nature machine intelligence

ARTICLES

https://doi.org/10.1038/s42256-021-00379-y

Check for updates

Modality attention and sampling enables deep learning with heterogeneous marker combinations in fluorescence microscopy

Alvaro Gomariz[1,2]✉, Tiziano Portenier[1], Patrick M. Helbling[2], Stephan Isringhausen[2], Ute Suessbier[2], César Nombela-Arrieta[2,4] and Orcun Goksel[1,3,4]

Fluorescence microscopy allows for a detailed inspection of cells, cellular networks and anatomical landmarks by staining with a variety of carefully selected markers visualized as colour channels. Quantitative characterization of structures in acquired images often relies on automatic image analysis methods. Despite the success of deep learning methods in other vision applications, their potential for fluorescence image analysis remains underexploited. One reason lies in the considerable workload required to train accurate models, which are normally specific for a given combination of markers and therefore applicable to a very restricted number of experimental settings. We herein propose 'marker sampling and excite'—a neural network approach with a modality sampling strategy and a novel attention module that together enable (1) flexible training with heterogeneous datasets with combinations of markers and (2) successful utility of learned models on arbitrary subsets of markers prospectively. We show that our single neural network solution performs comparably to an upper bound scenario in which an ensemble of many networks is naively trained for each possible marker combination separately. We also demonstrate the feasibility of this framework in high-throughput biological analysis by revising a recent quantitative characterization of bone-marrow vasculature in three-dimensional confocal microscopy datasets and further confirm the validity of our approach on another substantially different dataset of microvessels in foetal liver tissues. Not only can our work substantially ameliorate the use of deep learning in fluorescence microscopy analysis, but it can also be utilized in other fields with incomplete data acquisitions and missing modalities.

Deep neural networks have been largely successful in many computer vision problems[1,2]; they learn network model parameters in layers to produce feature maps called activations, which in turn arrive at a desired output that is often given as ground truth during a training phase. The learned parameters can then be deployed in an inference phase to make predictions on new input data. Countless advances in this field have occurred in a relatively short timeframe, especially in the use of supervised segmentation where the goal is the semantic partitioning of an input image, with the ground truth typically consisting of pixels that are annotated interactively by experts. For instance, UNet[3] is a well-known deep convolutional neural network (CNN) architecture with proven success on semantic segmentation in various biomedical domains. Nevertheless, some aspects of biological images still pose several practical challenges in the application of deep CNN architectures (hereafter also called models).

Fluorescence-based microscopy is a mainstay technology for the study of living or fixed tissues in biomedical research. It operates by detecting microscopic signals that emanate from inorganic molecules or genetically encoded proteins. Fluorescent dyes are often coupled to antibodies, which target structures or cells of interest within complex samples in a highly specific fashion, a process known as immunostaining. Fluorescent signals are registered and separately encoded as independent image channels due to their distinct spectral properties, thereby allowing the visualization of stained anatomical landmarks of interest. Herein we refer to these channels as markers (they are also called labels in the literature), which are analogous to the acquisition of modalities in other imaging fields, such as the specific imaging sequences for quantifying different tissue properties with ultrasound or magnetic resonance imaging.

The inherent nature of markers in bioimaging studies poses further limitations in the creation of datasets that can be processed by typical CNN frameworks. First, the number of markers that can be simultaneously imaged is limited due to the overlapping spectral profiles of different fluorochromes precluding their reliable separation in individual channels; thus, any detailed characterization of tissues and their pathological perturbations often requires the use of different permutations of a restricted number of markers, which in turn can only provide a limited level of insight into the biological structures studied. Moreover, sample availability is typically a limiting factor, and processing, immunostaining and image acquisition are laborious and time-consuming tasks, especially for whole-organ or three-dimensional imaging techniques. It is therefore not always technically feasible to increase the number of markers, although such additional sources of information would simplify image processing techniques. Finally, the process of immunostaining does not always work consistently, leading to cells and structures that are stained with variable intensity despite using the same markers. Altogether, these issues hinder the generation of a large number of datasets of images stained consistently with all combinations of possible markers. Fluorescence-based microscopy datasets thereby

[1]Computer-assisted Applications in Medicine, Computer Vision Lab, ETH Zurich, Zurich, Switzerland. [2]Department of Medical Oncology and Hematology, University Hospital and University of Zurich, Zurich, Switzerland. [3]Department of Information Technology, Uppsala University, Uppsala, Sweden. [4]These authors contributed equally: César Nombela-Arrieta, Orcun Goksel. ✉e-mail: alvaroeg@ethz.ch

often consist of heterogeneous combinations of markers, and with each combination often being limited in number of samples, applications of deep learning become strongly limited. Furthermore, typical supervised segmentation algorithms allow a trained model for only limited future applicability, that is when the exact same marker combination is used as in training. This limitation leads to the tradeoff that either a separate specific model is trained each time a new combination is desired and data is available, or a small set of intersection of markers is found in the data; either way neglecting large amounts of data and any possibility of using the models later with alternative marker combinations.

In a general image analysis framework, the problem settings above can be referred to as missing modalities; these are somewhat related to multitask learning (a field in which it is studied whether information should be learned jointly or separately[4]) and to domain adaptation (which aims to bring datasets from different sources into a common space to improve generalization performance[5]). It is agreed in both of these fields that using a unique model that shares certain amount of information is advantageous. Despite numerous advances in these fields, the presented missing marker or modality problem is, however, largely unexplored. Synthesis approaches to complete missing data have recently been proposed for both markers[6–8] and time-sequences[9] in fluorescence-based microscopy, as well as for modalities in magnetic resonance imaging[10–13]. Such modality synthesis is cumbersome and potentially suboptimal when the segmentation model could instead encode information across modalities with shared features. Methods that combine different modalities in shared feature spaces are proposed in ref. [14] as heteromodal image segmentation framework (HeMIS), as well as in refs. [15,16]. One would reasonably expect a multimodal network model to behave differently in the existence or absence of a particular modality. Such processes of conditioning the models explicitly are known as attention mechanisms. For example, soft attention mechanisms transform the activations of a model conditioned on the activations themselves[17,18]. Notably, squeeze and excitation[19] and similar modules[20–23] have been very successful and since been integrated into several different network architectures to improve their performance and the interpretability of extracted features.

