



Universiteit
Leiden

Morphometric Read-outs from Neurite Segmentation in a Cardiac Innervation Chip Using U-Net

Pepijn Baggen (s3193993)

Supervisors:

Dr. Lu Cao & Marthe Kaal MSc

Third supervisor:

Dr. Verena Schwach

BACHELOR THESIS— INFORMATICA

Leiden Institute of Advanced Computer Science (LIACS)

liacs.leidenuniv.nl

July 15, 2026

Abstract

Cardiac toxicity is among the leading causes of drug withdrawal [2, 18], yet the standard cardiotoxicity screens (flat cardiac cell cultures, or animal hearts) routinely miss the damage. A more faithful alternative is a microfluidic chip with separate compartments. Human stem-cell-derived sympathetic neurons are grown in one compartment, and their neurites grow through narrow channels into a separate compartment, where a small piece of human heart tissue (a cardiac microtissue) can be added on top of them. When the neurites are damaged, this network breaks apart in a dose-dependent way, which makes its shape a direct read-out of that damage; extracting that read-out, however, costs hours of manual neurite tracing per image.

This thesis presents an automated deep-learning workflow that turns unprocessed confocal image stacks into seven structural descriptors of the neurite network, with no manual step. A controlled ablation across seven U-Net variants, covering encoder width, attention gating, a cDice topology-aware loss, and an elastic-augmentation control, finds that with only 15 training images architecture is not the limiting factor: the seven variants differ by at most 1.2 Dice Similarity Coefficient (DSC) percentage points. A lightweight Attention U-Net (M5; 8M parameters) is chosen as the smallest variant that matches the best. Deployed as a 5-fold ensemble with Hann-windowed tiled inference, M5 attains in-distribution agreement with the training annotator at full-image DSC = 0.887 and centre-line cDice = 0.96 on the 15 annotated images (from 14 unique cultures). As a held-out generalisation check, on two independently re-traced acquisitions from a separate experimental batch the model achieves cDice = 0.84 (centre-line agreement at 2-pixel tolerance) and total fibre length within 6% of the manual measurement.

For context against established baselines, M5 ranks above Frangi vesselness, Otsu thresholding, and a default 2D nnU-Net v2 on each of the 15 in-sample images. The accompanying deliverable is a Streamlit desktop application that accepts raw Olympus `.oir` (microscopy) stacks and returns all seven morphometric read-outs through an interactive visualisation, turning a previously hours-long manual task into a single-click operation.

Acknowledgments

This thesis was conducted at the Leiden Institute of Advanced Computer Science (LIACS), Leiden University, under the supervision of Dr. Lu Cao, with Dr. V. Schwach (Applied Stem Cell Technologies (AST) group, University of Twente) as second supervisor. I thank the AST group for sharing the confocal images and for answering my biological questions throughout the project. Training was carried out on the ALICE high-performance computing facility at Leiden University [11].

Code and data availability. The source code (preprocessing, training, inference, post-processing, and the analysis application of Section 4.10) is published at <https://github.com/pepijn-baggen/neurite-segmentation>. The annotated imaging data are governed by the AST group.

Contents

1	Introduction	1
1.1	Research question	2
1.2	Scope and contributions	3
1.3	Thesis overview	4
2	Background	4
2.1	Biological relevance	4
2.2	Segmentation	5
2.3	Morphometrics	6
3	Related Work	7
3.1	Tracing neurites by hand: Fiji and SNT	7
3.2	Classical filters: fast but topologically fragile	7
3.3	Deep learning methods: U-Net with its variations	8
3.4	Topology-aware losses: do they help?	8
3.5	Deep learning for cardiotoxicity	8
4	Approach	9
4.1	Experimental setup and imaging	10
4.2	Image preprocessing	11
4.3	Ground-truth generation	11
4.4	Training-data preparation	13
4.5	Model architecture	13
4.6	Training procedure	14
4.7	Ablation study design	15
4.8	Inference procedure	15
4.9	Hyperparameter summary	18
4.10	The analysis application	20
5	Experiments	22
5.1	Dataset and quality control	22
5.2	Ablation study results	22
5.3	Analysis of ablation results	23
5.4	Model selection: M5 lightweight Attention U-Net	24
5.5	Full-image ensemble DSC	24
5.6	Comparison against a modern deep-learning baseline (nnU-Net v2)	26
5.7	Ensemble gain and post-processing sensitivity	28
5.8	Morphological metrics	30
5.9	External validation against independent SNT tracings	32
5.10	End-to-end deployment on an unseen multi-condition batch	33

6	Discussion and Conclusions	36
6.1	Summary of findings	36
6.2	Limitations	38
6.3	Further research	40
	References	43

1 Introduction

Cardiac toxicity, the damage that certain drugs cause to heart cells, is one of the dominant safety-related causes of drug attrition and post-market withdrawal [1, 18]. Yet the standard preclinical screens (flat layers of cardiac cells in a dish, or animal hearts whose pharmacology differs from the human) routinely miss the damage that later emerges in patients. Human stem-cell models of the heart offer a more faithful alternative [12, 15]. The system studied here was developed by the Applied Stem Cell Technologies (AST) group at the University of Twente. It is a microfluidic chip with separate compartments. Sympathetic neurons are grown in one compartment, and their neurites grow through narrow channels into a separate compartment, where a small piece of heart tissue (a cardiac microtissue) can later be added on top of them. The channels are thin enough that only the neurites can pass, so the neuron cell bodies and the heart tissue are never in the same compartment. This thesis images the neurite compartment, which in the dataset used here holds only the network of neurites.

The way these neurites are laid out is called their morphology, the read-out this thesis targets. Under cardiotoxic stress, the network fragments, retracts, and loses branching density. Neurite morphology is therefore, in principle, a quantitative read-out of cardiotoxicity, measurable directly from a microscopy image.

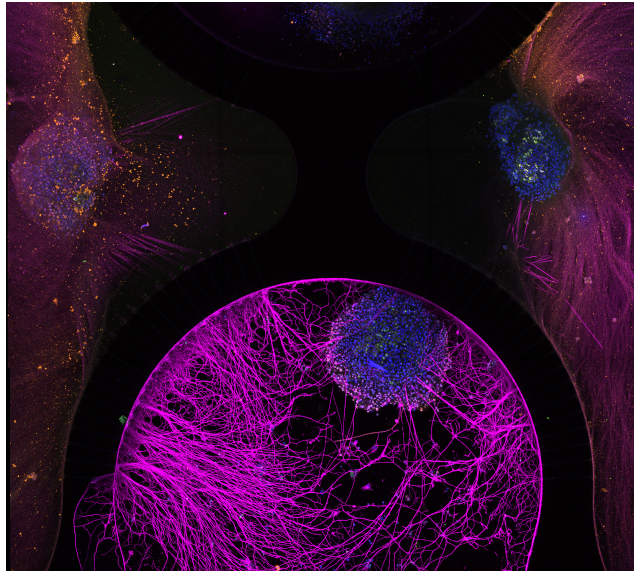


Figure 1. An image of a microfluidic cardiac innervation chip. Sympathetic neurons in the compartments (blue cell-body clusters) project neurites (magenta) through narrow microchannels into the neighbouring wells (microchannels are not visible here); the lower well holds the dense neurite network that this thesis targets to segment.

However, turning that read-out into a routine measurement is held back by the currently required human annotation. A single confocal acquisition of the neurite compartment contains tens of thousands of neurite pixels. An established tool for tracing them, the Simple Neurite Tracer (SNT) plugin for Fiji [13], still requires manual guidance: the researcher clicks along each fibre by hand, or places a region of interest around a bright neurite bundle from which

SNT semi-automatically follows the bright intensity ridge. The limitation is not the software’s quality. Even a rigorous, established tool like SNT remains prohibitively slow, on the order of several hours per image, and noticeably annotator-dependent, so it cannot support a routine, large-scale read-out. Classical filter-based segmenters such as the Frangi vesselness filter [3, 14] run in seconds, but they look at the image through a narrow local window: at every branching crossroad they hesitate, fragment, and lose the very topology that makes the read-out biologically meaningful. The result is a trade-off the field has not been able to improve: SNT tracing is accurate but does not scale, and the fast classical filters break exactly at the structures the read-out depends on.

A deep convolutional segmenter sidesteps this trade-off because it integrates evidence across receptive fields much larger than a single Hessian neighbourhood. The U-Net of Ronneberger et al. [22] has been the standard for this kind of small-data biomedical task for a decade, and later refinements have made it even better suited to sparse, thread-like structures such as vessels and neurites. The contribution of this thesis is to put a carefully selected, lightweight U-Net variant to work on neurite imagery from the chip and to surround it with the preprocessing, inference and morphometric analysis needed to turn raw microscopy stacks into structural numbers. The resulting system is then demonstrated on real data, including an eight-chip multi-condition NGF batch the network had never seen during training.

1.1 Research question

This thesis addresses the following research question:

To what extent can a U-Net automatically segment the sympathetic neurites of the cardiac innervation chip from fluorescence microscopy images, and what morphometric read-outs can be reliably extracted from the resulting segmentations?

It is decomposed into three sub-questions:

1. How does a trained U-Net compare to the two standard classical baselines (Frangi vesselness filtering and Otsu thresholding) and to a modern reference method (nnU-Net v2) on this dataset, measured by the Dice Similarity Coefficient (DSC)?
2. Among a small set of U-Net variants (encoder width, attention gates, topology-aware loss, elastic augmentation), how do these design choices affect segmentation performance, and which variant is most suitable for deployment on this task?
3. What morphometric read-outs can be extracted from the predicted masks, and how do those read-outs behave when the method is applied to an independent multi-condition chip batch?

Sub-question 3 is a behavioural test, not a cardiotoxicity verdict. An eight-chip multi-condition batch (control, low NGF, high NGF) was available for the deployment test, but with only $n = 2-3$ chips per condition the read-out is descriptive rather than statistically conclusive. Establishing neurite morphology as a cardiotoxic biomarker would need a properly powered, drug-exposed cohort with matched controls, which I flag as the main direction for future work.

1.2 Scope and contributions

The original research proposal specified a residual U-Net (Res-UNet) backbone [7]. During implementation the architectural focus shifted, since residual connections were designed primarily to ease optimisation in much deeper networks; for the present task, where the dominant difficulty is the sparseness of the foreground (5–7% of the image), attention gates [17] are the more directly relevant refinement. The ablation in Section 5 therefore takes the standard (non-residual) U-Net as the baseline and evaluates attention-gated variants against it.

This thesis makes the following contributions.

- **A full neurite-segmentation system** that takes raw Olympus `.oir` confocal stacks of the chips as input and returns binary neurite masks at the original 2048×2048 px resolution. It covers preprocessing, ground-truth generation, training, ensembled inference, and post-processing; a companion analysis tool wraps the inference pipeline (preprocessing, ensembled prediction, and morphometric extraction, everything except model training) in a single-click Streamlit interface for use by non-programmers.
- **A controlled ablation across a set of U-Net variants:** seven configurations covering every combination of encoder width and attention gating, plus the cLDice loss and elastic augmentation, all land within 1.2 DSC percentage points of one another. The lightweight Attention U-Net (M5) is selected for deployment because it matches the heaviest variant at one-quarter of the parameter count.
- **A reference-point benchmark against nnU-Net v2**, the standard framework for biomedical segmentation, on all five matched cross-validation folds. M5 ranks above nnU-Net v2 on every image of the 15-image set (Section 5.6). At this data scale the compact M5 model therefore outperforms the much larger self-configuring nnU-Net v2 in its 2D configuration.
- **A morphometric extraction layer** that converts the predicted masks into seven structural descriptors of the neurite network, including total fibre length, junction count, branchiness, area covered, and mean branch thickness.
- **A full deployment on an unseen eight-chip Nerve Growth Factor (NGF) dose-response batch** spanning three culture conditions, showing that the method runs without manual intervention on data it has never been trained on.

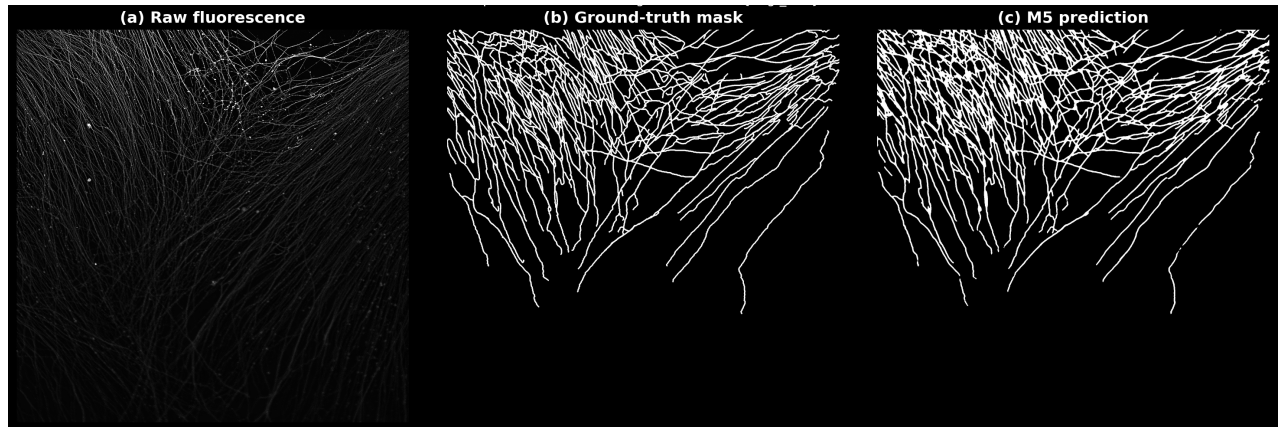


Figure 2. Preview of the final method on a representative image of the neurite compartment (image 007, 2048×2048 px). **(a)** Raw preprocessed fluorescence input. **(b)** Manual ground-truth neurite mask. **(c)** Automatic prediction by the method introduced in this thesis; agreement with the manual mask is $DSC = 0.903$. Full evaluation appears in Section 5.

1.3 Thesis overview

Section 2 reviews the biological and computational background. Section 3 surveys the related literature on biomedical segmentation and neurite analysis. Section 4 describes the method in full. Section 5 reports the ablation study, model selection, ensemble evaluation and morphometric extraction. Section 6 discusses findings, limitations, and directions for follow-up work.

2 Background

This section reviews the three ingredients the method draws on, in the order the rest of the thesis uses them: the biology that makes neurite morphology a meaningful read-out, the segmentation analysis that recovers the network from an image, and the morphometric analysis that turns a predicted mask into numbers.

2.1 Biological relevance

Human induced pluripotent stem cells (hiPSCs) are ordinary adult cells that have been reprogrammed back into a stem-cell state. From that state they can be turned into many human cell types, including sympathetic neurons and heart muscle cells.

The model used in this thesis is a microfluidic chip with separate compartments. The cell bodies of the sympathetic neurons sit in one compartment. Their neurites, the thin fibres that grow out of the neurons, grow through narrow channels into a separate compartment. These channels are so thin that only the neurites can pass through them, so the larger cell bodies stay behind in their own compartment. In that neurite compartment, a small piece of human heart tissue called a cardiac microtissue can be placed on top of the neurites [12, 15]. A cardiac microtissue is a small three-dimensional cluster of heart cells, which copies real heart tissue

more closely than a flat layer of cells in a dish. Because the channels block the cell bodies, the neuron cell bodies and the heart tissue are never in the same compartment. A further advantage of this model is that the cells are human, so the drug responses avoid the species differences that limit animal-based screens.

This thesis measures only the neurite compartment, the one the neurites grow into. In the images used here, no cardiac microtissue was placed in that compartment, so every image shows only the network of neurites, with no heart tissue and no neuron cell bodies in view. Sympathetic neurites are easily harmed by toxic drugs: when this happens they break apart, pull back, and lose their branching, a response that has been reported as dose-dependent in primary sympathetic-neuron cultures exposed to known cardiotoxic compounds. Total fibre length and junction density are therefore plausible quantitative read-outs for cardiotoxic damage, provided they can be measured reliably and at scale.

2.2 Segmentation

Recovering the neurite network from an image is a binary segmentation problem. This subsection sets out the computational pieces the method relies on: the segmentation task and its accuracy metric, the U-Net and the attention gates and loss functions that adapt it to sparse neurites, and the cross-validation protocol used to evaluate it on a small dataset.

Image segmentation and the Dice score. Image segmentation assigns a class label to every pixel. The binary variant addressed here labels each pixel as either neurite (foreground) or background. Performance is summarised by the Dice Similarity Coefficient,

$$\text{DSC} = \frac{2|P \cap G|}{|P| + |G|}, \quad (1)$$

where P is the predicted binary mask and G the ground-truth mask. DSC ranges from 0 (no overlap) to 1 (perfect overlap) and is robust to class imbalance, which makes it the standard metric for medical image segmentation.

