NeuroImage 249 (2022) 118906

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Towards a representative reference for MRI-based human axon radius assessment using light microscopy

Laurin Mordhorst[a,*], Maria Morozova[b,c], Sebastian Papazoglou[a], Björn Fricke[a], Jan Malte Oeschger[a], Thibault Tabarin[a], Henriette Rusch[c], Carsten Jäger[b], Stefan Geyer[b], Nikolaus Weiskopf[b,d], Markus Morawski[b,c], Siawoosh Mohammadi[a,b,*]

[a] Institute of Systems Neuroscience, University Medical Center Hamburg-Eppendorf, Hamburg, Germany
[b] Department of Neurophysics, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany
[c] Paul Flechsig Institute of Brain Research, Medical Faculty, Leipzig University, Leipzig, Germany
[d] Felix Bloch Institute for Solid State Physics, Leipzig University, Leipzig, Germany

ARTICLE INFO

Keywords:
Deep learning
MRI-based axon radius
Cross microscopy
Neuroanatomy
Axon radii distribution

ABSTRACT

Non-invasive assessment of axon radii via MRI bears great potential for clinical and neuroscience research as it is a main determinant of the neuronal conduction velocity. However, there is a lack of representative histological reference data at the scale of the cross-section of MRI voxels for validating the MRI-visible, effective radius (r_{eff}). Because the current gold standard stems from neuroanatomical studies designed to estimate the bulk-determined arithmetic mean radius (r_{arith}) on small ensembles of axons, it is unsuited to estimate the tail-weighted r_{eff}. We propose CNN-based segmentation on high-resolution, large-scale light microscopy (lsLM) data to generate a representative reference for r_{eff}. In a human corpus callosum, we assessed estimation accuracy and bias of r_{arith} and r_{eff}. Furthermore, we investigated whether mapping anatomy-related variation of r_{arith} and r_{eff} is confounded by low-frequency variation of the image intensity, e.g., due to staining heterogeneity. Finally, we analyzed the error due to outstandingly large axons in r_{eff}. Compared to r_{arith}, r_{eff} was estimated with higher accuracy (maximum normalized-root-mean-square-error of r_{eff}: 8.5 %; r_{arith}: 19.5 %) and lower bias (maximum absolute normalized-mean-bias-error of r_{eff}: 4.8 %; r_{arith}: 13.4 %). While r_{arith} was confounded by variation of the image intensity, variation of r_{eff} seemed anatomy-related. The largest axons contributed between 0.8 % and 2.9 % to r_{eff}. In conclusion, the proposed method is a step towards representatively estimating r_{eff} at MRI voxel resolution. Further investigations are required to assess generalization to other brains and brain areas with different axon radii distributions.

1. Introduction

The MRI signal generated by an ensemble of protons probing the local, microscopic environment in human brain tissue can contain information about microstructural tissue features such as the axonal radius (Alexander et al., 2010; Andersson et al., 2020; Assaf et al., 2008; Veraart et al., 2020). The axonal radius is a key to determine neuronal communication in the human brain because it is related to, e.g., the neuronal conduction velocity (Drakesmith et al., 2019; Schmidt and Knösche, 2019; Waxman, 1980). The estimation of the axonal radius and other microstructural features via biophysical modeling of the MRI signal (Alexander et al., 2019) is an active area of research because of its potential to partially replace or complement invasive ex-vivo histology with non-invasive, in-vivo, quantitative MRI approaches (Stikov et al., 2015; Weiskopf et al., 2021). However, before these models can be used, they need to be validated against a robust histological reference (Weiskopf et al., 2021).

The validation for the MRI-visible, effective radius (r_{eff}) is currently lacking a robust, histological reference for human brain tissue. Since r_{eff} is indicative of large, sparsely occurring axons, i.e., the tail of the axon radii distribution (Burcaw et al., 2015; Sepehrband et al., 2016; Veraart et al., 2020), large ensembles of axons need to be evaluated to representatively estimate r_{eff} for MRI voxels of a human MRI system (1 mm^3 or larger). Previous studies have compared estimates of the effective radius from diffusion-weighted MRI (dMRI) against small ensembles of axons in histological reference data of rats (Kakkar et al., 2018; Xu et al., 2014). However, representative histological validation of r_{eff}, i.e., using the same cross-sectional scale in histology and MRI, has only been attempted on perfusion-fixed rats for ultra-high-resolution MRI voxels ($\sim 100\ \mu m^3$) of a preclinical MRI system (Veraart et al., 2020). It is unclear whether the validation of MRI-based models on rats can be trans-

* Corresponding author.
E-mail address: laurin.mordhorst@gmx.de (L. Mordhorst).

https://doi.org/10.1016/j.neuroimage.2022.118906.
Received 1 June 2021; Received in revised form 6 January 2022; Accepted 11 January 2022
Available online 13 January 2022.

List of Symbols and Acronyms

Axon radii

r	individual axon radius

Axon radii ranges

$small$	axons with $r < 0.3\,\mu m$
$medium-sized$	axons with $0.3\,\mu m \leq r < 1.6\,\mu m$
$large$	axons with $r \geq 1.6\,\mu m$
$bulk$	$small$ and $medium-sized$ axons
$tail$	$large$ axons

Arithmetic mean axon radius

r_{arith}	arithmetic mean radius
ρ_{arith}	reference arithmetic mean radius
$\hat{r}_{arith}$	estimated arithmetic mean radius

MRI-visible, effective axon radius

r_{eff}	effective radius
ρ_{eff}	reference effective radius based on $\mathcal{E}_{eff}$
$\rho_{eff\downarrow}$	reference effective radius based on $\mathcal{E}_{eff\downarrow}$
$\rho_{eff\uparrow}$	reference effective radius based on $\mathcal{E}_{eff\uparrow}$
$\hat{r}_{eff}$	estimated effective radius
$\hat{r}_{eff,rng}$	estimated effective radius based on an axon radii distribution with erroneous radii in axon radii range $rng \in \{\text{small}, \text{medium}, \text{large}\}$

Axon radii distribution

$\mathcal{E}_{eff}$	axon radii distribution of ρ_{eff}
$\mathcal{E}_{eff\downarrow}$	axon radii distribution of $\rho_{eff\downarrow}$ with underrepresented $bulk$ (scaled according to $f_{\downarrow}$)
$\mathcal{E}_{eff\uparrow}$	axon radii distribution of $\rho_{eff\uparrow}$ with overrepresented $bulk$ (scaled according to $f_{\uparrow}$)

Bulk scaling factor

f	$bulk$ scaling factor
$f_{\downarrow}$	lower bound for $bulk$ scaling factor
$f_{\uparrow}$	upper bound for $bulk$ scaling factor
$f_{interp}(s)$	interpolated $bulk$ scaling factor
s	sweep variable

Subsections

S_{EM}	small-field-of-view EM subsection
S_{LM}	small-field-of-view lsLM subsection
S_{lsLM}	large-field-of-view lsLM subsection

Acronyms

CNN	convolutional neural network
COD	cause of death
CV	cross-validation
dMRI	diffusion-weighted MRI
EM	electron microscopy
lsLM	high-resoution, large-scale light microscopy
NMBE	normalized-mean-bias-error
NRMSE	normalized-root-mean-square-error
NRSD	normalized-residual-standard-deviation
PMD	post-mortem delay

lated to the human brain. As the tail of the axon radii distribution may vary between humans and other mammals (Biedenbach et al., 1986; Leenen et al., 1982), r_{eff} for humans may be shifted with respect to other species. This shift may be further reinforced by the reduced capability to resolve small axons in human MRI systems when compared to preclinical MRI systems (Drobnjak et al., 2016; Nilsson et al., 2017; Veraart et al., 2020). For human brain, the current gold standard for the validation of r_{eff} (Alexander et al., 2010; Horowitz et al., 2015; Innocenti et al., 2015; Veraart et al., 2020) stems from neuroanatomical studies (Aboitiz et al., 1992; Caminiti et al., 2009; Graf von Keyserlingk and Schramm, 1984; Liewald et al., 2014) of small ensembles of axons (100-1000 axons), aiming to evaluate the arithmetic mean radius (r_{arith}) on manually annotated electron microscopy images (EM). As r_{arith} is determined by the bulk of the axon radii distribution, it can be expected that estimates of r_{arith} are less sensitive to the ensemble size as compared to r_{eff}. For r_{eff}, however, small-ensemble estimates can strongly under- or overestimate r_{eff} (Mordhorst et al., 2021) of typical MRI voxels, because the tail of the axon radii distribution is insufficiently sampled.

Albeit high-resolution, large-scale light microscopy (lsLM) cannot resolve small axons as accurately as EM, an lsLM-based approach might be appropriate to generate a histological gold standard for the validation of MRI-based radius estimation in human brain tissue. Because of the large field-of-view of lsLM, covering cross-sections of $1\,mm^2$ or larger, it is possible to capture large ensembles of axons including 10^5 to 10^6 axons per section and thus sample the tail of the axon radii distribution more accurately. Moreover, lsLM has the advantage of being fast, cheap and simple to perform compared to EM. As the assessment of axon radii on large field-of-view microscopy data renders manual annotation infeasible, automated approaches, e.g., methods based on convolutional neural networks (CNN), are required. So far, CNN-based methods based on large two- or three dimensional scanning or transmission electron microscopy (SEM/TEM) sections have been trained on images of perfusion-fixed mice or rats (Abdollahzadeh et al., 2021; Zaimi et al., 2018). However, it is unlikely that the models generated in these studies translate well to immersion-fixed human brain tissue with higher tissue degradation.

In this study, we investigate the potential of lsLM and CNN-based segmentation to map the distribution of axon radii in a human corpus callosum specimen. We quantify the capability of the proposed method to estimate the MRI-visible r_{eff} and r_{arith}, which is commonly reported in neuroanatomical studies, by evaluating the estimation errors on six lsLM sections. While reference data for the frequency-weighted r_{arith} can be generated through manual annotation with reasonable effort, the tail-weighting of r_{eff} introduces the necessity to accurately capture the tail of the axon radii distribution and thus investigate larger ensembles of axons than can be realistically annotated. To address this challenge, we merge manually annotated radii from different sources into composite axon radii distributions, combining the accurate resolution of the bulk of axon radii in EM with representative sampling of the tail of the axon radii distribution on large-field-of-view lsLM subsections. Additionally, we investigate whether our method is capable of capturing anatomy-related, spatial variation of r_{arith} and r_{eff} in the presence of low-frequency image intensity variation, e.g., due to staining heterogeneity. Finally, we analyze the potential error due to individual, outstandingly large axons in r_{eff}.

2. Materials and methods

2.1. Ensemble mean axon radii

For a discrete axon radii distribution of B individual radii with $n_{(k)}$ axons with radius $r_{(k)}$ in bin k, the arithmetic mean radius can be defined as

$$r_{arith} = \sum_{k=1}^{K} w_{arith,(k)} \cdot r_{(k)}$$
$$w_{arith,(k)} = \frac{n_{(k)}}{B} \quad (1)$$

The MRI-visible, effective mean radius (r_{eff}) (Burcaw et al., 2015; Sepehrband et al., 2016; Veraart et al., 2020) can be estimated from the intra-axonal signal of dMRI. Clinical acquisition employs pulse-gradient spin echo dMRI sequences with wide pulses, i.e. using pulse widths

$\geq 10\,\mathrm{ms}$ (Burcaw et al., 2015). In the wide-pulse limit,

$$r_{\mathrm{eff}} = \sqrt[4]{\sum_{k=1}^{K} w_{\mathrm{eff},(k)} \cdot r_{(k)}} \text{with}$$

$$w_{\mathrm{eff},(k)} = \frac{n_{(k)}}{B} \cdot \frac{r_{(k)}^5}{\frac{1}{B}\sum_{j=1}^{K} n_{(j)} r_{(j)}^2}. \quad (2)$$

While r_{arith} is frequency-weighted ($w_{\mathrm{arith},(k)}$) and therefore determined by the bulk of the axon radii distribution, r_{eff} is weighted ($w_{\mathrm{eff},(k)}$) towards the *tail* of the axon radii distribution because $w_{\mathrm{eff},(k)}$ scales with the fifth power of $r_{(k)}$. Each radius $r_{(k)}$ denotes the radius of a circular approximation of the axonal body of a myelinated axon with equivalent area (West et al., 2016) (hereafter denoted as circular equivalent).

2.2. Axon radii ranges

Throughout this manuscript, we generated composite axon radii distributions by combining axon radii distributions from different sources at particular thresholds. As a consequence, we partitioned the axon radii distribution into three parts:

- *Large* axons ($r \geq 1.6\,\mu\mathrm{m}$) represent the *tail* of the axon radii distribution and therefore have a strong contribution towards the *tail*-weighted r_{eff}. The threshold was chosen so that the estimated r_{eff} was decreased by 50 % when axons above this threshold were removed from the pooled axon radii ensemble of the corpus callosum lsLM sections evaluated with a prototype of the proposed method.
- *Small* axons ($r < 0.3\,\mu\mathrm{m}$) are below the resolution limit of lsLM.
- *Medium-sized* axons ($0.3\,\mu\mathrm{m} \leq r < 1.6\,\mu\mathrm{m}$) constitute the *bulk* of the axon radii distribution together with *small* axons.

2.3. Data acquisition

Tissue preparation Four human white matter samples of four different subjects were used in this study: a corpus callosum (CC, male, 74 years, postmortem delay (PMD): 24 hours, cause of death (COD): multi organ failure), a corticospinal tract (CST, female, 89 years, PMD: 24 hours, COD: heart failure), an optic chiasm (OC, male, 59 years, PMD: 48 hours, COD: multi organ failure) and a sample obtained from the area dorsolateral of the olivary nucleus including the anterolateral system (AS, male, 81 years, PMD: 24 hours, COD: multi organ failure). Following standard procedures, blocks were immersion-fixed in 3 % paraformaldehyde and 1 % glutaraldehyde in phosphate-buffered saline at pH 7.4. Then, smaller blocks of 1 to 4 mm edge length were cut, contrasted with osmium tetroxide and uranyl acetate, dehydrated in graded acetones, embedded in Durcupan resin and cut into semi- ($\sim 500\,\mathrm{nm}$) and ultra-thin ($\sim 50\,\mathrm{nm}$) sections. Semi-thin sections were stained with 1 % toluidine blue for imaging with lsLM.

Microscopy In total, 13 lsLM images were acquired of semi-thin sections using a Zeiss AxioScan.Z1 (objective: 40×, numerical aperture: 0.95, resolution: $0.1112\,\mu\mathrm{m}$/pixel; resolution limit: 292 nm) (see Table 1). For $N = 6$ of the lsLM sections of the CC sample, *matching* EM sections were acquired, i.e., sections were cut within $100\,\mu\mathrm{m}$ proximity (see Fig. 1). For the latter EM sections, images were acquired using a Zeiss LEO EM 912 Omega TEM at 80 kV and digital micrographs were obtained with a dual-speed 2K-on-axis CCD camera-based YAG scintillator (TRS-Tröndle, resolution: $0.0043\,\mu\mathrm{m}$/pixel, resolution limit: 4 nm).

2.4. Axon radius estimation pipeline

Axon radius estimation was divided into three steps: semantic segmentation, instance segmentation and radius approximation (see Fig. 2). To perform semantic segmentation, i.e., to classify each pixel as either axon, myelin or background, we applied a CNN (see Section 2.5) in a sliding window manner (see Fig. 2a). To identify axon instances from individual pixels, we applied connected-component labeling (see Fig. 2b).

