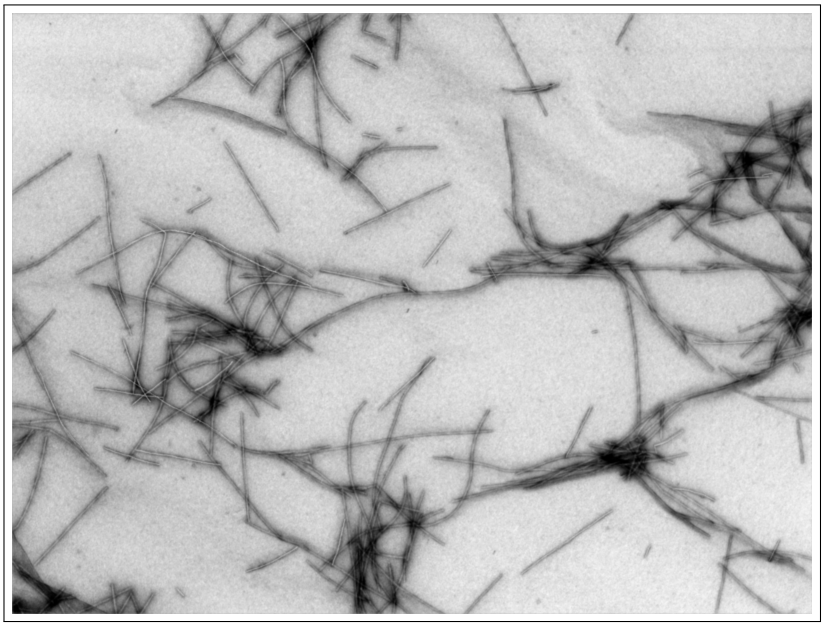
Imaging amyloid fibres using atomic force microscopy (AFM) or negative stain electron microscopy (EM) is often used to assess their structural characteristics [5, 6, 55], but manually measuring them is time-consuming, laborious, and potentially bias if the annotators unintentionally focus on easy to distinguish fibres, or fibres of specific sizes. Complicating matters further is the tendency of amyloid fibres to form dense clusters, making it difficult to discern the boundaries of any single fibre.

These challenges pose as a bottleneck in the research workflow, limiting the productivity and objectivity of quantitative analyses. The reliance on manual measurement lowers throughput and introduces potential inconsistencies and subjective bias. This thesis seeks to address this lack of automated methods for accurately segmenting and quantifying individual amyloid fibres from electron micrographs, especially in the presence of dense clustering and overlap.

Recent advances in machine learning (ML) have revealed the great potential of deep learning (DL)

Figure 1

Example of amyloid fibrils captured using negative stain electron microscopy.



Note. The image illustrates the dense clustering and overlapping of amyloid fibres, which pose challenges for measurements. The dimensions of the image are around 5.82×4.37 µm, at a resolution of around 1.766 nm per pixel.

models in a wide range of fields, including computer vision (CV) [25]. Deep learning has been used successfully for many applications that rely on the classification and segmentation of electron micrographs in life sciences and materials analysis [45]. In particular, a large amount of existing work on automated particle detection in microscopy images has focused on cryo-electron microscopy (cryo-EM), where the objective is to identify and extract particle instances to reconstruct 3D protein structures. Notable tools in this domain include Topaz [4], used in the popular CryoSPARC pipeline, CrYOLO [50], which employs the successful You Only Look Once (YOLO) object detection algorithm, among others [1, 17, 19, 45, 56]. Aside for being optimised for the peculiarities of cryogenic electron microscopy, these models typically expect to detect uniformly and regularly shaped particles, or in the case of CrYOLO's filament mode are specifically designed to omit fibre-dense regions and crossing fibres [51]. Additionally, in CryoEM, a lot of contamination disappears in heavy noise. Amyloid fibres present a wide variety of lengths, are highly anisotropic, and frequently overlap. Their regular clustering makes distinguishing

individual structures particularly difficult for CV models. Some segmentation approaches have been applied to simpler problems, such as segmenting isolated viral particles [30]. However, such applications are fundamentally different, as the particles are typically well separated, clearly visible, and isotropically shaped.

The problem of occlusion has also been highlighted in related domains [39]. In a recent study focused on measuring macerated wood cells from microscope images, deep learning models were found to ignore clumped and overlapping fibres, indicating that occlusion can significantly degrade segmentation performance when not explicitly addressed.

Deep learning models typically rely on large amounts of annotated data to generalise effectively [2]. Segmentation masks for amyloid fibrils are difficult and labour intensive to produce, requiring expert input, and often lack clarity due to overlapping fibre structures and contamination present in the samples. In the absence of such datasets, alternative strategies are needed to enable robust training and evaluation.

To address these challenges, we explore the feasibility of using deep learning models to automatically segment and quantify amyloid fibres in negative stain EM images, particularly in scenarios where fibrils are densely clustered, overlapping, and morphologically variable. Given the scarcity of annotated data, we generated synthetic datasets that allow control over fibre properties, including noise level, clustering behaviour, and morphological variation. This allowed us to evaluate model performance under controlled conditions and to leverage simulated data to train models with practical utility in segmenting real-world sample images.

To investigate these challenges, we sought to answer these questions:

1. Is it possible to accurately simulate images of amyloid fibres?

2. Can deep learning models accurately segment amyloid fibres in simulated EM images, particularly under overlap and clustering?

3. How do factors such as noise, contrast, and structural heterogeneity affect model performance?

4. To what extent can models trained solely on simulated data generalise to real electron micrographs?

5. Can the length of fibres be reliably extracted from the model segmentations?

Ultimately, the goal is to support the development of automated tools that can assist researchers in extracting meaningful structural information from complex EM data, thereby facilitating studies into biochemical mechanisms, such as those underlying neurodegenerative disease.

Thesis overview

This bachelor's thesis has been completed under the supervision of Dr. Daan M. Pelt and Drs. Djim de Ridder at the Leiden Institute of Advanced Computer Science (LIACS) and in collaboration with the Leiden Institute of Chemistry (LIC). The thesis is structured as follows:

Background and Related Work: This section explains the biological significance of amyloid fibres, the principles of electron microscopy, and introduces the fundamentals of deep learning as well as several deep learning architectures used by our models, namely Convolutional Neural Networks and Transformers. Finally, this section provides an overview of existing work in automated particle analysis in microscopy.

Methods: The Methods section describes the experimental design of this study. It describes the acquisition and characteristics of real-world electron microscopy data and details the development of our custom framework for simulating synthetic data. This section also introduces the selected models (YOLOv11-large, YOLOv11-nano, and SAMv2.1-tiny), specifies training and fine-tuning procedures, and explains how we evaluated the models and measured their performance.

Results: Here we present the findings of our experiments — both quantitative and qualitative. They include a benchmark of the models on the synthetic dataset, that analyses the impact of signal-to-noise ratio, clustering, and fibre count on segmentation accuracy. This is followed by an evaluation of the models' performance after being fine-tuned and tested on real microscopy images, and an analysis of the accuracy of fibre length measurements.

Discussion and Future Work: This section discusses the limitations of the current study, proposes potential avenues for future improvements that build upon our findings, and interprets the trends revealed by the results.

Background and Related Work

Biological Context

Amyloid fibres are insoluble, ordered aggregates of proteins that form when soluble proteins are secreted in a misfolded state[33]. They are characterised by a cross-β structure [42] in which misfolded proteins or their segments stack together to form stable beta-sheets [33]. Multiple layers of these beta-sheets assemble to create the core of the mature amyloid fibril. The resulting amyloid fibres are highly stable unbranched strings of varying lengths with a diameter of approximately 7 to 10 nm.

Amyloid fibres are associated with a wide spectrum of amyloid diseases, collectively known as amyloidoses [42]. These diseases often involve proteins that are necessary for normal biological processes [10, 36]. For example, α-synuclein, which is believed to play a role in neurotransmission, can misfold and aggregate in neurodegenerative disorders such as Parkinson's disease [3, 36]. Similarly, the amyloid plaques characteristic of Alzheimer's disease are formed from the Aβ peptide, which is a fragment of a larger precursor protein also involved in neuronal function [12, 14]. In each of these cases, the transformation of a functional protein into a stable aggregate is linked to cellular damage and disease progression, making the analysis of amyloid fibres critical for understanding these pathologies [11, 55].

The extent to which mature amyloid fibres are responsible for toxicity in various diseases is still in debate, but there are several mechanisms in which mature fibres can participate in disease [43]. One way in which amyloid fibres cause damage is by interacting with and disrupting cell membranes. Fibrils formed from several disease-related proteins have been shown to bind to and compromise biological membranes and liposomes [26, 29, 38]. These interactions can lead to a loss of cellular homeostasis or damage the integrity of the membrane, and ultimately trigger apoptosis and cause cell death.

In addition to direct toxicity, mature amyloid fibres can amplify and perpetuate disease progression [43]. This happens through a process known as secondary nucleation, in which the surfaces of mature amyloids act as templates that align proteins and accelerate their aggregation. Similarly, existing mature fibres have been observed to dissolve into smaller oligomers, which can spread disease, by interacting with cellular membranes [29].

Cells employ a number of defence mechanisms that disaggregate or clear protein aggregates, such as amyloid fibres [8, 27, 53]. This can be attributed to the fact that amyloids pose a threat to cell health [53]. Furthermore, disaggregation of amyloid fibres significantly extends the survival of diseased mice [32], and failure of the disaggregation mechanisms is linked to pathology [52]. These observations suggest that disaggregating fibrils is a viable strategy for slowing the progression of amyloid-related diseases. However, the precise dynamics of how amyloid fibres disassemble are not fully understood [5]. Therefore, developing methods to accurately measure the morphology of amyloid fibres can help study their disassembly mechanisms.

The need to accurately measure amyloid fibre morphology is further underscored by its impact on toxicity [55]. For example, shorter fibrils can be more toxic compared to longer ones, as they possess a larger surface area and more reactive ends per unit mass and have been shown to be more potent at disrupting cell membranes [6, 13, 55]. The larger surface can also help to seed new aggregation more effectively [13]. On the other hand, longer fibrils are more prone to fragmentation, a process that multiplies the number of shorter and potentially more toxic species [13]. The length distribution of a fibre population can therefore offer insights into the process of aggregation and the toxicity within a sample. Thus, the ability to accurately segment and measure the lengths of amyloid fibres can help assess amyloid fibres' toxicity and aggregation dynamics.

Electron Microscopy

Electron microscopy is a powerful imaging technique that uses a focused beam of electrons to obtain images of biological specimens or structures of materials [28]. Unlike optical microscopes, which are

limited by the wavelength of light, EM leverages the much shorter wavelength of electrons, allowing for a theoretical resolution of 0.1 nanometre. In transmission electron microscopy (TEM), electrons are travelling through an ultrathin specimen. As electrons pass through this specimen, some of them are scattered, producing a contrast against the background that is influenced by the density and composition of the sample.

Negative staining TEM is a widely used method for visualising biological structures in EM [18]. In this technique, the sample is soaked in a heavy metal stain such as uranyl acetate or phosphotungstic acid. The stain surrounds the biological material and leaves an imprint on the sample. The electron beam of a microscope that passes through the heavy material is diffracted significantly, producing a high contrast that outlines the specimen. This makes surface details visible without the need for complex preparation or sectioning and is often used to visualise the structural properties of macromolecules such as amyloid fibres [5, 6, 55].

Deep Learning

The following overview of the principles of deep learning principles, including Convolutional Neural Network and Transformer architectures, is based on the descriptions provided by Torralba, Isola, and Freeman in *Foundations of Computer Vision* [44].

In recent years, deep learning has revolutionised computer vision, outperforming classical computer vision methods, like edge detection or Fourier space analysis, and has been used successfully across a wide range of tasks and applications, including biomedical analysis [25]. The ability of neural networks (NNs) to learn complex relationships in raw unstructured data made it a key tool for biological imaging [7].

A neural network is a function approximator made up of layers of interconnected computational units called *neurons*, loosely inspired by brain neurons [44]. Each such artificial neuron performs a simple computation. It receives a vector of numerical inputs which may come from the input data or from neurons in the previous layer, and computes a weighted average of these inputs. It then adds a bias that shifts the neuron's activation value and applies a non-linear transformation, such as a hyperbolic

tangent or a sigmoid function. Formally, a neuron with a sigmoid activation function computes:

$$y = \sigma(w_1 x_1 + w_2 x_2 + \cdots + w_n x_n + b) \tag{1}$$

where y is the neuron's output (activation) value, σ is the sigmoid activation function, and w_i is the weight of input x_i.

The activation function is essential for the network's function as it introduces nonlinearity [44]. Without it, the operation of any number of linear layers would reduce to, and be equivalent to, a single linear transformation. This is similar to how biological neurons work in a thresholded fashion. The most common activation functions are the ReLU function, sigmoid, and tanh. The weights and biases are the network parameters, and their values are learnt from data. The training is done using a method called backpropagation. At each layer, gradients of the loss function (error between the network's output and the real ground truth) with respect to the parameters are calculated using the chain rule of derivatives. These gradients express how much each parameter contributes to the loss, and the parameters can be changed accordingly, iteratively improving the network's predictions.