U-Net. The U-Net [22] is a fully convolutional encoder–decoder network with skip connections. The encoder progressively downsamples the input through convolutional blocks and max-pooling, extracting features at several spatial scales. The decoder upsamples these features back to the input resolution using transposed convolutions. Skip connections concatenate encoder features directly to decoder features at matching resolutions, preserving spatial detail that would otherwise be lost during downsampling. A final 1×1 convolution followed by a sigmoid activation produces a pixel-wise probability map.

Attention gates. In a standard U-Net, every encoder feature map is passed to the decoder through the skip connections, background included. When the foreground is as sparse and thin as the neurite signal here, occupying only a few percent of the image, the decoder receives mostly irrelevant activation and has to learn to discard it. Attention gates are motivated by exactly this problem. Introduced by Oktay et al. [17] as part of the Attention U-Net, they

augment each skip connection with a soft spatial attention mechanism: a gating signal from the deeper decoder layer is combined with the corresponding encoder feature map to produce an attention coefficient $\alpha \in [0, 1]$ at every spatial position, and the encoder features are multiplied by these coefficients before concatenation. This suppresses background activations and directs decoder capacity towards the thin foreground, which is why such gating might help on a sparse segmentation task like ours. It also adds only a small number of parameters, making it inexpensive to adopt.

Loss functions. Standard binary segmentation training combines a weighted Binary Cross-Entropy (BCE) term and a Dice term, with a positive class weight w on the BCE term set proportional to the background-to-foreground ratio (≈ 15 in this dataset) to prevent the model from collapsing to an all-background prediction. The clDice loss [25] additionally penalises topological errors by computing a skeleton-based overlap between prediction and ground truth,

$$\mathcal{L}_{\text{clDice}} = 1 - 2 \cdot \frac{T_{\text{prec}} \cdot T_{\text{sens}}}{T_{\text{prec}} + T_{\text{sens}}}, \quad (2)$$

where T_{prec} measures how much of the predicted skeleton overlaps the ground-truth mask, and T_{sens} how much of the ground-truth skeleton is covered by the predicted mask. (Note that this clDice *loss* is distinct from the clDice *metric* used later in Section 5.9 to score centre-line agreement between a predicted mask and an independent reference tracing.)

Cross-validation. k -fold cross-validation partitions the dataset into k subsets, holds each out for validation once while training on the remaining $k - 1$ subsets, and combines the k resulting performance estimates as a mean and standard deviation. On small datasets, such as the $n = 15$ annotated images here, this is preferred over a single train/test split, whose result is dominated by the composition of the held-out set.

2.3 Morphometrics

Once the network is segmented, its biologically meaningful properties (total fibre length, number of branch points, how branchy the network is) are not in the filled mask itself but in its *skeleton*: the one-pixel-wide centre-line running along each fibre. Skeletonisation reduces a binary mask to that centre-line while preserving the topology of the original structure, and is implemented here through the medial-axis routine in scikit-image [27].

The skeleton is then read as a graph. A small 3×3 neighbour-count kernel is slid over the skeleton mask to classify each skeleton pixel: a pixel with exactly one foreground neighbour is an *endpoint*, a pixel with three or more is a *junction*, and adjacent junction pixels are merged into a single junction by 8-connected component labelling. The *branch segments* are then the connected components of the skeleton remaining after the junction pixels have been removed. These three primitives (endpoints, junctions, branch segments) are the foundation of the morphometric read-outs computed in Section 4; Figure 3 illustrates them on a synthetic Y-shaped skeleton.

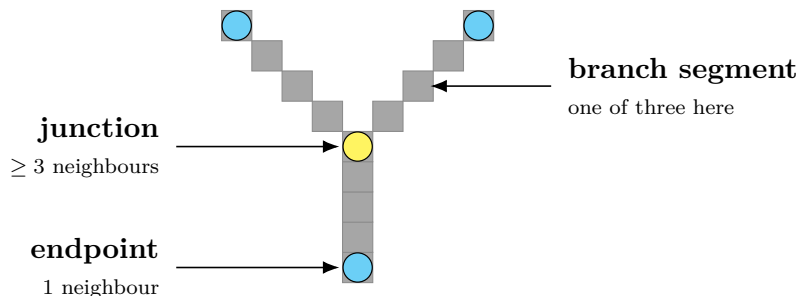


Figure 3. Skeleton-graph primitives on a synthetic Y-shaped skeleton. Each square is a skeleton pixel; the three highlighted classes follow from the 3×3 neighbour count. An endpoint has exactly one foreground neighbour (cyan), a junction has three or more (yellow), and the connected components remaining after the junction pixels are removed are the branch segments. The branchiness ratio reported in Section 5 is the junction count divided by the total skeleton length in micrometres.

3 Related Work

3.1 Tracing neurites by hand: Fiji and SNT

The prevailing way to quantify neurites in fluorescence microscopy is manual tracing in Fiji [23] with the Simple Neurite Tracer (SNT) plugin [13], the workflow this thesis adopts as its ground-truth source. It is accurate and reproducible within a single annotator, but slow, from tens of minutes to several hours per image, since the operator must guide and verify the trace of every branch by hand. A semi-automatic mode seeded from a region of interest reduces the clicking but not the supervision, and the manual correction it still requires introduces systematic variability between annotators. For any laboratory aiming to scale beyond a handful of images, SNT is the rate-limiting step of the annotation workflow.

3.2 Classical filters: fast but topologically fragile

To bypass the human bottleneck, classical neurite segmentation has long relied on vesselness filtering. The Frangi filter [3] computes Hessian eigenvalue ratios at multiple scales to highlight tubular structures, and has been adapted to fluorescence microscopy in tools such as NeuriteTracer [14]. The filter is fast, parameter-free in the data-driven sense, and works well on isolated, straight segments. It struggles, however, exactly where the biology becomes interesting: at the crossings, bifurcations and dense thickets where multiple neurites meet, the local Hessian signature transitions from tube-like to blob-like, and the resulting masks fragment at precisely the branch points that the downstream morphometric layer is asked to count. A second classical baseline is Otsu thresholding [19], which selects a global intensity cut-off and is the simplest reasonable filter to compare against. Both methods are used here as classical baselines (Section 5).

3.3 Deep learning methods: U-Net with its variations

The U-Net of Ronneberger et al. [22] demonstrated that an encoder-decoder architecture with skip connections could be trained successfully on fewer than forty annotated biomedical images, and it has been the standard of small-data biomedical segmentation. Subsequent work refines this template rather than replacing it. Residual connections [7] help optimisation in deeper variants. Attention gates [17] let the decoder weight task-relevant spatial regions on the skip connections without many additional parameters. Elastic-deformation augmentation, proposed alongside the original U-Net, continues to outperform more elaborate regularisers on small biological datasets because its deformations are physically plausible for the underlying tissue. The nnU-Net framework of Isensee et al. [8] later showed that a carefully tuned standard U-Net is competitive against a much bigger model, reinforcing the point that when annotated data are scarce, architecture is rarely the limiting factor.

3.4 Topology-aware losses: do they help?

Standard pixel-wise objectives such as BCE and Dice care only about how many pixels overlap; they do not directly reward keeping a fibre connected. In principle this is a concern for neurite analysis, because a single two-pixel gap can split a path in two and change the junction count without changing the Dice score by very much. Three different studies have proposed loss-level fixes. Mosinska et al. [16] compare predictions in a pretrained VGG feature space, on the grounds that broken paths produce mismatched deep-network activations. Hu et al. [5], refined by Stucki et al. [26], count topological errors directly using a persistent-homology loss based on Betti-number matching. Shit et al. [25] take the third route, the cDice loss, which defines a differentiable skeleton overlap and rewards the network when its predicted centre-line tracks the ground-truth centre-line.

Of the three, this thesis tests cDice: its differentiable skeleton aligns with the centre-line the morphometric layer later computes, and its gradient is dense at the pixel level rather than concentrated in a handful of topological events, which suits a small training set. Whether that translates into a measurable gain over BCE+Dice is left as an empirical question (Section 5).

3.5 Deep learning for cardiotoxicity

Grafton et al. [4] have shown that deep learning can detect cardiotoxicity from high-content imaging of hiPSC-derived cardiomyocytes: quantitative morphological features, including sarcomere organisation and contractility metrics, correlate significantly with the presence of known cardiotoxic compounds. Their work shows that automated image analysis can serve as a scalable, quantitative cardiotoxicity read-out, but the substrate and target differ from those addressed here: they train classifiers on flat (two-dimensional) cardiomyocyte cultures and read out sarcomere structure and beating behaviour (how strongly and how often the cells contract) at the well or sub-image level. The present work targets the *neurite*-network morphology in a cardiac innervation chip, which has a different imaging protocol, a different sample geometry, and a different annotation workflow.

4 Approach

The workflow splits into two phases that share preprocessing but otherwise serve different purposes. The *training phase* (Figure 4) takes the 15 annotated images, generates the ground-truth masks through manual SNT tracing followed by disk dilation, tiles them into patches, and trains seven U-Net variants under 5-fold cross-validation. It is run once on the ALICE cluster and saves one trained model per cross-validation fold (five checkpoints), which the inference phase later combines into an ensemble. The *inference phase* (Figure 5) then takes any previously unseen raw `.oir` stack, applies identical preprocessing, runs the trained model as a 5-fold ensemble with Hann-windowed tiled inference, and post-processes the predicted probability map into a clean binary mask, a skeleton, and the seven morphometric read-outs. Each stage of both phases is described in turn below.

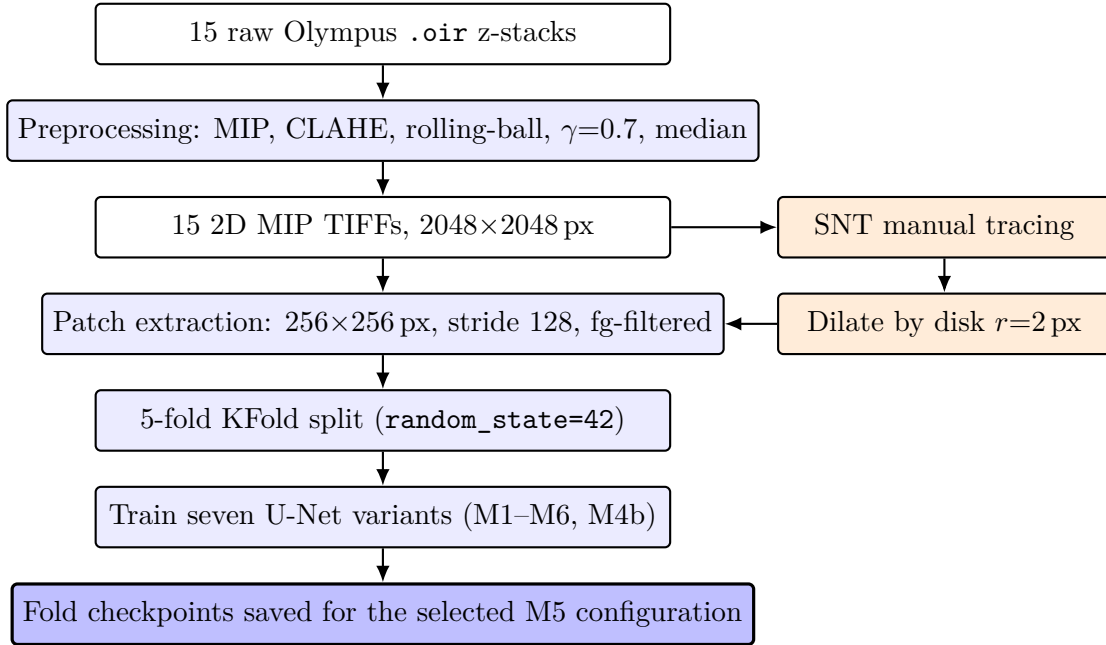


Figure 4. Training phase. The 15 raw `.oir` stacks are preprocessed and projected to 2D MIPs (left). In parallel, each MIP is traced in SNT and the resulting one-pixel skeleton is dilated by a disk of radius 2 px to produce the supervised target (right). The paired image/mask tiles drive a 5-fold cross-validated training of seven U-Net variants, from which the lightweight Attention U-Net (M5) is selected on the empirical ablation in Section 5. Only the fold checkpoints of M5 are carried forward to the inference phase (Figure 5).

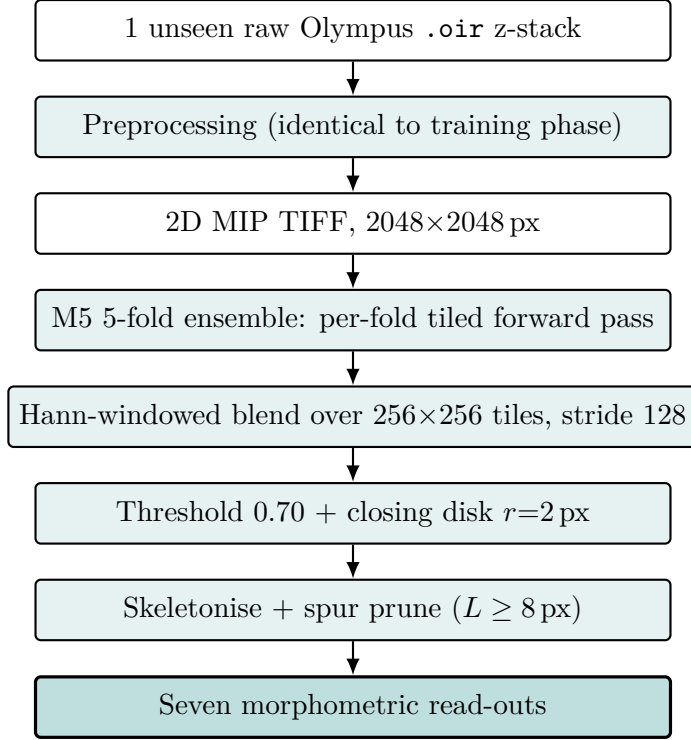


Figure 5. Inference phase. A previously unseen raw `.oir` stack is preprocessed identically to the training data and passed through the five M5 fold checkpoints in parallel. A 2D Hann window combines their per-tile predictions over overlapping 256×256 px tiles into one full-image probability map with no visible tile seams; this map is then thresholded, closed, and reduced to a spur-pruned skeleton. The same seven morphometric read-outs of Table 2 follow, with no human intervention. This is the workflow wrapped by the analysis application of Section 4.10 and exercised on the unseen NGF batch in Section 5.10.

4.1 Experimental setup and imaging

The neurite cultures analysed in this thesis grow on microfluidic chips, each chip containing several distinct wells (*compartments*) into which the cells and reagents are placed. Each compartment is imaged separately, so that a single chip can yield multiple spatially-registered acquisitions covering different parts of the experiment (e.g. upper, middle, lower compartments). The image data itself is acquired on an Olympus FV4000 confocal microscope and stored as proprietary `.oir` z-stacks, with a calibrated pixel size of $0.6215 \mu\text{m}/\text{px}$ ($10\times$ objective, optical zoom 1) confirmed across two independent metadata files. Each acquisition covers a 2048×2048 px field of view in xy and a per-acquisition z-stack of 14–20 slices (median 17) stepping through the depth of the sample; a maximum-intensity projection (MIP) collapses the z-stack to a single two-dimensional image that serves as the input to the segmentation pipeline. The relevant fluorescent channel isolates the neurite signal from the background of the neurite compartment.

4.2 Image preprocessing

The preprocessing steps and their parameters were chosen empirically, by visual inspection of the output.

Raw Olympus .oir z-stacks are converted to 8-bit grayscale TIFFs in two stages: a Fiji-macro stage that imports and projects each stack, and a per-image stage that enhances neurite contrast before segmentation. The aim throughout is to make faint distal neurites visible to the network without saturating the bright neurite bundles, which would otherwise dominate any global intensity adjustment.

Fiji-macro stage. The macro imports the stack through Bio-Formats with channel isolation, retaining only the neurite channel, and collapses the z-axis by maximum-intensity projection. A contrast stretch (0.35% saturation) and a first Contrast-Limited Adaptive Histogram Equalisation (CLAHE) pass (block size 127, 256 bins, slope 3.0) [28] are applied before saving. CLAHE equalises contrast locally rather than globally, so faint distal fibres are lifted within their own neighbourhood instead of being washed out by the much brighter neurite bundles elsewhere in the field.