Table 1
The dataset of human tissue samples. The following tissue *samples* were investigated: a corpus callosum (CC), a corticospinal tract (CST) an optic chiasm (OC) and an anterolateral system (AS). *Sections* were assigned exclusively to the *training* or *test* dataset.

			Sections		
Sample	Type	Size [mm^2]	Total	Training	Test
OC	lsLM	2.49	1	1	
CST	lsLM	12.71	1	1	
AS	lsLM	4.86	1	1	
CC	lsLM	0.34 to 9.16	10	4	6
CC	EM	0.01	6		6

For each axon instance, the circular equivalent radius was approximated (see Fig. 2c).

2.5. Semantic segmentation

Training data annotation For training of the CNN, we manually annotated 64 lsLM subsections of similar size ($70 \times 70\,\mu\mathrm{m}^2$ to $120 \times 120\,\mu\mathrm{m}^2$) originating from different sections of the four tissue samples: 46 CC subsections, 4 OC subsections, 4 CST subsections and 10 AS subsections. To avoid fitting to the test data, whole lsLM sections were used exclusively for training or testing (see Table 1). To cover a wide range of appearance in axon shape and image contrast, some subsections were only partially annotated, i.e., pixels were assigned an ignore label and were not considered during training. As large axons were expected to have particular relevance for r_{eff}, but occur with low frequency, we assigned higher priority to the annotation of these axons.

The manual annotation of individual axons followed the approach described in Zaimi et al., 2018: first, the myelin sheath was annotated, then the enclosed axonal body was filled. Remaining pixels were assigned a background label. At a later stage, we generated initial segmentations of the myelin sheaths using an early prototype of the CNN. Here, the procedure for the segmentation of myelin sheaths changed as follows: initial segmentations of myelin sheaths were refined, myelin sheaths of missed fibers were annotated and myelin sheaths of falsely segmented fibers were removed. Manual annotations were carried out using GIMP (The GIMP Development Team) or ITK-SNAP (Yushkevich et al., 2006).

The manual annotation was performed by a total of six raters (M. Morozova, B. Fricke, J.M. Oeschger, S. Papazoglou, T. Tabarin and L. Mordhorst). Each manually annotated subsection was crosschecked by a second rater. Initially, manual annotations were carried out in collaboration with two experts (i.e., M. Morawski and M. Morozova) who were furthermore consulted in case of doubt.

Network Architecture We used a CNN of the U-Net (Ronneberger et al., 2015); Yakubovskiy (2020)family (see Fig. 3a), i.e., we followed its general architecture of consecutive encoding and decoding paths with skip connections between shallow and deep layers processing features of the same spatial resolution. In U-Nets, the resolution is reduced after each encoder block while the number of channels is increased; this process is reversed along the decoding path. For the encoding path, we employed transfer learning, i.e., we used EfficientNet-B3 (Tan and Le, 2019) encoders pretrained on the ImageNet dataset (Deng et al., 2009). In the decoding path, we used two sequences of 3×3 convolutions with batch normalization (BN) and rectified linear activation units (Relu) (see Fig. 3b). The aforementioned sequences were framed with concurrent spatial and channel squeeze and excitation (scSE) (Roy et al., 2018) modules. While the encoding path decreased the spatial resolution by using one convolution with stride two in each encoder block, the decoding path increased spatial resolution using nearest neighbor interpolation as an initial step of each decoder block. Using skip connections between the encoding and

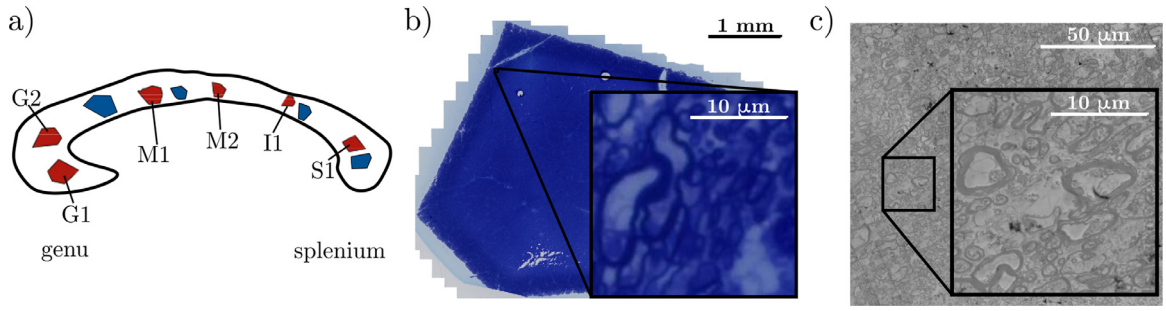

Fig. 1. The human corpus callosum sample. The schematic of the sample (a) highlights the regions used for training (blue) and testing (red). For each region, one large-scale light microscopy (lsLM) section was acquired. For the $N = 6$ test regions (red), *matching* lsLM and electron microscopy (EM) subsections were acquired: two sections from genu (G1, G2), two sections from midbody (M1, M2) and one section each from isthmus (I1) and splenium (S1). For section G1, the lsLM (b) and its *matching* EM section (c) are depicted as well as examples of subsections that were magnified to cover the same spatial extent ($20 \times 20\ \mu m^2$) at common resolution.

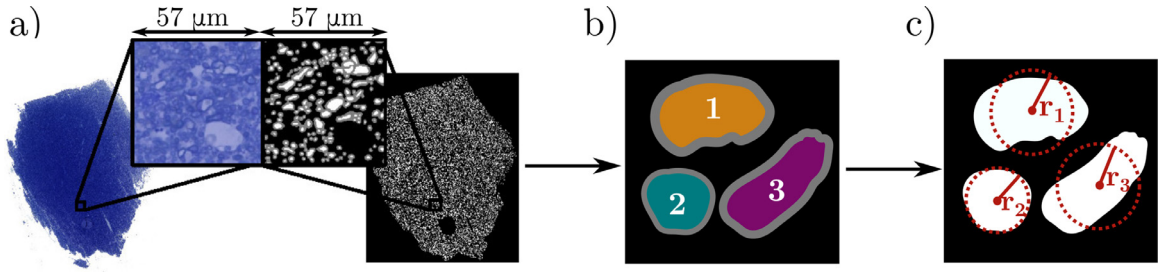

Fig. 2. The axon radius estimation pipeline. (a) Pixel-wise classifications as axon, myelin or background were obtained through application of a semantic segmentation network, i.e. a U-Net (Ronneberger et al., 2015) variant, in a sliding window manner. (b) Axon instances were identified through connected-component labeling. (c) Radii of axon instances were estimated as radii of circles with equivalent area (short: circular equivalent).

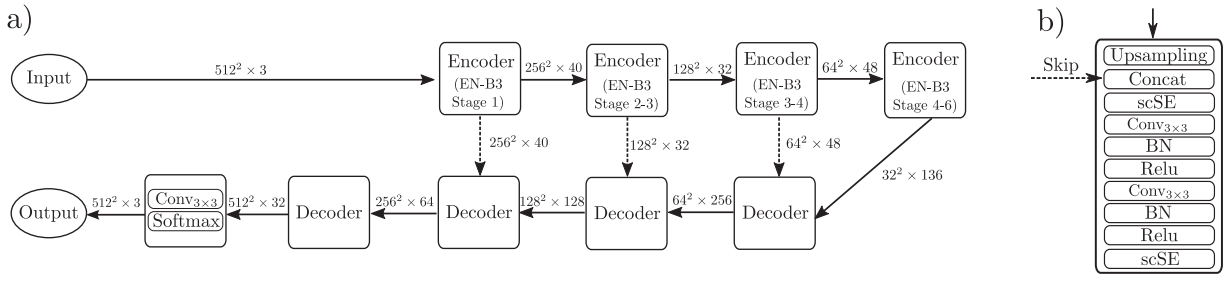

Fig. 3. The semantic segmentation network architecture. (a) Overview of the architecture: we used a variant of the U-Net (Ronneberger et al., 2015) architecture, following the approach of an encoding (top row) and decoding path (bottom row) with skip connections (dashed arrows) between encoding and decoding path. The encoding path consists of the first six stages of a pretrained EfficientNet-B3 (EN-B3) (Tan and Le, 2019) model. The decoding path used the fundamental decoder blocks illustrated in (b): each decoder was composed of upsampling by nearest neighbor interpolation, concatenation of the encoded features at same resolution, and two sequences of 3×3 convolutions ($Conv_{3\times3}$) with batch normalization (BN) and rectified linear activation units (Relu) framed by squeeze and excitation (scSE) (Roy et al., 2018) modules. The skip connections (dashed arrows) connected the intermediate features of the encoding path (after Efficient-Net B3 stages one, three, and four) with corresponding decoder outputs at the same resolution. The final outputs were obtained by applying $Conv_{3\times3}$ with a softmax activation to the output of the last decoder, yielding pixel-wise pseudo-probabilities for axon, myelin and background. Annotated numbers denote spatial resolution and the number of channels, e.g., $512^2 \times 3$ denotes a tensor with 512×512 pixels and 3 channels, which corresponds to the input and output size used during training.

decoding path, we concatenated the outputs of encoder blocks, i.e., the output of the pretrained EfficientNet-B3 after stages one, three, and four with decoder blocks processing features of the same resolution (after applying interpolation). Outputs were obtained using 3×3 convolution with softmax activation, yielding pixel-wise pseudo-probabilities for axon, myelin and background. In total, the network had ~ 3.8 million trainable parameters.

Input preprocessing Inputs were standardized per color channel with respect to the training dataset, i.e., we computed channel-wise mean and standard deviation across all pixels of the training dataset; then, for each input during training, we subtracted the channel-wise mean and divided by the channel-wise standard deviation.

Input augmentation The following augmentation steps were employed on-the-fly during training using (Jung et al., 2021): multiplication of hue, saturation and value in the hue-saturation-value color space by a randomly chosen factor; contrast adjustment; blurring with a Gaussian kernel; horizontal and vertical flipping with probability $p = 0.5$; affine transformation using rotation (by angles in $[-45°, 45°]$), scaling (by factors in $[0.8, 1.2]$) and shearing (by angles in $[-25°, 25°]$); staining augmentation (Macenko et al., 2009; Byfield).

Training We trained the model for 200 epochs, using pseudo-epochs of 150 randomly drawn training patches of 512×512 pixels. We used mini-batch gradient descent with a mini-batch size of 4, Nesterov momentum (0.95), an initial learning rate of 10^{-2} and a learning rate decay of $\gamma = 0.2$ every 50 epochs after initial 100 epochs to minimize a Lovàsz-softmax loss (Berman et al., 2018). All weights of the CNN were modified during training. The training phase took about 45 minutes on an NVIDIA Quadro RTX 6000 GPU. We used a framework (Falcon et al.,

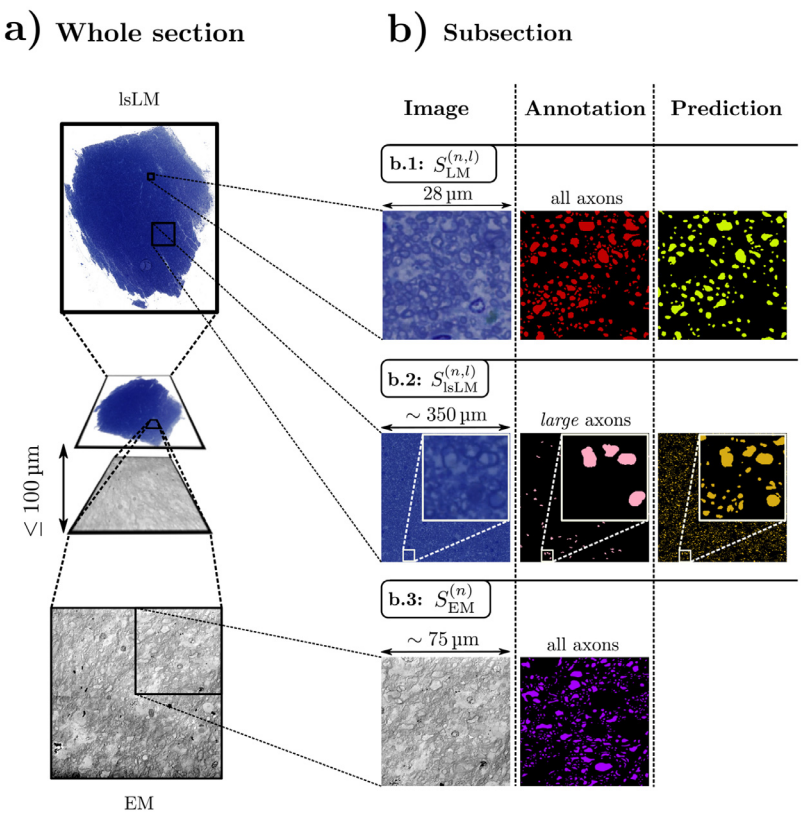

Fig. 4. Overview of test data. (a) *Matching* large-scale light microscopy (lsLM) and electron microscopy (EM) sections of the n-th test region, i.e., sections were cut within 100 µm distance. (b.1-3) lsLM and EM subsections originating from (a). Illustrations from left to right column show: microscopy subsection, axon mask from manual annotation, axon mask predicted by the proposed pipeline. (b.1) Small field-of-view lsLM subsection with all axons manually annotated (see Section 2.6.(a)). (b.2) Large field-of-view lsLM subsection with only *large* axons manually annotated (see Section 2.6.(b)). (b.3) Small field-of-view EM subsection with all axons manually annotated (see Section 2.6.(c)).

2019) based on PyTorch (Paszke et al., 2019) to carry out the training procedure.

Hyperparameter optimization To determine the above used initial learning rate, γ and the number of epochs, we carried out a grid search for the initial learning rate and γ using optuna (Akiba et al., 2019) in a 4-fold-cross-validation (CV) approach. CV splits were conducted at the level of entire lsLM training subsections. We considered the averaged dice score for axon and myelin as the target metric, which we evaluated every 10 epochs on entire subsections of the validation set of the particular CV fold. Each model was trained at least 150 epochs. To avoid overfitting, we stopped when the target metrics did not increase for three consecutive validation steps, i.e., 30 epochs. We then chose hyperparameters, i.e., learning rate, γ and the number of epochs, so that they optimized the mean of the above target metric across all CV folds.

2.6. Test dataset

To generate reference data for the evaluation experiments detailed in the following sections, we manually annotated multiple lsLM and one EM subsection for each test region $n \in \{1, \dots, N\}$ (see Fig. 4):

(a) To assess the axon segmentation performance of the semantic segmentation model (see Section 2.5), we manually annotated all axons on five lsLM subsections $S_{LM}^{(n,l)}$ (with $l \in \{1, \dots, L_{LM} = 5\}$) in small field-of-views of 28×28 µm^2 (see Fig. 4b.1). Only axonal bodies were manually annotated. Individual axons were manually annotated as follows: the outline of the axonal body was defined, then the enclosed region was filled. Manual annotations were crosschecked as described in Section 2.5.