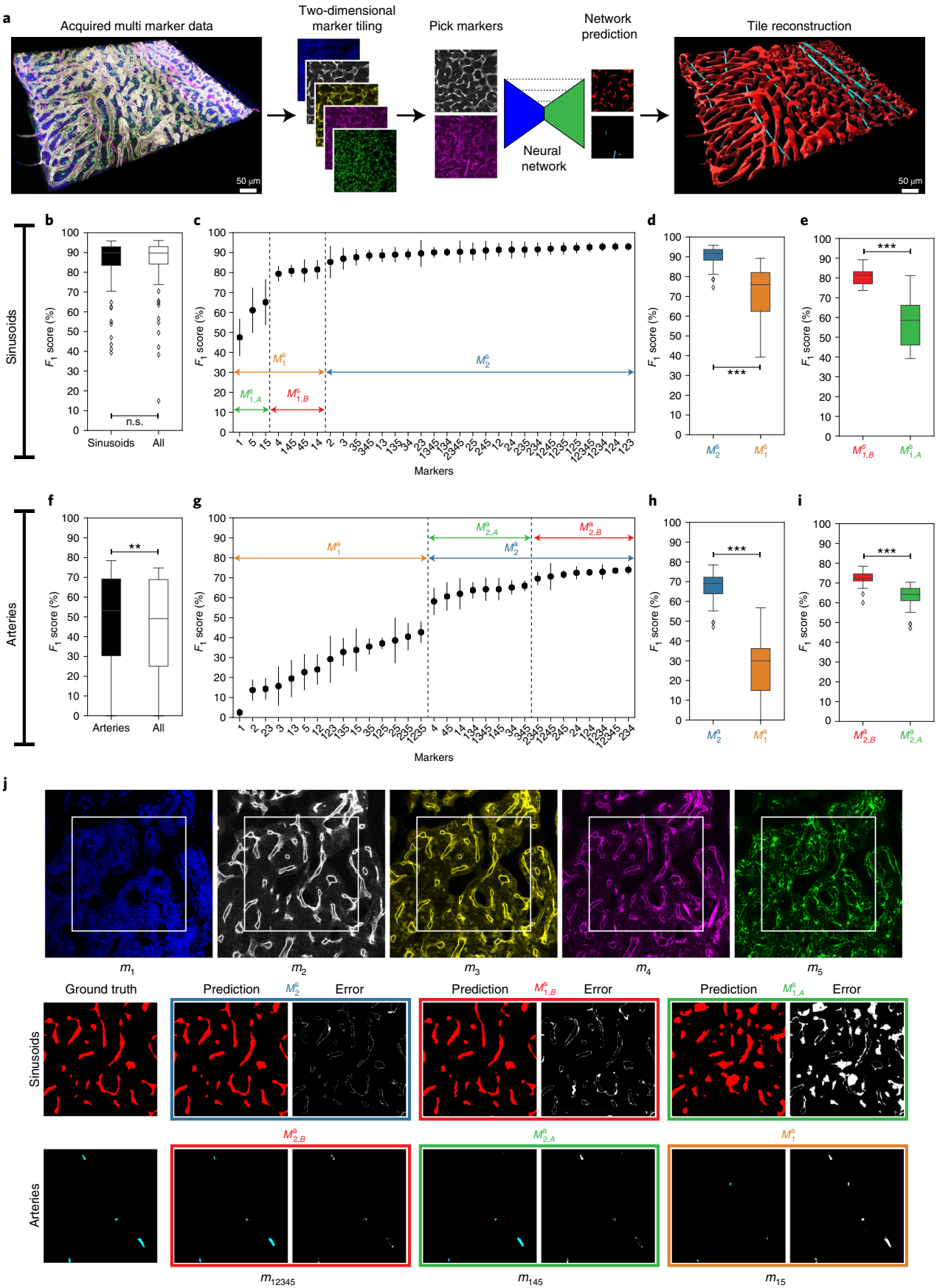
Building on these ideas, we herein devise a method that addresses the fundamental problem of multimodality heterogeneous sets and evaluate this on a three-dimensional microscopy image dataset of a bone-marrow vascular network[24] in which the annotations are divided into two vascular types, namely sinusoids and arteries. The variable size and morphology of the vasculature had hindered precise segmentation and thus a reliable vasculature characterization in earlier works[25–29]. In this experimental setting with five fluorescence-based microscopy imaging markers, we first evaluate multiple conventional baseline scenarios to analyse the effect of each possible marker combination on the performance of semantic (not to be confused with instance) vessel segmentation, which also serves as an upper bound for assessing our proposed methods. We then show that a marker sampling (MS) strategy enables a single CNN model to successfully perform in the presence of any marker combination while outperforming the current state-of-the-art HeMIS. We next present our novel marker excite (ME) soft attention module, which learns how to recalibrate the network activations as a function of available markers, showcasing this by training a single model that performs comparably to the upper bound scenario with an ensemble of 31 separate models individually trained and specialized for each of the marker combinations. We further demonstrate that our model can even outperform the upper bound by leveraging information shared across markers when the training dataset contains such practical variations. We next present a case study on the problem setting of ref. [24] to show the application of our method on an existing practical research question. We finally demonstrate the widespread applicability of our methods by applying them to a largely different and independently generated foetal liver fluorescence-based microscopy dataset, thereby providing the first quantitative characterization of microvascular networks in this organ across different embryonic stages.

Results

Segmenting fluorescence microscopy samples stained with multiple markers. Markers in fluorescence-based microscopy label specific biological structures—their efficient combination enables the visualization of distinct cellular/subcellular components or networks thereof. Manual annotation of these structures as classes is usually possible when the available markers accurately portray them. Meanwhile, image segmentation algorithms can target these classes by employing arbitrary combinations of markers, but their performance will largely vary depending on the combination employed. To be able to achieve such segmentation with different markers, we designed a neural network-based image processing pipeline applicable to tissue-wide fluorescence-based microscopy imaging (illustrated in Fig. 1a with details in the Methods) that processes markers as channels. In this work we employ a dataset with eight large samples divided into a number of two-dimensional patches (Supplementary Table 1). Each patch consists of two ground truth classes (sinusoids and arteries) and five markers (DAPI, endomucin, endoglin, collagen and CXCL12-GFP), denoted as $m_G, G \subseteq \{1,\ldots,5\}$. Hereafter we use multiple subscripts successively to indicate combinations of these markers. All of the CNN results that follow were validated with fourfold cross-validation and claims are made only when the respective null hypothesis can be rejected with a P-value of ≤ 0.05. More details are given in the Methods.

We build an extensive semantic segmentation baseline by training a distinct UNet separately for each of the 31 possible combinations of five markers. These also represent upper bound performance in the following sections given this architecture. We present results for models trained separately for two classes of vessels as, empirically (measured with F_1 score in Fig. 1b,f), this exhibited better performance than training them simultaneously. To analyse marker importance, in Fig. 1c–e,g–i we rank the marker combinations according to their mean F_1 score. Although segmentation accuracy

Fig. 1 | Segmentation of different classes with different UNet models trained with specific marker combinations. a, An illustration of our image segmentation pipeline in the presence of multiple markers. **b**,**f**, Comparisons of F_1 scores ($n=124$) evaluated separately on sinusoids (**b**) and arteries (**f**) for all CNN models (across marker sets and cross-validations) when separately trained to segment the class under evaluation as well as when trained to segment both classes simultaneously (All). n.s., not significant. **c**,**g**, F_1 scores (mean ± s.d.) for models trained with each possible combination of markers, sorted by ascending order of mean values. Each model was evaluated through fourfold cross-validation ($n=4$). We group markers as M_i^c on the basis of which models perform similarly for class c ($c=$ s for sinusoids and $c=$ a for arteries). The subscript i denotes different groups. **d**, For sinusoids, models with either m_2 or m_3 ($M_{2'}^s$, $n=96$) perform better than those without these markers ($M_{1'}^s$, $n=28$). **e**, Among the remaining $M_{1'}^s$ models, those with m_4 ($M_{1,B'}^s$, $n=48$) perform better than those without ($M_{1,A'}^s$, $n=48$). **h**, For arteries, models with m_4 ($M_{2'}^a$, $n=64$) perform better than those without m_4 ($M_{1'}^a$, $n=60$). **i**, Among $M_{2'}^a$, the ones with m_2 ($M_{2,B'}^a$, $n=32$) perform better than those without ($M_{2,A'}^a$, $n=32$). **j**, Visualization of the different markers employed (upper row) and the ground truth and segmentation predictions for models using sample marker combinations as examples of different aforementioned groups. The white squares within the marker images depict the size of their corresponding segmentation images. Error figures show false positive and false negative pixels. Significance is indicated with P-values of ≤ 0.05 (*), ≤ 0.01 (**) and ≤ 0.001 (***).