Training is carried out over multiple *epochs*. During each epoch, the model sees every training data point once and updates its internal parameters based on its error in predicting the label of this data point. This can be a number, a classification of the data point or a segmentation mask showing where objects are in an image. Typically, many epochs are needed for the model to learn, with performance improving gradually. A higher number of epochs allow the model to capture more patterns of the data, but too many can lead to overfitting, where the network becomes too tailored to the specific training data and performs poorly on new inputs.

Convolutional Neural Networks. Convolutional Neural Networks (CNNs) are a class of neural networks, made up of convolutional layers, specifically designed to process visual or spatial data, most commonly images and videos [44]. Their design is motivated by the observation that most images contain universal patterns such as edges, corners, textures and objects that recur across differing scales and

positions in the image. CNNs exploit this by using a operation known as *convolution*, which applies the same filter (or kernel) across all regions of the input. A convolutional layer performs a linear operation in which a filter slides across the input and calculates a weighted average of each region using the filter weights. This operation is repeated identically across the entire image, significantly reducing the number of learnable parameters and enforcing translational equivalence, which means that features can be detected regardless of where they occur in the image. A typical CNN stacks multiple convolutional layers, each followed by an activation function. Each convolutional layer outputs a feature map, where each channel represents specific features detected by a filter. To reduce the dimensionality of these features into more general features, CNNs often include pooling layers. A pooling operation (like max pooling) downsamples the signal by merging local regions by taking the maximum (or average) value. CNNs often end with fully connected layers or global pooling, which consolidate the learnt features into a single decision. However, for tasks that require spatial output, such as detecting positions of objects or segmenting which pixels belong to which object, the CNN maintains a spatial structure throughout.

Image classification is the application of a CNN (or any other computer vision model) to output a single label per image. Object detection involves both locating and classifying objects. The output of object detection is a bounding box that specifies the object's location and a label. Sometimes, such as when extracting structural features of fibres in images, it might be desirable to predict the specific pixels that belong to an object. In this case, the NN can detect objects and then also predict which pixels are likely to belong to which object. This task is called instance segmentation, or semantic segmentation if the model does not distinguish between different objects belonging to the same label.

Transformers. In recent years, transformer neural networks have emerged as powerful tools in domains requiring more complex reasoning and flexibility in handling different input modalities, including images, text, or user prompts, simultaneously [44]. Although originally developed for natural language processing, they have since been adapted for computer vision tasks with great success.

CNNs are excellent at extracting local patterns through filters, but they often struggle to integrate global context efficiently. Transformers address this limitation through a mechanism called *self-attention*,

which allows different parts of the input to interact with each other, regardless of spatial proximity. This allows them to model relationships between different regions of the image and to attend more to more informative parts of the input.

Transformers encode their input as tokens, and every token can be related to any other token. In this way, tokenised versions of different kinds of input, for example, both images and text simultaneously, can interact with each other, giving transformers great versatility for multimodal tasks.

Related Work

Deep learning has been used successfully in the segmentation of particles in electron microscopy. Most of these tools are specialised for particle picking in cryogenic electron microscopy with the objective of identifying a large number of instances of a particular particle from different angles and then using these various perspectives to reconstruct its 3D structure.

Two powerful techniques are CrYOLO [50] and Topaz [4] used by the SPHIRE and CryoSPARC particle picking pipelines, respectively. Topaz is designed for single-particle cryo-EM particle picking using positive unlabelled (PU) learning. A small set of hand-marked particles is treated as the positive class, while every remaining image window is left unlabelled without explicit negatives. A convolutional neural network is then trained with a specific PU loss that regularises the posterior of the model on the unlabelled pool, preventing the overfitting of earlier PU formulations. This allows it to achieve high precision with only a few annotated samples in a large dataset. CrYOLO, on the other hand, is based on the YOLO (You Only Look Once) object detection architecture, adapted for use in cryo-EM workflows. It is based on the first 13 layers of the original YOLO architecture, because further pooling into a coarser-grained grid removes the low-level information required for the picking of small particles. To further alleviate the problem of small particles, CrYOLO splits the image into a grid of smaller images, allowing it to maintain the resolution required to register small particles.

Both of these approaches, while powerful, are not suitable for our purpose. Both Topaz and CrYOLO purposefully try to avoid dense regions and patches with particle overlap or aggregate protein, and

instead focus on well-defined particles. This invalidates the vast majority of the data we have available, as amyloid fibres readily form heavy clusters.

There are more examples that showcase the effectiveness of the YOLO models in the extraction of structural features of particles. Qamar et al. (2024) [39] developed an instance segmentation pipeline for analysing macerated wood fibre and vessel cells based on the extra large variant of the YOLOv8 model. Their goal was to automate the measurement of cell shapes and sizes in optical microscopy across large fields of view that contain many overlapping translucent cells. Their approach demonstrated high recall and precision across hardwood cell types and could infer multiple shape-based features like fibre length and area. Importantly, the model was validated on a transgenic poplar line known for elongated fibres, and its results closely matched previously reported manual measurements. While optical microscopy images suffer from significantly less noise than electron micrographs and the cells are much more uniform in shape and appearance, there is remarkable similarity between the domains of measuring the structures of wood fibre cells and amyloid fibres.

Despite these successes, the authors reported persistent difficulties with dense clustering and overlapping cells, where the model failed to distinguish individual fibres and ignored particularly dense areas. This limitation highlights the challenge of instance segmentation in overlapping objects and reinforces the need for tackling the problem of overlapping objects and ambiguous object boundaries.

Meta's Segment Anything Model (SAM) has been shown to perform well in complex visual environments with a high density of diverse object, even under zero- or one-shot supervision [34]. It's transformer architecture may allow it to model the relationships between different objects more intricately and thus distinguish between close objects better. CryoSegNet presents a hybrid approach that combines a task-specific, attention-gated U-Net for initial particle picking whose detections guide SAM through its prompt encoder to accurate segmentation of protein particles from noisy cryo-EM micrographs [17]. While CryoSegNet uses SAM that was originally trained on mostly natural images and performs poorly when applied directly to cryo-EM data, its integration with a U-Net pre-segmentation step allows it to operate more effectively in this context. The resulting pipeline demonstrates a superior balance of

precision and recall compared to widely used tools such as CrYOLO and Topaz, achieving higher F1 scores and enabling the reconstruction of higher-resolution 3D density maps. These findings suggest that adapted use of foundation models like SAM, especially when paired with domain-specific architectures, holds significant promise for biological image analysis.

Methods

The primary objective of this thesis is to assess the effectiveness of deep learning models in segmenting amyloid fibres in EM images. In particular, our goal is to evaluate performance under conditions that typically hinder manual analysis, such as high levels of fibre overlap, clustering, and image noise. To remedy the limited availability of annotated data and ensure greater control over variables, we train and test the models on synthetic datasets generated to mimic the properties of real negative stain electron micrographs.

Real EM Data

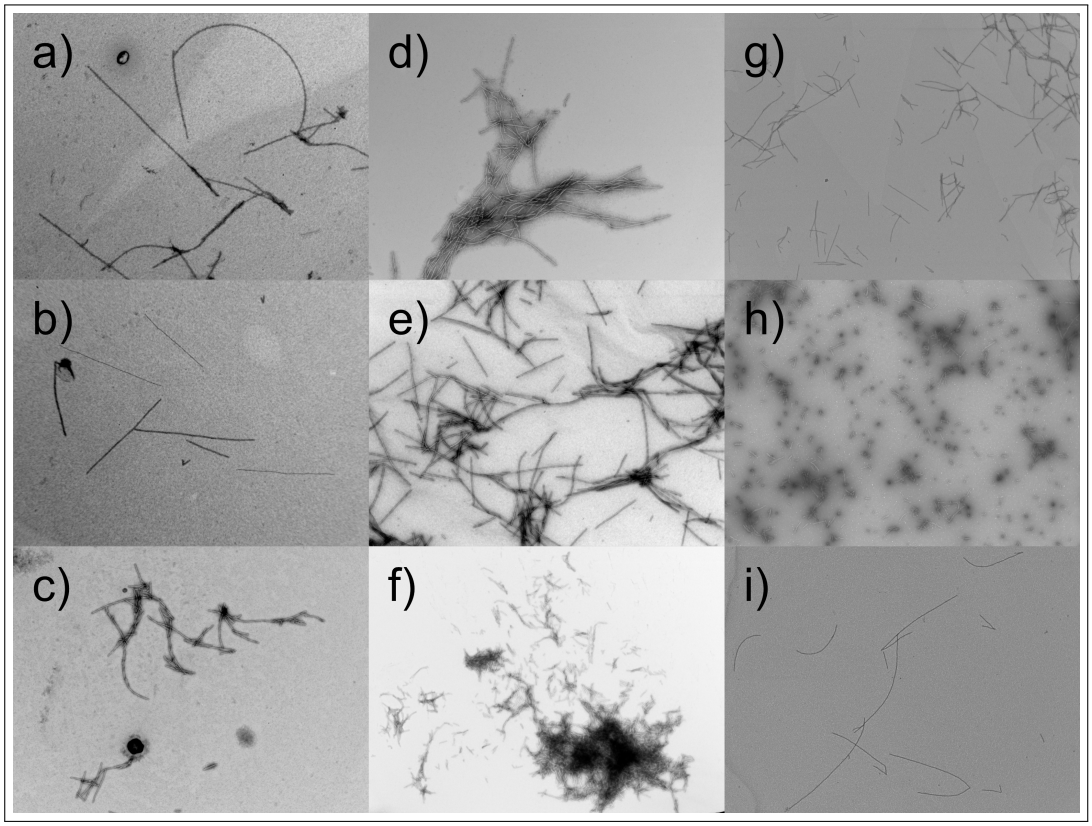
Data Acquisition. To evaluate generalisability on real data, a collection of 42, of which 30 were annotated, negative stain EM images was used. These were acquired in collaboration with the Leiden Institute of Chemistry and depict XG, FM, and F65 polymorphs of α-synuclein formed under in vitro conditions. Imaging was performed with a transmission electron microscope at a magnification of 10,000×. This means that a single pixel in the image corresponds to a distance of about 1.77 nanometres. Of the 42 images, 30 were considered representative of typical high-quality samples and were partially annotated, while 12 were selected to represent particularly challenging cases and were unlabelled. These included images with extreme fibril overlap, underexposure, contaminating material, damaged samples,

---

The complete source code for the data simulation framework, exact model training logs and setting, and evaluation code present in this thesis is publicly available on GitHub: https://github.com/adambujna/amyloid-fiber-segmentation.

Figure 2

The varying conditions present in amyloid fibre micrographs.



and other artefacts that complicate interpretation.

Data Characteristics. For synthetic data to be a meaningful training replacement, it must replicate the characteristics of real data. The images of the real dataset contain fibrils of varying lengths ranging in length from around 20 nm (comparable to their thickness), such as those in Figure 2h to several thousand nanometres, such as those in Figure 2a, which close to the size of the entire canvas. Most amyloid fibres are straight and rod-like; however, some are bent into a parabolic conformation. Owing to their uniform assembly mechanism, all fibrils appear 10 to 20 nm thick. The vast majority of images display clustering to a certain extent. Fibres tend to clump together along their axis, making their limits difficult to determine, and in some examples form complex tangles where individual fibres are indiscernible.

Another interesting detail arises as a result of negative staining. During staining, the heavy metal stain

adheres to the edges of objects, producing fibres with visible dark outer shells and lighter, apparently hollow cores. This effect is clearly visible in Figure 2b.

Signal-to-Noise-Ratio. To quantify the visual quality of the real data, signal-to-noise ratio (SNR) was estimated for 15 representative real EM images. The signal power P_{signal} was defined as the absolute difference between the mean background intensity and the mean fibre intensity.

$$P_{signal} = \mid \mu_{background} - \mu_{fibre} \mid \tag{2}$$

where $\mu_{background}$ is the mean background intensity, μ_{fibre} is the mean fibre intensity.