The per-image stage then applies a fixed sequence, grouped here by purpose. *Illumination correction:* rolling-ball background subtraction (radius 50 px) removes low-frequency illumination gradients, and gamma correction ($\gamma = 0.7$) redistributes intensity toward bright structures. *Noise suppression:* despeckling and outlier removal suppress isolated bright pixels. *Cross-acquisition normalisation:* a global contrast stretch, noise-floor suppression (subtract 15), and a second, per-image CLAHE pass bring intensity onto a common scale across acquisitions. A median filter (radius 1 px) finally closes the sequence. Each step can be inspected individually, making the chain reproducible and straightforward to adjust.

4.3 Ground-truth generation

Three approaches to ground-truth (GT) generation were evaluated before settling on the workflow used throughout this thesis (Figure 6).

Frangi filter pipeline. A semi-automated PyQt5 application applied Gaussian smoothing ($\sigma = 2.0$), CLAHE, Frangi vesselness filtering (scales 2–6, $\beta = 0.3$), hysteresis thresholding, morphological closing and shape/debris filtering. The resulting masks covered roughly 7% of the image area but fragmented heavily at neurite crossings, because the Hessian signature on which Frangi relies collapses precisely at branch points. Manually filling the resulting gaps (orange in Figure 6) proved impractical, and this approach was abandoned.

Manual brush painting in Krita. A small pilot was carried out by tracing neurites by hand with a brush tool in Krita [10], an open-source digital painting application. The resulting masks were biologically plausible, but annotating even a small section of an image took several hours; extending this to the full dataset would have demanded a prohibitive number of hours, so the approach was abandoned.

SNT manual tracing (final approach). The chosen workflow traces every image in the Simple Neurite Tracer plugin for Fiji [13]. The annotator first draws a region of interest (ROI) in Fiji around each bright bundle of neurites, where the fibres fan out; SNT then traces outward from the ROI boundary along each visible fibre. Tracing can proceed in either of two modes: an automatic mode, in which SNT propagates a path along the bright intensity ridge of the neurite using a Hessian-based ridge filter combined with an A* shortest-path search, or a manual mode, in which the annotator clicks waypoints along the fibre and SNT snaps each segment onto the nearest ridge between consecutive clicks. The two modes are mixed as needed within a single image: automatic propagation handles long, well-resolved fibres, while manual clicks resolve ambiguous distal branches and crossings where the ridge filter loses confidence. The output is a connected, topologically correct polyline for every fibre. The raw 1-pixel-wide tracings are then dilated with a disk-shaped structuring element of radius 2 px to produce masks with realistic neurite width, yielding 5–7% foreground coverage. This dilation matters: a one-pixel-wide target leaves no error tolerance, so a prediction offset by even one or two pixels is penalised as harshly as a structurally wrong prediction; dilating to a target width matching the physical extent of the neurite ($\sim 1\text{--}3\text{ }\mu\text{m}$, or $\sim 2\text{--}5\text{ px}$ at the present calibration) restores a biologically defensible matching tolerance. The downstream skeletonisation step undoes most of this dilation when the morphometric read-outs are extracted, so the wider mask serves only as a learning target.

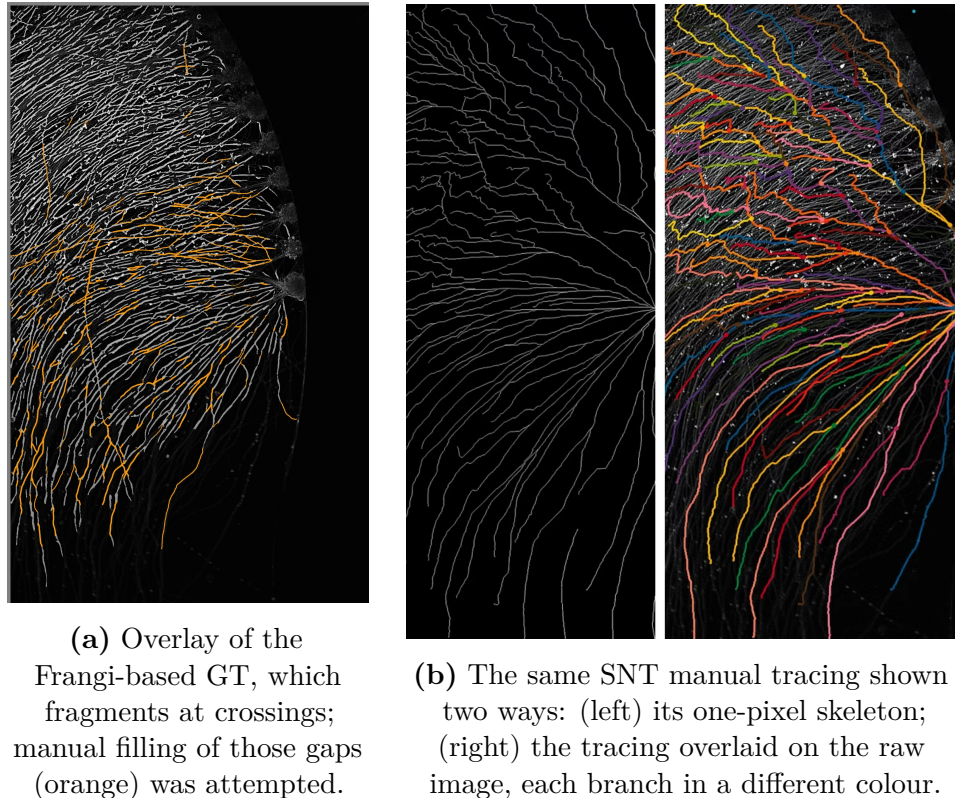


Figure 6. Side-by-side comparison of the two ground-truth strategies on the same image. Vesselness filtering preserves straight segments but breaks at the bifurcations and crossings; manual tracing in SNT yields a better-connected skeleton.

4.4 Training-data preparation

The 15 annotated image pairs (2048×2048 px, uint8 grayscale) are partitioned into 5-fold cross-validation splits and tiled into 256×256 px patches at 50% overlap. Because the foreground covers only 5–7% of each image, most random patches contain no neurite signal at all; training on those neurite-free patches would teach the network mostly that black pixels are background. The foreground filter retains only patches with $\geq 1\%$ foreground, and a small random sample (15%) of background-only patches is then re-introduced so that the model still encounters the true class prior at training time and does not collapse to an all-foreground prediction. The patches are augmented during training with elastic deformation [22], flips, 90° rotations, and photometric jitter. Per-stage hyperparameter values appear in Table 3; per-stage counts are visualised in Figure 7.

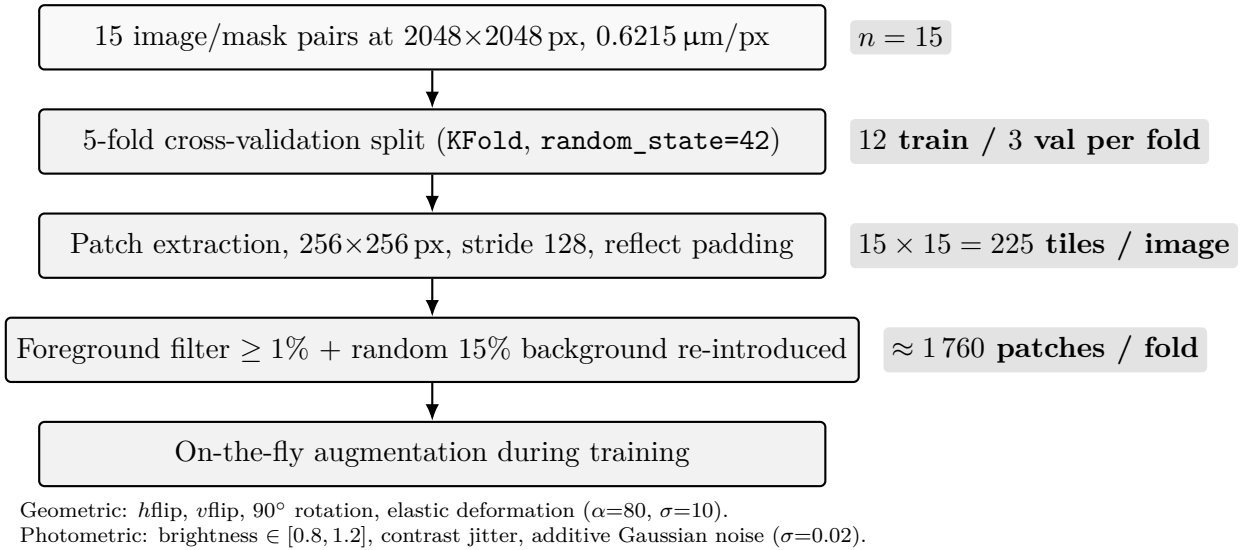


Figure 7. Training-data preparation, with effective sample counts at each stage (right column, bold). A dataset of 15 annotated images is partitioned at the image level into 5 folds. Within each fold, the 12 training images are tiled into 225 overlapping 256×256 px patches; the foreground filter and background-resampling step reduce this to roughly 1760 effective patches per fold. During training, augmentation then expands the per-epoch view of the data without increasing the static count.

4.5 Model architecture

The architectural starting point is the standard U-Net [22], augmented with the attention-gated skip connections of Oktay et al. [17] introduced in Section 2. This combination is a natural choice for sparse-foreground biomedical segmentation: attention gates suppress irrelevant background on the skip connections without substantial parameter overhead, which is the dominant difficulty when neurites occupy only 5–7% of the image. The combination is adopted here for that reason and because it provides the strongest established baseline for this class of problem without relying on architectural innovations the data cannot yet validate.

Concretely (Figure 8), the encoder comprises four convolutional blocks, each consisting of two 3×3 convolutions with batch normalisation and ReLU activation followed by 2×2

max-pooling; the decoder mirrors this structure using transposed convolutions for upsampling with attention-gated skip-connection concatenation, and a final 1×1 convolution with sigmoid activation produces a pixel-wise probability map in $[0, 1]$. Two encoder widths are compared in the ablation: a standard variant with channel widths $[64, 128, 256, 512]$ (approximately 31 M trainable parameters), and a 'lightweight' variant with $[32, 64, 128, 256]$ (approximately 8 M parameters). The attention-gate mechanism itself is unchanged from [17]; the mathematical details are recalled in Section 2.

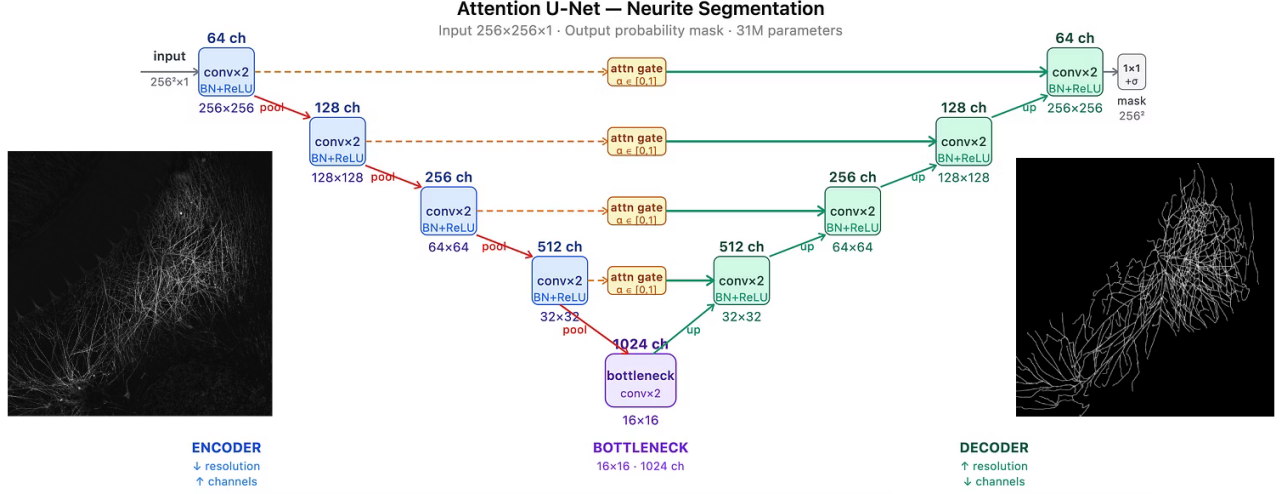


Figure 8. Attention U-Net used in this work. Encoder blocks at widths $\{32, 64, 128, 256\}$ (M5) or $\{64, 128, 256, 512\}$ (M1–M4b) downsample via max-pooling; the decoder upsamples via transposed convolution and concatenates each encoder feature map through an attention gate that re-weights it pixel-wise before fusion. A final 1×1 convolution with sigmoid activation produces the per-pixel neurite probability.

4.6 Training procedure

All models were trained on the ALICE high-performance computing cluster at Leiden University (NVIDIA RTX 2080 Ti and A100 GPUs) using PyTorch [20]. Each configuration used the Adam optimiser [9] with an initial learning rate of 10^{-4} , a cosine annealing schedule decaying to 10^{-6} , a batch size of 8, and a maximum of 200 epochs; training was terminated early when the validation DSC did not improve for 25 consecutive epochs.

The loss function for the standard configurations was a weighted combination of BCE and Dice loss,

$$\mathcal{L} = 0.5 \mathcal{L}_{\text{wBCE}} + 0.5 \mathcal{L}_{\text{Dice}}, \quad (3)$$

with positive class weight 15 to compensate for the $\approx 93\%$ background pixel ratio. For the clDice configurations, the total loss combines the topology-aware term defined in Section 2 with BCE and Dice. The mixture weights were copied of Shit et al. [25]; a weight sweep was outside the project’s time frame and is flagged in Section 6.2.

$$\mathcal{L} = 0.4 \mathcal{L}_{\text{wBCE}} + 0.4 \mathcal{L}_{\text{Dice}} + 0.2 \mathcal{L}_{\text{clDice}}. \quad (4)$$

4.7 Ablation study design

Seven model configurations were evaluated in a controlled ablation (Table 1). Each row introduces one design choice relative to the preceding row, so that any observed performance change can be attributed to a single intervention rather than to a composite of several. All seven configurations used identical 5-fold splits (random seed 42), enabling paired comparisons across architectures.

Model	Encoder	Attention	Loss	Elastic aug.	Params
M1 Baseline	[64, 128, 256, 512]	No	wBCE+Dice	No	31 M
M2 + Elastic	[64, 128, 256, 512]	No	wBCE+Dice	Yes	31 M
M3 + Attention	[64, 128, 256, 512]	Yes	wBCE+Dice	Yes	31 M
M4 + clDice	[64, 128, 256, 512]	No	wBCE+Dice+clDice	Yes	31 M
M4b + Attn+clDice	[64, 128, 256, 512]	Yes	wBCE+Dice+clDice	Yes	31 M
M5 Lightweight	[32, 64, 128, 256]	Yes	wBCE+Dice	Yes	8 M
M6 Light, no attn	[32, 64, 128, 256]	No	wBCE+Dice	Yes	7.8 M

Table 1. Ablation study configurations. M2–M5 each introduce one design choice relative to the row above; M4 and M4b branch from M2 (elastic augmentation only) to test loss-function variants independently of attention gates; M5 halves the encoder width of M3; M6 then removes the attention gates from M5, so that the M5 vs M6 pair and the M2 vs M3 pair together complete the (width $\times$ attention) factorial. The rightmost column lists the trainable-parameter count, which drops by a factor of four at the M3→M5 transition.

4.8 Inference procedure

Tiled inference with Hann-window blending. The full 2048×2048 px image is too large for a single forward pass, so inference proceeds tile-by-tile: the image is divided into overlapping 256×256 px tiles with stride 128 px (50% overlap) and reflect padding at the borders. This creates a problem however; averaging of overlapping predictions produces visible grid artefacts, because tile centres are covered by four overlapping tiles while edges are covered by only one. The standard fix, which dates back to the spectral analysis of windowed Fourier transforms [6] and has been widely adopted for tiled inference in biomedical segmentation networks [21], is to weight each tile’s probability map by a two-dimensional Hann (raised-cosine) window before accumulation, then divide out the accumulated weight map at the end. The result is a smooth, artefact-free probability map across the entire image. Figure 9 shows the one-dimensional cross-section of this blend; Figure 10 shows a representative full-image probability map produced by the procedure on image 007.