a
Acquired multi marker data
Two-dimensional marker tiling
Pick markers
Network prediction
Neural network
Tile reconstruction
50 µm
50 µm
Sinusoids
Arteries
b
c
d
e
f
g
h
i
F_1 score (%)
n.s.
Sinusoids
All
Arteries
**
***
Markers
M_1^s
M_2^s
$M_{1,A}^s$
$M_{1,B}^s$
M_1^a
M_2^a
$M_{2,A}^a$
$M_{2,B}^a$
j
m_1
m_2
m_3
m_4
m_5
Ground truth
Prediction
Error
m_{12345}
m_{145}
m_{15}

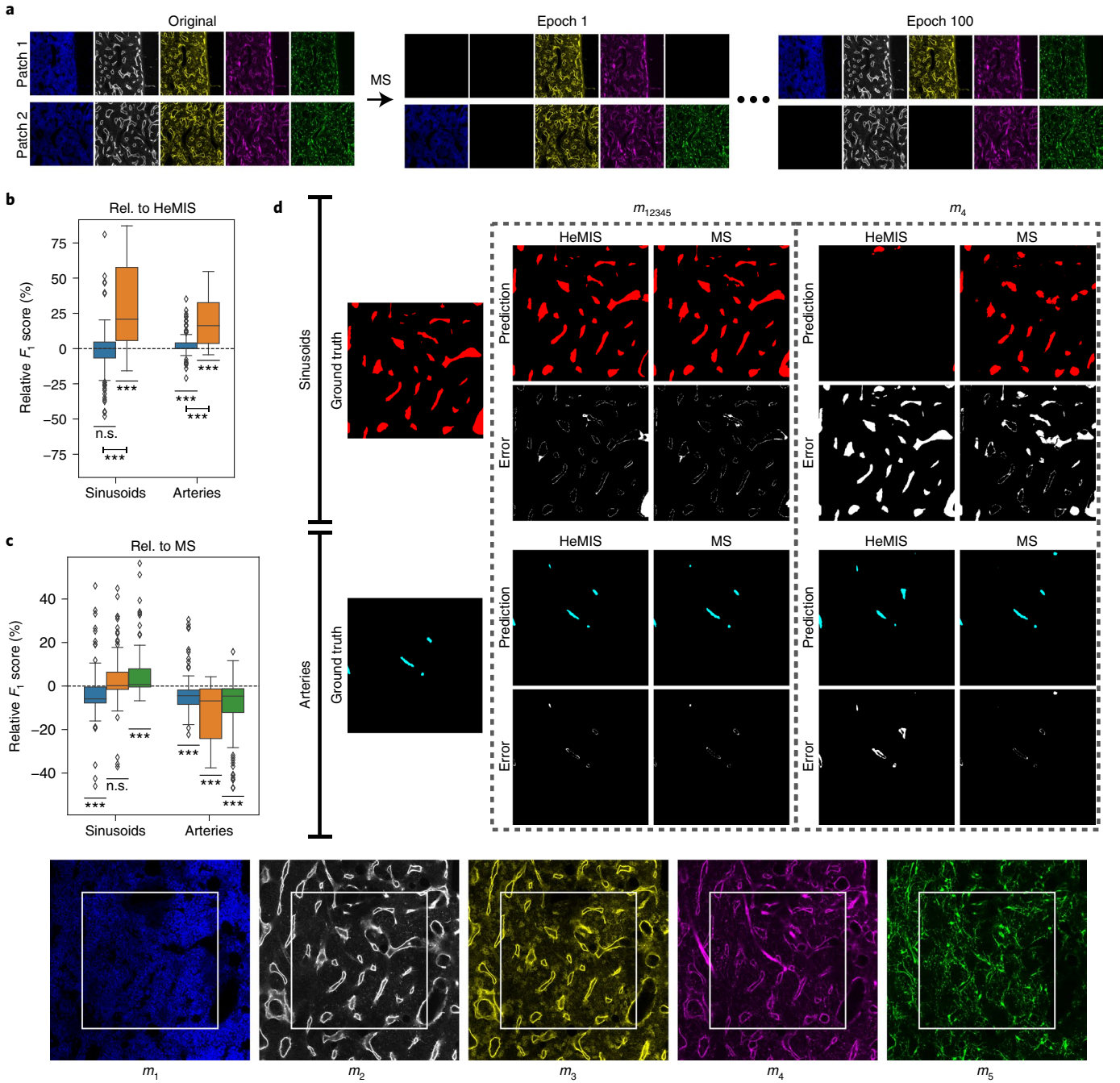


Fig. 2 | Segmentation with a single CNN model for all marker combinations. a, An illustration of the MS strategy. Every time a batch is fed into the network during training, its markers are randomly deleted, that is, set to blank (zero) images. **b**, A comparison between MZ (blue) and MS (orange) with respect to HeMIS. To this end, a paired test is employed to compare the F_1 score of each marker set between the model and HeMIS, which represents all differences ($n=124$) as box plots for sinusoids and arteries. **c**, The same representation of results ($n=124$) is employed to compare HeMIS-MS (blue), MS-DR (orange) and MS-VR (green) with respect to MS. **d**, Visual examples of the differences in segmentation between HeMIS and MS for different marker combinations (shown at the bottom). When all five markers are used, no significant differences are observed; however, the proposed MS performs much better than HeMIS when a subset is used (for instance, m_4 alone). Significance is indicated with respect to the baseline model of each graph (—) or between different models (|—|), with *P*-values of ≤0.05 (*), ≤0.01 (**) and ≤0.001 (***).

increases overall with more markers, it also highly depends on specific markers: Sinusoid segmentation in Fig. 1c–e has higher accuracy when either m_2 or m_3—which specifically label these structures—are present (blue). Without them, m_4 helps with segmentation (red), compared with the least helpful m_1 or m_5 (green). For arteries in Fig. 1g–i, having the arterial-specific m_4 marker seems essential (blue), and results can be further improved by adding m_2 (red). These observations are used later herein to interpret outcomes, for example, when some suitable markers are missing.

Marker sampling for segmenting with missing markers. The strategy described above allows one to determine the best combination of markers to segment a given class. Nevertheless, the need for a fixed marker combination would be restraining in different

applications due to the practical limitation on the number of markers that can be used simultaneously in sample preparation. It is also not a fail-safe method if individual markers fail during acquisition. Moreover, such a naive approach of training a distinct CNN for each combination would require $2^K - 1$ models for K markers (31 models in our study), that is, an exponential growth, which is prohibitive for training CNNs in reasonable time frames. Finally, retraining this many models becomes highly impractical as new samples are added to the training dataset. We propose a marker sampling (MS) approach to address this challenge (illustrated in Fig. 2a), a key component of which is a sampling layer at the input of a segmentation network, herein UNet. During training, we provide the network with all of the available markers while this sampling layer randomly selects a subset to be processed by the proceeding network. This single-model framework can consequently learn to generalize to any subset of markers at inference.