Noise power was calculated by finding the standard deviation of the background.

$$P_{noise} = \sigma_{background} \tag{3}$$

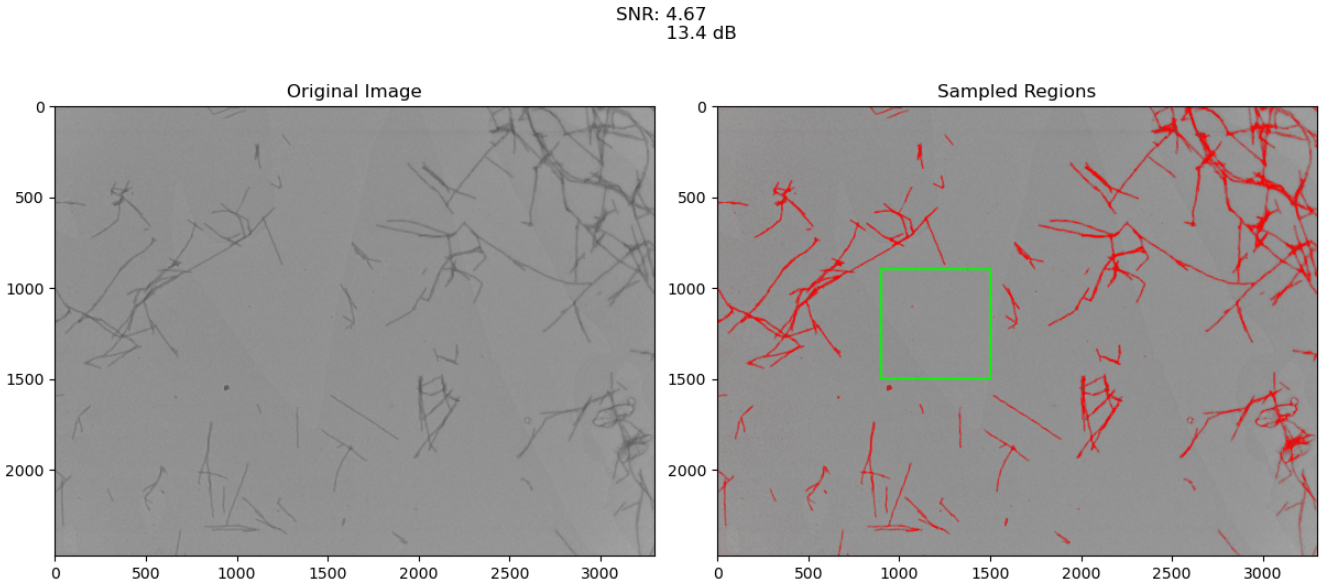
Finally, signal-to-noise ratio is the ratio of signal power to noise power.

$$\text{SNR} = \frac{P_{signal}}{P_{noise}} \tag{4}$$

In each image, empty regions were selected from which the background mean and standard deviations are calculated. To sample the signal strength, a signal threshold was used to separate the background. Then the mean of pixels below the threshold (keep in mind that the fibres are darker than the empty background) and pixels within a certain tolerance of this mean were considered to represent regions of signal. As seen in Figure 3 this strategy proved to be quite accurate in isolating fibre pixels in most images. Across the 15 images analysed, the **signal-to-noise ratios ranged** from **4.25 to 16.74** with a **mean** of **8.5** and a **standard deviation** of **4.53**. In this limited sample, brighter backgrounds tended to correlate with a higher SNR, though in some overexposed images the signal range was compressed, leading to detail loss as the stain left very thick dark spots that blended clusters into opaque black regions.

Figure 3

Sampled regions used for estimating SNR of an EM image.



Note. The background region, shown as a green box, and the signal (fiber) pixels, highlighted in red, are manually selected.

Contrast. In addition to providing estimates of signal clarity, the SNR analysis also yielded typical intensity values for background (approx. 135 to 220 in 0-255 images) and fibre regions (approx. 1.25 to 2 times darker than the background). These measurements informed the design of the synthetic data generator, which relies on background and signal intensity ratios rather than fixed absolute values. This approach allows the simulator to produce more realistic images by ensuring that overlapping fibres appear progressively darker, in line with the visual characteristics of real negative stain EM images.

Fibre Distribution. To better understand the spatial distribution of amyloid fibres in real EM images and to inform the design of our synthetic clustering methods, we extracted coarse fibre density maps from the raw micrographs. The goal was to visualise and approximate where the fibres are most concentrated.

Because objects appear darker than the background in negative stain electron micrographs, each image was first inverted so that fibres became bright on a dark background, giving us higher values for higher concentrations, which appears more intuitive. To suppress the background, which should represent a value of 0, the mode pixel intensity was subtracted from the image. Further, the image was

normalised such that the maximum intensity was scaled to 1. This produced a binary-like map where the background has a value of 0 and fibres a value close to 1.

To reveal more general spatial trends in the fibre distribution, a heavy Gaussian blur was applied using a large kernel (499×499 pixels on a 3296×2472 image) and a high standard deviation (σ = 1000). This removed any details while preserving the broad spatial structure, resulting in a smooth heatmap that visually approximated the fibre density across the image. The resulting maps provided a qualitative view of how fibres were spatially arranged within each micrograph, allowing us to assess how well our fibre placement algorithm corresponds to real samples.

Ground Truth. Ground-truth annotations for the real EM images were produced by researchers at the Leiden Institute of Chemistry by tracing a few points along each fibre. These points can be connected into rough skeletons. To turn these skeletons into usable binary masks, they were widened to a fixed width of 20 pixels. This is slightly thicker than most fibres, which are only about 9-12 pixels wide, but ensures that the entire fibre is included in the mask despite following the imperfect annotation. For many images, especially those with a higher number of fibres, the annotations were sparse, with many of the fibres unlabelled. As such, these ground-truth annotations do not constitute full segmentation masks, but rather served as partial training data for qualitative and limited quantitative assessment.

Simulating EM Data

To overcome limited availability and sparsity of annotated real EM data, we implemented a simulation framework capable of generating synthetic images of amyloid fibres under controlled and tunable conditions. This simulation software allows the generation of arbitrarily large datasets with fully known ground truth, making them particularly suitable for deep learning approaches that require extensive annotated data. In addition, the simulator provides precise control over key image properties, including fibre density, overlap, clustering severity, signal-to-noise ratio, and contrast. As such, it enables systematic investigation of the specific factors that influence segmentation performance. In the current implementation, all generated images match the resolution of the real EM images producing outputs

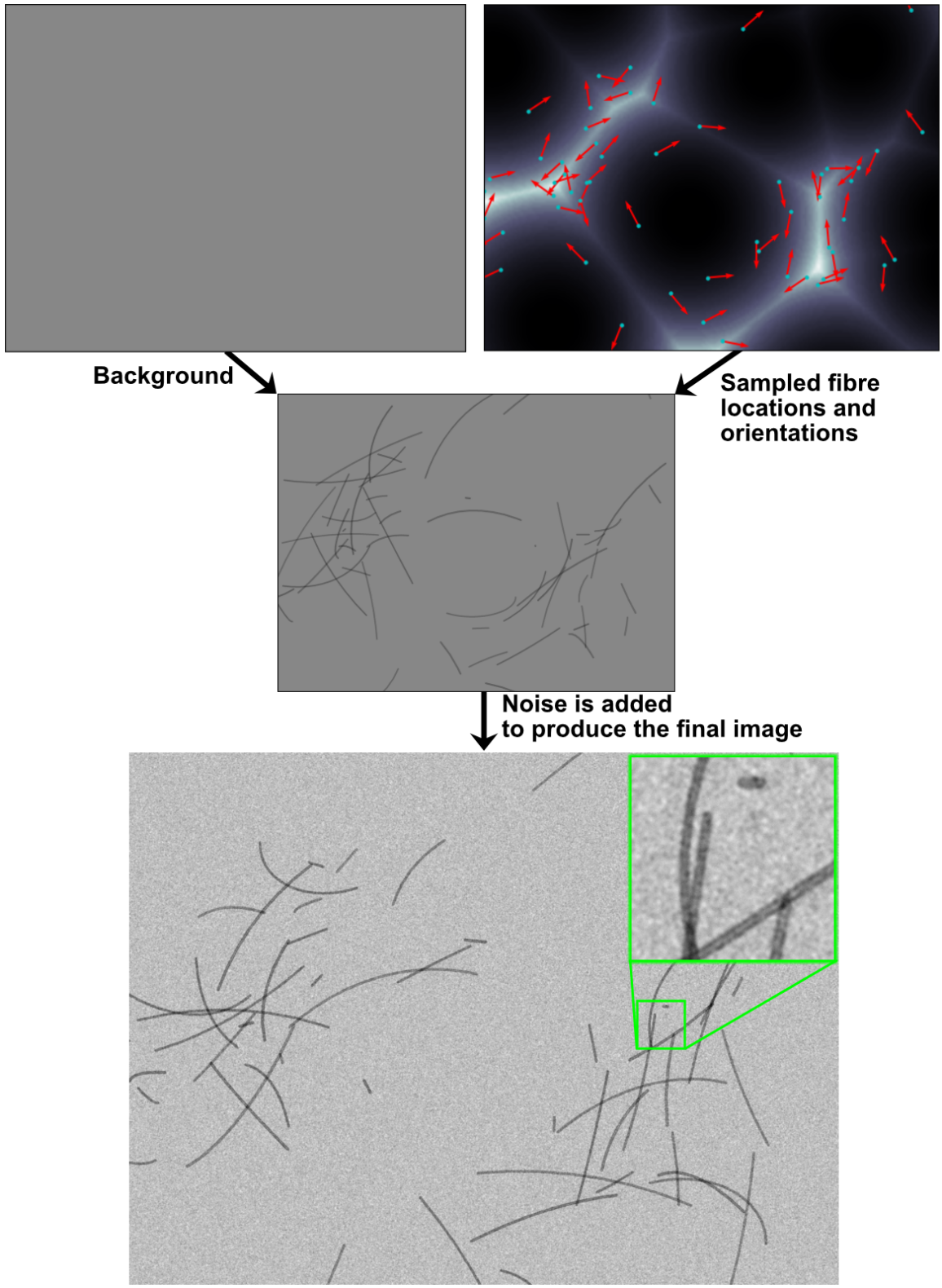
that are 3296×2472 pixels in size. Fibre length, curvature, and thickness are modelled in absolute pixel dimensions, which ensures consistency with real data but can be readily adjusted by scaling these parameters if needed.

Each synthetic image is initialised by creating a blank background with an intensity value that can be specified by the user or determined automatically based on default ranges derived from real data. After this, the fibres, which are segments of an ellipse, are placed in the image based on a distribution generated by a noise function. Most ellipses used to produce the fibres are very large and paired with how narrow the angular segments are they produce lines which are only slightly bent. It is possible for fibres to be made up of larger angular segments of smaller ellipses to simulate bent fibrils. The segment of an ellipse This distribution is a small section of a precomputed pseudorandom Worley noise map which simulates a cluster-like distribution. Finally, the image is post-processed by adding multistage random noise with a strength that ensures the desired signal-to-noise ratio. Each step of this image generation process is described in detail in the following sections.

Fibre Generation. Each synthetic fibril is rendered as a curved segment of an ellipse, designed to approximate the morphology of amyloid fibres. While it is true that not all fibres are bent, the simulation makes up for it by creating slightly bent small segments of very large or flat ellipses, which we presumed would be harder for the model to learn compared to perfectly straight lines. The ellipse is defined by two randomly sampled semiaxes, which determine the length and curvature of the fibre. To avoid unrealistically long or bent structures, the maximum arc angle is constrained on the basis of fibre length and never exceeds 180 degrees. The fibrils' starting coordinates and orientations are sampled from a distribution determined by the clustering algorithm (see Clustering Algorithm). The coordinates and orientation of the end of the fibre are also calculated in accordance to the general formulas for ellipses to allow for connecting multiple such fibre segments into more complex looking fibre morphologies. However, this is not present in the current implementation. After the fibre locations and structures are determined, the fibres are drawn onto the image one after another. The width of a fibre is randomly chosen between 9 and 11 pixels to provide some minor variation.

Figure 4

Synthetic image creation pipeline.



The intensity of the fibre can be chosen by the user or automatically sampled from a default range derived from real data. Simply setting the shade of the fibre in the image to the selected intensity would lose information at overlap boundaries and not accurately reflect real data, where overlaying multiple fibres causes the overlapped region to be thicker, which in turn diffracts more electrons and produces a darker shade than either fibre does individually. On the other hand, subtracting to reach the desired shade will quickly turn these intersections black after just a few overlaps. Instead, we presumed that the darkening on overlaps would happen on a multiplicative basis. If the stain of a fibre stops a portion of the electrons which, if uninterrupted, would yield background intensity, the next fibre underneath will stop a similar portion of the remaining electrons. Note that this is not necessarily physically or technologically accurate, but modelling fibres as multiplications of the background provides a more believable depiction of overlaps. As mentioned in Real EM Data Characteristics, staining causes fibres to have brighter hollow cores. This is simulated in the data by drawing onto each fibre a lighter inner mask that is narrower and represents the lighter central core. Its width is calculated as a fraction of the outer width, with slight random variation introduced to emulate natural heterogeneity.

Noising Images. To simulate the visual complexity of real negative stain EM images, each synthetic image includes controlled random noise. The strength of the signal that is generated on top of the image is governed by the background-to-signal intensity ratio and a target signal-to-noise ratio chosen by the user or sampled from defaults derived from real data. The difference between the background and fibre intensities, both chosen by the user or randomly, defines the effective signal as described by Equation 2. Given a target SNR, the required noise standard deviation σ_{noise} is computed as:

$$\sigma_{noise} = \frac{P_{signal}}{SNR} \tag{5}$$

This value represents the standard deviation of the noise needed to achieve the desired SNR between the fibre and the background. To introduce realistic texture and artefacts, noise is added in two stages. First, low-frequency noise is applied, using a Gaussian-blurred noise map to introduce broad variations across the image. This simulates uneven staining or imperfect focus commonly seen in EM images.

Second, high-frequency noise is added to introduce more fine-grained noise present in the real images.