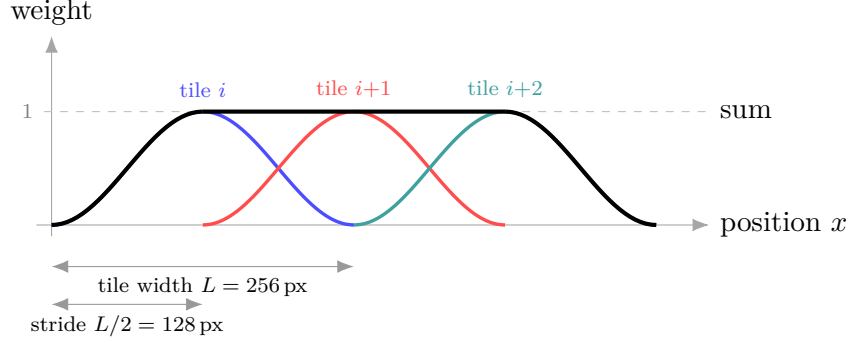


Figure 9. Hann-windowed tile blending, shown in a one-dimensional cross-section. Three consecutive tiles (coloured curves) carry a cosine bell that is zero at the tile boundary and unity at the centre. With a stride of half the tile width, the pixel-wise sum (black) is exactly constant in the interior and falls off only at the image boundary, so the inverse-weight normalisation that follows leaves no visible tile seams in the recovered probability map.

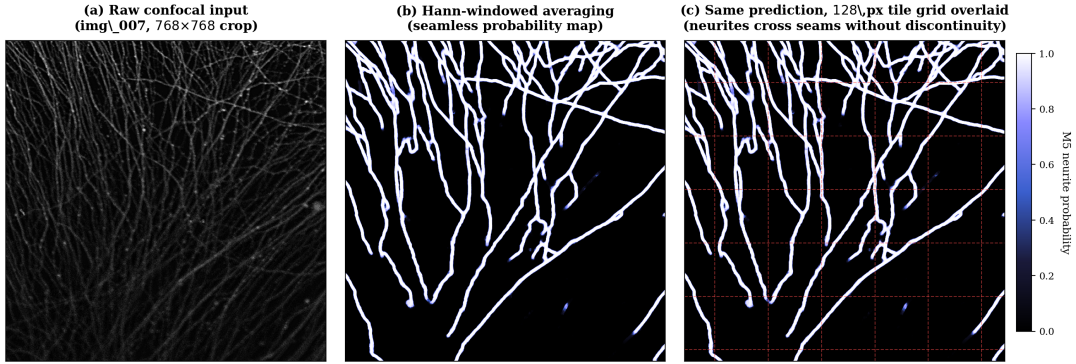


Figure 10. Hann-windowed tiled inference applied to image 007. The recovered probability map is artefact-free across the entire image field; no tile seams are visible at the 128-pixel stride boundaries.

For the 5-fold ensemble, inference is run independently with each fold’s best checkpoint and the five probability maps are averaged per pixel before thresholding. Ensembling reduces variance and tends to sharpen the probability boundary between neurite and background.

Post-processing. The averaged probability map is thresholded at $t = 0.70$ to produce a binary mask. The threshold was selected from a sensitivity sweep across $t \in [0.30, 0.95]$ in steps of 0.05 (Figure 11): the mean full-image DSC over the 15 images is essentially flat in the range $t \in [0.65, 0.75]$ (0.878–0.887), with $t = 0.70$ near the plateau peak (0.887) and at the minimum fold-to-fold standard deviation (0.021). The choice is therefore robust to perturbations of ± 0.05 . A morphological closing with a disk of radius 2 px then bridges small gaps left by local probability fluctuations: thin or faint fibres produce lower per-pixel probabilities than thick, well-resolved ones, so the threshold occasionally clips short sections along the centre of a real neurite, and closing restores continuity without measurably changing the overall foreground extent.

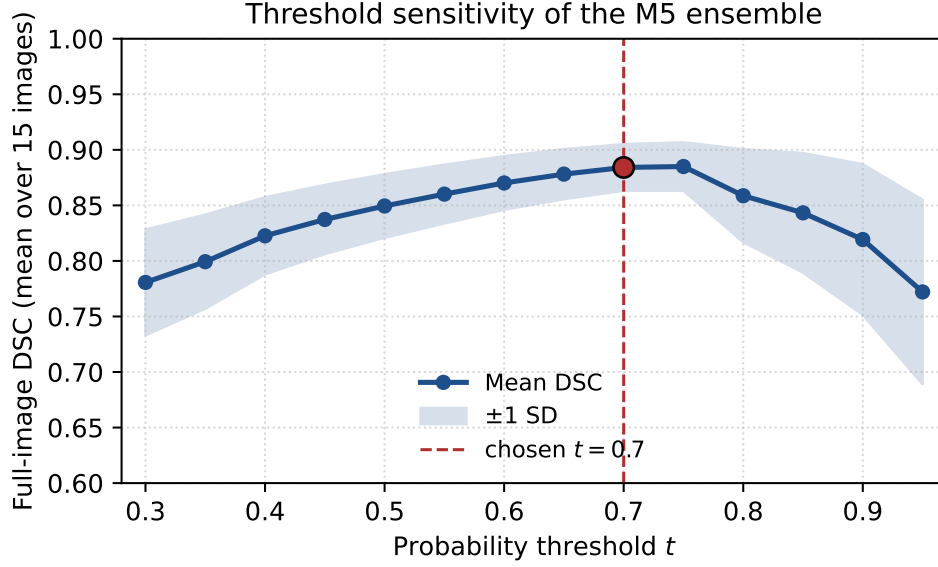


Figure 11. Threshold sensitivity of the M5 ensemble. Mean full-image DSC over the 15 images (blue line) with ± 1 SD band, as a function of the binarisation threshold t applied to the averaged probability map. The chosen $t = 0.70$ (red dashed) sits near the plateau peak; the curve is approximately flat in $t \in [0.65, 0.75]$ and degrades outside this range.

Skeletonisation and spur pruning. The closed binary mask is reduced to a one-pixel-wide centre-line by the `skeletonize` function of `scikit-image` [27]. The raw skeleton, however, contains a large number of short terminal stubs, pixels of width-induced noise that the thinning algorithm did not collapse, or genuine short side-branches whose biological significance is debatable. Without removing them, the junction count is inflated by exactly those spurious stubs, in much the same way that the spurious endpoints of any digital skeleton tend to dominate downstream topology metrics [24]. The skeleton is therefore spur-pruned: terminal branch segments (those with at least one endpoint and fewer than 8 pixels, corresponding to less than $\sim 5 \mu\text{m}$ of fibre in the present pixel calibration) are removed iteratively until convergence. Pruning operates on the connected components of the skeleton minus its junction pixels rather than pixel-by-pixel, so that only genuine stubs are removed and short sections of real branches are preserved.

Metric extraction. Seven morphometric descriptors are computed from the pruned skeleton and the binary mask, using the skeleton-graph primitives defined in Section 2. Their definitions are summarised in Table 2; the rationale for each is given in the caption.

Metric	Definition	Why it matters
Total fibre length	Skeleton pixel count $\times 0.6215 \mu\text{m}$	Primary size of the neurite network
Junction count	Number of 8-connected clusters with ≥ 3 skeleton neighbours	Branching complexity
Branch count	Skeleton components remaining once junction pixels are removed	Counts elementary fibre segments
Mean branch length	Total length $\div$ branch count	Sets a typical segment scale
Branchiness ratio	Junction count $\div$ total length (μm^{-1})	Size-invariant complexity index
Area covered	Foreground pixel count $\times$ pixel area; μm^2 and % of field of view	Network density per unit area
Mean branch thickness	Mean of $2\times$ distance-transform sampled at skeleton pixels	Average neurite diameter

Table 2. The seven morphometric descriptors extracted from each predicted mask. The first five are computed from the skeleton alone; the last two, area covered and mean branch thickness, additionally use the filled mask to characterise network density and fibre width. Branchiness, a ratio of junctions to fibre length, is invariant to both field of view and magnification, which makes it the most directly comparable descriptor across acquisitions; mean branch length, mean branch thickness, and area covered as a fraction of the field are likewise independent of the field of view. Endpoint counts are used internally for spur pruning but are not reported separately, since on a connected skeleton they are redundant with the branch count.

4.9 Hyperparameter summary

Table 3 consolidates every numerical hyperparameter used by the method, so that each design decision can be located in one place. Values were either fixed from prior work (e.g. Adam defaults, Hann tile blending), set proportional to a dataset property (e.g. $w_{\text{BCE}} = 15$ from the background/foreground ratio), or selected by sensitivity analysis on the training data (e.g. closing radius $r = 2$ px, threshold $t = 0.70$; Figure 11).

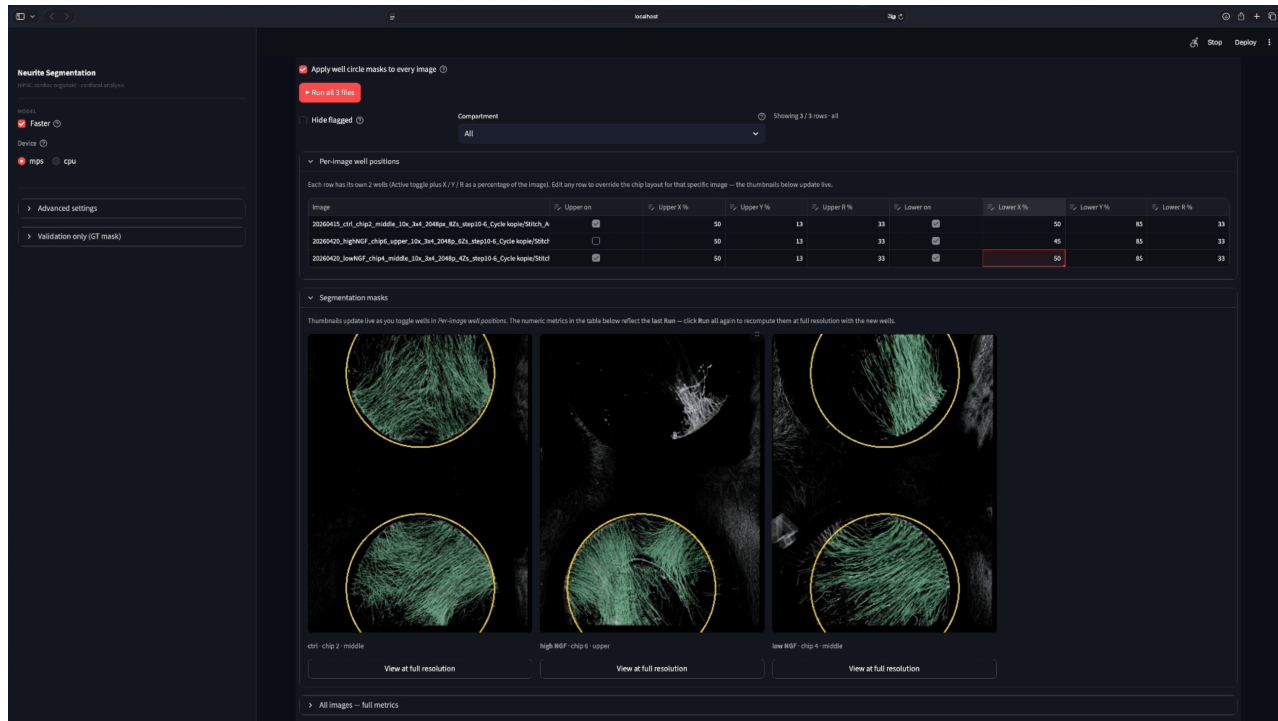
Stage	Hyperparameter	Value
Preprocessing	CLAHE block size	127 px
	CLAHE bins	256
	CLAHE slope	3.0
	Rolling-ball radius	50 px
	Gamma correction	$\gamma = 0.7$
	Median filter radius	1 px
GT generation	Tracing tool	SNT (Fiji)
	Dilation disk radius	2 px
Patches	Patch size	256×256 px
	Stride	128 px (50% overlap)
	Foreground filter	$\geq 1\%$ positive
	Background re-introduction	random 15% of all-bg patches
Augmentation	Elastic	$\alpha = 80, \sigma = 10$
	Brightness jitter	factor $\in [0.8, 1.2]$
	Gaussian noise	$\sigma = 0.02$
	Flips	horizontal, vertical
	Rotations	90°
Architecture	Encoder widths (M1–M4b)	[64, 128, 256, 512]
	Encoder widths (M5)	[32, 64, 128, 256]
	Trainable params (M5)	≈ 8 M
	Trainable params (M1–M4b)	≈ 31 M
Optimisation	Optimiser	Adam
	Initial learning rate	10^{-4}
	LR schedule	cosine annealing $\rightarrow 10^{-6}$
	Batch size	8
	Max epochs	200
	Early-stopping patience	25 epochs
Loss	Weighted BCE positive weight	$w = 15$
	BCE/Dice mix	$0.5 \mathcal{L}_{\text{BCE}} + 0.5 \mathcal{L}_{\text{Dice}}$
	clDice mix (M4, M4b)	(0.4, 0.4, 0.2) for BCE/Dice/clDice
Inference	Tile size	256×256 px
	Tile stride	128 px
	Blending window	2D Hann
	Ensemble	5-fold average
Post-process	Probability threshold	$t = 0.70$
	Morphological closing disk	$r = 2$ px
	Spur-prune length	≥ 8 px
Calibration	Pixel size	$0.6215 \mu\text{m}/\text{px}$

Table 3. Complete hyperparameter listing for the pipeline. Each stage maps to the corresponding subsection of Section 4.

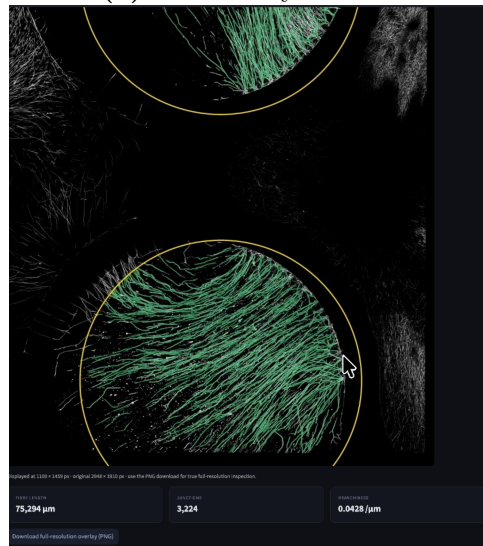
4.10 The analysis application

The method described above is, at its core, a research codebase: useful inside this thesis but ill-suited to the daily routine of a biology lab. To close that gap, the trained M5 ensemble and the full post-processing chain were packaged into a standalone desktop application built on Streamlit (<https://streamlit.io>). The defining feature is that the application performs inference on raw, unprocessed Olympus .oir z-stacks directly: a researcher drops a folder of raw images onto the window, the application runs the same preprocessing, ensembled M5 inference, post-processing and morphometric extraction described in Section 4, and the seven morphometric read-outs appear within minutes. The numbers it produces are identical to those reported in this thesis, but the user never has to write a line of code or touch the underlying scripts.

Five design choices make the application practical for routine use. First, it draws and stores user-defined region-of-interest (ROI) wells on the loaded image, so that all subsequent morphometric statistics are computed only inside the wells where the neurites actually grow and not on the empty plastic between them. Second, it parses the standardised .oir filenames to recover experimental metadata (condition, chip number, compartment, date, acquisition parameters) and propagates that metadata into every exported table, removing the manual bookkeeping that batch experiments otherwise demand. Third, each well in a multi-well chip can be flagged for exclusion through a per-well sidebar toggle. Upper and lower wells sometimes dry out during culture, and including such a well in the per-chip average distorts the dose-condition mean without indicating any biological response. Flagged wells are dropped from all downstream pooling. Their per-well predictions are still kept on disk for audit, and the flag is written to the exported master CSV as a dedicated column, so any later analysis can reproduce or revisit the exclusion. Fourth, once a folder of images has been processed, the application automatically pools condition-level summary statistics over the non-flagged wells and produces the figures and tables used to compare conditions (bar plots, radar plots, per-chip strip plots) without any further user effort. Fifth, batch processing of a folder runs every image through the full method in a single click and exports per-image masks, skeleton overlays, and a master CSV containing every metric for every image. Installation scripts for macOS, Linux and Windows are included, so setting it up in a new lab is straightforward.



(a) Main analysis view.



(b) ROI / well selection.

Figure 12. The Streamlit analysis application. (a) The main view places the raw confocal image alongside the M5 segmentation overlay and the extracted morphometric metrics, with threshold and post-processing parameters exposed as live sidebar controls and CSV/mask-TIFF/skeleton-overlay exports. (b) The ROI tool restricts the metrics to a named sub-region (here a single well of a multi-compartment chip acquisition), so the morphometric profile is computed only over the biologically relevant area of the image.