(b) To capture the *tail* of the axon radii distribution, we manually annotated *large* axons on three lsLM subsections $S_{lsLM}^{(n,l)}$ (with $l \in \{1, \dots, L_{lsLM} = 3\}$) in large field-of-views with an equivalent square area of 350×350 µm^2 (see Fig. 4b.2). Exhaustive manual annotation of all axons was considered infeasible due to the large field-of-view. Only axonal bodies were manually annotated. An early prototype of the proposed method was used as a guidance for the rater to detect and initially segment *large* axons. Then, the segmentation of detected axons was manually refined; missed axons were annotated; falsely detected axons were removed. Manual annotations were crosschecked as described in Section 2.5.

(c) To capture the *bulk* of the axon radii distribution, we manually annotated all axons on one *matching* EM subsection $S_{EM}^{(n)}$ in small field-of-views, ranging from equivalent square areas of 54×54 µm^2 to 87×87 µm^2 (on average: 75×75 µm^2) (see Fig. 4b.3). Outlines of axonal bodies were approximated as polygons by M. Morozova. To convert these polygons to axon segmentation masks, we assigned pixels inside polygons an axon label and classified remaining pixels as background.

2.7. Performance of the semantic segmentation network

To assess the capability of the semantic segmentation model (see Section 2.5) to segment axons, we considered the binary, pixel-wise classification task of discriminating between axon and background. As we evaluted the capability to segment axons, we did not consider myelin, i.e., we generated binary axon prediction masks and treated all non-axon pixels as background. We evaluated the axon segmentation performance both at the level of individual pixels and at the level of axon instances.

2.7.1. Pixel-wise segmentation performance

To quantify the axon segmentation performance at the level of individual pixels, we evaluated segmentation metrics (Eqs. (3) to (6)) on pairs of binary axon masks obtained through manual annotation and prediction using the semantic segmentation model on small-field-of-view subsections $S_{LM}^{(n,l)}$. From pixel-wise comparison of pairs of manually annotated and predicted axon masks, we determined the number of false negatives (|FN|), the number of false positives (|FP|), the number of true positives (|TP|) and the number of true negatives (|TN|). Finally,

we computed

$$\text{Balanced accuracy} = \frac{1}{2}\left(\frac{|\text{TP}|}{|\text{TP}| + |\text{FN}|} + \frac{|\text{TN}|}{|\text{TN}| + |\text{FP}|}\right), \tag{3}$$

$$\text{Recall} = \frac{|\text{TP}|}{|\text{TP}| + |\text{FN}|}, \tag{4}$$

$$\text{Precision} = \frac{|\text{TP}|}{|\text{TP}| + |\text{FP}|}, \tag{5}$$

and

$$\text{Dice} = \frac{2 \cdot |\text{TP}|}{2 \cdot |\text{TP}| + |\text{FP}| + |\text{FN}|} \tag{6}$$

for each of the $N \cdot L_{\text{LM}}$ subsections of all test regions and summarized each metric by the mean across subsections.

2.7.2. Instance-wise segmentation performance

We assessed the axon segmentation performance at the level of axon instances as a function of the axon radius. Two measures were considered: an instance-wise evaluation of the dice coefficient and a comparison of the number of undetected axons (i.e., false negatives) and falsely detected axons (i.e., false positives). For this analysis, we pooled all manually annotated axons over all small- and large-field-of-view lsLM subsections $S_{\text{LM}}^{(n,l)}$ and $S_{\text{lsLM}}^{(n,l)}$ across all test regions.

The instance-wise dice coefficient was assessed for pairs of manually annotated axons and their best-matching axon from the prediction following a similar approach as in Abdollahzadeh et al., 2019. The instance-wise dice coefficient was computed using Eq. (6) for pairs of predicted and manually annotated binary axon masks in which we considered only the respective two matching axons, whereas remaining pixels were considered to be background. For each manually annotated axon, the best-matching, predicted axon was determined in terms of the highest instance-wise dice coefficient. Manually annotated axons with no best-matching, predicted axon, i.e., the maximum instance-wise dice coefficient was zero, were considered to be false negatives. Then, we binned manually annotated axons by their radii (spacing: 0.1 μm) and computed the mean dice coefficient per bin. To disentangle the contribution of false negatives from the contribution of under- or oversegmentation of correctly detected axons towards the mean dice coefficients, we repeated the analysis without taking false negatives into account for the computation of the mean dice coefficients.

To compare over- and underdetection of axon instances as a function of the axon radius, we computed |FN| and |FP| (here: at axon instance level) per axon radius bin. While |FN| was immediately available from the computation of mean dice coefficients, we determined |FP| as the number of predicted axons that were not assigned as a best-matching axon to any manually annotated axon.

2.8. Error of estimated r_{arith} and r_{eff}

In this section, we evaluated different error metrics of estimates of r_{arith} and r_{eff}, i.e., $\hat{r}_{\text{arith}}^{(n,l)}$ and $\hat{r}_{\text{eff}}^{(n,l)}$, for axon radii distributions predicted on large-field of-view lsLM subsections $S_{\text{lsLM}}^{(n,l)}$. Corresponding reference values, i.e., $\rho_{\text{arith}}^{(n,l)}$ and $\rho_{\text{eff}}^{(n,l)}$, were generated from different axon radii distributions obtained through manual annotation of $S_{\text{lsLM}}^{(n,l)}$ and *matching* EM subsections $S_{\text{EM}}^{(n)}$ (see Fig. 5).

2.8.1. Error metrics

To assess the error of $\hat{r}_{\text{arith}}^{(n,l)}$ with respect to $\rho_{\text{arith}}^{(n,l)}$ (and the error of $\hat{r}_{\text{eff}}^{(n,l)}$ analogously), we considered three different error metrics. Using the residuals

$$\epsilon^{(n,l)} = \hat{r}_{\text{arith}}^{(n,l)} - \rho_{\text{arith}}^{(n,l)} \tag{7}$$

and denoting the m-th moment of ϵ (and others analogously) as

$$\langle \epsilon^m \rangle = \frac{1}{N \cdot L_{\text{lsLM}}} \sum_{n=1}^{N} \sum_{l=1}^{L_{\text{lsLM}}} (\epsilon^{(n,l)})^m,$$

we assessed accuracy in terms of the normalized-root-mean-square error

$$\text{NRMSE} = \frac{\sqrt{\langle \epsilon^2 \rangle}}{\langle \rho_{\text{arith}} \rangle}, \tag{8}$$

the bias in terms of the normalized-mean-bias-error

$$\text{NMBE} = \frac{\langle \epsilon \rangle}{\langle \rho_{\text{arith}} \rangle} \tag{9}$$

and the normalized-residual-standard-deviation

$$\text{NRSD} = \frac{\text{std}(\epsilon)}{\langle \rho_{\text{arith}} \rangle} = \frac{\sqrt{\langle (\epsilon - \langle \epsilon \rangle)^2 \rangle}}{\langle \rho_{\text{arith}} \rangle} = \frac{\sqrt{\langle \epsilon^2 \rangle - \langle \epsilon \rangle^2}}{\langle \rho_{\text{arith}} \rangle}. \tag{10}$$

Note, that NRMSE (see Eq. (8)) can be expressed in terms of Eqs. (9) and 10 using the following decomposition:

$$\text{NRMSE} = \sqrt{\text{NMBE}^2 + \text{NRSD}^2}. \tag{11}$$

2.8.2. Error of $\hat{r}_{eff}$

To assess the error of estimates of the *tail*-weighted r_{eff}, we compared estimates ($\hat{r}_{\text{eff}}^{(n,l)}$) obtained from predictions on large-field-of-view lsLM subsections $S_{\text{lsLM}}^{(n,l)}$ against reference values ($\rho_{\text{eff}}^{(n,l)}$) computed from composite reference axon radii distributions $\mathcal{E}_{\text{eff}}^{(n,l)}$. The *tail* of $\mathcal{E}_{\text{eff}}^{(n,l)}$ was sampled from manual annotations on $S_{\text{lsLM}}^{(n,l)}$. The *bulk* of $\mathcal{E}_{\text{eff}}^{(n,l)}$ was sampled from manual annotations on *matching* EM subsections $S_{\text{EM}}^{(n)}$ and rescaled according to a scaling factor $f^{(n,l)}$ to compensate for the smaller axon ensemble size of $S_{\text{EM}}^{(n)}$ as compared to $S_{\text{lsLM}}^{(n,l)}$ (see Fig. 5d). This composition of $\mathcal{E}_{\text{eff}}^{(n,l)}$ was motivated as follows: accurate representation of the *tail* required exhaustive manual annotation of the *tail* of the axon radii distribution of $S_{\text{lsLM}}^{(n,l)}$; the *bulk* of the axon radii distribution of $S_{\text{lsLM}}^{(n,l)}$ could not be sampled through manual annotation with reasonable effort due to the large ensemble size. Instead, we assumed that the *bulk* of the axon radii distribution could be representatively sampled from smaller axon ensembles annotated on $S_{\text{EM}}^{(n)}$.

To generate the reference axon radii distribution $\mathcal{E}_{\text{eff}}$ for one subsection, we determined its numbers of axons n_{eff} with radii $r_{(k)}$ per bin k using corresponding numbers of axons manually annotated on S_{EM} and S_{lsLM}, i.e., n_{EM} and n_{lsLM}. For the *bulk* of $\mathcal{E}_{\text{eff}}$, n_{eff} was obtained by rescaling n_{EM} according to f. For the *tail* of $\mathcal{E}_{\text{eff}}$, n_{eff} was equal to n_{lsLM}. Thus,

$$n_{\text{eff}} = \begin{cases} f \cdot n_{\text{EM}}, & \text{for } k \text{ with } r_{(k)} < 1.6\,\mu\text{m} \\ n_{\text{lsLM}}, & \text{for } k \text{ with } r_{(k)} \geq 1.6\,\mu\text{m} \end{cases}. \tag{12}$$

As there was no obvious choice of f, we determined a lower ($f_{\downarrow}$) and upper ($f_{\uparrow}$) bound for f (see Fig. 6).

$f_{\downarrow}$ was determined as the ratio between the number of lsLM-resolvable ($r \geq 0.3\,\mu\text{m}$) axons predicted on S_{lsLM} ($n_{\text{lsLM},r\geq 0.3\,\mu\text{m}}$) and the number of lsLM-resolvable axons manually annotated on S_{EM} ($n_{\text{EM},r\geq 0.3\,\mu\text{m}}$), i.e.,

$$f_{\downarrow} = \frac{n_{\text{lsLM},r\geq 0.3\,\mu\text{m}}}{n_{\text{EM},r\geq 0.3\,\mu\text{m}}}. \tag{13}$$

This choice of $f_{\downarrow}$ was due to the observation that the semantic segmentation network was more likely to miss axons than to falsely detect axons (see Section 3.1). Therefore, $f_{\downarrow}$ was likely to underrepresent the *bulk*. In contrast, we determined $f_{\uparrow}$ as the ratio of subsection areas of S_{lsLM} and S_{EM}, denoted as A_{lsLM} and A_{EM}:

$$f_{\uparrow} = \frac{A_{\text{lsLM}}}{A_{\text{EM}}}. \tag{14}$$

We assumed that $f_{\uparrow}$ would overrepresent the *bulk* because we expected a higher axon density in S_{EM} than in S_{lsLM} due to the lack of large non-fiber structures such as blood vessels in S_{EM}.

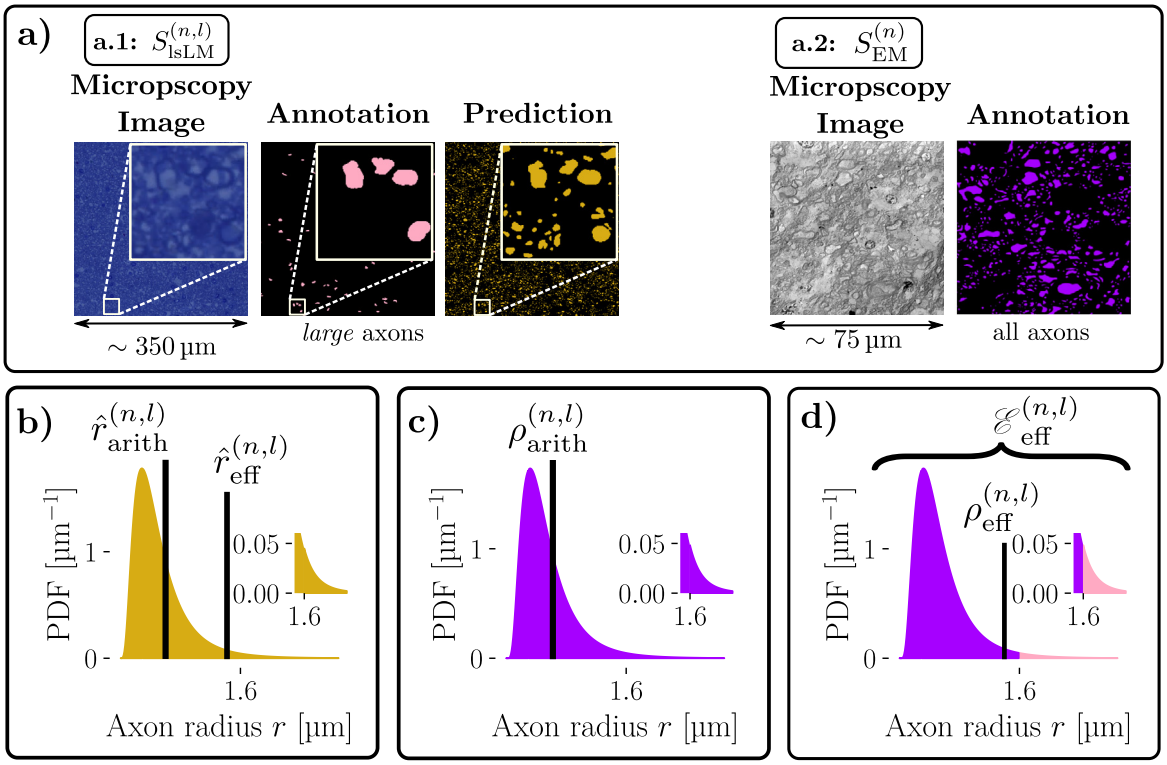

Fig. 5. Axon radii distributions and ensemble mean axon radii. (a) Subsections from *matching* lsLM and EM sections with axon annotation mask and its predicted counterpart: (a.1) lsLM subsections $S^{(n,l)}_{\text{lsLM}}$; (a.2) EM subsection $S^{(n)}_{\text{EM}}$. See Fig. 4 for context. (b) Estimates ($\hat{r}^{(n,l)}_{\text{arith}}$ and $\hat{r}^{(n,l)}_{\text{eff}}$) were obtained from the axon radii distribution of axons predicted (yellow) on $S^{(n,l)}_{\text{lsLM}}$. (c) Reference values ($\rho^{(n,l)}_{\text{arith}}$) were directly obtained from the axon radii distribution of axons manually annotated on $S^{(n)}_{\text{EM}}$ (purple). (d) Reference values ($\rho^{(n,l)}_{\text{eff}}$) were computed from composite axon radii distributions $\mathscr{E}^{(n,l)}_{\text{eff}}$, combining the *bulk* of the axon radii distribution (purple) of axons manually annotated on $S^{(n)}_{\text{EM}}$ with the *tail* of the axon radii distribution (pink) of axons manually annotated on $S^{(n,l)}_{\text{lsLM}}$. To enable combination of differently sized axon radii distributions, the axon radii distribution of axons manually annotated on $S^{(n)}_{\text{EM}}$ was rescaled by a scaling factor $f^{(n,l)}$ (see Section 2.8.2 for details). The tick on the x-axis denotes the threshold that partitioned the axon radii distribution into *bulk* ($r < 1.6$ µm) and *tail* ($r \geq 1.6$ µm) axons. The insets emphasize the *tail* of the axon radii distribution.