Adding noise in two stages with Gaussian blurs is tricky if the post-blur noise is supposed to deliver the desired SNR ratio. After applying a Gaussian blur to a noisy signal, the standard deviation of the new signal can be approximated as:

$$\sigma_{out} \approx \frac{\sigma_{in}}{\sigma_G \cdot 2\sqrt{\pi}} \tag{6}$$

where σ_G is the standard deviation of the Gaussian kernel. Accordingly, the input noise standard deviation σ_{in} can be scaled prior to application to compensate for the subsequent blurring. In our implementation, the high-level and low-level noise contribute to the equal noise equally (each one half of the final standard deviation), but the compensation is calculated independently for each due to the lower standard deviation of the Gaussian kernel of the low-level noise.

Clustering Algorithm. In real electron micrographs, amyloid fibres are not uniformly distributed and frequently form clustered tangles instead. The occlusion and intertwining that this clustering causes is the main factor complicating segmentation. Because of this, one of the main goals of this thesis is to investigate the impact of clustering on model performance. For this reason, the generator must realistically emulate this behaviour in synthetic data. Sampling fibre positions and orientations according to distributions generated by noise functions creates biologically realistic pseudorandom clusters. It is important to note that in this context, *'noise'* refers to stochastic distribution maps, and is distinct from usually unwanted image noise that degrades visual quality. In our implementation of clustering, noise functions produce regions of higher or lower spatial density. The degree of clustering can then be controlled by varying the intensity of these noise distributions.

To generate these distributions, we employ two types of procedurally generated noise: Simplex noise [37] and Cellular noise [54] (also known as Worley noise). Both are pseudorandom, which is a desirable property when modelling biological phenomena which are usually neither purely regular nor fully random. Simplex noise is a widely used method in procedural generation due to its ability to create randomly appearing structures connected by smooth natural gradients [15]. Its diffuse and patchy distributions resemble the densities we extracted from real EM images of amyloid fibres (see

Fibre Distribution). In contrast, Worley noise was explicitly designed to simulate biological patterns such as cells, scales, or vein networks [54]. It forms a distribution of clusters separated by voids and connected by thin paths. This offers a compelling analogue for the way fibres often arrange themselves into tangled tufts and align into strings.

The distributions are sampled from a large prebaked noise distribution. This avoids repeating the time-intensive noise generation for each image. The distribution is 5000×5000 pixels and for every fibre image a 206×153 patch is extracted from which the starting locations are sampled and then jittered by up to 1% of the image size. This overcomes the discrepancy between the large image size and the smaller noise patch, which is required for faster computation.

To control the amount of clustering, the 2D noise field is converted into a probability surface P:

$$P(x, y) = [\max(0.12, N(x, y))]^{\gamma} \tag{7}$$

where $N(x, y)$ is the noise value at (x, y), and γ is the chosen clustering amount.

Clipping the lowest valleys in the noise to 0.12 ensures that even the darkest parts of the noise retain a non-zero probability of fibre placement. This prevents completely empty regions and produces a low trickle of isolated fibres, which can also be seen even in very clustered real EM images.

Clustering is controlled by exponentiating the noise map. This increases contrast of the noise, thereby concentrating the sampling probability into more compact zones. Setting $\gamma = 0$ results in a flat probability map, producing fibres in a completely random distribution. Higher exponents create more tight clusters with realistic images being produced by clustering ranges of around 0 to 10.

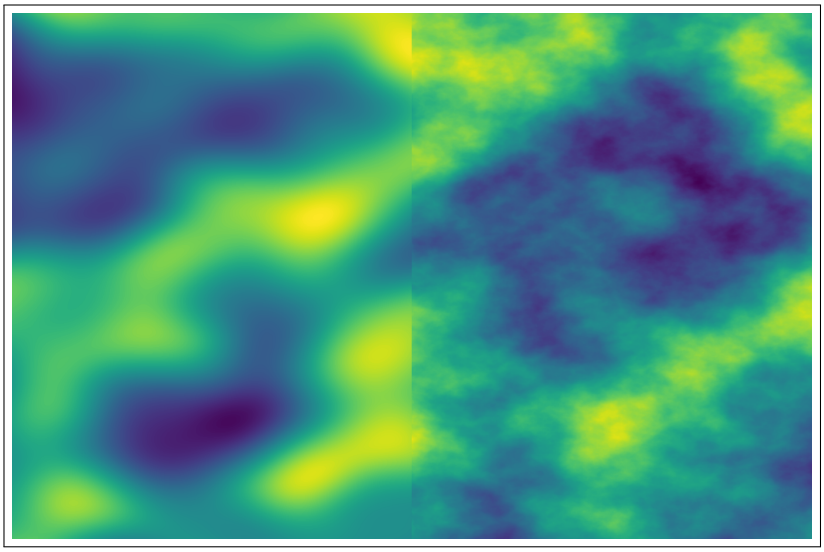
After normalising P so that $\sum_{x,y} P(x, y) = 1$, source pixels are chosen proportional to their value in P. In this way, the clustering exponent γ acts as a single continuous hyperparameter that moves the generator between sparse, random placements and concentrated tufts.

Simplex Noise. The first noise used was Simplex noise. Simplex noise is a noise algorithm introduced in 2001 by Ken Perlin [37] and, along with its predecessor Perlin noise, is often used in procedural graphics

and texture generation. Our implementation follows the formulation presented in Gustavson (2005)[16], and adapts it into Python naively by following the underlying mathematical structure.

Figure 5

Simplex Noise.



Note. The split image shows the first 2 octaves (left) and first 8 (right) octaves generated by one simplex noise function. Each additional octave adds finer-grained detail to the noise.

Simplex noise is a gradient noise algorithm. This means that instead of assigning a value to each location in the space directly, it generates an evenly spaced grid of gradients and computes the value of each point by interpolating between them. For 2D simplex noise, this grid is made up of unilateral triangles, and gradients are generated at each corner. For any given sample location, the algorithm first determines which simplex cell (a triangle of two dimensions) contains the point and then calculates the contribution of each of the cell's corners. The contribution is the product of the extrapolation of the gradient with a radially symmetric attenuation function designed to reach only the neighbouring 6 cells from each corner. This ensures that the value of any point will be determined by at most 3 corners (the corners of the triangle cell the point is in), which is desirable for fast and simple calculation.

Gradients at each corner are chosen pseudorandomly to allow for deterministic noise sampling without generating the entire noise. In our implementation, these gradients are selected from a fixed set of evenly spaced unit vectors pointing along the diagonals: [1, 1], [1, -1], [-1, 1], and [-1, -1]. Which gradient is

chosen is determined by a hashed permutation table which maps the starting coordinates of the sampled point's cell to a gradient index. Once the gradients are determined, the noise contribution from each corner is calculated by first forming a vector from the corner to the sample point and then computing the dot product between this vector and the gradient vector at the corner. The contributions of each corner to a point p_i are then weighted using a attenuation function $f(p_i)$ so that only the 3 corners of the current simplex contribute:

$$f(x_i) = 0.5 - r_x^2 - r_y^2 \tag{8}$$

where r_x^2 and r_x^2 are the distances of p_i from the current corner along the x and y axis, respectively. The attenuation function ensures that the contribution of each corner fades smoothly to zero as the distance increases, which ensures visually smooth noise fields.

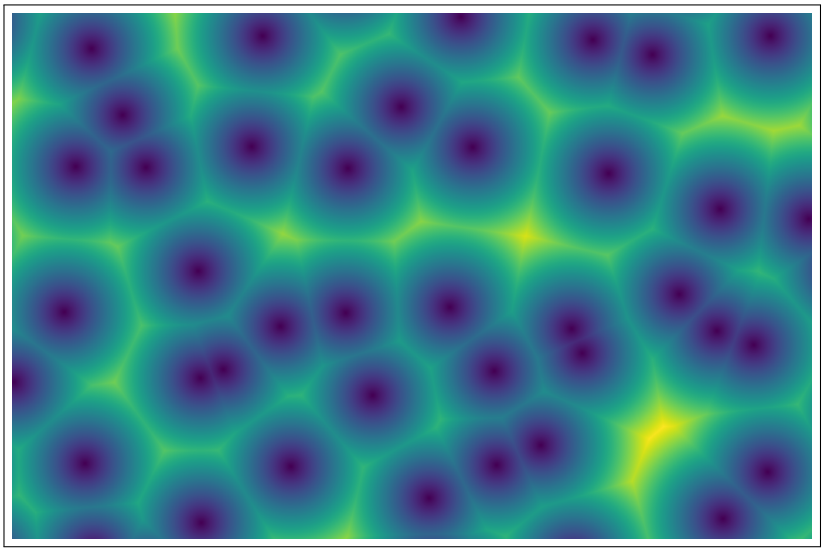
To introduce additional detail into the noise, multiple layers of noise known as octaves can be combined. Each octave is an independent instance of simplex noise but with doubled frequency and halved amplitude. Octaves at multiple scales are summed into the final noise where lower-frequency octaves capture coarse, large patterns, while higher-frequency octaves add finer detail and irregularity.

Cellular Noise. Cellular noise was introduced by Steven Worley with the intent of modelling natural textures such as skin cells or vein networks [54]. Its small and round peak regions connected by narrow strings noticeably resemble the tufts and strings of connected fibres seen in electron micrographs. Due to this, we have also implemented it as a source of fibre distributions based on the original formulation by Worley [54] and optimised based on the adaptive traversal strategy described by Jonchier et al. (2019) [22].

The algorithm defines a regular 2D grid of cells, each containing a single feature point whose exact position is determined procedurally using a multiplicative congruential generator (MCG). MCG is a hashing scheme used to generate pseudo-random sequences of numbers. Given a continuous query coordinate (x,y), the noise value is computed as the distance to the nearest feature point from the surrounding neighbourhood. This distance-based formulation naturally produces low-density zones near

Figure 6

Cellular Noise.



Note. Small sharp peaks and empty valleys make cellular noise a plausibly looking source of fibre distributions.

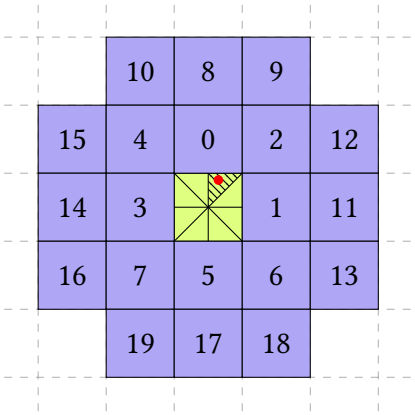
feature points and high-density strings and peaks in-between them.

To compute the distance efficiently, we first determine the cell containing the query point and calculate the local offset within that cell. The current cell is partitioned into eight directional slices, each associated with an optimal traversal order of neighbouring cells. These traversal orders prioritise cells most likely to contain the closest feature point, reducing the number of unnecessary evaluations. For axis-aligned neighbours, lower bound distances are precomputed analytically based on the relative offset of the query point within its cell. For example, for a point in the middle of a cell with a relative offset of (0.5, 0.5), the neighbour directly above it will have a lower bound distance of 0.5 and the neighbour one more above will have a lover bound of 1.5 (0.5 for the distance to the top border of the point's cell and 1 for the height of the cell in-between). If the currently known minimum distance is already lower than the bound for a neighbour, the algorithm safely skips evaluating that neighbour and all other cells in its row or column (which are at a diagonal from the current cell and will thus always have higher lower bounds than the neighbour directly above, below or to the side). This pruning reduces the computational load

required to compute the noise value.

Figure 7

Optimal Neighbour Traversal Order of Cellular Noise.



Note. The optimal traversal order from Jonchier et al. (2019) [22]. The currently sampled point is represented by a red dot and its corresponding slice in the current cell is striped. The number in each neighbouring cell is its index in the traversal order that is optimal for the given slice.

Direction Sampling. In real micrographs, fibres often align along their axes and stretch from clusters along thin strings in a radial fashion. For every chosen fibre position, we need an initial heading that mimics the orientations of fibres in real images. In cellular noise, we obtain this using a weighted average of the probability corresponding to the noise itself and its gradient.

1. The 2D gradient of the underlying distribution is evaluated at the starting location. Because fibres in real EM micrographs tend to lie along thin ridges strands that run between neighbouring clusters, we rotate the gradient by 90 degrees. The resulting angle thus points tangent to the ridge rather than up towards it or down towards the empty region.

2. Within a square window of 23 pixels centred on the starting location, the same probability map P is used a second time as an angular weight. Each pixel in the window corresponds to the starting direction that joins it to the starting location. This makes fibres more likely to point towards nearby regions of high fibre density while the limits of the window prevent each fibre from pointing towards the global maximum in the largest cluster.

The two angles are blended to get the starting direction θ inversely according to the confidence carried by the gradient. This ensures that a fibre with a low gradient farther away from a ridge or a cluster is more likely to point towards other fibres in a dense region:

$$\theta = \arctan\left(\frac{w * \sin(\theta_{grad}) + (1-w) * \sin(\theta_{noise})}{w * \cos(\theta_{grad}) + (1-w) * \cos(\theta_{noise})}\right), \quad \text{with } w = \frac{\Delta P}{\Delta P + 1} \in [0, 1] \tag{9}$$

where ΔP is the gradient of P, θ_{grad} is the angle perpendicular to ΔP, θ_{noise} is the angle chosen according to P, and w is the gradient weight.