5 Experiments

5.1 Dataset and quality control

The AST group at the University of Twente supplied an initial pool of 33 raw `.oir` acquisitions, each containing four fluorescence channels and a per-acquisition z-stack depth in the range 14–20 slices (median 17). A manual quality check retained 15 acquisitions (numbered 1–15) that were in-focus across the entire 2048×2048 px field and rich enough in neurite signal to support tracing; the rest were discarded as out-of-focus, neurite-poor, or showing acquisition artefacts (bleach lines, sample drift). Because annotated images were scarce, two of the retained acquisitions are independently-selected z-slices of the same physical culture, traced separately in SNT; they are kept as two file entries, so the pool of $n = 15$ entries is drawn from 14 unique cultures. The implications for paired-test independence are discussed in Section 6.2.

All selected images were preprocessed and SNT-traced as described in Section 4. To compensate for the small image count, the training procedure relies on dense patch extraction ($\approx 1\,760$ patches per fold) and extensive on-the-fly augmentation (elastic, geometric, photometric), so that each epoch exposes the network to far more effective examples than the unique image count alone suggests.

5.2 Ablation study results

Table 4 summarises the mean and standard deviation of the best validation DSC across all five folds for each configuration.

Model	Mean DSC	Std	Parameters
M1 Baseline	0.706	0.041	31 M
M2 + Elastic	0.716	0.031	31 M
M3 + Attention	0.718	0.032	31 M
M4 + clDice	0.716	0.030	31 M
M4b + Attn+clDice	0.709	0.043	31 M
M5 Lightweight	0.716	0.032	8 M
M6 Light, no attn	0.717	0.030	7.8 M

Table 4. 5-fold cross-validation results (patch-level DSC, 256×256 px patches). Best mean DSC and lowest parameter count in **bold**. All models trained for up to 200 epochs with early stopping (patience 25).

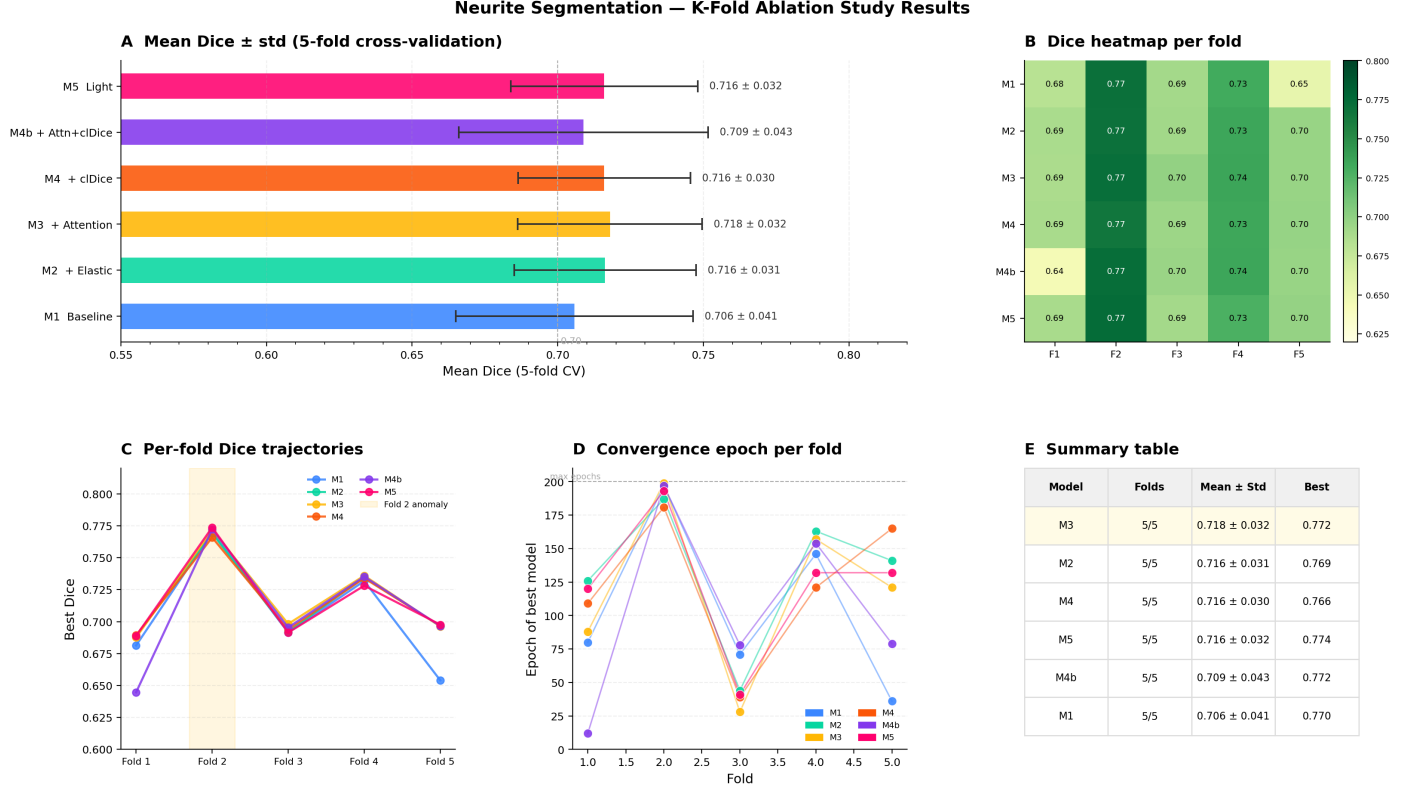


Figure 13. Ablation study dashboard. (A) Mean DSC $\pm$ std per model. (B) Fold-wise DSC heatmap, showing the consistent fold-2 advantage. (C) Per-fold convergence trajectories. (D) Convergence epoch per fold. M5 matches M3 at one-quarter of the parameter count.

5.3 Analysis of ablation results

All seven configurations land in a narrow DSC range of 0.706–0.718, a spread of just 1.2 percentage points (Figure 13), despite a fourfold difference in parameter count between the heaviest and lightest variant. Two clean comparisons within the now-complete (width $\times$ attention) factorial grid sharpen the interpretation. The attention-gate contribution is isolated by the M2 vs M3 pair at wide encoder width ($0.716 \rightarrow 0.718$, $+0.002$ DSC) and by the M6 vs M5 pair at narrow encoder width ($0.717 \rightarrow 0.716$, -0.001 DSC); both contrasts are well inside fold-to-fold noise, so attention gates contribute essentially nothing on this dataset. The width contribution is isolated by the M2 vs M6 pair without attention ($0.716 \rightarrow 0.717$, $+0.001$ DSC) and by the M3 vs M5 pair with attention ($0.718 \rightarrow 0.716$, -0.002 DSC); halving the encoder width therefore costs nothing measurable either. Two of the standard refinements one would expect to matter at this image regime, namely the attention gates of Oktay et al. [17] and a fourfold parameter expansion of the encoder, are not contributing at this data size.

What these seven near-identical numbers say is that the U-Net variants are already strong enough for this task; the dataset, not the architecture, is what limits the headline number. With only $n = 14$ unique cultures generating roughly 1760 patches per fold, no model in the configuration ‘family’ has enough varied examples to pull ahead of the others. Therefore,

within-family architectural choices that change the parameter count fourfold or add attention gates produce no measurable benefit.

The clDice configurations (M4, M4b) also do not improve over their non-clDice counterparts; M4b in fact shows the highest variance (0.043). At the single mixture weight tested (0.4, 0.4, 0.2 for BCE, Dice, clDice) the additional term did not improve over BCE+Dice; a weight sweep was not performed (Section 6.2). My guess is that clDice’s soft-skeleton signal is largely redundant at this DSC level with what dilated-mask supervision already provides, since the dilation radius itself is the dominant source of topology-preserving signal on this dataset (Section 3).

A separate fold-level pattern is worth noting. Fold 2 (image 006, image 009, and image 014) consistently achieves a validation DSC of about 0.77, several points higher than the other folds (which sit in the range 0.69–0.71). These three images appear to be particularly representative of the training distribution. The pattern is itself useful, because it illustrates why a fixed train/test split would have produced a misleading evaluation here: a test set drawn from fold 2 would have inflated the reported DSC by roughly five percentage points relative to a test set drawn from a harder fold.

5.4 Model selection: M5 lightweight Attention U-Net

M5 is selected as the final model on two grounds. First, its mean patch-level DSC of 0.716 ± 0.032 is comparable to both the heaviest variant M3 (0.718 ± 0.032 , four times the parameters) and the matching no-attention variant M6 (0.717 ± 0.030 , 0.2 M fewer parameters): all three sit within one fold-to-fold standard deviation of one another. Second, its 8 M parameters represent one-quarter of the 31 M used by M3, reducing inference cost and making future fine-tuning on an expanded dataset substantially cheaper. M6 could be defended as a smaller alternative. However, the difference between M5 and M6 (-0.001 DSC) is well below the within-fold noise. Moreover, because M6 was trained later in the study, the deployed-ensemble checkpoints in the software already exist for M5. Therefore M5 kept as the deployment choice.

5.5 Full-image ensemble DSC

The 5-fold ensemble of M5 was applied to all 15 full-resolution 2048×2048 px images using Hann-windowed tiled inference, thresholded at $t = 0.70$ and closed with a disk of radius 2 px. Three reference baselines were scored on the same images and appear alongside M5 in Table 5: Otsu intensity thresholding [19], a Frangi vesselness filter [3] with hysteresis thresholding (the abandoned GT pipeline of Section 4, reused here as a non-deep baseline with $\sigma = 2.0$, scales 2–6 px, $\beta = 0.3$), and nnU-Net v2 (discussed in Section 5.6). All four methods share the same closing radius and the same full-image evaluation.

Image	Otsu $\uparrow$	Frangi $\uparrow$	nnU-Net $\uparrow$	M5 $\uparrow$	$\Delta_{\text{M5-nn}}$
image 001	0.562	0.594	0.648	0.877	+0.229
image 002	0.466	0.532	0.582	0.900	+0.318
image 003	0.392	0.639	0.719	0.911	+0.192
image 004	0.391	0.516	0.759	0.876	+0.117
image 005	0.351	0.569	0.705	0.873	+0.168
image 006	0.424	0.574	0.731	0.888	+0.157
image 007	0.472	0.622	0.657	0.903	+0.246
image 008	0.590	0.646	0.690	0.902	+0.212
image 009	0.492	0.593	0.854	0.861	+0.007
image 010	0.107	0.182	0.381	0.838	+0.457
image 011	0.509	0.624	0.857	0.909	+0.052
image 012	0.375	0.616	0.774	0.895	+0.121
image 013	0.406	0.579	0.835	0.879	+0.044
image 014	0.360	0.568	0.763	0.887	+0.124
image 015	0.365	0.586	0.827	0.906	+0.079
Mean $\pm$ SD	0.418 ± 0.109	0.563 ± 0.108	0.719 ± 0.123	0.887 ± 0.020	+0.168

Table 5. Per-image full-image DSC across all four methods, evaluated against the dilated SNT ground truth, ordered worst to best by mean DSC. Otsu is global intensity thresholding with closing disk $r = 2$ px; Frangi uses hysteresis thresholding with the same closing; nnU-Net v2 is the framework’s auto-configured 2D model trained fold-by-fold on the identical splits; M5 is the 5-fold ensemble (threshold 0.70, same closing). $\Delta_{\text{M5-nn}}$ is the per-image M5–nnU-Net margin. Higher DSC is better.

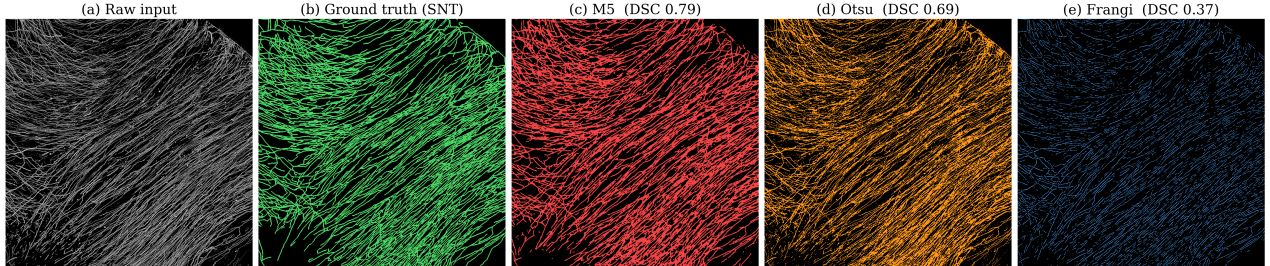


Figure 14. Qualitative comparison of the segmenters on the held-out CTRL acquisition (Section 5.9; upper well, representative 2200×2200 px crop, not used for training), the manual tracing shown at the 2 px dilation used throughout. **(a)** Raw preprocessed input. **(b)** Manual SNT tracing (ground truth). **(c)** M5 ensemble; the predicted mask tracks the manual tracing across the field. **(d)** Otsu thresholding; the global threshold captures the bright fibres but bleeds into the diffuse background between them. **(e)** Frangi vesselness; more selective but fragmented at crossings and faint distal branches. DSCs are computed against the dilated ground truth on the crop shown.

The improvement from both classical baselines to M5 is large: +0.469 DSC over Otsu and +0.324 over Frangi. M5 ranks above each baseline on every image: the paired Wilcoxon signed-rank test reaches its maximum statistic $W = W_{\text{max}} = 120$ at $n = 15$ ($p = 3.1 \times 10^{-5}$, exact two-sided) against both baselines. Image 009 and image 011 are two z-slices of the same

culture; the result holds when these are collapsed to one entry ($W = 105$, $p = 6.1 \times 10^{-5}$). The Otsu-to-Frangi gap of +0.145 DSC confirms that the lead is not an artefact of comparing against a particularly weak baseline; a modern classical vesselness filter still loses by a wide margin. One calibration asymmetry is worth surfacing here rather than only in Limitations: M5’s threshold and closing radius were sensitivity-tuned on this dataset, whereas Frangi was run with parameters inherited from the abandoned GT-generation pipeline and Otsu has no free parameter. A fraction of the +0.324 DSC gain over Frangi therefore reflects calibration rather than capacity, mirroring the shortcoming of the nnU-Net comparison (Section 5.6).

The patch-level DSC of 0.706–0.718 in the ablation and the full-image DSC of 0.887 here measure different things. The ablation uses 256×256 px patches filtered to contain $\geq 1\%$ foreground; that metric quantifies performance on the foreground-dense parts of the image. The full-image metric is computed over the entire 2048×2048 px field, which is roughly 95% background, and background pixels are easy. The patch number is therefore the appropriate one for model selection, the full-image number for deployment. All baselines in Table 5 are scored at the same full-image level, so the cross-method comparison is internally consistent.

The one image that scores lower than the rest under every method is image 010 (M5: 0.838; Frangi: 0.182; Otsu: 0.107; nnU-Net: 0.381). Visual inspection reveals a sparse, low-contrast neurite network that is inherently difficult to segment regardless of method. Excluding it raises the M5 mean only marginally to 0.890 ± 0.014 , and the Wilcoxon test against either classical baseline still reaches the maximum $W = W_{\max} = 105$ at $p = 1.2 \times 10^{-4}$ ($n = 14$, exact two-sided), so the claim “M5 beats both classical baselines on every image” holds whether or not the outlier is included.

Figure 14 shows this contrast visually on a held-out CTRL acquisition with its blind SNT tracing alongside: M5 performs best (DSC 0.79 on the crop shown); Here Otsu also performs reasonably well, but introduces substantial noise (0.69); Frangi fragments at crossings and faint distal branches (0.37). These are the same failure modes Table 5 captures numerically on the in-sample images.

5.6 Comparison against a modern deep-learning baseline (nnU-Net v2)

The classical baselines settle whether M5 beats rule based methods; they leave open the more interesting question of whether a heavier, fully self-configuring deep model does better. To answer it, nnU-Net v2 [8] was trained fold-by-fold on the identical 5-fold splits (`random_state=42`), using the framework’s auto-configured 2D plan (9-stage PlainConvUNet, 1280×1024 px patches) for 100 epochs, past the convergence point M5 reached at roughly 70–100 epochs. Inference used nnU-Net’s own sliding-window prediction with default test-time augmentation. The per-image scores occupy the *nnU-Net* column of Table 5.