As the accuracy for each marker combination may vary widely (as seen in Fig. 1c,g), we herein report comparative improvements in a paired test manner; that is when we test a hypothesis in relation to an alternative method (indicated in the figures as 'Rel. to'), we calculate the relative metric difference separately for each marker combination experiment and report the distribution and statistical significance of such differences (see the Methods for details). Figure 2b shows a comparison between our MS method and the state-of-the-art HeMIS. We also train our single CNN without MS, that is, the network always uses all markers during training without sampling. When performing inference on a subset of markers, we simply set the missing input to zero; we refer to this simplistic baseline as marker zero (MZ). The results show that our MS model vastly outperforms both HeMIS and MZ, indicating that training a network simply with randomly sampled marker subsets generalizes across the possible input combinations better than the other two approaches. Surprisingly, MZ does not perform much differently to HeMIS, suggesting that the latter does not provide any advantage for this task over a standard UNet architecture. Based on this, learning shared features can be considered comparable with marker-specific representations for this purpose. For completeness, we also evaluate incorporating the MS strategy into the HeMIS architecture (Fig. 2c) and find that there is no improvement compared with UNet-based MS, thus concluding that any advantage originates from our MS strategy.

It could be beneficial to normalize the output of the sampling layer across channels to keep its signal magnitude constant. In fact, MS can be interpreted as a variant of dropout regularization of neural networks[30], in which it is applied only to the input layer, and not only during training, but also during inference (deterministically via available markers). We investigate such relation by training two variants of our model: MS with dropout (MS-DR), which scales the intensities of the available images by a constant dropout ratio, and MS with a variable ratio (MS-VR), which scales the intensities by the ratio of available markers in each sample. Among these, MS is found to be the best-performing model (Fig. 2c), indicating that our improvements cannot be matched by such hand-crafted normalization in the sampling layer.

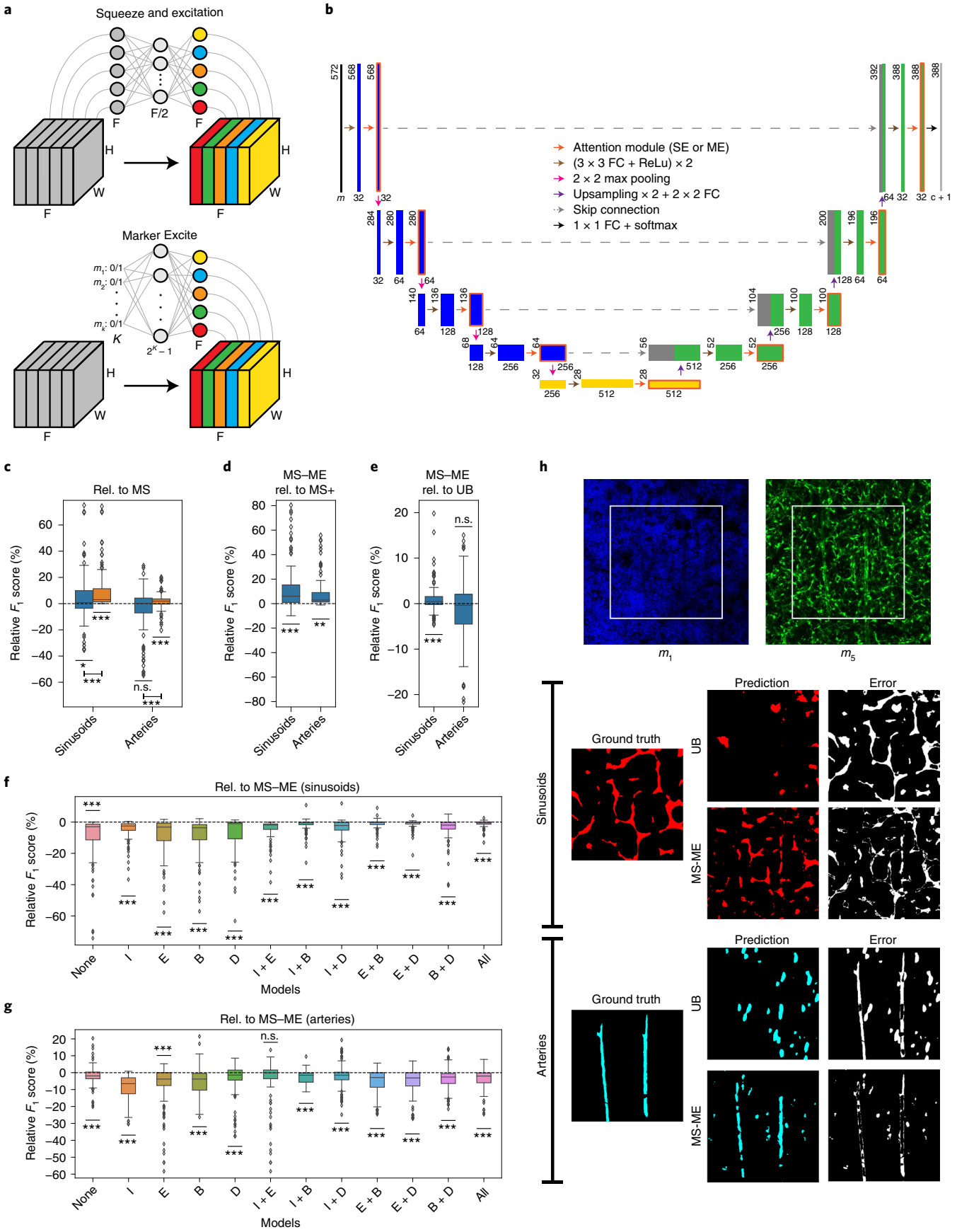
In summary, although HeMIS, MZ and MS all perform well when all markers are present (m_{12345}), our proposed MS generalizes considerably better when some markers are missing at the time of segmentation. Indeed, for instance, with m_4 alone our proposed MS can already segment satisfactorily (Fig. 2d), which the baselines are unable to.

Marker excite for learning attention to markers. As demonstrated above, MS can effectively generalize across marker combinations using a traditional neural network backbone such as UNet to learn shared features for distinct markers. We herein also propose marker excite (ME)—a novel attention module for a further boost by learning a weighting of deep features as a function of available markers. Note that existing attention modules such as squeeze and excitation learn such weighting as a function of layer activations, whereas our ME module instead learns attention to marker availability, which is provided explicitly as an extra input into the network in the form of a one-hot encoded symbolic vector (Fig. 3a). We integrate ME modules at different layers of a UNet as shown in Fig. 3b. We attain our ultimate model MS–ME by integrating our ME approach with the previously presented MS strategy of prepending a sampling layer. We compare this below also using the conventional attention by replacing ME with SE, a baseline referred hereafter as MS–SE.

Figure 3c shows that MS–ME yields improved overall accuracy with respect to MS as well as MS–SE, especially for sinusoids. A major advantage of attention modules is that they increase the model complexity only marginally (MS–ME having merely 0.64% more parameters than the UNet architecture it was based on). To demonstrate that the presented improvement does not originate from inflation of model size, we conducted an additional experiment with a much larger UNet baseline (MS+) of over 20% more parameters and show that the clear improvement from our proposed approach persists (Fig. 3d). We also performed an ablation study by placing ME attention at different network layers (Fig. 3f,g), concluding that placing ME after every convolutional block (outlined in orange in Fig. 3b) is the optimal configuration (namely, MS–ME) for our task. An analysis of the influence of the proposed ME modules reveals that the recalibration effect is stronger for activations of higher resolution (Extended Data Fig. 1).