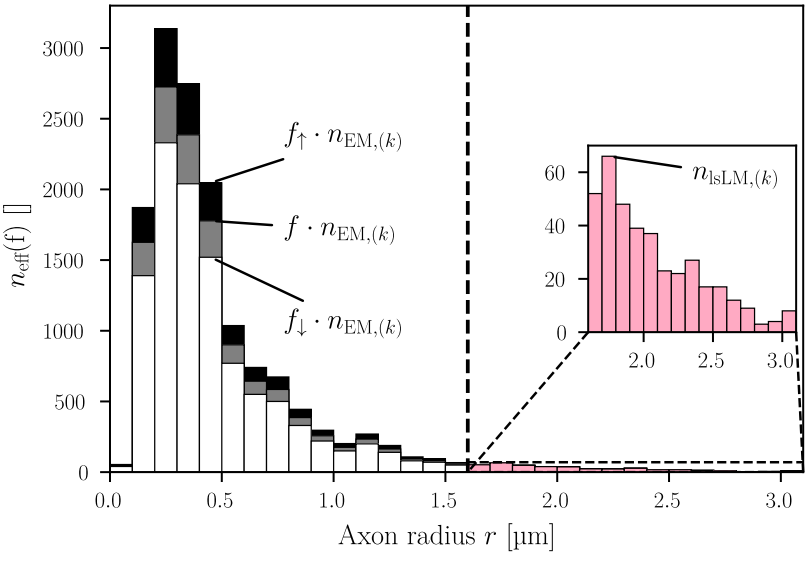

Fig. 6. Generation of the reference axon radii distribution $\mathcal{E}_{\text{eff}}(f)$ for one subsection S_{lsLM}. Depicted are the numbers of axons $n_{\text{eff}}(f)$ with radii $r_{(k)}$ per bin k of the reference axon radii distribution $\mathcal{E}_{\text{eff}}(f)$. For the *bulk* of $\mathcal{E}_{\text{eff}}(f)$ ($r < 1.6$ µm; left of dashed line), we used corresponding, rescaled numbers of axons manually annotated on S_{EM}, i.e., $n_{\text{eff}}(f) = f \cdot n_{\text{EM}}$. For the *tail* of $\mathcal{E}_{\text{eff}}(f)$ ($r \geq 1.6$ µm; right of dashed line), we used the axons manually annotated on S_{lsLM}, i.e., $n_{\text{eff}}(f) = n_{\text{lsLM}}$. Colors of bars for the *bulk* of $\mathcal{E}_{\text{eff}}(f)$ illustrate different values of f: (white) lower bound $f_\downarrow$; (black) upper bound $f_\uparrow$; (gray) an intermediate case.

An over- or underrepresentation of the *bulk* of the axon radii distribution leads to an error in ρ_{eff}. Due to the *tail*-weighting of r_{eff}, we hypothesized that using a reference axon radii distribution with overrepresented *bulk* ($\mathcal{E}_{\text{eff}\uparrow}$) would lead to an underestimation of r_{eff}, whereas using a reference axon radii ensemble with underrepresented *bulk* ($\mathcal{E}_{\text{eff}\downarrow}$) would lead to an overestimation of r_{eff}, i.e., $\rho_{\text{eff}\uparrow} < r_{\text{eff}} < \rho_{\text{eff}\downarrow}$. Therefore, we assessed the error metrics of $\hat{r}_{\text{eff}}$ with respect to both reference values $\rho_{\text{eff}\downarrow}$ and $\rho_{\text{eff}\uparrow}$ and used the maximum absolute value per error metric as an upper bound for the true error. Moreover, we assessed the dynamic range of errors by investigating the error metrics of $\hat{r}_{\text{eff}}$ with respect to reference values obtained based on scaling factors in the range between $f_\downarrow$ and $f_\uparrow$. To this end, we computed reference values $\rho_{\text{eff}}(f_{\text{interp}})$ for

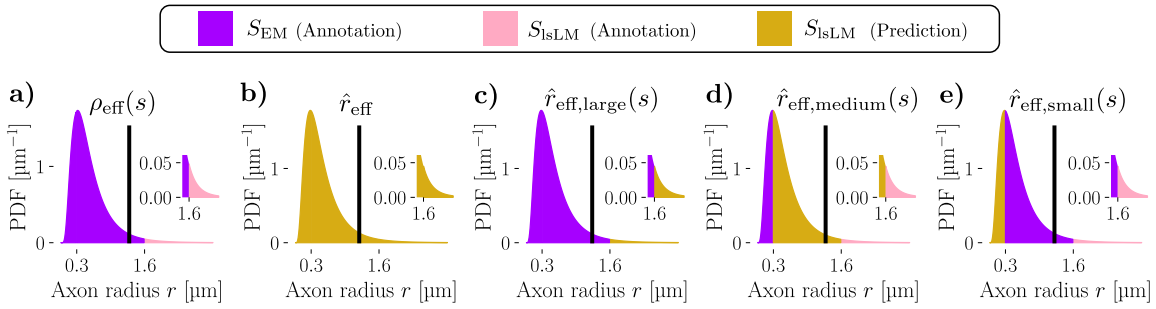

Fig. 7. Schematic of axon radii distributions used to assess the error of $\hat{r}_{eff}$ for one subsection S_{lsLM}. (a-b) Reference axon radii distribution (purple and pink) $\mathcal{E}_{eff}(s)$ and predicted (yellow) axon radii distribution as described in Section 2.8.2, Fig. 5b and Fig. 5d. (c-e) Axon radii distributions generated to assess the error of $\hat{r}_{eff}$ due to predicted axon radii in distinct axon radii ranges. These axon radii distributions used the predicted axon radii distribution (b) in the *large* (c), *medium-sized* (d) and *small* (e) axon radii range and axon radii of $\mathcal{E}_{eff}(s)$ (see (a)) in the remaining ranges. Axon radii distributions in (c-e) partially relied on axon radii of $\mathcal{E}_{eff}(s)$ (see (a)), thereby inheriting a depencency on the sweep variable s, which determined the scaling of the *bulk* of $\mathcal{E}_{eff}(s)$ as described in Section 2.8.2. Vertical bars (a-e) mark values of r_{eff} computed from the respective axon radii distributions. The ticks on x-axes denote the two thresholds that partitioned the axon radii distribution into *small* ($r < 0.3$ µm),*medium-sized* (0.3 µm $\leq r < 1.6$ µm) and *large* ($r \geq 1.6$ µm) axons. The insets emphasize the *tail* of the axon radii distribution.

interpolated scaling factors $f_{interp} \in [f_{\downarrow}, f_{\uparrow}]$. To assess the error metrics as a function of $f_{interp}^{(n,l)}$ across all $N \cdot L_{lsLM}$ subsections with varying $f_{\downarrow}^{(n,l)}$ and $f_{\uparrow}^{(n,l)}$, we parameterized

$$f_{interp}^{(n,l)}(s) = f_{\downarrow}^{(n,l)} + s \cdot (f_{\uparrow}^{(n,l)} - f_{\downarrow}^{(n,l)}) \tag{15}$$

in terms of a common sweep variable $s \in [0, 1]$. Then, we computed NRMSE(s), NMBE(s) and NRSD(s) of $\hat{r}_{eff}$ with respect to $\rho_{eff}(s)$ across all $N \cdot L_{lsLM}$ subsections.

2.8.3. Contribution of axon radii ranges towards the error of $\hat{r}_{eff}$

In Section 2.8.2, we assessed the error of $\hat{r}_{eff}^{(n,l)}$ based on axon radii distributions with erroneous (i.e., predicted) axon radii across the entire range of axon radii. To distinctively assess the contribution of erroneous axon radii of individual ranges towards the error of $\hat{r}_{eff}^{(n,l)}$, we computed $\hat{r}_{eff,large}^{(n,l)}(s)$, $\hat{r}_{eff,medium}^{(n,l)}(s)$ and $\hat{r}_{eff,small}^{(n,l)}(s)$ from composite axon radii distributions with predicted axon radii in the particular range, i.e., in the range of *small, medium-sized* and *large* axon radii; the remaining ranges used axon radii of $\mathcal{E}_{eff}^{(n,l)}(s)$ (see Fig. 7).

We assessed errors of $\hat{r}_{eff,large}^{(n,l)}(s)$, $\hat{r}_{eff,medium}^{(n,l)}(s)$, $\hat{r}_{eff,small}^{(n,l)}(s)$ with respect to $\rho_{eff}^{(n,l)}(s)$ across all $N \cdot L_{lsLM}$ subsections of all test regions in terms of NRMSE(s) NMBE(s), NRSD(s) (see Eqs. (8) to (10)).

2.8.4. Error of $\hat{r}_{arith}$

To assess the error of estimates of the *bulk*-determined r_{arith}, we compared estimates ($\hat{r}_{arith}^{(n,l)}$) obtained from predictions on large-field-of-view lsLM subsections $S_{lsLM}^{(n,l)}$ against reference values ($\rho_{arith}^{(n,l)}$) obtained from manual annotations on *matching* EM subsections $S_{EM}^{(n)}$ (see Fig. 5c). The choice of an EM-based reference was due to its accurate representation of the *bulk* of the axon radii distribution, including *small* axons below the resolution limit of lsLM. As only one EM subsection $S_{EM}^{(n)}$ existed per test region, we used the same reference $\rho_{arith}^{(n,l)}$ for all L_{lsLM} subsections per region, i.e., $\rho_{arith}^{(n,1)} = ... = \rho_{arith}^{(n,L_{lsLM})}$. Note, that the generation of $\rho_{arith}^{(n,l)}$ was simplified in comparison to the approach used for $\rho_{eff}^{(n,l)}(s)$ in Section 2.8.2: instead of computing $\rho_{arith}^{(n,l)}$ in analogy to $\rho_{eff}^{(n,l)}(s)$ from composite axon radii distributions $\mathcal{E}_{eff}^{(n,l)}(s)$ combining EM- and lsLM-based axon radii distributions, we calculated $\rho_{arith}^{(n,l)}$ exclusively from EM-based axon radii distributions. The motivation for this simplification was as follows: first, EM accurately captures the bulk of the axon radii distribution that determines r_{arith}; second, we avoided the dependency of $\rho_{arith}^{(n,l)}$ and derived error metrics for $\hat{r}_{arith}$ on s.

We assessed errors of $\hat{r}_{arith}^{(n,l)}$ with respect to $\rho_{arith}^{(n,l)}$ across all $N \cdot L_{lsLM}$ subsections of all test regions in terms of NRMSE, NMBE and NRSD (see Eqs. (8) to (10)).

2.9. Sensitivity of $\hat{r}_{arith}$ and $\hat{r}_{eff}$ to variation of the image intensity

We assessed whether the influence of spatially varying intensity, e.g. introduced by staining heterogeneity, affected the capability of our method to map anatomy-related, spatial variation of $\hat{r}_{arith}$ and $\hat{r}_{eff}$ across whole lsLM sections. For qualitative analysis, we generated spatially smoothed maps of $\hat{r}_{arith}$ and $\hat{r}_{eff}$ by computing the average of randomly positioned subsections (equivalent square area: 350×350 µm^2) and visually compared the patterns of the spatially smoothed maps to those of the corresponding lsLM images. For quantitative analysis, maps of $\hat{r}_{arith}$, $\hat{r}_{eff}$ and the image intensity were generated similar to those above but sampled on an equally spaced grid (grid pixel area: 350×350 µm^2). To obtain a scalar value for the image intensity, we applied gray scale conversion. Then, grid pixels of sections with similar axon radii distribution (G1, G2, M1, M2) were pooled and the correlation between image intensity and mapped radii was computed. As visual inspection suggested that *small* axons were particularly difficult to resolve in strongly stained areas, the above experiments were performed with and without considering *small* axons to test this hypothesis.

2.10. Sensitivity of r_{eff} to outstandingly large axons

To evaluate how much r_{eff} is affected by outstandingly large axons, we investigated how r_{eff} changed as a function of a varying threshold τ when only axons with $r < \tau$ were considered for the computation of r_{eff}. τ was chosen to cover the whole range of observed axon radii for a given axon radii distribution. In particular, we assessed the worst case in which the largest individual axon was missed. To exclude estimation errors from this experiment, we considered only reference data, i.e., the reference axon radii distributions generated in Section 2.8.2. To rather over- than underestimate the sensitivity to outstandingly large axons, we used reference axon radii distributions $\mathcal{E}_{eff\downarrow}^{(n,l)}$ with underrepresented *bulk*. Furthermore, to carry out this analysis at a scale as close as possible to the cross-sectional size of typical voxels of a human MRI system (1 mm^2 or larger), we computed $\rho_{eff\downarrow}^{(n)}$ from pooled axon radii distributions, combining axon radii distributions of all L_{lsLM} subsections per test region, yielding $\mathcal{E}_{eff\downarrow}^{(n)} = \bigcup_{l=1}^{L_{lsLM}} \mathcal{E}_{eff\downarrow}^{(n,l)}$ for the n-th test region. Thereby, we obtained $\rho_{eff\downarrow}^{(n)}$ from the largest axon ensembles available for each test region based on combined areas of about 0.37mm^2 ($\approx L_{lsLM} \cdot (350\,\mu m)^2$).

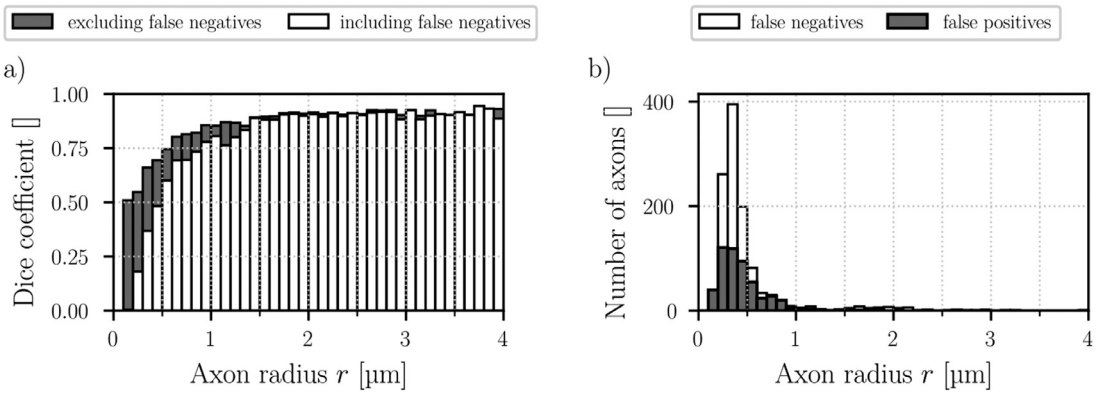

Fig. 8. Instance-wise segmentation metrics as a function of the axon radius. (a) Mean instance-wise dice coefficients: each bar denotes the mean instance-wise dice coefficient computed from manually annotated axons and their predicted counterparts. Two different cases were considered: with (white bars) and without (gray bars) inclusion of undetected axons (false negatives), i.e., manually annotated axons with no matching, predicted axon. False negatives contributed to the mean dice coefficient with a dice coefficient of zero. (b) Number of false negatives (white bars) and falsely detected axons (false positves; gray bars) as a function of the axon radius. The bin width in (a-b) is 0.1 µm.