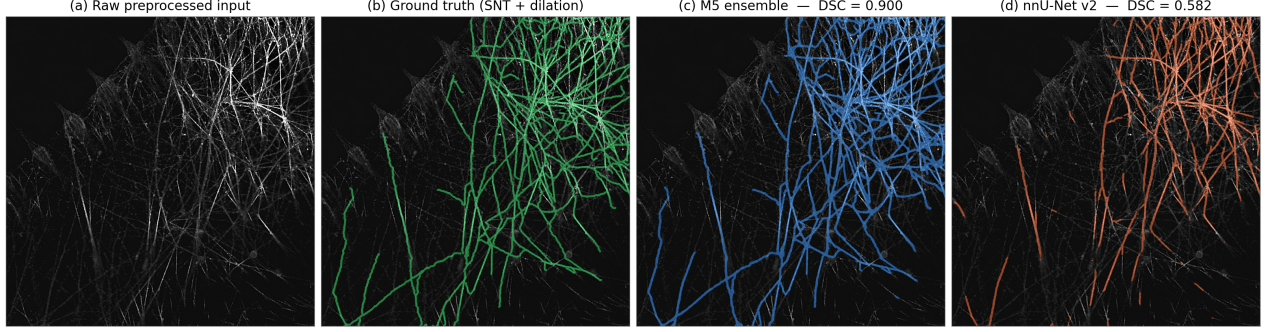


Figure 15. Qualitative comparison on image 002 (centre crop, 1024×1024 px of the full 2048×2048 px field), the image with the largest M5–nnU-Net gap ($\Delta = +0.318$ DSC). (a) Raw preprocessed input. (b) Ground truth (SNT tracing, disk dilation $r = 2$ px). (c) M5 overlay; the mask follows the network across the field. (d) nnU-Net overlay; it recovers the bright fibre cores but loses thinner distal branches.

M5 ranks above nnU-Net on every held-out image, with a median per-image margin of $\Delta = +0.157$ DSC (95% confidence interval (CI) $(+0.117, +0.229)$, paired Cohen’s $d_z = +1.45$; Δ is M5 minus nnU-Net). A paired Wilcoxon signed-rank test returns $W = W_{\max} = 120$ at $p = 3.1 \times 10^{-5}$ (unchanged when the duplicate-culture pair is collapsed; Section 6.2). The smallest margin ($\Delta = +0.007$) sits well inside M5’s fold-to-fold SD of 0.020, so on that one image the two are essentially tied. The honest summary is that M5 ranks above nnU-Net on all 15 images, with the margin concentrated on the harder ones.

The gap is unlikely to be one of architectural capacity: if it were, nnU-Net’s deeper encoder (≈ 28 M parameters vs. M5’s 8 M) would help rather than hurt, and the ablation (Section 5) already showed capacity to be inert on this dataset. A simpler and fairer explanation is that nnU-Net’s default 2D configuration operates outside its design envelope. Its 1280×1024 px patches yield on the order of 24 gradient updates per epoch on $n = 12$ images, against M5’s $\approx 1\,760$ augmented 256×256 px patches, which is a roughly hundred-fold richer gradient signal per epoch. In addition, M5’s post-processing was tuned on this dataset and matched to the GT dilation radius, while nnU-Net ran at defaults. So the honest headline is that default 2D nnU-Net is not competitive in this regime, not that M5’s architecture is intrinsically superior. These caveats, together with the 100-epoch training budget and the 2D-only evaluation, are detailed in Section 6.2.

The practical implication is that at the data scale of a really small size ($n \approx 15$ traced images), a compact U-Net wrapped in a domain-tailored preprocessing, tiling and post-processing chain can outperform a big generic self-configuring framework.

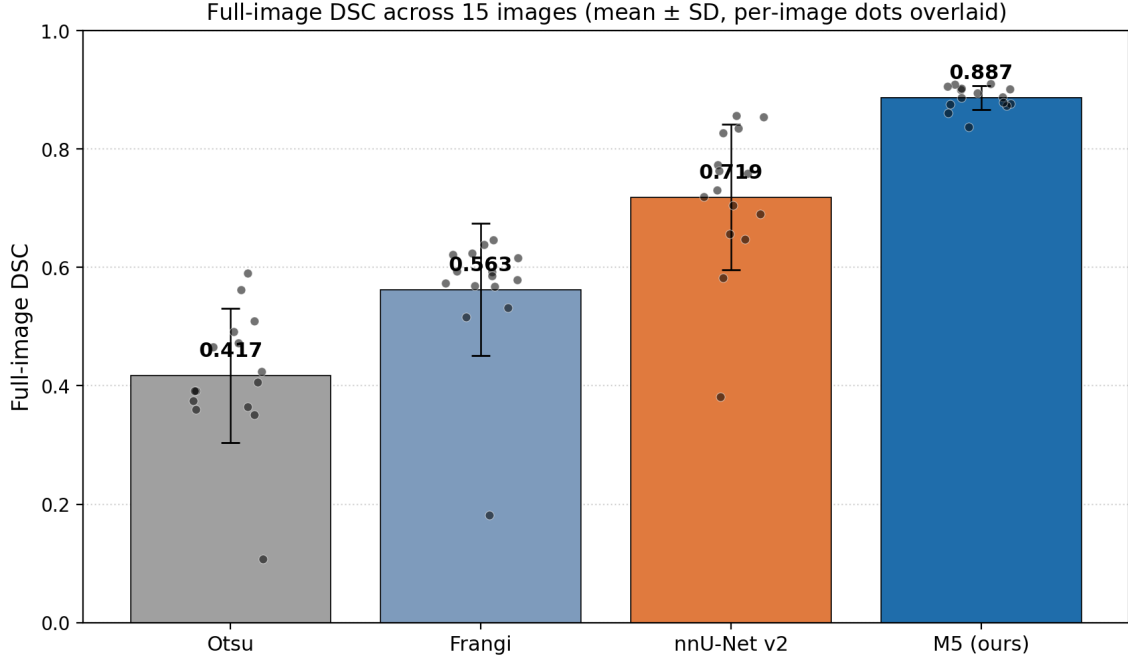


Figure 16. Full-image DSC across 15 images for the four methods (mean $\pm$ SD; per-image dots overlaid). Mean DSC climbs from Otsu (0.42) through Frangi (0.56) and nnU-Net v2 (0.72) to M5 (0.887). M5 also has the narrowest spread (SD 0.020 vs. 0.11–0.12), which is why the paired Wilcoxon attains $W_{\max}$ against each baseline. The single outlier on each non-M5 method is image 010 (Section 5.7).

5.7 Ensemble gain and post-processing sensitivity

Two further sanity checks confirm that the headline DSC of 0.887 is driven by the trained network rather than by the choice of ensembling strategy or post-processing hyperparameters.

Single-fold vs. 5-fold ensemble. Evaluating each fold’s checkpoint independently gives a mean single-fold DSC of 0.818 ± 0.038 . Averaging the five checkpoints into the deployed ensemble raises this to 0.887 ± 0.020 and roughly halves the fold-to-fold spread (SD $0.038 \rightarrow 0.020$), so the ensemble is both more accurate and more stable than any single fold. It is therefore the configuration carried into the external validation of Section 5.9.

Closing radius and spur-prune length. Tables 6 and 7 sweep the two post-processing hyperparameters around their chosen values. The closing radius costs at most 0.008 DSC across $r \in [0, 2]$, which is well within fold-to-fold noise. We chose $r = 2$ deliberately, even though $r = 0$ scores marginally higher on DSC. Matching the GT dilation radius produces a more topologically reliable binary mask for downstream skeletonisation, which is what actually matters in deployment. The spur-prune length L is similarly inconsequential within $[4, 16]$ px. Total fibre length varies by only $\approx 3.3\%$ across this range, much smaller than the natural inter-image variability of $\approx 38\%$, so the metric is robust to the choice. Both sweeps used

$L = 8$ px as the default. At the present pixel calibration this corresponds to roughly five micrometres, below which terminal stubs are unlikely to represent biologically meaningful side branches at the imaging resolution used here. The choice is empirical for this dataset rather than inherited from a specific reference.

Closing radius r (px)	0	1	2	3	4
M5 mean DSC $\uparrow$	0.892	0.890	0.887	0.874	0.862
SD	0.021	0.021	0.021	0.020	0.021

Table 6. Sensitivity of full-image DSC to the morphological closing radius (spur prune fixed at $L = 8$ px, threshold $t = 0.70$). The chosen $r = 2$ is bolded. DSC declines only gently across the small radii, and the 0.005 gap between the unclosed mask ($r = 0$, 0.892) and $r = 2$ (0.887) sits well inside the 0.021 fold-to-fold SD, so the choice among small radii is within measurement noise. We fixed $r = 2$ to match the GT dilation radius, giving a closed binary mask that is more topologically reliable under downstream skeletonisation (Section 4).

Spur-prune length L (px)	4	8	12	16
Mean fibre length (μm)	23 073	22 763	22 519	22 314
SD across images	8 702	8 640	8 590	8 545

Table 7. Sensitivity of total fibre length to the spur-prune length (closing fixed at $r = 2$ px). The chosen $L = 8$ px is bolded.

5.8 Morphological metrics

Figure 17 illustrates how the morphometric descriptors are extracted from the skeleton, using a representative region of the held-out CTRL acquisition. The M5 prediction and SNT ground-truth skeletons are shown side by side, with their morphological primitives (neurite centre-line, junctions, endpoints) colour-coded. Table 8 reports the four primary morphometric descriptors extracted from the M5 ensemble skeletons for all 15 images, at the calibrated pixel size given in Section 4.1. The final row summarises the population statistics over the 15 files as stored.

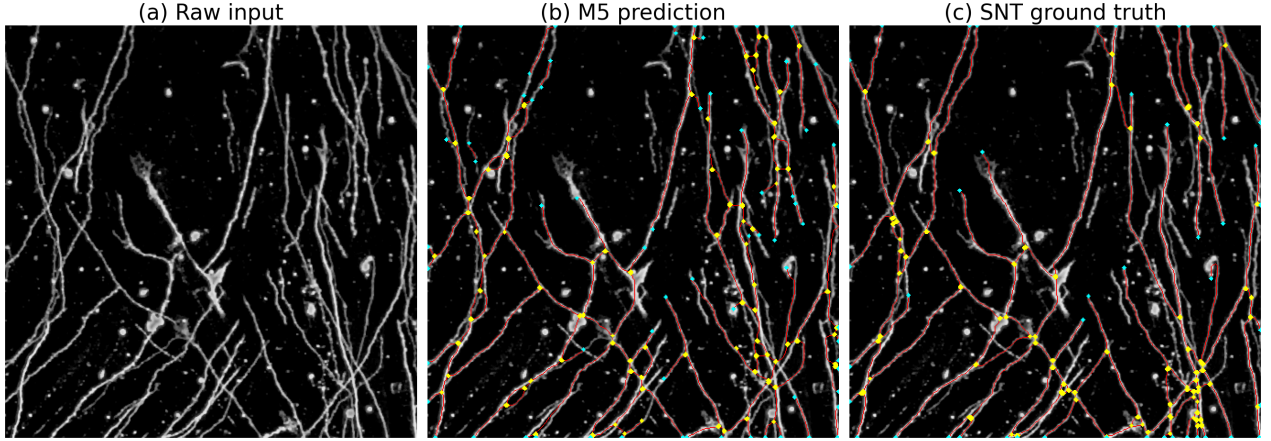


Figure 17. Skeleton-derived metric extraction on the held-out CTRL acquisition, shown on a representative 450×450 px region. **(a)** Raw preprocessed input. **(b)** M5 prediction skeleton and **(c)** SNT ground-truth skeleton, both colour-coded red = neurite centre-line, yellow = junction (≥ 3 neighbours), cyan = endpoint (1 neighbour); the junction and endpoint markers are dilated for visibility. In this region the predicted and ground-truth skeletons track the same fibres with junctions in matching locations (86 vs 79 junctions in the crop). Over the full image the total fibre length agrees closely (330.2 vs 340.9 mm, $\sim 3\%$), whereas junction counts differ more (11 739 vs 7 707), consistent with the higher fold-to-fold variability of count-style metrics reported in Table 9. The region shown is a representative well-agreeing patch and is not claimed to be representative of junction agreement across the whole image.

Image	Length (μm)	Junctions	Branches	Mean branch (μm)
image 001	29 241	1093	1759	16.6
image 002	26 484	852	1386	19.1
image 003	23 815	1030	1629	14.6
image 004	19 460	718	1175	16.6
image 005	23 773	877	1419	16.8
image 006	21 899	793	1306	16.8
image 007	39 526	1122	1823	21.7
image 008	32 782	1085	1744	18.8
image 009	33 160	1067	1712	19.4
image 010	7 451	148	290	25.7
image 011	33 160	1067	1712	19.4
image 012	16 633	722	1181	14.1
image 013	13 319	475	782	17.0
image 014	15 960	604	984	16.2
image 015	13 957	491	795	17.6
Mean $\pm$ SD	23 375 $\pm$ 8 743	810 $\pm$ 277	1313 $\pm$ 434	18.0 $\pm$ 2.8

Table 8. Morphological metrics per image (0.6215 $\mu\text{m}/\text{px}$, threshold 0.70, closing disk $r = 2$ px, spur prune ≥ 8 px). image 009 and image 011 are two z-slices of the same culture and yield near-identical predictions, hence the identical metrics here; their statistical independence is discussed in Section 6.2.

Metric reliability across model realisations. The standard deviation reported here is a property of the segmenter, not of the metric itself: it is the spread of a read-out across the five folds of the model, not the natural variability of the quantity being measured. Two uncertainties feed into a morphometric output and should not be conflated. The first is biological, the natural variation of neurite morphology from one image to the next, which is the inter-image SD already reported in Table 8. The second is model-realisation uncertainty: had the 5-fold ensemble drawn on a different subset of folds, how much would the extracted metric move? Table 9 and Figure 18 address this by running each fold’s post-processing chain independently and reporting the standard deviation across folds, per image and per metric. This fold-to-fold spread is not annotation error, which would need an inter-annotator comparison, and not biological spread; it is deployment risk, telling a future user how much the read-out depends on which subset of data the model happened to be trained on. Total fibre length is the most stable (CV = 8.5%). Count- and ratio-style metrics (junctions, branches, mean branch length, branchiness ratio) carry roughly twice the relative uncertainty (CV $\in$ 10.6–20.8%), because each false-positive blob or false-negative gap near a branch crossing toggles the junction count by one or two.

Metric	Mean fold SD	Mean CV
Total fibre length (μm)	2 111	8.5%
Branchiness ratio (μm^{-1})	0.0040	10.6%
Mean branch length (μm)	2.03	12.3%
Junction count	179	19.1%
Branch count	322	20.8%

Table 9. Fold-to-fold variability of the five skeleton-derived morphometric descriptors, computed as the mean across 15 images of the per-image standard deviation across the 5 fold checkpoints. Length-style metrics are stable ($\text{CV} \approx 8\text{--}12\%$); count-style metrics carry approximately twice the relative uncertainty because each false-positive blob or false-negative gap near a junction toggles the junction count by one or two.

Complementary morphometrics. The two remaining descriptors, area coverage and mean branch thickness, are spatial-extent measures derived directly from the binary mask of the M5 ensemble rather than from the skeleton. They therefore do not admit the same per-fold decomposition, because the per-fold masks are averaged into a single probability map before thresholding. Across the 15 training images they evaluate to a branchiness ratio of $0.035 \pm 0.005 \mu\text{m}^{-1}$, area coverage of $8.5 \pm 3.2\%$ of the field of view, and mean branch thickness of $6.3 \pm 0.6 \mu\text{m}$. The dataset outlier image 010 also shows the lowest branchiness ratio (0.020 versus the training-set mean of 0.035), consistent with its sparse, poorly developed network.