Ground Truth. Because the simulated fibres are programmatically drawn onto the image canvas, the precise pixel-wise ground truth for each fibre is available at generation time. During rendering, each fibre is stored as a binary mask identifying which pixels belong to it. These masks are then used to produce segmentation annotations in one of two formats. For YOLO models, COCO-style polygon annotations, contours are extracted from each fibre mask using OpenCV's contour detection and simplified with OpenCV's polygon contour approximation which uses the Ramer–Douglas–Peucker algorithm to reduce the contour into fewer points. These polygons are then normalised to match the dimensions of the image and written into text files following the segmentation format following ultralytics documentation [47]: `<class index> <x1> <y1> <x2> <y2> ... <xn> <yn>`. Alternatively, for lossless saving of segmentation masks compatible with the Segment Anything Model (SAM), the binary masks are saved as compressed run-length encoding (RLE) strings into a JSON structure following the structure of the SA-1B designed for SAM [31].

Models

To evaluate segmentation of amyloid fibres in electron micrographs, we selected three deep learning models that represent state-of-the-art (SOTA) approaches in instance segmentation. These were the YOLO v11-large, the YOLO v11-nano, and the SAM v2.1-tiny. These models were chosen to capture a spectrum of model sizes and architectures. The number of learnable weights of each model is shown in Table 1 and determines the complexity of each model, with more complex models with more trainable

parameters being capable of capturing more intricate patterns.

Table 1
Comparison of models.

Model	**Num. Trainable Parameters (in Million)**
YOLOv11-large	25.3
YOLOv11-nano	2.6
SAMv2.1-tiny	38.9

Note. A comparison of the complexities of the three models we evaluated. A higher parameter count generally means higher model capacity.

The YOLO family of models initially introduced in Redmon et al. (2016) [40] are a family of single-stage object detectors and segmenters [23]. This means that the entire object detection and classification, and in later versions instance segmentation pipeline, is performed by a single convolutional neural network. This streamlines training and speeds up inference. Since its introduction, YOLO has undergone a series of improvements, including the introduction of spatial attention mechanisms [23]. We include the large and nano variants of the YOLOv11 model to compare the effect of model capacity on performance and generalisability. This is crucial as larger models could suffer from overfitting when detecting simple geometrical shapes such as fibres.

The Segment Anything Model (SAM), developed by Meta AI, is a foundation segmentation model designed to generalise easily across new domains [24]. Despite being significantly larger than both YOLO models, the SAM v2.1-tiny with almost 40 million learnable parameters is the most lightweight variant of the original model. SAM uses an encoder-decoder transformer architecture. It uses a vision transformer to turn the input image into tokens, and a prompt encoder which uses the positions of points to create tokens. These tokens are then used by a transformer mask decoder, which updates the tokenized embeddings iteratively according to their interactions. It also creates an output mask token which is updated as a part of this process until the final output mask is formed.

For all models, we used pre-trained weights provided by Ultralytics [21] and Meta [9]. The YOLO models are trained on the COCO-seg dataset comprising 330,000 images of everyday objects [46]. The

SAM model is trained on the Segment Anything dataset with 11 million segmented photographs [31]. We hope that using pre-trained weights speeds up training, as models maintain some universal features from previous large-scale training.

Training

Synthetic Dataset. Our models were trained and evaluated on a large, synthetically generated dataset of images containing simulated fibres. A total of 3200 images were generated and partitioned into three sets:

1. 2000 training images
2. 200 validation images
3. 1000 test images

The generation process was automated using our image generation framework. All images were generated at a high resolution of 2472×2472 and matching the resolution of images in the real EM datasets. Each generated image was defined by a set of stochastically determined parameters which controlled its simulated imaging conditions.

Fibre Count. The quantity of fibres in each image was sampled from a normal (Gaussian) distribution with a mean (μ) of 45 and a standard deviation (σ) of 15. This allowed for unlimited variation in fibre density while ensuring realistic fibre counts for images with mixed fibre lengths (some real images contain hundreds of tiny fibre specs).

SNR. To simulate different levels of image clarity, the signal-to-noise ratio of each image was a real number randomly sampled from a continuous uniform distribution over the interval [2.5, 12.5]. Lower values correspond to images with higher noise.

Background and Fiber Intensity. The intensity of the background of each image was selected from a discrete uniform distribution and was an integer between 30 and 220 (both included). The intensity of

the fibres, was then derived from the background colour to be dimmer. The contrast between the fibres and the background (or background to signal ratio) was sampled from a continuous uniform distribution over the range [1.2, 1.8]. The intensity of fibres C_{fibre} is given by:

$$\left\lfloor \frac{C_{bg}}{\text{BGSR}} \right\rfloor \tag{10}$$

where C_{bg} is the intensity of the background and BGSR is the chosen background to signal ratio. The strength of the noise added was calculated to maintain the chosen signal-to-noise ratio as described in Equation 5.

Clustering. Lastly, each image had a randomly chosen clustering parameter γ controlling the amount of fibre clustering as described in Clustering Algorithm. Clustering levels ranged between 0 and 10 (both included). A γ of 0 is equivalent to completely random fibre positions and orientations, while 10 is equivalent to heavy overlap with most fibres generally confined in one or two large clusters.

All images were further resized to a scale of 1024×1024 bilinear interpolation. This was done to match the expected input dimensions of the SAM model and to ensure fast training and inference. Despite this being a relatively high resolution for DL computer vision models, some detail is inevitably lost through this transformation. Annotations of the synthetic images were saved in the formats of the COCO [47] and SA1B [31] dataset as described in the section on simulated data Ground Truth.

Real Data. To test model generalisability we utilised a small partially annotated dataset of 30 real EM images. The 30 images of the real dataset were split into two sets:

1. Training set of 25 images
2. Test set of 5 images

Images in each set were then resized from 3296×2472 pixels to 1365×1024 pixels, and each image was tiled into two overlapping horizontal tiles of size 1024×1024 pixels to match the model input dimensions. The horizontal overlap between the two tiles of each image is 683 pixels. It is important to

note, that the tiling was performed after splitting the data into training and test sets and so there was no leak of information from the testing images into the training dataset.

The annotation masks were extracted as described in the chapter on real data Ground Truth and saved in the formats of the COCO and SA1B datasets. This task was facilitated by the utilities implemented in our simulation framework, which automatise the process of saving binary masks corresponding to an image into these two formats.

Hardware. All models were trained and evaluated on an Ubuntu 22.04 system equipped with the following specifications:

GPU: 2x NVIDIA GeForce RTX 5090 Graphics Cards with 32GB of VRAM each.

CPU: AMD EPYC 9354 32-Core Processor.

RAM: 234GB of system memory.

Software: CUDA 12.8 and cuDNN 8.9.7 were used for GPU acceleration.

Training Parameters.

YOLOv11-large. The YOLOv11-large segmentation model was trained on the synthetic train dataset of 2000 images on both GPUs simultaneously.

The model was configured to train for a maximum of 1000 epochs with a batch size of 24. To avoid overfitting and save computational resources, we enabled the early stopping mechanism with a patience of 50 epochs. This mechanism monitors the loss on the validation set and halts training if no improvement is observed for 50 consecutive epochs. The training process stopped automatically at epoch 835 with the best performance on the validation set being achieved on epoch 785. The training took approximately 6.5 hours. As our dataset contains only one object class, we enabled single class training. It simplifies the classification head of the model to a binary foreground/background task, which can improve focus on object localisation and segmentation. We also explicitly set mask overlapping to False. In the YOLO training suite, this ensures that when two masks overlap, their corresponding segmentation masks are

kept separate and the smaller object is not overlaid on the larger. Except for this, the training settings, learning rates, optimisers, etc. followed the default settings described in Ultralytics documentation [48].

The weights from epoch 785, which achieved the best performance on the validation set were used as our weights for the baseline YOLOv11-large model.

YOLOv11-nano. The training settings for the nano YOLOv11 were nearly identical to those we used for training the YOLOv11. The early stopping mechanism in this case followed a patience of only 20 epochs to accommodate the faster convergence of the smaller model. The batch size was changed 28. An important difference is that the nano model was only trained using one GPU. The early stopping was triggered after epoch 415 with the lowest validation loss being achieved on epoch 395. The training took approximately 2.3 hours.

The weights from epoch 395, which achieved the best performance on the validation set were used as our weights for the baseline YOLOv11-nano model.

SAMv2.1-tiny. The SAMv2.1-tiny was trained for 1000 epochs and model weights were evaluated on the synthetic validation set every 50 epochs. The data loader was configured to normalise the dataset with a mean value of 0.482 and a standard deviation of 0.213 for each colour channel. These values correspond to the mean pixel intensity and the standard deviation of the pixel intensity in our training dataset.

During training, a large set of data augmentation techniques provided by the SAM training framework was applied to improve generalisation and artificially increase the amount of unique training images. These included random horizontal flipping and random affine transformations. The latter introduced variation through random rotations of up to 25 degrees and shear transformations of up to 20 degrees. The original implementation of SAM's data augmentation utilities skips random affine transformations if by the means of one of the transformations an object is edited out of the image. Because our training images contain a large number of fibres, at least one fibre will virtually always be occluded by a transformation, which leads SAM to skip random affines for almost all images. To redress this, we modified the data augmentation code to skip the transform only if all objects are removed from the image through the

transformation. Additionally, colour space augmentations were applied. Especially colour jittering is applied which changes the brightness by 0.1 to 0.2. Then, contrast and saturation are increased by 0.03 to 0.08.

The AdamW optimiser was used. The learning rate of the image encoder was 3.0×10^{-6} and decayed following a cosine schedule to 3.0×10^{-7}. The rest of the model's parameters used a higher learning rate, starting at 5.0×10^{-7} and decaying with a cosine schedule to 5.0×10^{-7}.

Evaluating every 50 epochs, the model achieved the best results in terms of F1 score after 950 epochs and the weights from epoch 950 were used as our baseline SAMv2.1-tiny model.

Fine-Tuning on Real Data. To adapt the models to the specific characteristics of real-world images, we conducted fine-tuning of each model. The fine-tuning dataset was derived from 25 real EM images. To make these images suitable for the training pipeline, each of the 25 images was partitioned into two smaller overlapping tiles. This process resulted in a fine-tuning dataset composed of 50 different training tiles.

All models were initialised with the weights from their best-performing checkpoints from the synthetic data training phase. The models were trained for a fixed duration of 60 epochs on the 50 real images. All other configuration and hyperparameters including batch size, optimisers, learning rates, and data augmentations for each respective model remained identical to the settings used during the first training phase.

Evaluation. We conducted a systematic analysis of model performance on the held-out synthetic test set. To generate predictions for evaluation, the YOLO models were applied directly to the test images. For the Segment Anything Model (SAM), which requires prompting, a different approach was necessary. As some does not detect objects itself, to evaluate its mask generation quality, we prompted SAM with a single point sampled from each ground-truth mask. This methodology was chosen because SAM's default automatic mask generation, which applies a 32×32 grid of prompt points across the entire image, is not suitable for this task as the chance of a larger number of thin fibres being randomly picked is minimal. We must acknowledge that this approach provides an advantage to SAM by supplying the

location of each ground-truth object. We collected the masks predicted by each model for all 1000 test images with known parameters. For each test image, we computed a greedy matching strategy, where a predicted mask was successfully matched to a ground-truth mask if their intersection-over-union (IoU) reached at least 0.5, and the ground mask did not have a better fit in terms of IoU among the predicted masks.

These successful matches were classified as true positives (TP). False positives (FP), then, were defined as the number of unmatched predictions, and false negatives (FN), on the other hand, we defined as the count of unmatched ground-truth fibres. These counts were used to calculate standard performance metrics, including precision, denoting the accuracy of predictions, and recall, which measures what portion of ground truth fibres were detected successfully. Last, we calculated a harmonic measure of these two metrics called the F1 score for a balance between these two distinct measures of performance.

Performance metrics were analysed as a function of the parameters used to generate the synthetic test images. The results from the test set were aggregated and binned to measure the mean F1-score, precision, and recall across different levels of:

- **Signal-to-Noise Ratio**: Grouped into bins of [2.5, 5), [5, 7.5), [7.5, 10), and [10, 12.5].
- **Clustering**: Grouped into bins of [0, 2], (2, 4], (4, 6], (6, 8], and (8, 10].
- **Fiber Count**: Grouped into bins of [0, 30], (30, 60], (60, 90], and (90, 120].

These mean metrics of each bin and over the whole dataset were averaged on a per-image basis instead of the total number of TP, FP and FN. This is because each image is an individual data point which represents unique image conditions and it is impossible to determine the properties such as clustering or overlap of individual fibres. Additionally, this ensures that images with higher fibre counts are weighted equally to other images.

In the same process, we analysed the 5 annotated real test images split into 10 tile data points. This analysis is not as extensive, as we do not have complete knowledge about the properties of these images,

as we do with simulated images, and has limited informative value as it is performed on a very small set that is only partially annotated.