Table 2
Pixel-wise segmentation metrics. Each value in the table denotes a mean value of the corresponding *metric* (see Eqs. (3) to (6)) over all manually annotated small-field-of-view subsections $S_{\mathrm{LM}}^{(n,l)}$ (see Section 2.6.(a)).

Metric	Value
Balanced accuracy	0.85
Dice	0.77
Precision	0.82
Recall	0.74

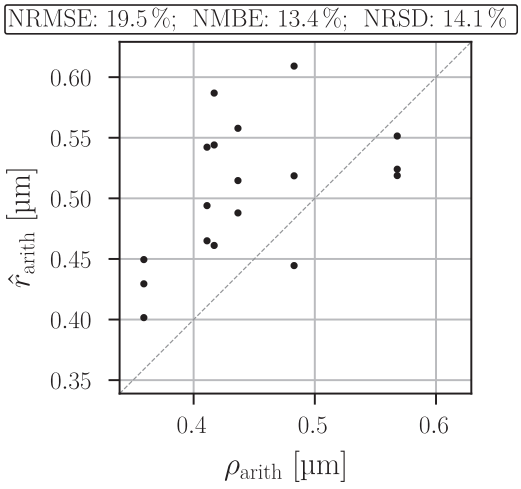

Fig. 9. Error of estimated arithmetic mean radii. Depicted are comparisons of lsLM-based estimates $\hat{r}_{\mathrm{arith}}$ (y-axis) against EM- based reference values ρ_{arith} (x-axis) of the arithmetic mean radius r_{arith}. Each point corresponds to one of $N \cdot L_{\mathrm{lsLM}} = 18$ distinct lsLM subsections. The dashed line represents the line of unity. Accuracy (NRMSE), bias (NMBE) and residual standard deviation (NRSD) were computed over all subsections using Eqs. (8) to (10).

3. Results

3.1. Segmentation performance

Table 2 lists pixel-wise segmentation metrics: balanced accuracy, dice, precision and recall as defined in see Eqs. (3) to (6). Higher precision than recall indicates that the number of false negatives was larger than the number of false positives.

Fig. 8 shows segmentation metrics evaluated at the level of axon instances as a function of the axon radius. The mean dice coefficient increased as a function of the axon radius in the range from 0.0 µm to 1.4 µm (Fig. 8a). For larger axons, the mean dice coefficient varied only little and was always higher than 0.88, regardless of whether false negatives were considered or not to compute the mean dice coefficient. In contrast, mean dice coefficients of smaller axons were determined by the large fraction of false negatives, indicated by the difference between gray and white bars. The number of false negatives per axon radius was mostly higher than the number of false positives, in particular for axons with $r < 1$ µm (Fig. 8b).

3.2. Error of $\hat{r}_{arith}$ and $\hat{r}_{eff}$

Fig. 9 shows estimates of r_{arith} (i.e., $\hat{r}_{\mathrm{arith}}$) against reference values (ρ_{arith}) and denotes accuracy, bias and random error in terms of NRMSE, NMBE and NRSD as defined in Eqs. (8) to (10). $\hat{r}_{\mathrm{arith}}$ deviated from the line of unity, yielding an NRMSE of 19.5 % (see Fig. 9). NMBE and NRSD contributed with similar magnitude (~ 14%) to the NRMSE (see Eq. (11) for a decomposition of NRMSE into NMBE and NRSD).

Fig. 10 shows NRMSE(s), NMBE(s) and NRSD(s) of different estimates of r_{eff} (i.e., $\hat{r}_{\mathrm{eff}}$, $\hat{r}_{\mathrm{eff,large}}(s)$, $\hat{r}_{\mathrm{eff,medium}}(s)$, $\hat{r}_{\mathrm{eff,small}}(s)$) with respect to reference values $\rho_{\mathrm{eff}}(s)$. Thereby, Fig. 10 is based on repeated comparison between estimates and reference values as illustrated for r_{arith} in Fig. 9, but shows only the above error metrics as a function of s. Here, s determined the scaling of the *bulk* of axon radii distributions $\mathcal{E}_{\mathrm{eff}}(s)$, interpolating between lower ($f_{\downarrow}$) and upper bound ($f_{\uparrow}$) scaling factors (see Eq. (15)). Generally, NMBE varied as a function of s, whereas the NRSD was less dependent on s (see Fig. 10, center and bottom row). NMBE and NRSD translated into NRMSE (see Fig. 10, top row), yielding an overall NRMSE between 7.2 % to 8.5 % (see Fig. 10a). The overall NRSD (7.1 % to 7.3 %) was predominantly determined by an s-independent contribution of *large* axons (6.9 %) and complemented by a smaller, s-dependent contribution of *medium-sized* axons (1.4 % to 2.1 %) (see Fig. 10a-c, bottom row). In contrast, the overall NMBE (-3.7 % to 4.8 %) was predominantly determined by a strong s-dependent contribution of *medium-sized* axons (-1.3 % to 6.5 %) and a smaller, s-independent contribution of *large* axons (-2.9 %) (see Fig. 10a-c, center row). *Small* axons

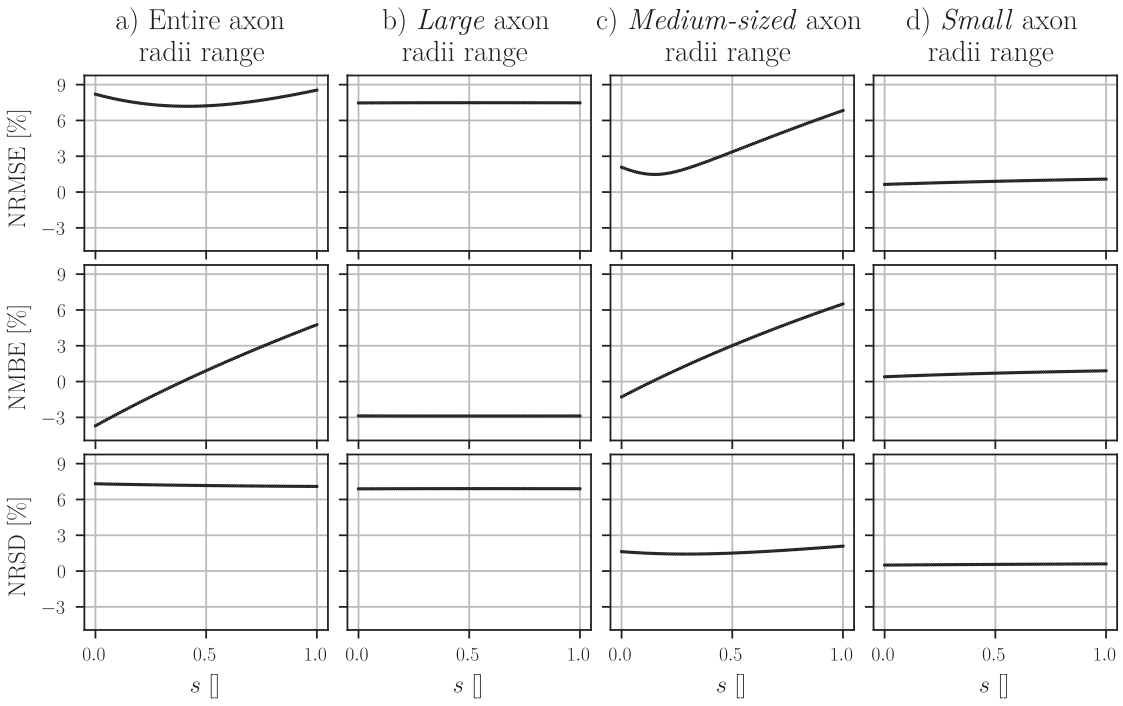

Fig. 10. Error of estimated effective radii. Rows show three error metrics for different estimates of the MRI-visible, effective axon radius r_{eff}: (top row) accuracy as evaluated by NRMSE (see Eq. (8)); (center row) bias as evaluated by NMBE (see Eq. (9)); (bottom row) the residual standard deviation NRSD (see Eq. (10)). Each column depicts the aforementioned errors with respect to reference values $\rho_{\text{eff}}(s)$ due to erroneous axons in distinct axon radii ranges: (a) entire axon radii range, estimating overall errors; (b) *large* axon radii range ($r \geq 1.6$ µm); (c) *medium-sized* axon radii range (0.3 µm $\leq r < 1.6$ µm); (d) *small* axon radii range ($r < 0.3$ µm). Errors in (a-d) are shown as a function of a sweep variable s, which determined the scaling of the *bulk* of reference axon radii distributions $\mathcal{E}_{\text{eff}}(s)$. These reference axon radii distributions $\mathcal{E}_{\text{eff}}(s)$ were used to compute reference values $\rho_{\text{eff}}(s)$ in (a-d) and estimates of r_{eff} ($\hat{r}_{\text{eff,large}}(s)$, $\hat{r}_{\text{eff,medium}}(s)$ and $\hat{r}_{\text{eff,small}}(s)$) in (b-d). Here, $s = 0$ and $s = 1$ correspond to using lower ($f_{\downarrow}$) and upper ($f_{\uparrow}$) bounds of the scaling factor (see Eq. (15)). Error metrics were evaluated over $N \cdot L_{\text{lsLM}} = 18$ lsLM subsections. Note, that NRMSE combines NMBE and NRSD as described in Eq. (11).

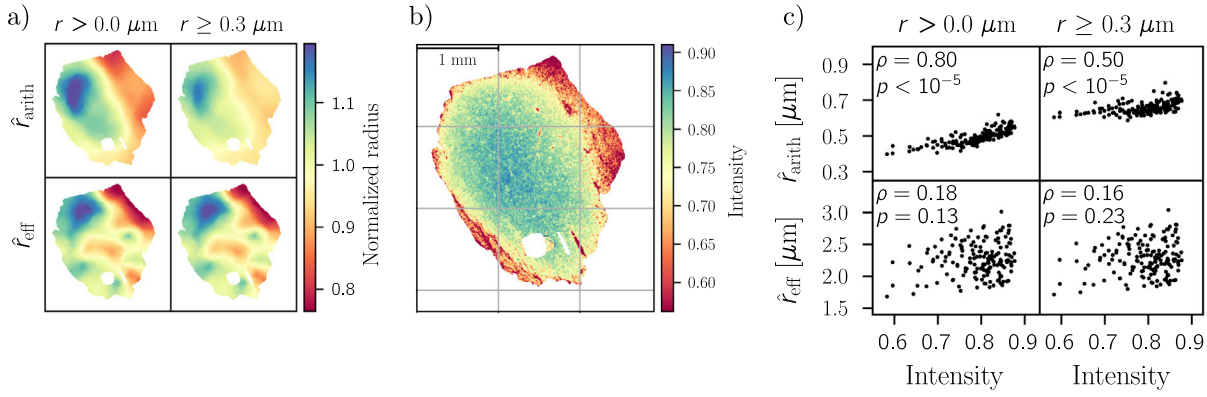

Fig. 11. Sensitivity of estimates of the arithmetic mean radius r_{arith} and the MRI-visible effective axon radius r_{eff} to variation of the image intensity. Depicted are: spatially smoothed maps of estimates $\hat{r}_{\text{arith}}$ and $\hat{r}_{\text{eff}}$ (a), the lsLM image of section M1 (b) adjusted to illustrate the correlation with maps of $\hat{r}_{\text{arith}}$ (a), and scatter plots between ensemble mean axon radii ($\hat{r}_{\text{arith}}$ and $\hat{r}_{\text{eff}}$) and lsLM image intensities (c). The correlation plots (c) pool across four sections (G1, G2, M1, M2). The p-values have been multiplied by the number of sections to correct for multiple comparisons (Pearson's ρ is the correlation coefficient).

had a small s-dependent NMBE (0.4 % to 0.9 %) and small errors overall, i.e., NRMSE was at most 1.1 % (see Fig. 10d).

3.3. Sensitivity of $\hat{r}_{arith}$ and $\hat{r}_{eff}$ to variation of the image intensity

The spatial variation of $\hat{r}_{\text{arith}}$ resembled the image intensity distribution of the corresponding lsLM section (see Fig. 11a, top row and Fig. 11b). In contrast, maps of $\hat{r}_{\text{eff}}$ had a high local heterogeneity, which was not observed in the image intensity distribution of the corresponding lsLM section (see Fig. 11a, bottom row and Fig. 11b). These observations were supported by a strong correlation (Pearson's $\rho = 0.80$, $p < 10^{-5}$) between $\hat{r}_{\text{arith}}$ and the image intensity, which was reduced when *small* axons were discarded (Pearson's $\rho = 0.50$, $p < 10^{-5}$) (see Fig. 11c, top left and top right). In contrast, $\hat{r}_{\text{eff}}$ did not show a significant correlation with the image intensity (see Fig. 11c, bottom row).

3.4. Sensitivity of r_{eff} to outstandingly large axons

ρ_{eff} increased nonlinearly as a function of τ but with decreasing slope (see Fig. 12a). For large τ, there is a step-wise dependence between $\rho_{\text{eff}\downarrow}$

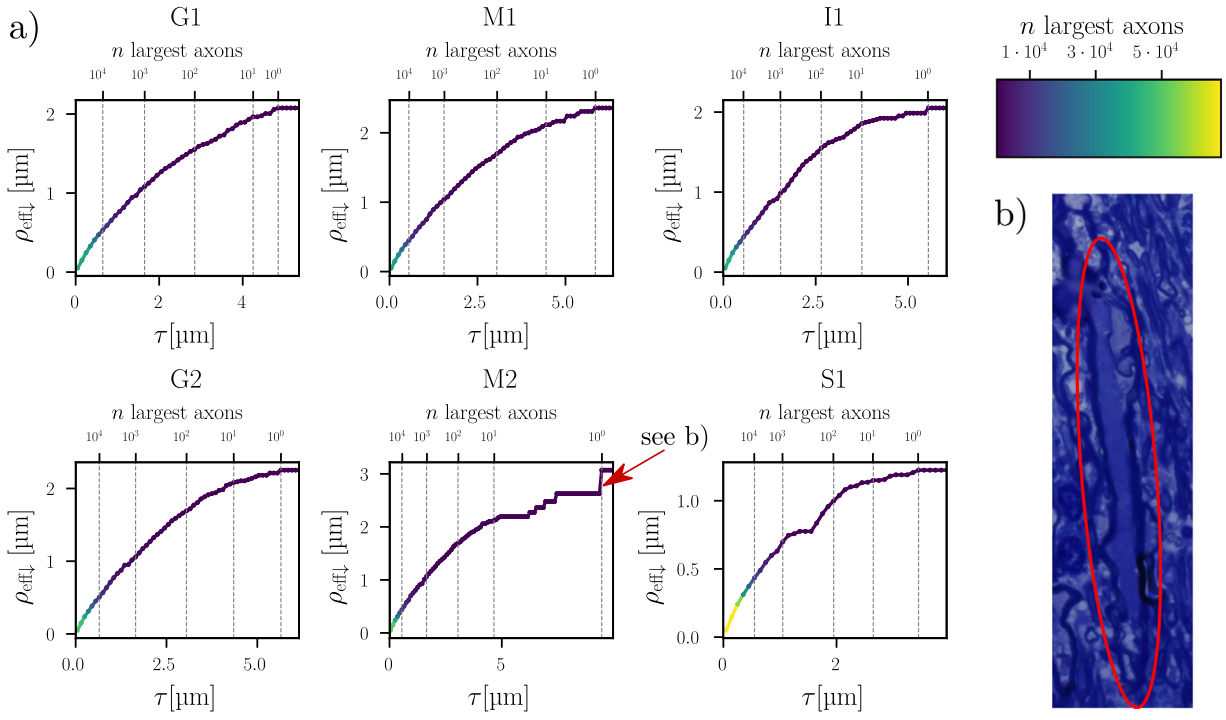

Fig. 12. Sensitivity of the MRI-visible, effective axon radius r_{eff} to outstandingly large axons. (a) Values of $\rho_{\text{eff}\downarrow}$ for the $N = 6$ test regions of the corpus callosum sample when considering only axons with radius of $r < \tau$ for the computation of $\rho_{\text{eff}\downarrow}$. The line colors indicate the order of axons sorted by their radius in descending order according to the colorbar in the top right. For orientation, powers of 10 are marked on top of the plots. (b) Extracted lsLM subsection showing the largest axon ($r = 9.46$ µm) observed across all regions. The elongated shape of this axon is likely due to the axon being oriented almost parallel to the cutting plane of the two-dimensional section. When discarding this axon, $\rho_{\text{eff}\downarrow}$ decreased from 3.07 μm to 2.63 μm, i.e., a decrease of 14.3 %. For the remaining regions, the decrease of $\rho_{\text{eff}\downarrow}$ ranged from 0.8 % to 2.9 %. The total number of axons ranged from $3.7 \cdot 10^4$ to $6.8 \cdot 10^4$. $\rho_{\text{eff}\downarrow}$ used the lower bound ($f_\downarrow$) of a scaling factor, which determined the scaling of the *bulk* of reference axon radii distributions $\mathcal{E}_{\text{eff}\downarrow}$ of $\rho_{\text{eff}\downarrow}$ (see details in Sections 2.8.2 and 2.10).