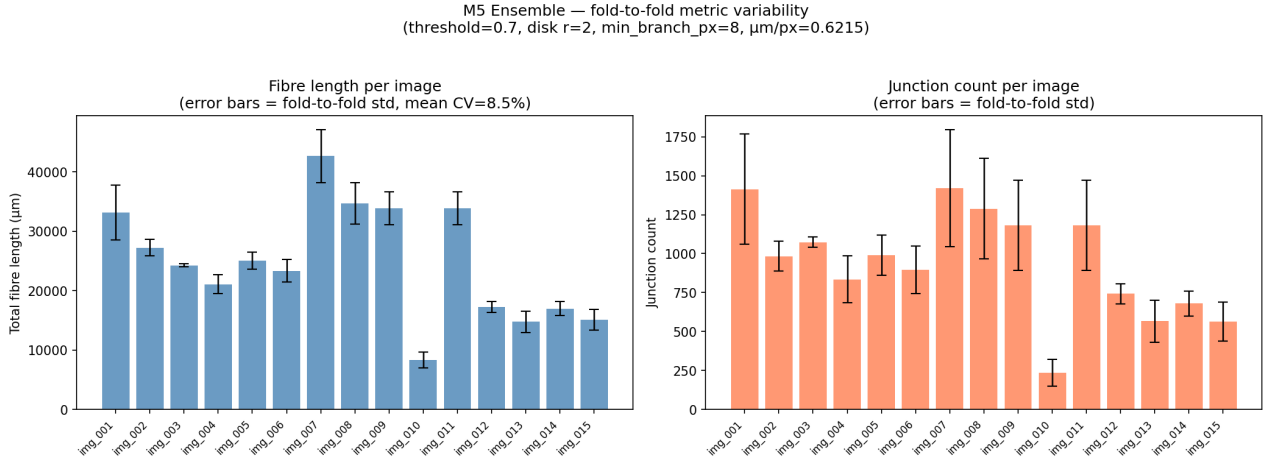


Figure 18. Fold-to-fold metric variability across the 15 images. Error bars show the standard deviation across the five fold checkpoints for each image. Left: total fibre length (mean $\text{CV} = 8.5\%$). Right: junction count (mean $\text{CV} = 19.1\%$). The outlier image 010 shows the highest relative uncertainty, consistent with its sparse network.

5.9 External validation against independent SNT tracings

Every number so far compares M5 to the same SNT tracings it was trained on. That controls for overfitting within the training distribution, but it cannot rule out a drop on genuinely new

images. To test this, two confocal images from the deployed NGF batch (Section 5.10) were retraced from scratch in SNT, one CTRL and one LOW. This batch was never seen in training, and both images came from the middle compartment to hold compartment effects fixed. No retraining or re-tuning was done. Both tracings came from the annotator who labelled the training set, so this probes image-distribution shift but not annotator style; a second annotator would be better, and we flag it as the principal follow-up (Section 6).

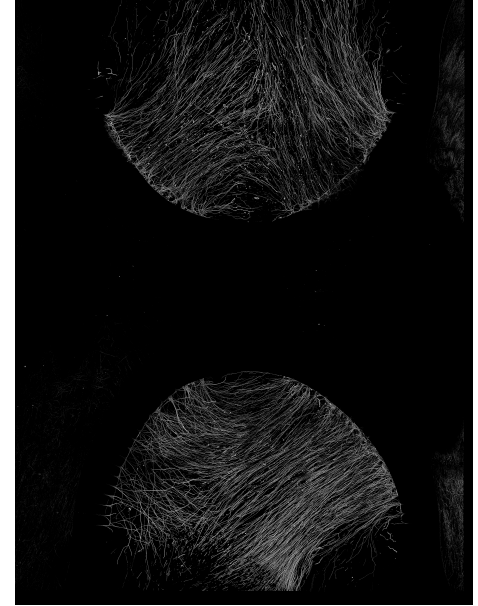
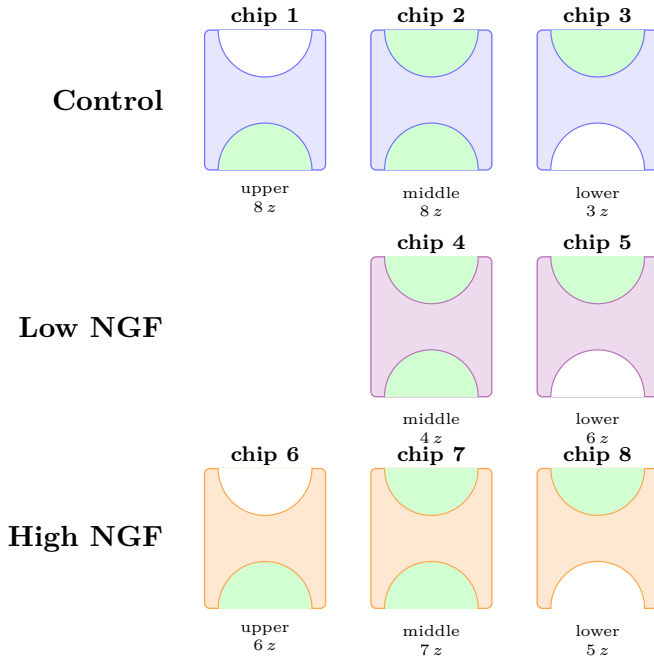
Before scoring, the prediction was restricted to the wells visible in each acquisition, drawn as an operator would mark them with the tool of Section 4.10. The manual ground truth is traced only inside the wells, while the raw prediction covers the whole stitched field; without this step, valid out-of-well signal would count as a false positive.

Two numbers were then computed. Centre-line agreement (clDice at a 2-pixel tolerance, equal to the GT dilation radius) reached a mean of 0.84 (CTRL 0.84, LOW 0.85): offsets within the painted width are forgiven, wider drift is not. Total fibre length, the most stable read-out, landed within 6% of the manual measurement on both images (+2.4% CTRL, -5.6% LOW). The same clDice on the 15 in-sample images against their cross-validation targets is 0.96 ± 0.02 . The $0.96 \rightarrow 0.84$ drop is therefore the generalisation number, scored consistently on one metric, and 6% on the length agreement. Within the limits of $n = 2$ re-annotated images, this is the strongest available evidence that the network has learned transferable features of the neurite signal rather than memorising the 15 training images.

5.10 End-to-end deployment on an unseen multi-condition batch

Late in the project, the AST group at Twente provided an additional, previously unavailable dataset: a batch of 8 chips spanning three culture conditions with a known applied dose of Nerve Growth Factor (NGF). NGF is a neurotrophic factor that promotes neurite outgrowth, so it changes the network in the opposite direction to a toxic insult. The batch therefore serves as a sensitivity check: it tests whether the read-outs track a controlled change in neurite density. It is not a toxicity assay. The batch let us test the method in two ways at once. The first is a generalisation check on acquisitions from a separate experimental batch. The second is a first look at whether the read-outs respond to a controlled biological change, a question the single-condition 15-image training set could not address.

The trained M5 ensemble was applied in full (preprocessing, inference, post-processing, metric extraction) with no manual annotation, parameter tuning, or retraining. The batch comprises three control chips, two low-NGF chips and three high-NGF chips, each imaged in a single 3×4 stitched compartment containing two wells. To limit the oxygen-exposure gradient across a compartment, the analysis keeps both wells of the middle compartments but only the inner well of the upper and lower compartments. Each retained well is counted as an independent read-out, giving control ($n = 4$ wells), low NGF ($n = 3$) and high NGF ($n = 4$). All chips were acquired on the same microscope at $10\times$ in a 3×4 stitched field. The two acquisitions taken with $2\times$ binning were resampled to the training-set pixel calibration ($0.6215 \mu\text{m}/\text{px}$) so that every well shares one scale, while z-stack depth still varies per chip (3–8 slices). Figure 19 summarises the layout.



Example acquisition. CTRL chip 2, middle-compartment imaging: two wells visible in a single stitched acquisition.

Figure 19. Experimental layout of the unseen NGF batch and a representative acquisition. **Left:** eight chips assayed across three culture conditions, control (3 chips, blue), low NGF (2 chips, purple) and high NGF (3 chips, orange). Each chip schematic shows the two wells visible in a single imaging compartment; green half-moons indicate the wells retained for per-well analysis (excluded the oxygen-exposed wells). The compartment imaged (upper/middle/lower) and chip-specific z-stack depth (3–8 slices) appear below each chip. The per-well analysis keeps both wells of a middle compartment and the inner well of an upper or lower compartment, mitigating the oxygen-exposure gradient. **Right:** the actual CTRL middle-chip acquisition, illustrating the two-well field that each chip schematic represents. Imaging modality and magnification were held identical to the training set; the two $2\times$ -binned acquisitions were resampled to the training pixel calibration, after which the imaged compartment and z-stack depth remain the sources of acquisition heterogeneity discussed in Section 6.2.

Across the retained wells, control records the highest absolute total fibre length ($200\,701 \pm 27\,192\,\mu\text{m}$). Once normalised to well area, control also has the highest fibre density ($50\,675 \pm 5\,253\,\mu\text{m}/\text{mm}^2$), with the two NGF groups falling below it and close to each other (low NGF $39\,278 \pm 16\,182$, high NGF $39\,110 \pm 8\,588\,\mu\text{m}/\text{mm}^2$). Density per mm^2 is the more interpretable cross-condition read-out, since absolute length also scales with how much network the imaged well contains. The branch-architecture metrics move in a consistent direction from control through low to high NGF: mean branch length increases ($13.5 \rightarrow 14.6 \rightarrow 14.9\,\mu\text{m}$) while branchiness falls ($0.040 \rightarrow 0.037 \rightarrow 0.035\,\text{j}/\mu\text{m}$). These two trends are the same signal seen from two sides, and both fit NGF shifting the network toward sparser, longer-running neurites. Mean branch thickness varies only modestly ($5.8\text{--}6.1\,\mu\text{m}$), with no consistent ordering. Table 10 and Figure 20 give the full profile.

Table 10. Group-level morphometric read-outs across NGF conditions for all seven morphometric metrics, plus fibre length per mm^2 as the primary density read-out. Values are mean $\pm$ SD over wells (population SD), each retained well counted once (Section 5.10). Total fibre length, junction count and branch count scale with the amount of network in the imaged well; the density-normalised read-outs (fibre / mm^2 , area covered, branchiness, mean branch length and thickness) are the more interpretable cross-condition comparisons.

Metric	ctrl ($n = 4$)	low NGF ($n = 3$)	high NGF ($n = 4$)
Total fibre (μm)	$200\,701 \pm 27\,192$	$149\,644 \pm 60\,154$	$161\,636 \pm 42\,120$
Fibre / mm^2 ($\mu\text{m}/\text{mm}^2$)	$50\,675 \pm 5\,253$	$39\,278 \pm 16\,182$	$39\,110 \pm 8\,588$
Junctions	$7\,996 \pm 1\,479$	$5\,824 \pm 3\,278$	$5\,826 \pm 2\,142$
Branches	$14\,975 \pm 2\,677$	$10\,951 \pm 5\,878$	$11\,142 \pm 3\,712$
Mean branch length (μm)	13.5 ± 1.0	14.6 ± 2.1	14.9 ± 1.6
Branchiness ($\text{j}/\mu\text{m}$)	0.040 ± 0.003	0.037 ± 0.007	0.035 ± 0.006
Area covered (%)	29.6 ± 3.8	23.2 ± 9.7	22.5 ± 5.1
Mean branch thickness (μm)	5.9 ± 0.2	6.1 ± 0.3	5.8 ± 0.4

No three-group difference reaches $\alpha = 0.05$ under a Kruskal–Wallis test (fibre/ mm^2 and area $p = 0.32$, total length $p = 0.41$, junctions $p = 0.52$, branchiness and mean branch length $p = 0.70$, thickness $p = 0.65$). At $n = 3$ –4 wells per group the test is heavily underpowered. A null result is therefore the expected outcome when group means differ by less than the within-group variability, and it is not evidence against a true effect. The within-group spread is itself uneven. Control is the tightest group (fibre density $\pm 5\,253 \mu\text{m}/\text{mm}^2$, 10% relative), while low NGF is the widest ($\pm 16\,182$, 41% relative). Low NGF spans a shallow 4-slice binned chip and a full-resolution 6-slice one, so part of that dispersion is still acquisition geometry rather than biology. Resampling the two binned acquisitions to the training calibration removed the earlier pixel-size error, but a confirmatory study with $n \geq 5$ chips per dose and matched acquisition would be needed to resolve the rest.

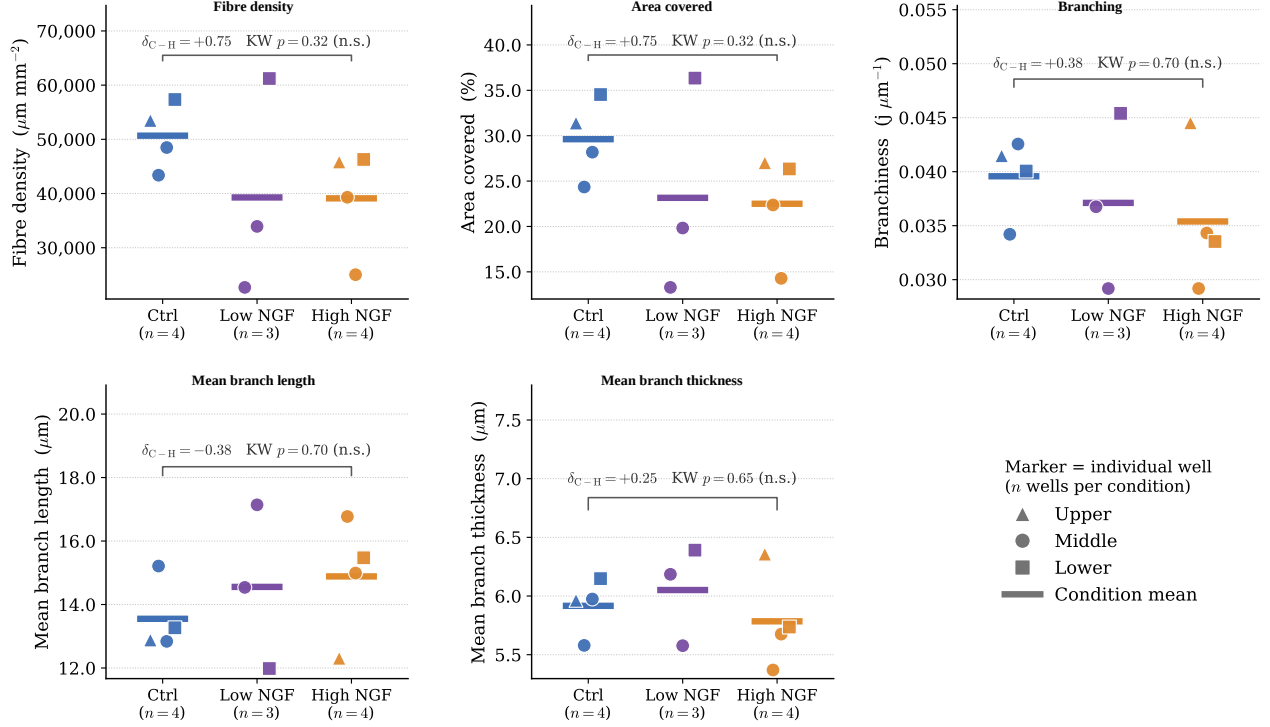


Figure 20. Size-independent (intensive) morphometric read-outs across NGF conditions, with each retained well plotted as an independent point (control $n = 4$, low NGF $n = 3$, high NGF $n = 4$). Marker shape encodes the imaged compartment (upper / middle / lower) and the thick bar is the per-condition mean over wells. Each panel reports Cliff's δ between control and high NGF (a non-parametric effect size in $[-1, 1]$; 0 means the groups fully overlap, +1 that every control well exceeds every high-NGF well) and the omnibus Kruskal-Wallis p ; no metric reaches significance. Fibre density and area covered carry the same signal (control densest), while mean branch length and branchiness shift monotonically toward longer, less-branched segments from control to high NGF.

6 Discussion and Conclusions

6.1 Summary of findings

This thesis set out to determine whether a U-Net can automatically segment the sympathetic neurites from fluorescence microscopy images, which of several plausible variants is the smallest that still matches the rest of the family on this task, and whether the resulting masks support a reliable morphometric read-out. The three goals are reached, with the qualifications listed in Section 6.2.

Sub-question 1 (U-Net versus baselines). The trained U-Net outperforms all three reference baselines. Deployed as a 5-fold ensemble with Hann-windowed tiled inference, the lightweight Attention U-Net (M5) reaches a mean full-image DSC of 0.887 on the 15 images, against 0.418 for Otsu, 0.563 for Frangi, and a median per-image margin of +0.157 over nnU-Net v2 [8] (Section 5.6). M5 ranks above every baseline on every image: a paired Wilcoxon

test gives $W = W_{\max} = 120$ ($p = 3.1 \times 10^{-5}$) and survives collapsing the image 009/image 011 duplicate (Section 6.2).

The nnU-Net comparison is informative because the ablation found the seven U-Net configurations within 1.2 DSC pp of one another, which points to the data, not the architecture, as the limit. The gap against nnU-Net therefore cannot be about architectural capacity. Its default 2D configuration runs outside the setting it was designed for. A calibration asymmetry reinforces this, as M5’s post-processing was tuned on this dataset while Otsu has no free parameter and Frangi inherited its parameters from the abandoned GT pipeline (Section 4).

The picture is consistent: at $n \approx 15$, the machinery around the U-Net matters more than the U-Net itself. A compact 8 M-parameter M5 trained on 14 unique cultures beats the far larger nnU-Net v2, in its default 2D configuration, on every in-sample image.