Fibre Length Measurement Error. The length of a fibre was determined using its mask. It's segmentation mask was reduced to a 1-pixel wide skeleton using the using the Zhang-Suen thinning algorithm [57]. The algorithm follows an iterative erosion process that thins down a shape to its skeleton. The algorithm repeatedly passes over the shape, removing outer-layer pixels from its boundary in each pass. A pixel is only removed if doing so does not interrupt the structure of the shape or shorten the shape. This process continues until no more pixels can be removed without damaging the structure, leaving a centred, 1-pixel-thick line representing the line following the object's shape. This algorithm is ran on a fibre's mask and once the mask has a thickness of just one pixel, it's nonzero pixels are summed which gives the length of the masks center line.

For every correctly matched true positive, the length of the predicted fibre and the ground-truth fibre was calculated using the method mentioned above. The relative length error for that instance was then computed as:

$$\text{Length Error} = \frac{|L_{pred} - L_{gt}|}{L_{gt}} \tag{11}$$

where L_{pred} is the length of the mask predicted for a ground truth fibre of length L_{gt}.

The relative length error ensures that shorter fibres contribute equally to longer fibres. For each image, the length error was aggregated as the median relative length error to ensure that a small number of incorrectly segmented outliers does not skew the distribution.

Results

The main goal of this thesis was to assess whether deep learning models can accurately segment amyloid fibrils with significant particle overlap, clustering, and noise. Secondly, we sought to explore whether models trained on synthetic data generalise to real EM images and to what extent are they able to accurately extract the length of fibres.

To investigate these questions, we evaluated three state-of-the-art segmentation architectures: YOLOv11-large (a high-capacity CNN), YOLOv11-nano (a lightweight version of YOLOv11), and Segment Anything Model (SAM) v2.1-tiny (A general purpose segmentation transformer).

Our experimentation followed two stages. Each model was first trained and tested on a large set of 2000 synthetic EM images that mimic a range of challenging imaging conditions, which allowed systematic performance evaluation in a controlled setting. Each model was then fine-tuned and tested on a small dataset of 25 manually annotated real micrographs. The results of each stage are described in the sections that follow.

Synthetic Benchmark

To assess baseline model performance under controlled conditions, we conducted a large benchmark on a synthetic dataset comprising 1000 images. The images contained a variable number of simulated amyloid fibres, had varying signal-to-noise ratios, and features varying levels of clustering. Each image was generated with known ground truth segmentations, SNR, clustering level, number of fibres, and colours. This setup enables precise evaluation of model performance on "perfect" data.

Because each image represents unique imaging conditions, we collected the metrics on a per-image basis. For each image we computed the image's precision: proportion of predicted fibres that matched a ground truth fibre, recall: proportion of ground truth fibres that were correctly detected, and F1-score: a harmonic mean of precision and recall.

Final results for each model were obtained by averaging these metrics across all 1000 synthetic images. This ensures equal weighting across different fibre densities, noise levels, and clustering levels.

Impact of Signal-to-Noise Ratio. As seen in Table 2, model performance was remarkably stable across varying levels of signal-to-noise ratio. All three models achieved high F1 scores across the full SNR range. Notably, YOLOv11-large performed the best across all metrics in images with a lower SNR, but the differences are very subtle.

Overall, all models are very resilient to varying noise levels and capable of segmenting fibres accurately even under low-contrast conditions.

Table 2
Model performance across SNR bins.

	SNR	F1-score (mean)	Precision (mean)	Recall (mean)
YOLOv11-large	[2.5, 5.0]	**0.889**	**0.900**	**0.879**
	(5.0, 7.5]	0.885	0.894	0.876
	(7.5, 10.0]	0.879	0.889	0.869
	(10.0, 12.5]	0.880	0.889	0.872

	SNR	F1-score (mean)	Precision (mean)	Recall (mean)
YOLOv11-nano	[2.5, 5.0]	0.827	**0.826**	0.830
	(5.0, 7.5]	0.815	0.808	0.825
	(7.5, 10.0]	0.821	0.820	0.824
	(10.0, 12.5]	**0.828**	0.825	**0.832**

	SNR	F1-score (mean)	Precision (mean)	Recall (mean)
SAMv2.1-tiny	[2.5, 5.0]	0.885	0.885	0.885
	(5.0, 7.5]	0.878	0.878	0.878
	(7.5, 10.0]	0.898	0.898	0.898
	(10.0, 12.5]	**0.903**	**0.903**	**0.903**

Impact of Clustering. The impacts of varying levels of clustering are shown in Table 3. Clustering had a pronounced impact on model performance as it puts fibres in close proximity and overlap, which makes resolving individual fibres more difficult. As the degree of clustering increased, model performance decreased substantially, particularly for the SAMv2.1-tiny and YOLOv11-nano models.

Clustering's effects are the most pronounced in the SAMv2.1-tiny and YOLOv11-nano models. SAMv2.1-tiny, which achieved near-perfect performance under low clustering conditions, experienced a sharp drop in F1 score from 0.971 in the lowest clustering bin to 0.781 in the highest clustering bin. Similarly, YOLOv11-nano saw its F1-score decline from 0.887 to 0.724 over the same clustering range.

In contrast, YOLOv11-large demonstrated comparatively strong resilience to increasing clustering.

While its F1 score also decreased with clustering, the drop was much more moderate from 0.905 in the lowest clustering bin to 0.845 in the highest. Notably, YOLOv11-large heavily outperformed the other models under high clustering, despite achieving much lower performance that SAMv2.1-tiny under low clustering and only mildly outperforming its lightweight counterpart YOLOv11-nano in the lowest clustering range.

Table 3
Model performance across clustering levels.

	Clustering	F1-score (mean)	Precision (mean)	Recall (mean)
YOLOv11-large	[0, 2]	0.902	**0.914**	**0.892**
	(2, 4]	**0.903**	0.915	**0.892**
	(4, 6]	0.886	0.895	0.877
	(6, 8]	0.867	0.877	0.858
	(8, 10]	0.845	0.852	0.839

	Clustering	F1-score (mean)	Precision (mean)	Recall (mean)
YOLOv11-nano	[0, 2]	**0.887**	**0.893**	**0.882**
	(2, 4]	0.870	0.871	0.869
	(4, 6]	0.820	0.814	0.828
	(6, 8]	0.773	0.765	0.783
	(8, 10]	0.724	0.709	0.743

	Clustering	F1-score (mean)	Precision (mean)	Recall (mean)
SAMv2.1-tiny	[0, 2]	**0.971**	**0.971**	**0.971**
	(2, 4]	0.941	0.941	0.941
	(4, 6]	0.891	0.891	0.891
	(6, 8]	0.822	0.822	0.822
	(8, 10]	0.781	0.781	0.781

Impact of Fibre Count. Increasing the number of fibres in the image resulted in a pronounced decline in segmentation performance across all models (Table 4). This effect was even more pronounced than for clustering and can be attributed to the fact that fibre count directly governs the overall degree of overlap at all levels of clustering. That is, images with a high number of fibres exhibit more frequent and extensive occlusions compared to other images of the same clustering level, making individual object

boundaries more difficult to resolve.

For all models, the highest F1 scores were observed in the lowest fibre count bin of less than 30 fibres, with performance steadily deteriorating as fibre count increased. SAMv2.1-tiny began with an impressive F1 score of 0.953 but declined to only 0.626 in the highest bin of 90-120 fibres per image. YOLOv11-nano showed a similar trend, dropping from 0.878 to 0.601. Similarly to the effects of clustering, YOLOv11-large, despite achieving weaker results than SAMv2.1-tiny in the easiest bin with the lowest fibre count, markedly outperformed both other models at high fibre counts with its F1 score decreasing from 0.9 to 0.793. This is a decrease of only around 12%, compared to the drastic decline in the performance of SAMv2.1-tiny and YOLOv11-nano of more than 34% and 31%, respectively, across the same data ranges.

Table 4
Model performance across fibre count bins.

	Fibre Count	F1-score (mean)	Precision (mean)	Recall (mean)
YOLOv11-large	(0, 30]	**0.900**	**0.913**	**0.888**
	(30, 60]	0.883	0.893	0.874
	(60, 90]	0.857	0.863	0.850
	(90, 120]	0.793	0.785	0.802

	Fibre Count	F1-score (mean)	Precision (mean)	Recall (mean)
YOLOv11-nano	(0, 30]	**0.878**	**0.884**	**0.874**
	(30, 60]	0.820	0.816	0.827
	(60, 90]	0.751	0.739	0.765
	(90, 120]	0.601	0.598	0.604

	Fibre Count	F1-score (mean)	Precision (mean)	Recall (mean)
SAMv2.1-tiny	(0, 30]	**0.953**	**0.953**	**0.953**
	(30, 60]	0.888	0.888	0.888
	(60, 90]	0.820	0.820	0.820
	(90, 120]	0.626	0.626	0.626

Summary and Model Comparison. The results of testing on the synthetic dataset demonstrate several clear trends. First, noise alone is not responsible for lowering performance. All models maintained F1 scores above 0.80 throughout the entire SNR range (Table 2), indicating that the networks with

extensive data augmentation learn features that are robust to noisy signal.

When fibres become overlapped (Table 3 and Table 4) performance deteriorates markedly, especially notably for SAMv2.1-tiny and YOLOv11-nano. On the other hand, YOLOv11-large is distinctly more robust to increasing fibre overlap, suggesting that the additional capacity of the large backbone compared to YOLOv11nano and dense fully convolutional structure show a better ability to resolve close and overlapping objects.

Overall, SAMv2.1-tiny excels when fibres are well separated and reaches near perfect performance, but its reliance on prompts for global masking may make it vulnerable to dense and overlapping masks, or its design may be too focused on general complete segmentation of images instead of object detection. On the other hand, YOLOv11-large offers the best worst-case performance and degrades gracefully under heavy overlap, albeit at the cost of slightly lower peak performance compared to SAMv2.1-tiny. YOLOv11-nano provides a lightweight compromise and achieves competitive results, but due to its lower capacity it shares many of the limitations of SAM under heavier object density.

Table 5
Model performance on the synthetic test dataset.

Model	**F1-score (mean)**	**Precision (mean)**	**Recall (mean)**
YOLOv11-large	0.883	**0.893**	0.873
YOLOv11-nano	0.823	0.819	0.828
SAMv2.1-tiny	**0.891**	0.891	**0.891**

Over the whole dataset, as shown in Table 5, SAMv2.1-tiny achieved the highest mean F1 score across all synthetic conditions (0.891), slightly outperforming YOLOv11-large (0.883) and exceeding YOLOv11-nano (0.823). YOLOv11-large obtained the highest overall precision (0.893), which is a desirable trait in scientific analysis as false positives contaminate results. This is despite the fact that SAM was informed of fibre locations through point prompts. YOLOv11-nano performed the weakest in all metrics, but remained competitive given its considerably smaller size and complexity.

Finetuning on Real Microscopy Images

To assess how well the models generalise to real-world conditions, we fine-tuned each model using sparsely annotated electron micrographs. We exposed each model to 25 real images to familiarise them with the visual characteristics, noise patterns, and artefacts of real EM data. These images include background inhomogeneities, carbon film edges, broken film, and amorphous contamination that are not simulated in the synthetic dataset.

Following 60 epochs of fine-tuning, we evaluated the models on the annotated fibres in a separate set of 5 held-out real images. This evaluation aimed to determine the models' ability to transfer their learning on synthetic datasets to segment amyloid fibres in real images, after they were only minimally exposed to real images. The results convey the robustness of the models to domain shift from simulated to experimental data. Note, however, that these images, just as the real training images, are only partially annotated, which means that the results are only representative of the annotated fibres. Moreover, the training did not contain any empty background examples.

Table 6
Performance of fine-tuned models on five real test images.

Model	F1-score (mean)	Precision (mean)	Recall (mean)
YOLOv11-large (fine-tuned)	0.469	0.406	0.565
YOLOv11-nano (fine-tuned)	0.481	0.421	0.583
SAMv2.1-tiny (fine-tuned)	**0.601**	**0.601**	**0.601**

Note. Although the test set was only comprised of 5 images, each image contains 9 to 298 fibres and collectively, the test set contains several hundred annotated fibres.

The results of evaluating the fine-tuned models on five held-out real electron micrographs are summarised in Table 6. After being fine-tuned on only 25 annotated real images, all three models achieved reasonable performance, indicating some degree of successful domain adaptation from synthetic to experimental data.

Among the models, SAMv2.1-tiny achieved the best overall performance with an F1 score of 0.601. It outperformed both YOLO-based models by a substantial margin, suggesting that SAM's varied pre-

training and inherent segmentation capabilities provided robustness in difficult low-data conditions.

YOLOv11-nano slightly outperformed YOLOv11-large in this setting, achieving a higher F1 score (0.481 vs. 0.469) and slightly better precition (0.421 vs. 0.406) and recall (0.583 vs. 0.565). This contrasts with their performance on synthetic data, where YOLOv11-large was consistently superior. Given the small dataset, the high capacity of YOLOv11-large could not be fully leveraged, preferring the simpler relationships learned by the smaller model.

A qualitative comparison of the fine-tuned models' predictions on the five real test images is shown in Figure 8. These examples reveal several features of the segmentation performance in real-world conditions. All three models were often able to resolve tightly clumped fibres that even human annotators might find difficult to separate, such as the fibres in the central clump in Figure 8b and Figure 8g.