We also compare our proposed MS–ME model with an upper bound presented in Fig. 1c,g. Note that the upper bound consists of 31 individual models, each of which separately trained and specialized to the availability of a specific marker combination, whereas MS–ME is a single model aiming to achieve well on whatever marker combination may be available at a given time. Figure 3e shows that across all markers combinations our MS–ME is not significantly

Fig. 3 | Results with attention modules. a, An illustration of the attention modules: conventional squeeze and excitation (top) and proposed ME (bottom). These modules learn weights (coloured circles) that are employed to recalibrate the activations (coloured stacks) of a feature map (in grey) with width *W*, height *H* and *F* activations. Squeeze and excitation learns such weights as a function of spatially pooled feature maps, whereas our proposed ME learns from a symbolic one-hot encoded vector, indicating marker availability (see the Methods for details). **b**, Schematics of our network architecture based on UNet; the feature maps calibrated by an attention module are outlined in orange. The numbers below each block indicate the number of activations in that feature map, whereas the numbers to their left specify their width and height. Fully convolutional layers are denoted as FC. For ablation experiments, we refer to parts of the network as encoder (blue), bottleneck (yellow), decoder (green) and input (black). I, input; E, encoder; B, bottleneck; D, decoder. **c**–**e**, Relative F_1 scores ($n = 124$) of MS-SE (blue) and MS–ME (orange) compared pair-wise with MS (**c**); similarly, MS–ME is compared with MS+ (**d**) and an upper bound (**e**). **f**,**g**, An ablation study for placing ME attention at different network layers coloured in **b**. Results for sinusoids (**f**) and arteries (**g**) are presented relative to the proposed MS–ME (equivalent to E + B + D). I, input; E, encoder; B, bottleneck; D, decoder. **h**, A visualization of a sample output with marker combination m_{15} (which is among the worst, according to Fig. 1c,g), where it is seen that our multitask single model MS–ME exhibits superior F_1 scores than the dedicated upper bound thanks to the ME attention module, and by potentially leveraging further information from other marker combinations in the shared network body. Significance is indicated with respect to the baseline model of each graph (—) or between different models (|—|) with *P*-values of ≤0.05 (*), ≤0.01 (**) and ≤0.001 (***).


a
Squeeze and excitation
F
F/2
F
H
W
F
H
W
F
Marker Excite
m_1: 0/1
m_2: 0/1
m_k: 0/1
K
$2^K - 1$
F
H
W
F
H
W
F
b
Attention module (SE or ME)
(3 × 3 FC + ReLu) × 2
2 × 2 max pooling
Upsampling × 2 + 2 × 2 FC
Skip connection
1 × 1 FC + softmax
c
Rel. to MS
Relative F_1 score (%)
Sinusoids
Arteries
d
MS–ME
rel. to MS+
Relative F_1 score (%)
Sinusoids
Arteries
e
MS–ME
rel. to UB
Relative F_1 score (%)
n.s.
Sinusoids
Arteries
f
Rel. to MS–ME (sinusoids)
Relative F_1 score (%)
None
I
E
B
D
I + E
I + B
I + D
E + B
E + D
B + D
All
Models
g
Rel. to MS–ME (arteries)
Relative F_1 score (%)
n.s.
None
I
E
B
D
I + E
I + B
I + D
E + B
E + D
B + D
All
Models
h
m_1
m_5
Sinusoids
Ground truth
Prediction
Error
UB
MS-ME
Arteries
Ground truth
Prediction
Error
UB
MS-ME

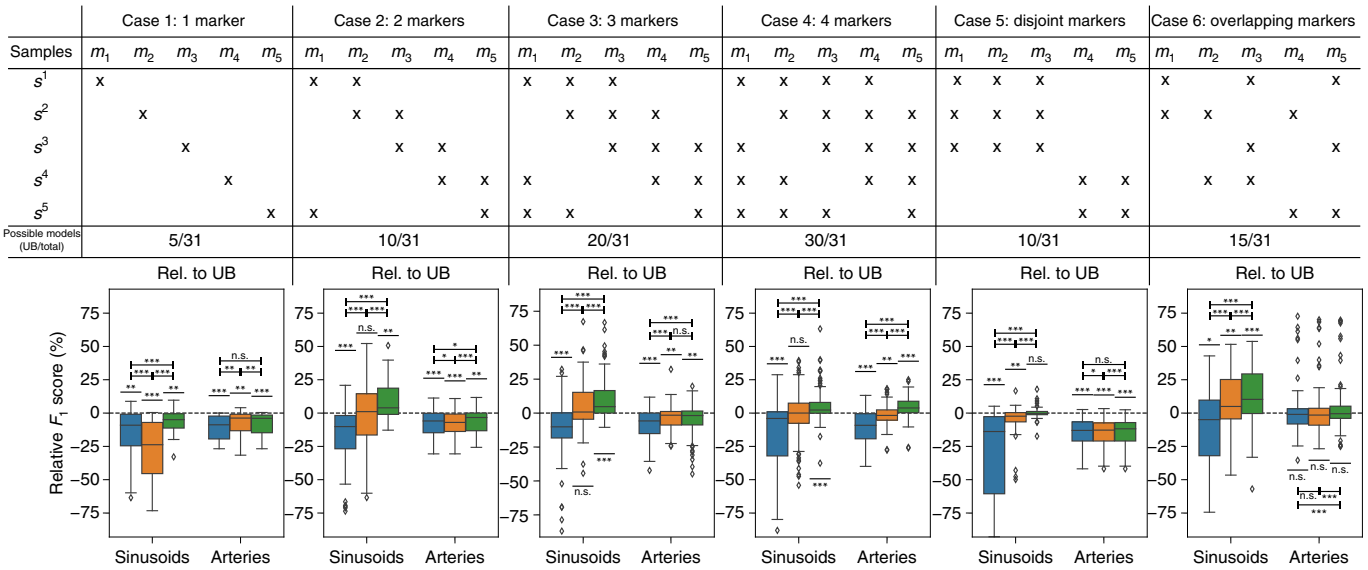


Fig. 4 | Comparison of proposed CNN models to an upper bound when training with heterogeneous combinations of markers. The upper table represents the markers available (denoted by crosses) for each of the five training samples in the different simulated cases. The number of marker combinations for which an upper bound model can be created as compared with our models (as explained in the text) in each case is shown as an extra row (possible models). In the lower table, the three most representative models proposed in this work (MZ (blue), MS (orange) and MS–ME (green)) are compared for the segmentation of sinusoids and arteries across the six different cases. Significance is indicated with respect to the baseline model of each graph (—) or between different models (|—|), with *P*-values of ≤0.05 (*), ≤0.01 (**) and ≤0.001 (***).

different than UB, and even slightly superior for sinusoid segmentation potentially thanks to leveraging additional information shared across combinations. A sample qualitative result is provided in Fig. 3h for a marker combination (m_{15}) known to be suboptimal (based on Fig. 1c,g) with a specialized training of a dedicated upper bound model. Even in this case, MS–ME still performs somewhat satisfactorily, whereas the upper bound fails completely.