and τ due to the sparse occurrence of large axons. Compared to other regions, the influence of the largest axon on $\rho_{\text{eff}\downarrow}$ was particularly strong in region M2: $\rho_{\text{eff}\downarrow}$ decreased by 14.3 % when the largest axon was discarded (see region M2 in Fig. 12a). The largest axon was much larger than other axons across all regions and its elongated shape suggested that this axon was oriented almost parallel to the cutting plane, i.e., its axon radius was strongly overestimated by the circular equivalent approximation (see Fig. 12b). The influence of the largest axon was smaller for the remaining regions: $\rho_{\text{eff}\downarrow}$ decreased by 0.8 % to 2.9 % when discarding the largest axon (see Fig. 12a). For region S1, there is a plateau between $1.3\ \mu\text{m} \lesssim \tau \lesssim 1.6\ \mu\text{m}$, indicating that no axons were sampled in this axon radii range. As the axon radii distributions $\mathcal{E}_{\text{eff}\downarrow}$ of $\rho_{\text{eff}\downarrow}$ used the *bulk* of the axon radii distribution (i.e., $r < 1.6\ \mu\text{m}$) from *matching*, small-field-of-view EM subsections S_{EM} (see Section 2.8.2), it seems that axons with $r \gtrsim 1.3\ \mu\text{m}$ were not representatively sampled in S_{EM} of region S1.

Discussion

We investigated the potential of CNN-based segmentation on high-resolution, large-scale light microscopy (lsLM) sections to narrow the scale gap between histological reference data and MRI voxels for the validation of diffusion MRI-based effective axon radius (r_{eff}) estimation in human brain tissue. The proposed pipeline accurately estimates r_{eff} in a human corpus callosum on sections spanning several cross-sections of typical voxels of human MRI systems (1 mm^2 or larger) and is thus a promising candidate for the validation of MRI-based r_{eff} estimation in the human brain. However, the arithmetic mean radius (r_{arith}), which is commonly reported in neuroanatomical studies, is less accurately estimated.

Estimation error of r_{arith} and r_{eff}

To assess the estimation error of r_{eff} representatively for cross-sections of MRI voxels (1 mm^2 or larger) of a human MRI system, sufficient sampling of the *tail* of the axon radii distribution is required. Therefore, we investigated large ensembles of axons representing at least 10,000 axons per sample. To address the challenge of generating reference data for r_{eff} on large ensembles of axons, we captured the *tail* ($r \geq 1.6\ \mu\text{m}$; also denoted as *large* axons) of the axon radii distribution by exhaustive manual annotation and complemented the *tail* with the *bulk* ($r < 1.6\ \mu\text{m}$) from closeby-cut, small-field-of-view EM sections. To compensate for the smaller ensemble size in EM, we rescaled the axon radii distribution according to a scaling factor. As the true scaling factor was unknown, we estimated lower and upper bound scaling factor and assessed the estimation error of r_{eff} for scaling factors in the so-defined scaling factor range to estimate an upper bound and the dynamic range of different error metrics.

Across the entire range of axon radii, we conclude higher suitability of the proposed method to estimate r_{eff} than r_{arith} due to higher accuracy (maximum normalized-root-mean-square-error: 8.5 % versus 19.5 %) and lower bias (maximum absolute normalized-mean-bias-error: 4.8 % versus 13.4 %). Assessment of individual ranges revealed that erroneous, *large* axons predominantly determine the estimation accuracy of r_{eff} followed by *medium-sized* axons ($0.3\ \mu\text{m} \leq r < 1.6\ \mu\text{m}$). A decomposition of the accuracy into bias and residual standard deviation revealed that the residual standard deviation was predominantly determined by *large* axons and had a small dynamic range. The bias, however, had a large dynamic range due to the scaling factor-dependent bias of *medium-sized* axons. Since the true scaling factor is unknown, the true bias cannot be quantified for *medium-sized* axons. *Small* axons ($r < 0.3\ \mu\text{m}$) below the resolution limit of lsLM introduced only a minor overestima-

tion, even when they were neglected altogether for estimating r_{eff} (see Appendix A). Thus, the potential of lsLM to sample the *tail* of the axon radii distribution in large field-of-views outweighs its limited capability to resolve *small* axons for mapping r_{eff}.

While we assessed the presented pipeline with particular focus on the ensemble mean radii of segmented axons, i.e., r_{arith} and r_{eff}, we employed pixel-wise optimization to train the semantic segmentation model. We evaluated the commonly used dice coefficient per axon instance and found a reflection of the better suitability to estimate r_{eff}: larger axons were better segmented.

Mapping anatomy-related, spatial variation across whole sections

Toluidine blue staining introduces low-frequency variation of image intensity across lsLM sections. In a spatial correlation analysis, we identified this variation as a confounding factor for mapping r_{arith} but not for mapping r_{eff}. In the light of moderate errors, spatial variation of r_{eff} seems anatomy-related. As r_{arith} was particularly confounded when *small* axons were taken into account, *small* axons seem particularly prone to staining effects. The inaccurate resolution of *small* axons may explain the observed overestimation of r_{arith}. For r_{eff}, the correlation with the image intensity was hardly affected by inclusion or rejection of *small* axons which underlines their minor contribution towards r_{eff}.

Sensitivity of r_{eff} to outstandingly large axons

Due to the *tail*-weighting of r_{eff}, individual, outstandingly large axons may strongly contribute towards r_{eff} and thus strongly decrease estimation accuracy in case of erroneous segmentation. We assessed this potential source of error by discarding the largest axon for the computation of r_{eff} in axon ensembles representing at least 35,000 axons.

The strongest contribution (14.3% in region M2) of an individual axon was due to an outlier. Across the remaining regions, the contribution was smaller (0.8% to 2.9%), but still notable, considering that these axons represented only 0.001 % to 0.003 % of the axon ensembles. For the outlier-region M2, the largest axon ($r = 9.46\,\mu m$) was oriented almost parallel to the cutting plane, resulting in an elongated shape. Thus, circular equivalent approximation may largely overestimate axon radii and bias the estimation of r_{eff}. To avoid such outliers, axon radii may be estimated based on the minor axes of ellipsoids fitted to the axon areas.

The investigated lsLM subsections (area: $\sim 0.37\,mm^2$) were smaller than the cross-section of a typical MRI voxel ($1\,mm^2$ or larger). In the latter, we expect reduced potential of individual axons to bias r_{eff} due to the larger axon ensemble size.

Limitations and future directions

Although the proposed method accurately estimated r_{eff} for different axon radii distributions sampled across the corpus callosum, further investigation is required to assess how well the model generalizes and how well the overall method translates to other brain areas.

Recent, automated methods for large-scale axon segmentation used different acquisition techniques and segmentation algorithms (Abdollahzadeh et al., 2021; Zaimi et al., 2018), which, however, were trained on perfusion-fixed brain tissue of mice or rats. The method of Abdollahzadeh et al., 2021 is tailored towards three-dimensional data and is therefore not immediately comparable to the proposed method. Although the two-dimensional method of Zaimi et al., 2018 employs a similar approach of subsequent, U-Net-based (Ronneberger et al., 2015) semantic and instance segmentation, our method differs, e.g., in details of the U-Net architecture and by employing transfer learning. In comparison, the method of Zaimi et al., 2018 yielded slightly higher metrics for segmentation of axons in human tissue transmission electron microscopy (TEM) data, e.g. a dice coefficient of 0.81 as compared to 0.77 in our lsLM-based approach, and higher metrics for other mammals in both scanning electron microscopy (SEM) and TEM data (e.g., mean dice coefficient > 0.9). However, due to the different microscopy and tissue preparation techniques, immediate comparison between these results is difficult. In a future study, a segmentation model could be trained using the framework of Zaimi et al., 2018 with the presented data to benchmark the aforementioned method against the proposed method for mapping r_{eff}. In the present study, the primary focus was to assess the feasibility of generating reference data for r_{eff} using automated axon radius segmentation on large-field-of-view lsLM sections.

Estimates of individual axon radii from two-dimensional cross-sections can be biased for axons that are non-orthogonally oriented to the cutting plane (Abdollahzadeh et al., 2019; Andersson et al., 2020; Lee et al., 2019). While approximating axon radii based on the minor axes of fitted ellipsoids may underestimate axon radii, the circular equivalent approximations may overestimate individual axon radii. This bias has been reported to be similar for individual circular equivalent and minor axis approximations in terms of absolute deviation from the along-axis median (~10%) in the corpus callosum of sham-operated rats (Abdollahzadeh et al., 2019). However, further investigations are required to assess how this bias translates towards estimation accuracy of r_{eff}. In our analyses, we identified a potential bias of r_{eff} based on circular equivalent radii caused even by individual axons that were oriented almost parallel to the cutting plane (see Section 3.4). For the two-dimensional reference used in this work, estimates of r_{eff} based on minor axis radii approximations yielded similar accuracy (maximum normalized-root-mean-square-error: 8.1 %) and bias (maximum absolute normalized-mean-bias-error: 5.2 %) as estimates of r_{eff} based on circular equivalent radii (see Appendix B).

Two-dimensional cross-sections cannot capture along-axon variation of the axon radius, given that strong along-axon variation of the axon radius has been reported at the level of individual axons in the corpus callosum of mice (Lee et al., 2019), rats (Abdollahzadeh et al., 2019) and monkeys (Andersson et al., 2020). However, at the ensemble level, good agreement between axon radii distributions estimated from two and three dimensions has been reported for an ensemble of 54 large axons (arithmetic mean radius: $1.35\,\mu m$) within a section of the monkey corpus callosum (Andersson et al., 2020). In fact, the aforementioned study concluded that it may be feasible to compensate the incapability to capture along-axon variation by sufficient in-plane sampling. Following this hypothesis, our proposed method may complement three-dimensional microscopy studies of small ensembles of axons with large-ensemble sampling in two dimensions.

To assess the estimation error of r_{arith}, we compared lsLM-based estimates against EM-based references from close-by cut sections. This choice of reference has two limitations: first, the EM-based axon ensembles were smaller, i.e., covering only 5 to 10 % of their lsLM-based counterparts; second, spatial misalignment arised from section-to-section distance and unknown in-section location. However, the latter section-to-section distance may not render the choice of reference data unsuitable, given that previous studies have reported good agreement between axon radii distributions across comparable distances (Andersson et al., 2020). Furthermore, we assumed that representative estimation of the frequency-weighted r_{arith} is enabled by accurate resolution of frequently occurring axons rather than by a large ensemble size or exact spatial alignment. Consequently, we regarded EM as a more suitable reference than lsLM because EM can resolve all frequently occurring axons, including *small* axons below the resolution limit of lsLM. Indeed, we found *small* axons to be particularly prone to variation of the image intensity in lsLM which in turn led to systematic overestimation of r_{arith}. While the residual standard deviation observed for r_{arith} may partially be due to the choice of unrepresentative reference data, the systematic overestimation of r_{arith} seems to reflect a bias of the proposed pipeline.

To assess the estimation error of r_{eff}, we computed reference values from composite axon radii distributions, combining the *bulk* of axon radii from EM-based, manual annotations with the *tail* from lsLM-based, manual annotations. In one particular region, the assumption that EM can accurately capture the *bulk* of axon radii in small field-of-views

seemed to be violated for larger axons of the *bulk* of the axon radii distribution. For this region, the reference value of r_{eff} may have been less accurate than in other regions.

We have limited analyses to the definition of the MRI-visible, effective radius r_{eff} measured with diffusion MRI in the wide-pulse limit (Burcaw et al., 2015; Sepehrband et al., 2016; Veraart et al., 2020). However, our pipeline can also be used to estimate r_{eff} in the short-pulse limit (Burcaw et al., 2015; Sepehrband et al., 2016) with lower accuracy (maximum normalized-root-mean-square-error: 10.6 %) and higher bias (maximum absolute normalized-mean-bias-error: 9.2 %) (see Appendix C). The lower performance for short-pulse estimates of r_{eff} is likely to the fact that r_{eff} in the short-pulse limit is less *tail*-weighted than r_{eff} in the wide-pule limit. Consequently, the decreased segmentation performance for axons of the *bulk* of the axon radii distribution becomes more relevant for r_{eff} in the short-pulse limit.

The manual annotation of microscopy slides is prone to errors and inter-observer variability, in particular in the presence of staining and tissue degradation due to the immersion-fixation used in this study. Employing strategies that address noisy and uncertain manual annotations, e.g. by design of specific loss functions may improve axon segmentation accuracy (Karimi et al., 2020) and thus radius estimation accuracy.

Conclusion

The presented pipeline is a step towards mapping the MRI-visible, effective radius (r_{eff}) by combining high-resolution, large-scale light microscopy (lsLM) with deep learning. As the two-dimensional lsLM sections span the cross-sectional scale of typical MRI voxels (1 mm^3 or larger), the proposed method may complement three-dimensional microscopy studies of small ensembles of axons with large-ensemble sampling in two dimensions. Since the pipeline is based on the fast, cheap and simple to perform lsLM measurement, it can easily be used beyond the realm of MRI-based radius models, e.g., to generate a representative, neuroanatomical atlas of the ensemble of large axons across the human corpus callosum. However, before this can be done the generalization to different brains is yet to be demonstrated.

Code and data availability

The source code and training data used in this study will be made publicly available upon publication of this study on https://github.com/quantitative-mri-and-in-vivo-histology/ls_axon_segmentation.