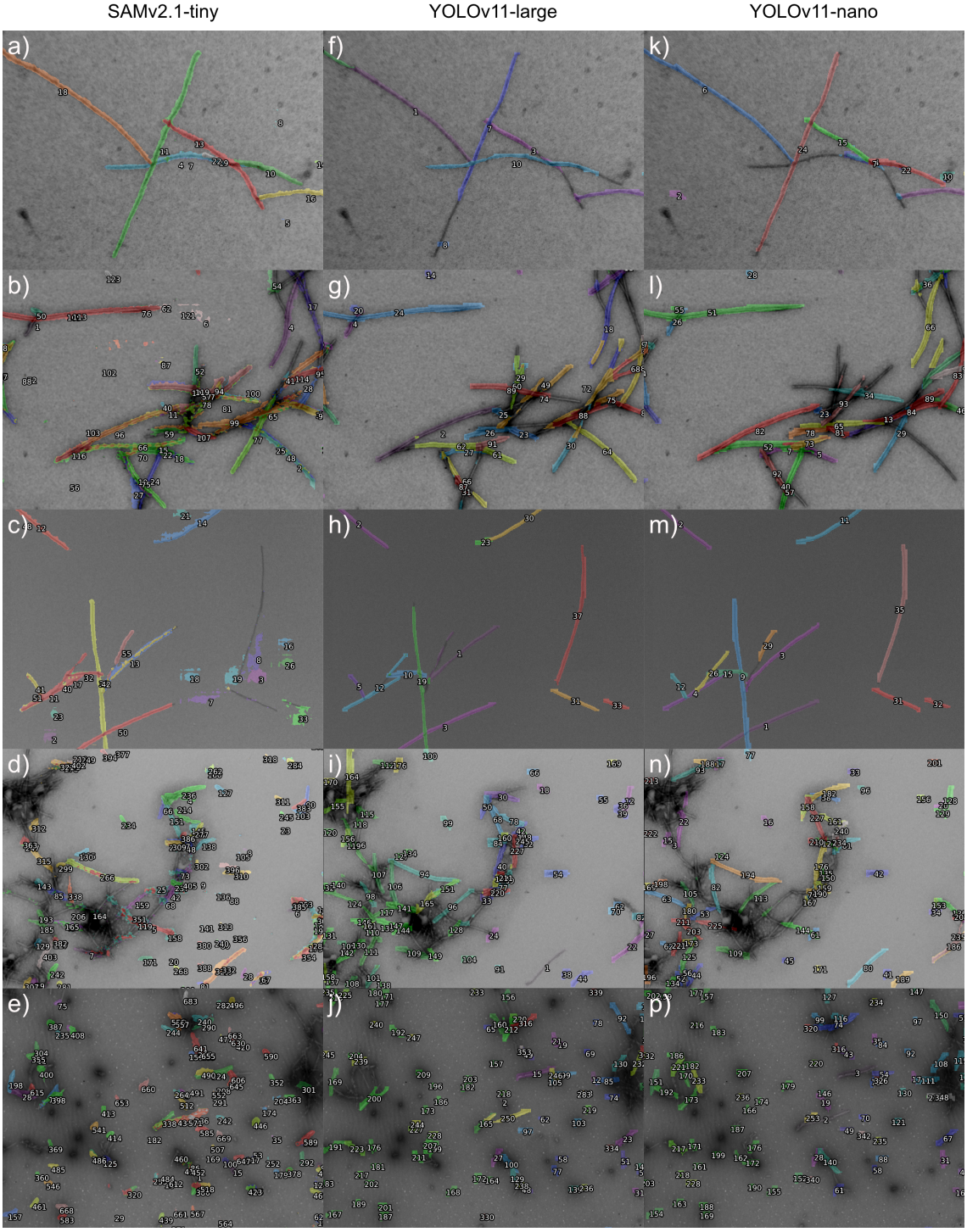
However, we also observed a high false negative rate in all models among the tiny dot-like fibres in the bottom image. This issue may stem from the lack of such extremely short fibres in the synthetic training data, or from a lack of positive-unlabeled (PU) learning in the training setup, which could otherwise mitigate learning bias from incomplete annotations. These tiny fibres are relevant for downstream analyses such as length distribution estimation. Their absence in the predicted masks thus poses a limitation for quantitative applications. False positive rate is higher in some images, like the first and fourth image, than the others, which may also suggest that the small training dataset had a bias.

Conversely, in most fibres that are clearly visible and do not suffer from a loss of signal due to overexposure are successfully detected and segmented somewhat successfully even in the presence of moderate overlap. For example, in the third image in Figure 8, all models successfully segmented almost all fibres.

SAMv2.1-tiny often produces remarkably clean and detailed segmentation masks, but its predictions are occasionally sometimes to over-segmentation in regions with stronger background noise, as seen by the background regions erroneously segmented as fibres in Figure 8b and Figure 8c. In contrast, when SAM does not exhibit this sensitivity, such as in Figure 8a, the resulting masks are exceptionally clean, with sharp contours and minimal noise. This dichotomy highlights SAM's potential for high-fidelity

Figure 8

Predicted segmentation masks of five real test images.



Note. This figure shows instance segmentation masks predicted by the three fine-tuned models. Each row corresponds to a different test image and each mask is overlaid on the image with a unique colour.

segmentation which was leveraged in previous works [17].

Despite these occasional artifacts, an important improvement over earlier prototype models is that none of the fine-tuned models erroneously segment contaminants (except for the treacherous fibrous contaminant in the third image which SAM, however, successfully ignores). Preliminary models during early experimentation frequently produced masks in regions such as carbon edges or at the edges of ice crystals or holes in the film. The fine-tuning process on even a small set of real images appears to have significantly improved robustness against this common failure.

Measuring Length. To evaluate whether we can accurately measure individual fibres, we compared the lengths of annotated fibres with the lengths of predicted fibres for correctly identified fibres (true positives). The results show that the length measurement errors range from approximately 5% to 33% with an average of around 15% as shown in Table 7. While this level of error is relatively high for the precise measurement of any single fibre, it is important to consider the wide range of fibre lengths present in the dataset. At the population level, the length distributions of fibres would reliably be captured even if single measurements have moderate errors.

Table 7
Precision of fibre length measurements of the three models. The mean value represents the mean of median length errors of 5 test images and the minimum and maximum value are the median length errors of the most and least accurately measured image.

Model	Mean Median Error	Min. Error	Max. Error
YOLOv11-large (fine-tuned)	17.3%	**4.8%**	32.9%
YOLOv11-nano (fine-tuned)	15.7%	6.6%	33.3%
SAMv2.1-tiny (fine-tuned)	**14.5%**	5.8%	**25%**

These errors are larger for some images and lower for others, but importantly, both images with predominantly short and long fibres achieve low errors in the single digits of percent.

Compared to each other, all models achieve similar precision among the fibres they segment correctly. Mind that this only reflects how accurately the models segment true positives, not all fibres in an image. Additionally, true positives are gated with an IoU of 0.5, which means that the error of incorrect

segmentation masks is not considered in the data, as they do not have any matching ground truths to compare to. SAMv2.1-tiny achieved the highest precision of length measurements while maintaining superior precision and recall.

Although individual fibre length measurements are not perfect, the models consistently capture fibre lengths with sufficient accuracy to preserve overall distribution patterns. This level of precision supports reliable downstream analyses that depend on fibre length, particularly when considering the natural diversity of fibre sizes within the dataset.

Discussion and Future Work

In this study, we investigated the effectiveness of deep learning models in segmenting amyloid fibres in EM images. Our approach shows promising generalisation capabilities from simulated datasets to real data, and accurate measurements on real EM data. Despite this, some limitations and potential improvements remain. These include domain-specific techniques for model training, increasing the quality of simulated data, and exploring some factors which we did not track in our experiments. In this section, we discuss these limitations in detail, outline potential enhancements for future work, and reflect on the broader implications of our findings for EM image analysis.

One limitation of our current study is the absence of custom data augmentation during training. We only used common augmentation strategies which are pre-packaged with the YOLO and SAM models training scripts rotation. While these are helpful, there are specialized augmentation techniques for increasing the contrast in EM images and similar domains such as those implemented in Gyawali et al. (2024) [17]. Incorporating such methods in future work could further improve the model's performance, particularly in distinguishing the edges of individual fibrils. Similarly, we did not perform any hyperparameter optimization and instead mostly relied on the default settings provided by the training framework. Changing model parameters, including modifications to internal architecture such as those employed by the CrYOLO [50] model could further increase performance.

A further limitation stems from the fact that the models we used were pre-trained on natural image

datasets, such as the COCO dataset used to train YOLOv11 [49], or the SA1B dataset which was used to train the SAM models [31]. While this transfer learning approach accelerates training and often yields better results, it may not provide features that are optimised for the characteristics of EM images. A promising direction for future work would be to pre-train models from scratch or through self-supervised learning on large-scale EM datasets. This strategy has already been successfully applied in tools like CrYOLO [50] and can help models train more efficiently and ensure better generalisation.

While noise did not appear to significantly affect model accuracy in our experiments, our experience suggests that variations in contrast and exposure in images can influence apparent segmentation performance in real electron microscopy (EM) images. We did not explicitly investigate the impact of contrast, which may remain an unexplained variable. However, our simulation framework is capable of adjusting and recording these parameters, making it feasible to systematically explore their effects in future work.

In this work, we used ground truth segmentations to prompt the Segment Anything Model (SAM). While this allowed us to assess its mask quality in isolation, it does not reflect a complete end-to-end segmentation pipeline. Additionally, we only prompted SAM with a single point for each object to ensure a competitive environment with the other models and this does not show the full potential of SAM. Pairing SAM with an object detection model, a zero-shot object proposal network [20], or a lightweight segmenter capable of generating initial masks [17], has been shown to greatly increase its utility. Given that SAM produces high-quality masks even with our simple prompts, such a combined approach could enable fully automated and more accurate fibril segmentation in EM images.

A recent paper presented by Rühle et al. (2021) [41] introduces a workflow for automated segmentation of agglomerated non-spherical particles from SEM images using artificial neural networks trained on simulated data without requiring large manually annotated datasets. Their approach makes use of generative adversarial networks (GANs), specifically cycle-consistent adversarial networks (CycleGANs), to translate segmentation masks into realistic SEM images with deep features which are difficult for NNs to distinguish from real data. This unsupervised image-to-image translation bridges the gap between

synthetic and real images by producing training data that is visually, and more importantly for machine learning applications, statistically closer to real EM micrographs.

Integrating this concept with our existing simulation framework could significantly improve the realism of our synthetic EM images. By combining our simulator which can generate detailed ground truth masks and controlled variations with GAN-based image translation, we could produce synthetic EM images with authentic noise patterns, contaminations, and structural complexities of real EM images. This enhanced realism could enable models trained on synthetic data to generalise much more robustly to real data. Hence, future work could explore the synergy between simulation-based mask generation and domain adaptation through GANs.

Despite these limitations, we achieved consistently good results overall. All models clearly demonstrate generalisation capabilities and can not only segment but accurately measure fibres in a variety of conditions despite extremely limited data. SAM2 stands out with near-perfect segmentation results on clear synthetic images. Additionally, it generalises to real data by far the best, likely due to its inherent boundary discovering abilities. When predicting with SAM, it is prompted with a single point from each ground truth mask. It is possible that SAM's superior performance comes from this infallible object detection and that the YOLO segmentation models are bottlenecked by their object detection mechanism. However, its performance tends to decline in images featuring clustering and overlapping fibres, where the prompt points may become ambiguous or confusing. One promising direction to address this is to integrate SAM2 with an object detection system capable of providing bounding boxes or preliminary masks for individual fibre objects. By constraining the segmentation task to more focused regions, the model could avoid confusion caused by dense overlaps better and focus on refining the masks of specific objects instead. On the other hand, even providing it with a single point for each object already gives away the position of each true object, which may have lead to SAM having an unfair advantage, especially in the unfamiliar domain of real EM data.

The results also clearly indicate that larger models demonstrate a superior ability to represent complex relationships resulting from fibre overlaps even in relatively sterile synthetic data. This can be seen by

the lagging performance of the nano version of the YOLO model compared to its larger counterpart. This is despite the fact that a small model can be advantaged in detecting simple objects, like amyloid fibres, due to its lower tendency to overfit. In our case, the increased capacity of the larger models is clearly needed instead to capture the spatial relationships in fibre clusters, which are challenging for the simpler model to represent effectively.

The last interesting observation is that while SAM generally outperformed YOLOv11-large on clearer images, the latter significantly outperformed SAM in scenarios involving large fibre overlaps. In fact, under conditions of very strong overlap, SAM's performance was comparable to that of the much smaller YOLOv11-nano model. This suggests that despite SAM's high parameter count and strong generalisation, it may struggle with complex scenes of this nature, or that a fully convolutional architecture is more efficient at capturing these spatial relationships.