Training with heterogeneous panels of markers. We next study the scenario with an incomplete training set, that is with acquisitions of heterogeneous combinations and different number of markers—a typical setting in the field in practice. We used subsets of our fully annotated dataset for this purpose. Training a separate network of all combinatorial test settings would be computationally prohibitive, thus we artificially ablate the data to create a number of case studies (Fig. 4) and thereby emulate various practical or extreme scenarios.

In cases 1 to 4 we studied settings in which training data contain a fixed number of markers per sample, each with a different marker combination. Case 5 simulates a common scenario where two different sets of samples are prepared with two different staining protocols. Case 6 studies a practical scenario with samples available from different studies, which have so far been largely unusable for machine learning due to their heterogeneity.

Although our proposed CNN models are by construction applicable for the above scenarios, any baseline segmentation methods such as UNet require an overlapping set of markers. These baseline models are therefore trainable only for a few intersections of marker combinations, and they discard any potentially useful information outside of such common intersections. For instance, the marker combination m_{23} in case 3 is only trainable using samples s1 or s2, and in case 1 it is not even trainable.

In all cases any marker combination is trainable with our proposed models; thus, for numerous test combinations, our proposed models are inherently superior by design as the baselines cannot accommodate test combinations unseen during training. For many other tests, the baselines would be using a small intersection subset of samples, again at a major disadvantage and also forcing us to retrain a network for each of these subsets. We herein compared our methods (Fig. 4) to the upper bound models only for marker combinations that are achievable by the baseline in training. As this clearly biases results towards the upper bound, we also report herein the ratio of marker combinations subject to such an evaluation, that is, available in the training set. Small ratios in most cases demonstrate that our contribution not only exhibits superior F_1 scores, but also the ability to include and utilize datasets hitherto impossible.

The superiority of MS over MZ and of MS–ME over MS is emphasized with increasing number of markers available for individual samples, as there are more combinations from which information can be leveraged through sampling and attention; thus, MS and MS–ME do not offer any advantage when all samples have only a single marker (case 1). Remarkably, when considering both sinusoids and arteries, MS–ME outperforms the upper bound in cases 3, 4 and 6, where there is more heterogeneity of markers across samples. Thus, best relative gains of MS–ME are observed with more heterogeneity and number of markers across samples. We see that MS–ME not only requires a single model and is applicable for several marker combinations previously unattainable, but it also presents a superior segmentation performance, even over the upper bound for several cases.

A practical application on segmenting bone-marrow vasculature. As a showcase of the proposed methods in practice, we herein revise our segmentation results from ref. [24], where the use of traditional morphological image processing (MIP) algorithms had restricted our analyses both in terms of the number of samples included as well as in the accuracy of quantification. As indicated in Supplementary Table 2, the herein proposed MS–ME (1) permits full automation of the analysis, (2) substantially increases F_1 score by $47.3 \pm 9.6\%$ and (3) almost quadruples the number of samples that can be successfully processed, allowing the inclusion of 35 samples that could not be previously analysed due to marker heterogeneity (Fig. 5b) or insufficient image quality for MIP. Segmentation examples with MS–ME can be seen in Extended Data Fig. 2.

We separately quantify the vascular volumes in two distinct regions of the bone marrow, namely diaphysis and metaphysis (Fig. 5a).

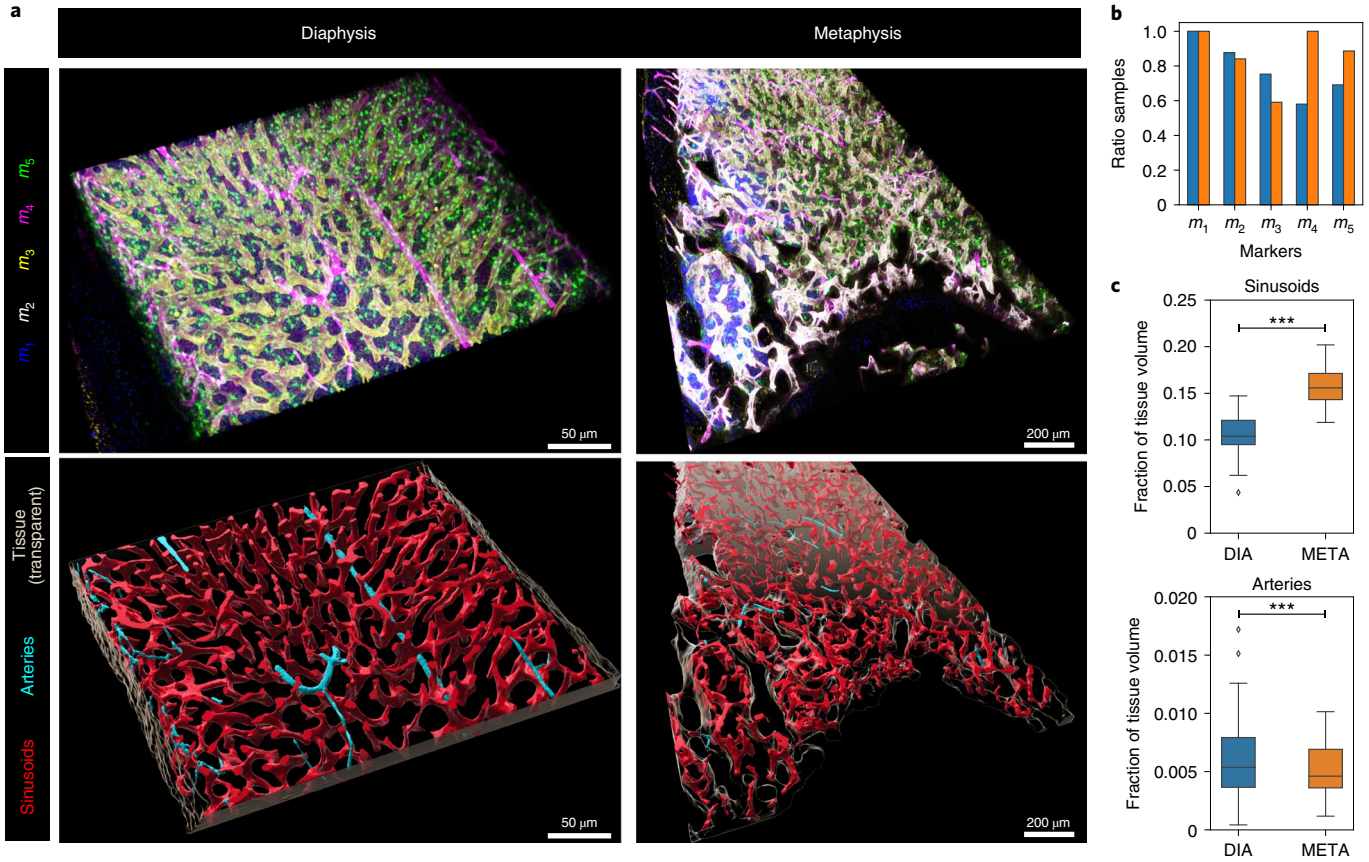


Fig. 5 | Vasculature segmentation and quantification in bone marrow with MS–ME. a, A visual example of bone-marrow images for diaphysis (left) and metaphysis (right). Each sample has a marker combination (displayed as a maximum intensity projection in the top row) that is employed by our MS–ME to segment arteries and sinusoids (bottom row). The tissue mask is shown as a transparent three-dimensional mask. Image visualization with Imaris (Bitplane AG). **b**, The fraction of quantified images that contain each of the markers for both sinusoids (blue) and arteries (orange). The number of images is different due to the quality standards defined in the Methods. **c**, The fraction of bone-marrow volume occupied by the different vascular structures for diaphysis (DIA; $n=47$ for sinusoids and $n=29$ for arteries) and metaphysis (META; $n=14$ for sinusoids and $n=6$ for arteries) regions. Significance is indicated by *P*-values of ≤0.05 (*), ≤0.01 (**) and ≤0.001 (***).