Sub-question 2 (which U-Net variant). The ablation across seven configurations, covering every combination of encoder width and attention gating plus the clDice loss and elastic augmentation, leaves all seven inside a 1.2-point DSC band. M3 attains the highest mean (0.718) but is comparable to both M5 (0.716) and M6 (0.717) at this sample size, all three within one fold-to-fold standard deviation; M5 is selected because it matches the heaviest variant at one-quarter of the parameter count. The clean within-grid contrasts (M2 vs M3 and M5 vs M6 for the attention effect; M2 vs M6 and M3 vs M5 for the width effect) each move performance by at most 0.002 DSC, which is well inside fold-to-fold noise. The architectural choices the literature would expect to matter, attention gates and a fourfold expansion of the encoder, are not contributing at this data scale. The route to a higher headline number is a larger and more varied training set, not a heavier model. The dataset size was fixed by constraints beyond the project’s control, but the available images were used to their fullest: extensive augmentation together with dense, 50%-overlapping patch extraction expands the 15 images into about 1 760 training patches per fold.

Sub-question 3a (morphological metrics). The seven read-outs fall into two groups by how much they move between model folds. Length-style metrics are stable. Total fibre length (CV = 8.5%), branchiness ratio (CV = 10.6%), and mean branch length (CV = 12.3%) vary less between folds than they do between images, so each can be reported as a single number. Count-style metrics are about twice as noisy. Junction count (CV = 19.1%) and branch count (CV = 20.8%) should be reported with a fold-to-fold range, not as a bare number, because one false-positive blob or one missed gap near a junction changes them by one or two. All seven are extracted without manual steps, so a whole batch runs as a single script (Section 5.10). The answer is therefore qualified: length-style read-outs are reliable as single numbers, while junction and branch counts should be read as a value with a range, not an exact count.

Generalisation to previously unseen acquisitions. Beyond the cross-validation results above, the method was tested against fresh manual SNT tracings on two previously-unseen confocal images from the deployed NGF batch, traced from scratch in a separate session and blind to the M5 prediction (Section 5.9). The predicted skeletons follow the manually traced ones closely: well-restricted centre-line agreement at a 2 px tolerance is clDice = 0.84

and total fibre length lands within 6% of the manual measurement on both images. Within the limits of $n = 2$ re-annotated images, this is the strongest available indication that the network has learned features of the neurite signal that transfer across acquisitions, rather than memorising the specific 15 training images. The new tracings share an annotator with the training set, so the result demonstrates image-distribution robustness but does not by itself rule out annotator-style bias.

Sub-question 3b (multi-condition deployment). Using the read-outs as an actual cardiotoxicity classifier would need drug-exposed cultures with matched controls, which are not provided here. As a step in that direction, the trained M5 ensemble was run in full on an independent batch of 8 chips across three NGF conditions (Section 5.10). It processed every chip without manual intervention and returned a complete metric profile for each condition. No group difference is statistically significant at $n = 3$ –4 wells per group, but a test this small returns non-significance whether or not a real effect exists. The tightest group, control, shows about a 10% relative SD on fibre density, which I read as an encouraging noise floor for a properly powered follow-up, though we have not run a formal power calculation. The practical takeaway: the method runs unattended on a multi-condition held-out batch, but it does not yet confirm a dose-response signal.

Descriptive ordering as a post-hoc hypothesis. The clearest trend is in the branch-architecture metrics. From control to low to high NGF, mean branch length rises and branchiness falls steadily, a shift toward fewer but longer branches, while control has the highest fibre density and both NGF groups sit below it. This is only a descriptive ordering: at $n = 3$ –4 wells per condition, no Kruskal–Wallis p falls below 0.32, and with groups this small a clean-looking order can come from noise alone. So it should be read as a hypothesis for a properly powered follow-up, not a confirmed dose-response. The read-outs most likely to hold up in deployment are the ones normalised to the imaging extent (fibre density, area covered, branchiness), since drug screens in hiPSC cardiac models report values relative to a matched control rather than as absolute counts [15], and the method reports all three automatically.

6.2 Limitations

A number of caveats apply.

Circular evaluation against the training target. Every metric reported here (DSC, skeleton length, junction count) is scored against the same ground truth the model trained on, and SNT tracings are themselves imperfect: they depend on where the annotator clicked and how faint distal branches and crossings were resolved. If the annotator consistently misses a structure, the model is rewarded for missing it too. So the DSC of 0.887 measures agreement with one annotator’s tracings, not absolute segmentation quality. This is a general problem in supervised medical image segmentation, and I did not have time to measure inter-annotator agreement.

Small dataset. Confocal imaging of sparse-foreground neurite networks through significant tissue depth makes good acquisitions hard to obtain in quantity, and the 15 annotated images are the most this slow, low-throughput imaging could realistically yield. Dense patch extraction (≈ 1760 patches per fold) and heavy training augmentation partially compensate, but there is no substitute for more annotated images. The same limit bounds the ablation. The seven configurations span only 1.2 percentage points of mean DSC (0.706–0.718). Fold-to-fold variation, most visibly the persistent fold 2 advantage, accounts for more variance than any architecture tested, and no pairwise claim of architectural superiority would survive a paired statistical test on these scores. Choosing M5 on Occam’s razor is therefore an engineering decision, not a claim that lightweight encoders generalise better.

nnU-Net comparison caveats. The benchmark against nnU-Net v2 (Section 5.6) is informative, but it runs outside the conditions nnU-Net is designed for. Its default schedule is 1000 epochs; the 100-epoch run here matches M5’s convergence point, not nnU-Net’s, so longer training could narrow the gap on harder images. Its auto-configured 2D plan uses 1280×1024 px patches at effective batch size 2, on the order of two dozen gradient updates per epoch on 12 training images, well outside its tuned regime. M5’s threshold and closing radius were also tuned on this dataset and set to the GT dilation radius ($r = 2$ px), so its binary output lines up with the scored GT’s geometry in a way nnU-Net cannot. Part of the gap is therefore calibration and geometry-matching, not capacity. I did not run a single-variable ablation (for example retraining nnU-Net with 256×256 px patches, or re-evaluating after a matched closing) to separate these factors, so the breakdown above is reasoned, not measured.

Hyperparameters tuned on the evaluation set. Four choices that affect the headline DSC were fixed on the same 15 images the number is reported on: the threshold ($t = 0.70$), closing radius ($r = 2$ px), spur-prune length ($L = 8$ px), and GT dilation radius ($r_{\text{GT}} = 2$ px; see *Dilated-mask supervision* below). The sensitivity curves are flat over a wide range: ± 0.05 in t and ± 1 px in r cost at most 0.008 DSC, and $L \in [4, 16]$ px shifts total fibre length by under 3.3%. The optimism is therefore small, but 0.887 is still a within-set summary, not a strictly held-out estimate.

Dilated-mask supervision and its measurement bias. The model is trained to predict the dilated mask, not the 1-pixel skeleton. A skeleton is under 1% of a foreground that is itself only 5–7% of the image, too sparse for stable optimisation, and a one-pixel offset would be penalised as heavily as a structural error. The $r = 2$ px dilation raised patch-level DSC from 0.581 to 0.718 (Section 4). The cost is that it changes what is measured: the dilated mask is the neurite plus a 2-pixel buffer, and true neurite width varies with focal depth, expression and sample thickness, so a fixed radius introduces an image-dependent error. The principled fix would set the radius from the microscope point-spread function in physical units, but that needs a bead-calibration acquisition I did not have. Since inference-time skeletonisation undoes most of the dilation, and cIDice (M4/M4b) was the topology-aware alternative I tested, I kept the empirical 2-pixel setting.

NGF deployment: power and acquisition heterogeneity. The NGF batch (Section 5.10) covers 8 chips across three conditions. Treating the oxygen-mitigated wells as independent read-outs gives $n = 3\text{--}4$ per group, with within-group SDs of the same order as the between-group differences and no Kruskal–Wallis test reaching $\alpha = 0.05$. The descriptive ordering therefore shows self-consistency, not a confirmed biological response. Two caveats temper even that. First, the two wells kept from a middle-compartment chip share that chip’s dose, culture and acquisition, so they are sub-samples, not fully independent replicates. This is the same independence concern as the duplicate training z-slices in Section 5: the effective sample size is optimistic, and a stricter chip-level analysis leaves nothing significant either. Second, total fibre length still scales with how much network each well contains, which is why the comparison leans on extent-normalised read-outs (fibre density, area covered, branchiness, mean branch thickness).

No biological calibration of the read-outs. The read-outs have not been calibrated against an independent biological measurement. There is no published reference value for total fibre length or junction count in this model, and the thesis includes no healthy-versus-damaged comparison. The assumption that network fragmentation reflects neurotoxic damage rests on primary sympathetic-neuron cultures and the broader hiPSC-cardiomyocyte cardiotoxicity literature [4]; the link from these specific metrics to a cardiotoxicity verdict in this model remains an untested hypothesis.

6.3 Further research

Dataset expansion and a properly powered dose-response study. The single most impactful next step is more data. More annotated images give the segmenter more to train on, which improves generalization and shrinks its fold-to-fold variability. Extending the eight-chip NGF batch (Section 5.10) to $n \geq 5$ chips per dose would raise the statistical power of the test, turning the present descriptive ordering into a significant, confirmable signal.

Inter-annotator reliability. Re-annotating a small subset of the images with a second annotator would give an inter-annotator agreement score, the first independent estimate of the highest DSC attainable on this dataset. Without it, any further architectural gain is hard to judge: we cannot tell how close the current DSC already sits to that ceiling.

Architecture and 3D information. With a larger dataset, architectural choices that show no measurable benefit now, residual connections [7], deeper encoders, and transformer-based backbones, would have more room to help. The current method also discards depth by collapsing each z-stack to a maximum-intensity projection; a 3D or 2.5D U-Net could use that depth to reduce the MIP artefacts that overlapping neurites at different focal planes create. A different projection could help too: Extended Depth of Focus (EDF) in place of MIP may recover faint neurites better in preprocessing.

From morphometric read-outs to a screening pipeline. Validating the metrics against compounds with known dose-response relationships, and pairing each prediction with an explanation a screener can inspect, would close the loop from image to screening decision and directly answer sub-question 3.

References

- [1] David Cook, Dearg Brown, Robert Alexander, Ruth March, Paul Morgan, Gemma Satterthwaite, and Menelas N. Pangalos. Lessons learned from the fate of AstraZeneca’s drug pipeline: a five-dimensional framework. *Nature Reviews Drug Discovery*, 13(6):419–431, 2014.
- [2] Péter Ferdinandy, István Baczkó, Péter Bencsik, Zoltán Giricz, Anikó Görbe, Pál Pacher, Zoltán V. Varga, András Varró, and Rainer Schulz. Definition of hidden drug cardiotoxicity: paradigm change in cardiac safety testing and its clinical implications. *European Heart Journal*, 40(22):1771–1777, 2019.
- [3] Alejandro F. Frangi, Wiro J. Niessen, Koen L. Vincken, and Max A. Viergever. Multi-scale vessel enhancement filtering. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 130–137. Springer, 1998.
- [4] Fiona Grafton, Jaclyn Ho, Sara Ranjbarvaziri, Mohsen Anwar, Pasqualina Lange, Lin Pang, Mohan Krishnan, Tien-Lin Lee, Sui Q. Lee, Yu-Cheng Wei, Liang Kang, Vaidehi Mishra, Yu Liu, Atishay Khire, Daniel J. Kowalewski, Hugh Tang, Mohit Sangha, Maria Pang, Robert Schwartz, et al. Deep learning detects cardiotoxicity in a high-content screen with induced pluripotent stem cell-derived cardiomyocytes. *eLife*, 10:e68714, 2021.
- [5] Saumya Gupta, Yikai Zhang, Xiaoling Hu, Prateek Prasanna, and Chao Chen. Topology-aware uncertainty for image segmentation. In *Advances in Neural Information Processing Systems 36*, NeurIPS 2023, page 8186–8207. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023.
- [6] Fredric J. Harris. On the use of windows for harmonic analysis with the discrete Fourier transform. *Proceedings of the IEEE*, 66(1):51–83, 1978.
- [7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.
- [8] Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021.
- [9] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. Preprint at arXiv:1412.6980 (2014).

- [10] Krita Foundation. Krita: Free and open-source digital painting application. <https://krita.org>, 2024. Accessed 2025-05-24.
- [11] Leiden University. ALICE high performance computing facility. <https://wiki.alice.universiteitleiden.nl>, 2026. Accessed 2026-02-27.
- [12] Yonatan R. Lewis-Israeli, Aaron H. Wasserman, Mitchell A. Gabalski, Brett D. Volmert, Yixuan Ming, Kristen A. Ball, Weiyang Yang, Jinyun Zou, Guangming Ni, Natalia Pajares, et al. Self-assembling human heart organoids for the modeling of cardiac development and congenital heart disease. *Nature Communications*, 12(1):5142, 2021.
- [13] Mark H. Longair, Dean A. Baker, and J. Douglas Armstrong. Simple Neurite Tracer: open source software for reconstruction, visualization and analysis of neuronal processes. *Bioinformatics*, 27(17):2453–2454, 2011.
- [14] Erik Meijering, Matthieu Jacob, Jean-Christophe F. Sarria, Pascal Steiner, Harald Hirling, and Michael Unser. Design and validation of a tool for neurite tracing and analysis in fluorescence microscopy images. *Cytometry Part A*, 58A(2):167–176, 2004.
- [15] Richard J. Mills, Benjamin L. Parker, Gregory A. Quaife-Ryan, Holly K. Voges, Elise J. Needham, Antonio Bornot, Mei Ding, Henrik Andersson, Magnus Polla, David A. Elliott, et al. Drug screening in human PSC-cardiac organoids identifies pro-proliferative compounds acting via the mevalonate pathway. *Cell Stem Cell*, 24(6):895–907, 2019.
- [16] Agata Mosinska, Pablo Marquez-Neila, Mateusz Kozinski, and Pascal Fua. Beyond the pixel-wise loss for topology-aware delineation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3136–3145, 2018.
- [17] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y. Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention U-Net: Learning where to look for the pancreas. In *Medical Imaging with Deep Learning (MIDL)*, 2018.
- [18] Igbo J. Onakpoya, Carl J. Heneghan, and Jeffrey K. Aronson. Post-marketing withdrawal of 462 medicinal products because of adverse drug reactions: a systematic review of the world literature. *BMC Medicine*, 14:10, 2016.
- [19] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979.
- [20] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pages 8024–8035, 2019.
- [21] Nicolas Pielawski and Carolina Wählby. Introducing Hann windows for reducing edge-effects in patch-based image segmentation. *PLOS ONE*, 15(3):e0229839, 2020.

- [22] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241. Springer, 2015.
- [23] Johannes Schindelin, Ignacio Arganda-Carreras, Erwin Frise, Verena Kaynig, Mark Longair, Tobias Pietzsch, Stephan Preibisch, Curtis Rueden, Stephan Saalfeld, Benjamin Schmid, Jean-Yves Tinevez, Daniel James White, Volker Hartenstein, Kevin Eliceiri, Pavel Tomancak, and Albert Cardona. Fiji: an open-source platform for biological-image analysis. *Nature Methods*, 9(7):676–682, 2012.
- [24] Wei Shen, Xiang Bai, Rong Hu, Hongyuan Wang, and Longin Jan Latecki. Skeleton growing and pruning with bending potential ratio. *Pattern Recognition*, 44(2):196–209, 2011.
- [25] Suprosanna Shit, Johannes C. Paetzold, Anjany Sekuboyina, Ivan Ezhov, Alexander Unger, Andrey Zhylka, Josien P. W. Pluim, Ulrich Bauer, and Bjoern H. Menze. clDice: a novel topology-preserving loss function for tubular structure segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16560–16569, 2021.
- [26] Nico Stucki, Johannes C. Paetzold, Suprosanna Shit, Bjoern H. Menze, and Ulrich Bauer. Topologically faithful image segmentation via induced matching of persistence barcodes. In *International Conference on Machine Learning (ICML)*, pages 32698–32727, 2023.
- [27] Stéfan van der Walt, Johannes L. Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D. Warner, Neil Yager, Emmanuelle Gouillart, Tony Yu, and the scikit-image contributors. scikit-image: image processing in Python. *PeerJ*, 2:e453, 2014.
- [28] Karel Zuiderveld. Contrast limited adaptive histogram equalization. In Paul S. Heckbert, editor, *Graphics Gems IV*, pages 474–485. Academic Press, 1994.