Ethics

For samples used in this study, the entire procedure of case recruitment, acquisition of the patient's personal data, the protocols and the informed consent forms, performing the autopsy and handling the autopsy material have been approved by the responsible authorities (Approval #205/17-ek).

Declaration of Competing Interest

The Max Planck Institute for Human Cognitive and Brain Sciences has an institutional research agreement with Siemens Healthcare. NW was a speaker at an event organized by Siemens Healthcare and was reimbursed for the travel expenses.

Credit authorship contribution statement

Laurin Mordhorst: Conceptualization, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft. **Maria Morozova:** Investigation, Writing – review & editing. **Sebastian Papazoglou:** Investigation, Supervision, Writing – review & editing. **Björn Fricke:** Investigation. **Jan Malte Oeschger:** Investigation, Writing – review & editing. **Thibault Tabarin:** Conceptualization, Investigation. **Henriette Rusch:** Investigation, Writing – review & editing.

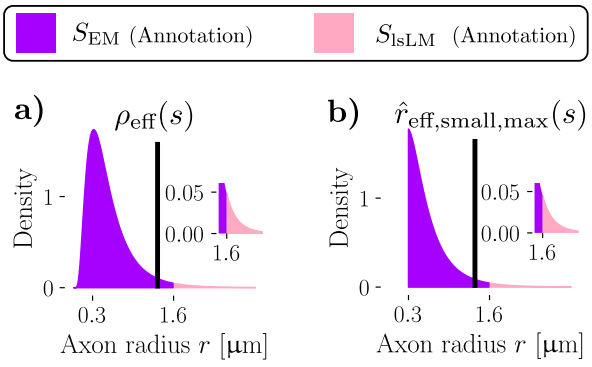

Fig. A.13. Assessment of the error of $\hat{r}_{\text{eff}}$ due to undetected *small* axons for one subsection S_{lsLM}. (a) The reference axon radii distribution $\mathcal{E}_{\text{eff}}(s)$, combining *bulk* from S_{EM} (purple) and *tail* from S_{lsLM} (pink). (b) Axon radii distribution of (a) with *small* axons neglected altogether. The sweep variable s determined the scaling of the *bulk* of $\mathcal{E}_{\text{eff}}(s)$ as described in Section 2.8.2. Vertical bars (a-b) mark values of r_{eff} computed from the respective axon radii distributions. The ticks on x-axes denote the two thresholds that partition the axon radii distribution into *small* ($r < 0.3$ μm), *medium-sized* (0.3 μm $\leq r <$ 1.6 μm) and *large* ($r \geq 1.6$ μm) axons. The insets emphasize the *tail* of the axon radii distribution.

Carsten Jäger: Resources, Writing – review & editing. **Stefan Geyer:** Funding acquisition. **Nikolaus Weiskopf:** Funding acquisition, Writing – review & editing. **Markus Morawski:** Funding acquisition, Methodology, Resources, Writing – review & editing. **Siawoosh Mohammadi:** Conceptualization, Funding acquisition, Methodology, Writing – review & editing, Supervision.

Acknowledgment

The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement No. 616905.

This work was supported by the German Research Foundation (DFG Priority Program 2041 "Computational Connectomics", [MO 2397/5-1; MO 2249/3-1; GE 2967/1-1], by the Emmy Noether Stipend: MO 2397/4-1) and by the BMBF (01EW1711A and B) in the framework of ERA-NET NEURON and the Forschungszentrums Medizintechnik Hamburg (fmthh; grant 01fmthh2017).

We are grateful to Dr. René Werner and Amra Hot for insightful discussions.

Appendix A. Upper bound of the error due to undetected, *small* axons in $\hat{r}_{\text{eff}}$

The error due to *small* axons using $\hat{r}^{(n,l)}_{\text{eff,small}}(s)$ as assessed in Section 2.8.3 used predicted, *small* axon radii and axon radii of the reference axon radii distributions $\mathcal{E}^{(n,l)}_{\text{eff}}(s)$ for the remaining axons. Due to the resolution limit of lsLM, we expected the proposed pipeline to miss *small* axons and consequently expected estimates ($\hat{r}^{(n,l)}_{\text{eff,small}}(s)$) to overestimate reference values ($\rho^{(n,l)}_{\text{eff}}(s)$). We aimed to assess error metrics for the worst-case scenario, i.e, an upper bound for the overestimation of $\rho^{(n,l)}_{\text{eff}}(s)$ due to undetected, *small* axons. To achieve this, we repeated the experiment described for $\hat{r}^{(n,l)}_{\text{eff,small}}(s)$ in Section 2.8.3 with neglected *small* axon radii, yielding $\hat{r}^{(n,l)}_{\text{eff,small,max}}(s)$ (see Fig. A.13).

Results Fig. A.14 shows accuracy (NRMSE(s); see Eq. (8)), bias (NMBE(s); see Eq. (9)) and the residual standard deviation (NRSD(s); see Eq. (10)) of $\hat{r}^{(n,l)}_{\text{eff,small,max}}(s)$ with respect to reference values $\rho_{\text{eff}}(s)$. All errors varied almost linearly as a function of s. The NRMSE was between 1.6 % to 1.8 % (see Fig. A.14a). The NMBE had a larger maximum absolute value and a larger dynamic range (1.5 % to 1.6 %) than the NRSD (~0.6%) (see Fig. A.14b-c). When compared to the errors due to *small*

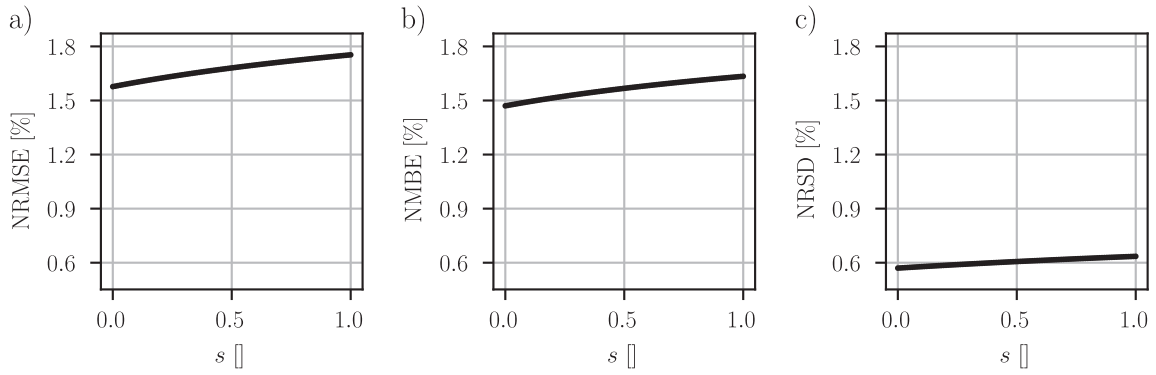

Fig. A.14. Upper bound of the error due to undetected, *small* axons in estimated effective axon radii. Depicted are three different error metrics of estimates $\hat{r}_{\text{eff,small,max}}(s)$ of the MRI-visible, effective axon radius r_{eff} with respect to reference values $\rho_{\text{eff}}(s)$: (a) the accuracy as evaluated by NRMSE (see Eq. (8)); (b) the bias as evaluated by NMBE (see Eq. (9)); (c) the residual standard deviation as evaluated by NRSD (see Eq. (10)). The axon radii distribution of $\hat{r}_{\text{eff,small,max}}(s)$ neglected *small* ($r < 0.3$ µm) axon radii altogether, thereby simulating a potential incapability of large-scale light microscopy (lsLM) to detect *small* axons. All errors (a-c) are shown as a function of a sweep variable s, which determined the scaling of the *bulk* of reference axon radii distributions $\mathcal{E}_{\text{eff}}(s)$. These reference axon radii distributions $\mathcal{E}_{\text{eff}}(s)$ were used to compute both reference values $\rho_{\text{eff}}(s)$ and estimates $\hat{r}_{\text{eff,small,max}}(s)$. Here, $s = 0$ and $s = 1$ correspond to using lower ($f_\downarrow$) and upper ($f_\uparrow$) bounds of the scaling factor (see Eq. (15)). Error metrics were evaluated over $N \cdot L_{\text{lsLM}} = 18$ lsLM subsections. Note, that NRMSE combines NMBE and NRSD as described in Eq. (11).

axons observed in Section 3.2 (using estimates $\hat{r}_{\text{eff,small}}$), all errors were increased, i.e. the maximum NRMSE increased from 1.1 % to 1.8 %. However, all errors, remained much smaller than corresponding errors of *medium-sized* (maximum NRMSE: 6.8 %) or *large* (maximum NRMSE: 7.5 %) axons observed in Section 3.2.

Appendix B. Error of $\hat{r}_{\text{eff}}$ for minor axis approximations of axon radii

Throughout the manuscript, we used the circular equivalent approximation for individual axon radii (see Section 2.1). Here, we assessed the error of $\hat{r}_{\text{eff}}$ with individual axon radii approximated from minor axes of ellipsoids fitted to axonal areas (short: minor axis radii).

Methods To fit ellipsoids to axonal areas, we used an implementation (van der Walt et al., 2014) of the non-iterative least-squares approach described in Halir and Flusser, 1998. Axon radii were then computed by halving the length of minor axes of fitted ellipsoids. The assessment of the error of $\hat{r}_{\text{eff}}$ for minor axis radii was carried out analogously to the procedure described in Section 2.8.2. In particular, we used the same procedure to generate reference axon radii distributions $\mathcal{E}^{(n,l)}_{\text{eff}}(s)$. However, to compute $\hat{r}^{(n,l)}_{\text{eff}}$ and $\rho^{(n,l)}_{\text{eff}}(s)$, we determined minor axis radii for associated axons of $\mathcal{E}^{(n,l)}_{\text{eff}}(s)$ and the axon radii distribution predicted on $S^{(n,l)}_{\text{lsLM}}$.

Results Fig. A.15 shows accuracy (NRMSE(s); see Eq. (8)), bias (NMBE(s); see Eq. (9)) and the residual standard deviation (NRSD(s); see Eq. (10)) of $\hat{r}_{\text{eff}}$ with respect to reference values $\rho_{\text{eff}}(s)$ based on minor axis approximations of individual axon radii. The NRMSE was between 5.8 % to 8.1 % (see Fig. A.15a). The NMBE had a smaller maximum absolute value and a larger dynamic range (-3.4% to 5.2 %) than the NRSD (5.7 % to 6.2 %) (see Fig. A.15b-c). When compared to the errors for circular equivalent-based $\hat{r}_{\text{eff}}$ observed in Section 3.2, the maximum NRMSE and the maximum absolute NMBE were comparable (NRMSE: 8.1 % versus 8.5 %; NMBE: 5.2 % versus 4.8 %), whereas the maximum NRSD was slightly lower (6.0 % versus 7.3 %). However, the dynamic ranges of errors for minor axis-based $\hat{r}_{\text{eff}}$ were similar to those for circular equivalent-based $\hat{r}_{\text{eff}}$.

Appendix C. Error of $\hat{r}_{\text{eff}}$ in the short-pulse limit

The MRI-visible, effective mean radius can be estimated from the intra-axonal signal and is determined by the pulse-length of the specific sequence. Throughout the manuscript, we used the definition of the effective radius in the wide-pulse limit as defined in Eq. (2). In the short-pulse limit,

$$r_{\text{eff,SP}} = \sqrt{\sum_{k=1}^{K} w_{\text{eff,SP},(k)} \cdot r_{(k)}} \text{with}$$
$$w_{\text{eff,SP},(k)} = \frac{n_{(k)}}{B} \cdot \frac{r^3_{(k)}}{\frac{1}{B}\sum_{j=1}^{K} n_{(j)} r^2_{(j)}} \tag{C.1}$$

can be analogously defined (Burcaw et al., 2015; Sepehrband et al., 2016). To assess the error of estimates of $r_{\text{eff,SP}}$, i.e., $\hat{r}_{\text{eff,SP}}$, we repeated the analysis in Section 2.8.2 for $\hat{r}_{\text{eff,SP}}$ with respect to reference values $\rho_{\text{eff,SP}}(s)$ computed from reference axon radii distributions $\mathcal{E}_{\text{eff,SP}}(s)$.

Results Fig. A.16 shows accuracy (NRMSE(s); see Eq. (8)), bias (NMBE(s); see Eq. (9)) and the residual standard deviation (NRSD(s); see Eq. (10)) of $\hat{r}_{\text{eff}}$ with respect to reference values $\rho_{\text{eff}}(s)$. The NRMSE was between 5.3 % to 10.6 % (see Fig. A.16a). The NMBE had a larger absolute maximum value and a much larger dynamic range (-2.5 % to 9.2 %) than the NRSD (5.6 % to 5.9 %) (see Fig. A.16b-c). When compared to the errors for wide-pulse $\hat{r}_{\text{eff}}$ observed in Section 3.2, maximum NRMSE and maximum absolute NMBE were higher (NRMSE: 10.6 % versus 8.5 %; NMBE: 9.2 % versus 4.8 %), whereas the maximum NRSD was lower (5.9 % versus 7.3 %). Furthermore, the dynamic ranges of both NRMSE and NMBE for short-pulse estimates $\hat{r}_{\text{eff,SP}}$ were higher than those for wide-pulse estimates $\hat{r}_{\text{eff}}$. Thus, the s-dependent error introduced by scaling of the *bulk* of $\mathcal{E}_{\text{eff}}(s)$ has larger impact for short-pulse $r_{\text{eff,SP}}$ due to the weaker *tail*-weighting of $r_{\text{eff,SP}}$.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.neuroimage.2022.118906

References

Abdollahzadeh, A., Belevich, I., Jokitalo, E., Sierra, A., Tohka, J., 2021. DeepACSON automated segmentation of white matter in 3D electron microscopy. Communications Biology 4 (1), 1–14. doi:10.1038/s42003-021-01699-w.

Abdollahzadeh, A., Belevich, I., Jokitalo, E., Tohka, J., Sierra, A., 2019. Automated 3D Axonal Morphometry of White Matter. Sci Rep 9 (1), 6084. doi:10.1038/s41598-019-42648-2.

Aboitiz, F., Scheibel, A.B., Fisher, R.S., Zaidel, E., 1992. Fiber composition of the human corpus callosum. Brain Research 598 (1), 143–153. doi:10.1016/0006-8993(92)90178-C.