Our revision of the previous analysis[24] indicates that the volume occupied by sinusoids is significantly lower than we previously reported (Extended Data Fig. 3). The results in Fig. 5c also confirm the previously observed trend (with more probability of the tests being correct, given the higher number of samples and superior F_1 score achieved with our method) that they are more abundant in the metaphysis (15.65 ± 2.55%) than in the diaphysis (10.69 ± 2.16%). The differences can be explained by the higher accuracy of our new method. Furthermore, our methods allow for an accurate segmentation of much rarer arterial networks which, to the best of our knowledge, have only been quantified in very limited tissue regions in past studies[31]. We find that these vessels occupy 0.63 ± 0.40% of the volume in the diaphysis and 0.52 ± 0.32% in the metaphysis.

Marker sampling and excite for the characterization of foetal liver vasculature. To further demonstrate the ability to generalize across samples and the applicability of our proposed methods on significantly different biological tissues in which different marker combinations are available, we provide a comprehensive quantitative analysis of a murine foetal liver fluorescence-based microscopy dataset. The segmentation label (vessels) and the five different markers employed for this dataset, largely divert from those employed in bone marrow and are illustrated in Fig. 6a, denoted as $m_G, G \subseteq \{a, \ldots, e\}$ (details in the "Segmentation of foetal liver vasculature" section).

Evaluating the upper bound for each of the possible marker combinations shows that superior F_1 scores are always achieved when m_b is available (Fig. 6b). If all markers are available at training time, our models behave similarly as with the bone-marrow dataset; that is, MS is superior to MZ and MS–ME is superior to MS in terms of the F_1 score evaluated across all possible marker combinations at test time (Fig. 6c). We further evaluate the segmentation performance when the training samples have different combinations of markers by artificially assigning a random marker combination to each of the samples. The results in Fig. 6d again show MS–ME's superior F_1 scores relative to MS and MZ. In this case, MZ has a comparable F_1 score to that of MS, probably because the ablation strategy employed in this setting already creates enough training examples with the different combinations of markers, rendering sampling unnecessary. The F_1 scores of all three models are comparable to that of the upper bound in this setting, corroborating results from some training settings in Fig. 4 for the bone-marrow dataset.

We next follow the analysis strategy of the previous section to demonstrate the potential of our framework in the characterization of different biological samples. We employ MS–ME to segment the liver vasculature in 15 foetal liver samples at two different embryonic stages denominated E13.5 and E17.5 as illustrated in Fig. 6e, each having different marker combinations (Fig. 6f). The results in Fig. 6g reveal isometric vascular growth during these embryonic stages, as no significant differences in vascular occupancy between