Conclusion

This thesis investigated the application of deep learning for the automated segmentation and measurement of amyloid fibres in electron micrographs, with a focus on overcoming the challenges posed by clustering and overlap, and image noise. We developed a synthetic data generation framework, systematically evaluated state-of-the-art models, and demonstrated their potential to serve as tools of biological image analysis.

First and foremost, this work demonstrates that it is possible to simulate images of amyloid fibres to a degree sufficient for training effective deep learning models. Additionally, our framework has the ability to programmatically control and replicate key characteristics observed in real electron micrographs, including fibre morphology, intensity at overlaps, and complex clustering patterns governed by procedural noise. Furthermore, the framework incorporates signal-to-noise ratios and contrast levels derived directly from measurements of real data. The ultimate validation of this approach is the successful generalization of models trained on this synthetic data to real-world images, confirming that the simulation was functionally accurate. While the simulation does not capture all real-world complexities, it establishes a

foundation for this domain.

We found that deep learning models can achieve high accuracy in segmenting amyloid fibres in simulated images. The Segment Anything Model v2.1-tiny, in particular, achieved near-perfect F1 scores on images with well-separated fibres. However, performance of all models degraded with increasing fibre overlap and clustering. The YOLOv11-large model proved to be the most robust to increasing fibre overlap, maintaining significantly better performance than both SAM and its lightweight YOLOv11-nano counterpart. Our experiments revealed that all models were surprisingly resilient to a lower signal-to-noise ratio, with all models maintaining high scores even in images with high noise. Clustering and a high fibre count both independently led to a pronounced decline in accuracy across all models, as both higher density and clustering directly increase the frequency of occlusions that make individual fibres difficult to distinguish.

The models' performance declines severely when they are applied to real data, but they are able to generalise to a certain extent. SAM was especially able to generalise compared to the other models when applied to real data. This shows the success of SAM's general segmenting capabilities and generalisation performance delivered by a model designed from the ground up for few- and zero-shot segmentation. Despite the challenges in transferring learning to real data, the lengths of correctly picked fibres can be reliably extracted from the predicted segmentation masks with only a moderate degree of error. The mean median length measurement error across the models was around 15% and was significantly lower for some images. This level of precision may be insufficient for exact measurements of a single fibre, but it can be accurate enough to capture length distributions reliably, supporting downstream quantitative analyses.

In summary, this work validates the use of deep learning for analysing challenging EM images of amyloid fibres. We established that performance is more sensitive to overlap than to image noise and that models can successfully be trained on simulated data with these challenges in mind and transition to real-world applications. In this way, this research supports the development of fully automated analysis pipelines that can accurately pick out amyloid fibres from EM micrographs.

References

[1] Adil Al-Azzawi et al. “AutoCryoPicker: an unsupervised learning approach for fully automated single particle picking in Cryo-EM images”. In: *BMC bioinformatics* 20 (2019), pp. 1–26.

[2] Jayme Garcia Arnal Barbedo. “Impact of dataset size and variety on the effectiveness of deep learning and transfer learning for plant disease classification”. In: *Computers and electronics in agriculture* 153 (2018), pp. 46–53.

[3] Jacob T Bendor, Todd P Logan and Robert H Edwards. “The function of α-synuclein”. In: *Neuron* 79.6 (2013), pp. 1044–1066.

[4] Tristan Bepler et al. “Positive-unlabeled convolutional neural networks for particle picking in cryo-electron micrographs”. In: *Nature methods* 16.11 (2019), pp. 1153–1160.

[5] Joseph George Beton et al. “Cooperative amyloid fibre binding and disassembly by the Hsp70 disaggregase”. In: *The EMBO Journal* 41.16 (2022), e110410.

[6] Eva Bystrenova et al. “Amyloid fragments and their toxicity on neural cells”. In: *Regenerative Biomaterials* 6.2 (2019), pp. 121–127.

[7] Travers Ching et al. “Opportunities and obstacles for deep learning in biology and medicine”. In: *Journal of the royal society interface* 15.141 (2018), p. 20170387.

[8] Eduardo P De Mattos et al. “Protein quality control pathways at the crossroad of synucleinopathies”. In: *Journal of Parkinson’s Disease* 10.2 (2020), pp. 369–382.

[9] facebookresearch. *sam2*. 01/2024. URL: https://github.com/facebookresearch/sam2 (visited on 30/06/2025).

[10] Douglas M Fowler et al. “Functional amyloid–from bacteria to humans”. In: *Trends in biochemical sciences* 32.5 (2007), pp. 217–224.

[11] Céline Galvagnion et al. “Lipid vesicles trigger α-synuclein aggregation by stimulating primary nucleation”. In: *Nature chemical biology* 11.3 (2015), pp. 229–234.

[12] Julia M George et al. “Characterization of a novel protein regulated during the critical period for song learning in the zebra finch”. In: *Neuron* 15.2 (1995), pp. 361–372.

[13] JE Gillam and CE MacPhee. “Modelling amyloid fibril formation kinetics: mechanisms of nucleation and growth”. In: *Journal of Physics: Condensed Matter* 25.37 (2013), p. 373101.

[14] Michel Goedert. “The neurofibrillary pathology of Alzheimer’s disease”. In: *The Neuroscientist* 3.2 (1997), pp. 131–141.

[15] Simon Green. “Implementing improved perlin noise”. In: *GPU Gems* 2.409-416 (2005), p. 5.

[16] Stefan Gustavson. “Simplex noise demystified”. In: *Linköping University, Linköping, Sweden, Research Report* 1.2 (2005), p. 6.

[17] Rajan Gyawali et al. “Accurate cryo-EM protein particle picking by integrating the foundational AI image segmentation model and specialized U-Net”. In: *bioRxiv* (2024), pp. 2023–10.

[18] M.A. Hayat. *Principles and techniques of electron microscopy : biological applications*. eng. 2nd ed. London: Arnold, 1981. ISBN: 0713128305.

[19] Ayelet Heimowitz, Joakim Andén and Amit Singer. “APPLE picker: Automatic particle picking, a low-effort cryo-EM framework”. In: *Journal of structural biology* 204.2 (2018), pp. 215–227.

[20] Yanning Hou et al. “Enhancing Zero-Shot Anomaly Detection: CLIP-SAM Collaboration with Cascaded Prompts”. In: *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*. Springer. 2024, pp. 46–60.

[21] Glenn Jocher, Jing Qiu and Ayush Chaurasia. *Ultralytics YOLO*. Version 8.0.0. 01/2023. URL: https://github.com/ultralytics/ultralytics (visited on 26/06/2025).

[22] Théo Jonchier, Alexandre Derouet-Jourdan and Marc Salvati. “Implementation of Fast and Adaptive Procedural Cellular Noise”. In: *Journal of Computer Graphics Techniques* 8.1 (2019), pp. 35–44.

[23] R Khanam and M Hussain. “Yolov11: An overview of the key architectural enhancements.” In: *arXiv preprint arXiv:2410.17725* (2024).

[24] Alexander Kirillov et al. *Segment Anything*. 2023. arXiv: 2304.02643 [cs.CV]. URL: https://arxiv.org/abs/2304.02643.

[25] Yann LeCun, Yoshua Bengio and Geoffrey Hinton. "Deep learning". In: *nature* 521.7553 (2015), pp. 436–444.

[26] Chang-Chun Lee, Yen Sun and Huey W Huang. "How type II diabetes-related islet amyloid polypeptide damages lipid bilayers". In: *Biophysical Journal* 102.5 (2012), pp. 1059–1068.

[27] Kimberly Jia Yi Low et al. "Molecular mechanisms of amyloid disaggregation". In: *Journal of Advanced Research* 36 (2022), pp. 113–132.

[28] JEOL Ltd. *Transmission Electron Microscope (TEM) | Science Basics*. URL: https://www.jeol.com/products/science/em.php.

[29] Ivo Cristiano Martins et al. "Lipids revert inert Aβ amyloid fibrils to neurotoxic protofibrils that affect learning in mice". In: *The EMBO journal* 27.1 (2008), pp. 224–233.

[30] Damian J Matuszewski and Ida-Maria Sintorn. "Reducing the U-Net size for practical scenarios: Virus recognition in electron microscopy images". In: *Computer methods and programs in biomedicine* 178 (2019), pp. 31–39.

[31] Meta. *SA-1B Dataset*. 04/2023. URL: https://ai.meta.com/datasets/segment-anything/.

[32] Maria Nagy et al. "Extended survival of misfolded G85R SOD1-linked ALS mice by transgenic expression of chaperone Hsp110". In: *Proceedings of the National Academy of Sciences* 113.19 (2016), pp. 5424–5428.

[33] David L. Nelson, Michael M. Cox and Aaron A. Hoskins. *Lehninger principles of biochemistry*. eng. Eighth edition. New York, NY: Macmillan Learning, 2021. Chap. 4. ISBN: 9781319228002.

[34] Lucas Prado Osco et al. "The segment anything model (sam) for remote sensing applications: From zero to one shot". In: *International Journal of Applied Earth Observation and Geoinformation* 124 (2023), p. 103540.

[35] A Osterhaus et al. “a-Synuclein in Lewy bodies”. In: *Nature* 388.1 (1997), pp. 839–840.

[36] Daniel Otzen and Roland Riek. “Functional amyloids”. In: *Cold Spring Harbor perspectives in biology* 11.12 (2019), a033860.

[37] Ken Perlin. “Chapter 2: Noise Hardware”. In: *SIGGRAPH 2002 Course 36 Notes Real-Time Shading Languages*. Department of Computer Science and Electrical Engineering - University of Maryland, Baltimore County, 2001. URL: https://userpages.cs.umbc.edu/olano/s2002c36/ch02.pdf.

[38] Laura Pieri et al. “Fibrillar α-synuclein and huntingtin exon 1 assemblies are toxic to the cells”. In: *Biophysical journal* 102.12 (2012), pp. 2894–2905.

[39] Saqib Qamar et al. “Segmentation and characterization of macerated fibers and vessels using deep learning”. In: *Plant Methods* 20.1 (2024), p. 126.

[40] Joseph Redmon et al. “You only look once: Unified, real-time object detection”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 779–788.

[41] Bastian Rühle, Julian Frederic Krumrey and Vasile-Dan Hodoroaba. “Workflow towards automated segmentation of agglomerated, non-spherical particles from electron microscopy images using artificial neural networks”. In: *Scientific reports* 11.1 (2021), p. 4942.

[42] Jean D Sipe et al. “Amyloid fibril protein nomenclature: 2012 recommendations from the Nomenclature Committee of the International Society of Amyloidosis”. In: *Amyloid* 19.4 (2012), pp. 167–170.

[43] Kevin W Tipping et al. “Amyloid fibres: inert end-stage aggregates or key players in disease?” In: *Trends in biochemical sciences* 40.12 (2015), pp. 719–727.

[44] A. Torralba, P. Isola and W.T. Freeman. *Foundations of Computer Vision*. Adaptive Computation and Machine Learning series. MIT Press, 2024. ISBN: 9780262378666.

[45] Kevin P Treder et al. “Applications of deep learning in electron microscopy”. In: *Microscopy* 71.Supplement_1 (2022), pp. i100–i115.

[46] Ultralytics. *COCO*. 03/2025. URL: https://docs.ultralytics.com/datasets/segment/coco/#what-pretrained-models-are-available-for-coco-seg-and-what-are-their-performance-metrics.

[47] Ultralytics. *Instance Segmentation Datasets Overview*. 03/2025. URL: https://docs.ultralytics.com/datasets/segment/#ultralytics-yolo-format.

[48] Ultralytics. *Train*. 06/2025. URL: https://docs.ultralytics.com/modes/train/#augmentation-settings-and-hyperparameters.

[49] Ultralytics. *YOLO11*. 02/2025. URL: https://docs.ultralytics.com/models/yolo11/#usage-examples.

[50] Thorsten Wagner et al. "SPHIRE-crYOLO is a fast and accurate fully automated particle picker for cryo-EM". In: *Communications biology* 2.1 (2019), p. 218.

[51] Thorsten Wagner et al. "Two particle-picking procedures for filamentous proteins: SPHIRE-crYOLO filament mode and SPHIRE-STRIPER". In: *Biological Crystallography* 76.7 (2020), pp. 613–620.

[52] Conrad C Weihl et al. "Loss of function variants in DNAJB4 cause a myopathy with early respiratory failure". In: *Acta neuropathologica* 145.1 (2023), pp. 127–143.

[53] Anne Wentink and Rina Rosenzweig. "Protein disaggregation machineries in the human cytosol". In: *Current opinion in structural biology* 83 (2023), p. 102735.

[54] Steven Worley. "A cellular texture basis function". In: *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*. 1996, pp. 291–294.

[55] Wei-Feng Xue et al. "Fibril Fragmentation Enhances Amyloid Cytotoxicity". In: *Journal of Biological Chemistry* 284.49 (2009), pp. 34272–34282.

[56] Jingrong Zhang et al. "PIXER: an automated particle-selection method based on segmentation using a deep neural network". In: *BMC bioinformatics* 20 (2019), pp. 1–14.

[57] Tongjie Y Zhang and Ching Y. Suen. "A fast parallel algorithm for thinning digital patterns". In: *Communications of the ACM* 27.3 (1984), pp. 236–239.