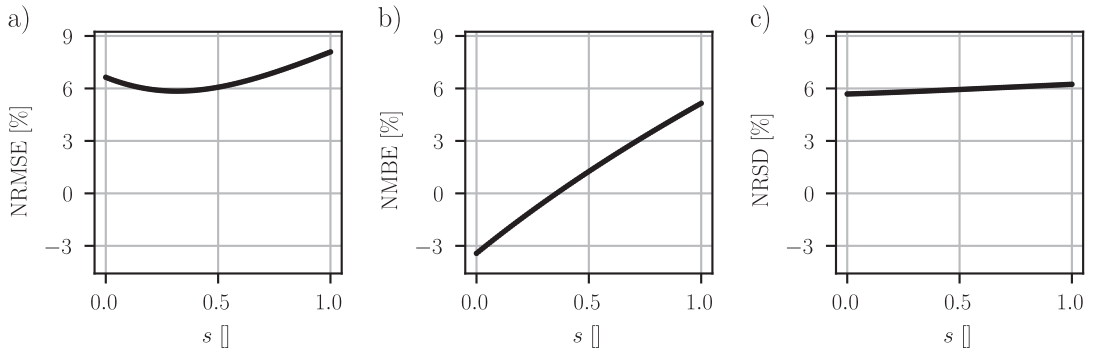

Fig. A.15. Error of estimated effective axon radii based on minor axis axon radii approximations. Depicted are three different error metrics of estimates $\hat{r}_{\text{eff}}$ of the MRI-visible, effective axon radius r_{eff} with respect to reference values $\rho_{\text{eff}}(s)$: (a) the accuracy as evaluated by NRMSE (see Eq. (8)); (b) the bias as evaluated by NMBE (see Eq. (9)); (c) the residual standard deviation as evaluated by NRSD (see Eq. (10)). Both $\hat{r}_{\text{eff}}$ and $\rho_{\text{eff}}(s)$ were estimated using minor approximations of individual axon radii. All errors (a-c) are shown as a function of a sweep variable s, which determined the scaling of the *bulk* of reference axon radii distributions $\mathcal{E}_{\text{eff}}(s)$. These reference axon radii distributions $\mathcal{E}_{\text{eff}}(s)$ were used to compute reference values $\rho_{\text{eff}}(s)$. Here, $s = 0$ and $s = 1$ correspond to using lower ($f_\downarrow$) and upper ($f_\uparrow$) bounds of the scaling factor (see Eq. (15)). Error metrics were evaluated over $N \cdot L_{\text{lsLM}} = 18$ large-scale light microscopy (lsLM) subsections. Note, that NRMSE combines NMBE and NRSD as described in Eq. (11).

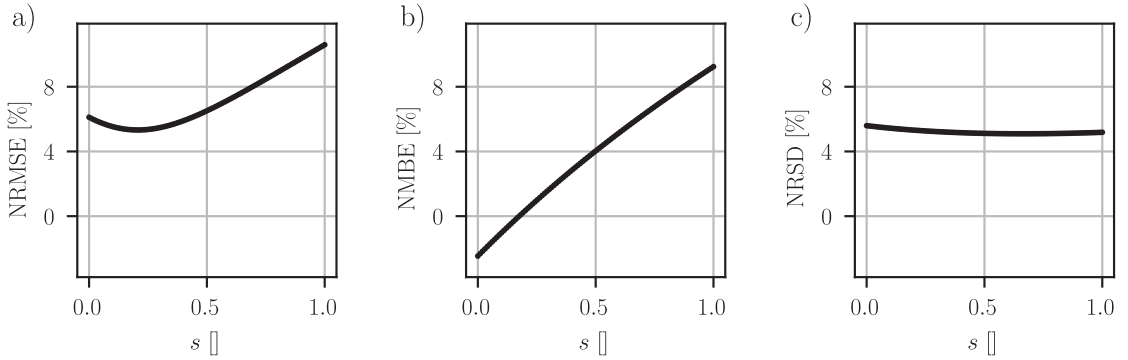

Fig. A.16. Error of estimated effective axon radii in the short-pulse limit. Depicted are three different error metrics of estimates $\hat{r}_{\text{eff,SP}}$ of the MRI-visible, effective axon radius in the short-pulse limit $r_{\text{eff,SP}}$ with respect to reference values $\rho_{\text{eff,SP}}(s)$: (a) the accuracy as evaluated by NRMSE (see Eq. (8)); (b) the bias as evaluated by NMBE (see Eq. (9)); (c) the residual standard deviation as evaluated by NRSD (see Eq. (10)). All errors (a-c) are shown as a function of a sweep variable s, which determined the scaling of the *bulk* of reference axon radii distributions $\mathcal{E}_{\text{eff,SP}}(s)$. These reference axon radii distributions $\mathcal{E}_{\text{eff,SP}}(s)$ were used to compute reference values $\rho_{\text{eff,SP}}(s)$. Here, $s = 0$ and $s = 1$ correspond to using lower ($f_\downarrow$) and upper ($f_\uparrow$) bounds of the scaling factor (see Eq. (15)). Error metrics were evaluated over $N \cdot L_{\text{lsLM}} = 18$ large-scale light microscopy (lsLM) subsections. Note, that NRMSE combines NMBE and NRSD as described in Eq. (11).

Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M., 2019. Optuna: A Next-generation Hyperparameter Optimization Framework. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Association for Computing Machinery, New York, NY, USA, pp. 2623–2631.

Alexander, D.C., Dyrby, T.B., Nilsson, M., Zhang, H., 2019. Imaging brain microstructure with diffusion MRI: Practicality and applications. NMR Biomed 32 (4), e3841. doi:10.1002/nbm.3841.

Alexander, D.C., Hubbard, P.L., Hall, M.G., Moore, E.A., Ptito, M., Parker, G.J.M., Dyrby, T.B., 2010. Orientationally invariant indices of axon diameter and density from diffusion MRI. NeuroImage 52 (4), 1374–1389. doi:10.1016/j.neuroimage.2010.05.043.

Andersson, M., Kjer, H.M., Rafael-Patino, J., Pacureanu, A., Pakkenberg, B., Thiran, J.-P., Ptito, M., Bech, M., Bjorholm Dahl, A., Andersen Dahl, V., Dyrby, T.B., 2020. Axon morphology is modulated by the local environment and impacts the noninvasive investigation of its structure–function relationship. Proc Natl Acad Sci USA 117 (52), 33649–33659. doi:10.1073/pnas.2012533117.

Assaf, Y., Blumenfeld-Katzir, T., Yovel, Y., Basser, P.J., 2008. AxCaliber: A Method for Measuring Axon Diameter Distribution from Diffusion MRI. Magn Reson Med 59 (6), 1347–1354. doi:10.1002/mrm.21577.

Berman, M., Triki, A.R., Blaschko, M.B., 2018. The Lovász-Softmax Loss: A Tractable Surrogate for the Optimization of the Intersection-Over-Union Measure in Neural Networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 4413–4421.

Biedenbach, M.A., DeVito, J.L., Brown, A.C., 1986. Pyramidal tract of the cat: Axon size and morphology. Exp Brain Res 61 (2), 303–310. doi:10.1007/BF00239520.

Yushkevich, P.A., Piven, J., Hazlett, H.C., Smith, R.G., Ho, S., Gee, J.C., Gerig, G., 2006. User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. NeuroImage 31 (3), 1116–1128. doi:10.1016/j.neuroimage.2006.01.015.

Zaimi, A., Wabartha, M., Herman, V., Antonsanti, P.-L., Perone, C.S., Cohen-Adad, J., 2018. AxonDeepSeg: Automatic axon and myelin segmentation from microscopy data using convolutional neural networks. Sci Rep 8 (1), 3816. doi:10.1038/s41598-018-22181-4.

Byfield, P., 2021. StainTools. https://github.com/Peter554/StainTools.

Burcaw, L.M., Fieremans, E., Novikov, D.S., 2015. Mesoscopic structure of neuronal tracts from time-dependent diffusion. NeuroImage 114, 18–37. doi:10.1016/j.neuroimage.2015.03.061.

Caminiti, R., Ghaziri, H., Galuske, R., Hof, P.R., Innocenti, G.M., 2009. Evolution amplified processing with temporally dispersed slow neuronal connectivity in primates. Proc. Natl. Acad. Sci. U.S.A. 106 (46), 19551–19556. doi:10.1073/pnas.0907655106.

Deng, J., Dong, W., Socher, R., Li, L., Li, K., Fei-Fei, L., et al., 2009. ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 248–255.

Drakesmith, M., Harms, R., Rudrapatna, S.U., Parker, G.D., Evans, C.J., Jones, D.K., 2019. Estimating axon conduction velocity in vivo from microstructural MRI. NeuroImage 203, 116186. doi:10.1016/j.neuroimage.2019.116186.

Falcon, W. A., et al., 2019. PyTorch Lightning. URL https://github.com/PyTorchLightning/pytorch-lightning.

Halir, R, Flusser, J., 1998. Numerically stable direct least squares fitting of ellipses (Vol. 98„ 125–132.

Horowitz, A., Barazany, D., Tavor, I., Bernstein, M., Yovel, G., Assaf, Y., 2015. In vivo correlation between axon diameter and conduction velocity in the human brain. Brain Struct Funct 220 (3), 1777–1788. doi:10.1007/s00429-014-0871-0.

Jung, A. B., Wada, K., Crall, J., Tanaka, S., Graving, J., Reinders, C., Yadav, S., Banerjee, J., Vecsei, G., Kraft, A., Rui, Z., Borovec, J., Vallentin, C., Zhydenko, S., Pfeiffer, K., Cook, B., Fernández, I., De Rainville, F. c.-M., Weng, C.-H., Ayala-Acevedo, A., Meudec, R., Laporte, M., et al., 2021. Imgaug. URL https://github.com/aleju/imgaug.

Innocenti, G.M., Caminiti, R., Aboitiz, F., 2015. Comments on the paper by Horowitz et al. (2014). Brain Struct Funct 220 (3), 1789–1790. doi:10.1007/s00429-014-0974-7.

Kakkar, L.S., Bennett, O.F., Siow, B., Richardson, S., Ianuş, A., Quick, T., Atkinson, D.,

Phillips, J.B., Drobnjak, I., 2018. Low frequency oscillating gradient spin-echo sequences improve sensitivity to axon diameter: An experimental study in viable nerve tissue. Neuroimage 182, 314–328. doi:10.1016/j.neuroimage.2017.07.060.

Drobnjak, I., Zhang, H., Ianuş, A., Kaden, E., Alexander, D.C., 2016. PGSE, OGSE, and sensitivity to axon diameter in diffusion MRI: Insight from a simulation study. Magn Reson Med 75 (2), 688–700. doi:10.1002/mrm.25631.

Graf von Keyserlingk, D., Schramm, U., 1984. Diameter of axons and thickness of myelin sheaths of the pyramidal tract fibres in the adult human medullary pyramid. Anat Anz 157 (2), 97–111.

Karimi, D., Dou, H., Warfield, S.K., Gholipour, A., 2020. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. Medical Image Analysis 65, 101759. doi:10.1016/j.media.2020.101759.

Lee, H.-H., Yaros, K., Veraart, J., Pathan, J.L., Liang, F.-X., Kim, S.G., Novikov, D.S., Fieremans, E., 2019. Along-axon diameter variation and axonal orientation dispersion revealed with 3D electron microscopy: Implications for quantifying brain white matter microstructure with histology and diffusion MRI. Brain Struct Funct 224 (4), 1469–1488. doi:10.1007/s00429-019-01844-6.

Leenen, L., Meek, J., Nieuwenhuys, R., 1982. Unmyelinated fibers in the pyramidal tract of the rat: A new view. Brain Research 246 (2), 297–301. doi:10.1016/0006-8993(82)91179-9.

Liewald, D., Miller, R., Logothetis, N., Wagner, H.-J., Schüz, A., 2014. Distribution of axon diameters in cortical white matter: An electron-microscopic study on three human brains and a macaque. Biol Cybern 108 (5), 541–557. doi:10.1007/s00422-014-0626-2.

Macenko, M., Niethammer, M., Marron, J.S., Borland, D., Woosley, J.T., Guan, X., Schmitt, C., Thomas, N.E., 2009. A method for normalizing histology slides for quantitative analysis. In: 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro. IEEE, Boston, MA, USA, pp. 1107–1110.

Mordhorst, L., Morozova, M., Papazoglou, S., Fricke, B., Oeschger, J.M., Rusch, H., Jäger, C., Morawski, M., Weiskopf, N., Mohammadi, S., 2021. Human Axon Radii Estimation at MRI Scale. In: Proceedings of the 2021 German Workshop on Medical Image Computing. Springer, Wiesbaden, pp. 180–185.

Nilsson, M., Lasič, S., Drobnjak, I., Topgaard, D., Westin, C.-F., 2017. Resolution limit of cylinder diameter estimation by diffusion MRI: The impact of gradient waveform and orientation dispersion. NMR in Biomedicine 30 (7), e3711. doi:10.1002/nbm.3711.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. Advances in Neural Information Processing Systems 32.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Springer International Publishing, Cham, pp. 234–241. doi:10.1007/978-3-319-24574-4_28.

Roy, A.G., Navab, N., Wachinger, C., 2018. Concurrent Spatial and Channel 'Squeeze & Excitation' in Fully Convolutional Networks. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. Springer International Publishing, Cham, pp. 421–429.

Schmidt, H., Knösche, T.R., 2019. Action potential propagation and synchronisation in myelinated axons. PLOS Computational Biology 15 (10), e1007004. doi:10.1371/journal.pcbi.1007004.

Sepehrband, F., Alexander, D.C., Kurniawan, N.D., Reutens, D.C., Yang, Z., 2016. Towards higher sensitivity and stability of axon diameter estimation with diffusion-weighted MRI. NMR Biomed 29 (3), 293–308. doi:10.1002/nbm.3462.

Stikov, N., Campbell, J.S.W., Stroh, T., Lavelée, M., Frey, S., Novek, J., Nuara, S., Ho, M.-K., Bedell, B.J., Dougherty, R.F., Leppert, I.R., Boudreau, M., Narayanan, S., Duval, T., Cohen-Adad, J., Picard, P.-A., Gasecka, A., Côté, D., Pike, G.B., 2015. In vivo histology of the myelin g-ratio with magnetic resonance imaging. Neuroimage 118, 397–405. doi:10.1016/j.neuroimage.2015.05.023.

Tan, M., Le, Q., et al., 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In: Proceedings of the 36th International Conference on Machine Learning. PMLR, pp. 6105–6114. URL https://proceedings.mlr.press/v97/tan19a.html

The GIMP Development Team,. GIMP - The GNU Image Manipulation Program. URL https://www.gimp.org/.

van der Walt, S., Schönberger, J.L., Nunez-Iglesias, J., Boulogne, F.c., Warner, J.D., Yager, N., Gouillart, E., Yu, T., et al., 2014. Scikit-image: Image processing in Python. PeerJ 2, e453. doi:10.7717/peerj.453.

Veraart, J., Nunes, D., Rudrapatna, U., Fieremans, E., Jones, D.K., Novikov, D.S., Shemesh, N., 2020. Noninvasive quantification of axon radii using diffusion MRI. eLife 9, e49855. doi:10.7554/eLife.49855.

Waxman, S.G., 1980. Determinants of conduction velocity in myelinated nerve fibers. Muscle & Nerve 3 (2), 141–150. doi:10.1002/mus.880030207.

Weiskopf, N., Edwards, L.J., Helms, G., Mohammadi, S., Kirilina, E., 2021. Quantitative magnetic resonance imaging of brain anatomy and in vivo histology. Nat Rev Phys 3 (8), 570–588. doi:10.1038/s42254-021-00326-1.

West, K.L., Kelm, N.D., Carson, R.P., Does, M.D., 2016. A revised model for estimating g-ratio from MRI. Neuroimage 125, 1155–1158. doi:10.1016/j.neuroimage.2015.08.017.

Yakubovskiy, P., 2020. Segmentation_models.pytorch. URL https://github.com/qubvel/segmentation_models.pytorch.

Xu, J., Li, H., Harkins, K.D., Jiang, X., Xie, J., Kang, H., Does, M.D., Gore, J.C., 2014. Mapping mean axon diameter and axonal volume fraction by MRI using temporal diffusion spectroscopy. Neuroimage 103, 10–19. doi:10.1016/j.neuroimage.2014.09.006.