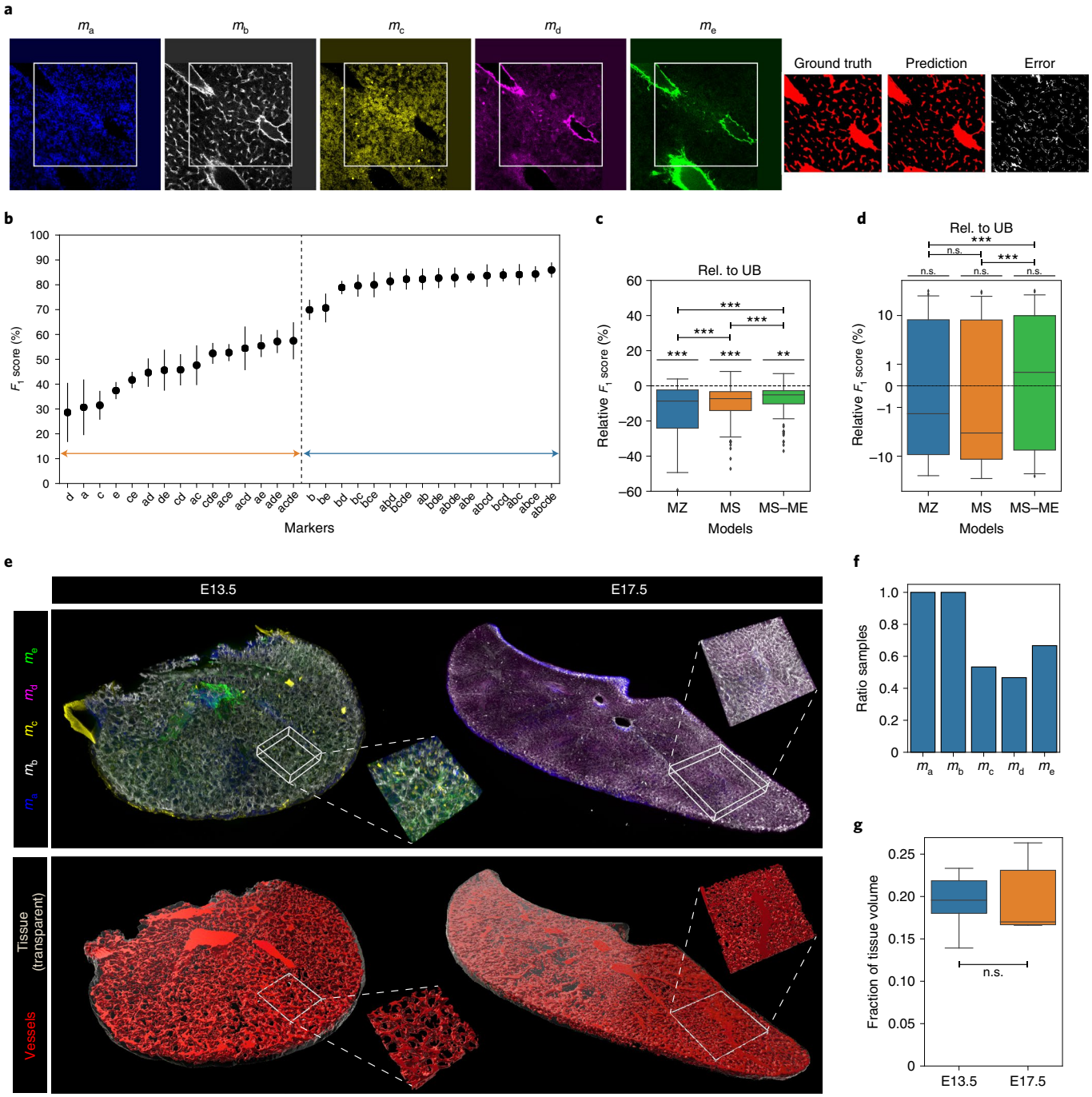


Fig. 6 | Study of our proposed models in the segmentation of foetal liver vasculature with different marker combinations. a, An example of labelled patch that illustrates the markers available in this dataset with their corresponding segmentation output as ground truth, model prediction and error. **b**, The F_1 score achieved when training upper bound models specialized on each of the possible marker combinations. The segmentation quality is observed to be superior when m_b is available (region above blue arrow) as compared to marker combinations without it (region above orange arrow). **c**,**d**, Segmentation F_1 scores of the different proposed models relative to upper bound when all markers are available at training time (**c**) and they are artificially ablated as described in the main text (**d**). A symmetrical logarithmic scale (linear between −2 and 2, and logarithmic elsewhere) is used in **d** to emphasize the differences near the upper bound line. **e**, An example visualization of two foetal liver samples corresponding to embryonic stages E13.5 (left) and E17.5 (right). The maximum intensity projections of the volumes for the available markers are displayed in the top row, whereas their corresponding vessels segmented with MS-ME together with the tissue mask are shown in the bottom row. **f**, The ratio of availability of the five different markers across the samples employed for quantification of vascular occupancy. **g**, The fraction of tissue volume occupied by vasculature as segmented by MS-ME, compared across foetal liver samples in the different embryonic stages E13.5 and E17.5. Significance is indicated with respect to the baseline model of each graph (—) or between different models (|—|), with P-values ≤0.05 (*), ≤0.01 (**) and ≤0.001 (***).

timepoints is observed, with a vascular ratio of 19.4 ± 3.60% across all samples.

Discussion

In this work we formalize the widespread challenges concerning segmentation of structures in fluorescence-based microscopy datasets with different markers, and address them with unprecedented accuracy and efficiency. We first studied the contributions of all marker combinations by naively training a distinct CNN model for each combination. The segmentation results were ranked for each of the possible markers employed, of which even the worst combinations could produce meaningful results. Albeit useful as an upper bound performance, training this number of models is not feasible in practice. We hence proposed a number of methods that allow for a practicable workflow with a single model that can operate with any combination of markers both during inference (application-time) and training.

During our preliminary experiments, the fact that a simple UNet adaptation MZ could perform as well as HeMIS—the state-of-the-art deep learning approach for missing modalities—indicated to us that the solution to this challenge was potentially not in the design of novel network architectures, but rather in adapting suitable training strategies that can inherently address such problem. Accordingly, we devised MS as a sampling strategy, the results of which confirmed our hypothesis, providing a drastic improvement at inference for segmenting different structures with any marker combination. We also tested normalization strategies for the sampled markers, but none helped: MS-DR may have failed because it scales image intensities by a constant factor at training time, which makes sense in the original Dropout idea that samples weights as a regularization technique, but sampling in our missing marker setting intrinsically occurs at inference too. MS-VR, which scales intensities by the ratio of available markers, did not improve the results either. The reason could be that such scaling is reasonable as regularization in the case of marker-independent features, whereas the employed network may learn different features for each marker.

Results with our novel ME method show that attention mechanisms do not have to be limited to layer activations, but using other sources of information as weak labels can also boost outcomes substantially with only a slight increase in model complexity. When our two contributions were combined as MS–ME the best segmentation results were obtained, performing as well as an ensemble of 31 specialized upper bound models. Furthermore, when training samples had different markers, MS–ME was applicable to combinations that were inaccessible with typical CNNs, such as UNet or HeMIS, whereas the segmentation accuracy of our MS–ME surpassed even an upper bound setting, especially when marker heterogeneity was high. This result, as a far-reaching consequence, shows that the solution we originally devised to accommodate missing markers can easily help to improve outcomes even in traditional learning scenarios, potentially by leveraging and incorporating complementary information from multiple sources.

When applied to the bone-marrow vasculature dataset, our MS–ME-based approach showed a dramatic increase in accuracy, speed and the number of samples that could be processed, indicating the importance of chosen image analysis techniques in scientific studies. We therewith increased the probability of correctness for the tests employed in the characterization reported in ref. [24], allowing us to confirm the previously hinted tendency of sinusoids that occupy larger volumes of the metaphysis than the diaphysis. Application of the same analysis pipeline to a dataset depicting foetal liver vasculature with a different marker set further allowed us to quantitatively describe the vascular occupancy in different embryonic stages. Our proposed techniques will be instrumental in imaging studies to accurately study the morphometric and structural features of microvascular networks in hematopoiteic organs, to in turn gain key deeper insight on their described fundamental roles in the pathophysiology of blood-forming tissues. Beyond this, vascular segmentation in large multidimensional datasets remains a widespread and largely unresolved challenge for many groups using advanced microscopy technologies. The methods presented may reveal as extremely powerful tools for studies aimed at uncovering novel mechanisms regulating tissue and cellular dynamics in almost any biological tissue, from animal models, to human samples, to model organoids grown in vitro[25].

Despite MS–ME exhibiting superior segmentation results without compromises to the number of parameters or inference speed (Supplementary Table 4), some limitations exist. For instance, although MS–ME can be applied to previously unseen combinations of markers, quantitative quality assessment is only possible if labelled examples exist featuring the respective combination. Moreover, although not observed in our study, our MS strategy may amplify existing imbalances of markers in the training set, increasing the danger of overfitting to specific combinations.

These challenges reveal a number of interesting questions for future works. For instance, novel uncertainty estimation methods for deep neural networks[32,33] can be employed as proxy for assessment of segmentation quality on previously unseen combinations. Furthermore, although overfitting to specific markers can be naively addressed with a validation set containing as many markers as possible, investigating how to adapt the dropout rate of MS to different experimental settings may lead to solutions with superior results. Finally, our MS–ME model may enable novel possibilities in the context of transfer learning[34] that enables seamless application of trained networks to distinct biological tissues with new markers. For example, fine-tuning only the few parameters in the ME module as opposed to expensive, classical fine-tuning strategies may yield competitive results if the core network already learned suitable general purpose features.

A major challenge and limitation for utilizing deep learning in fluorescence-based microscopy has been the difficulty in establishing standardized staining protocols that would enable more homogeneous marker combinations to train supervised models. With our methods proposed herein, a single model is shown to perform comparably or superior to a number of individual problem-specific models that would be infeasible in practice due to the exponential growth in model parameters and training time with an increasing number of markers. Furthermore, the versatility of our methods enables them to be easily applied to different network architectures for tasks beyond semantic segmentation, such as classification[35,36], detection[37,38] or instance segmentation[39,40]. These contributions can facilitate the sharing and exchange of trained CNNs across labs in the field as well as a faster adoption of neural network solutions in routine laboratory work at, for example, microscopy facilities.

Methods

Dataset composition and notation. The dataset under evaluation (S) consists of eight samples $s^i, i \in \{1, \ldots, 8\}$ and each sample s^i is composed of J patches $p_j^i,\ j \in \{1, \ldots, J\}$. Samples are prepared with a set of different markers $k \in \{1, \ldots, K\}$, where K is the total number of markers. Marker combinations are denoted m_G, where G is a non-empty subset of available markers, that is, $G \subseteq \{1, \ldots, K\}$, denoted in subscript as a successive sequence (for example, m_{12345}) to indicate the combination of all markers for $K = 5$ in our study. Each s_i was manually annotated for C classes $c \in \{1, \ldots, C\}$. Each p_j^i thus consists of $|G|$ input images $x_k \in \mathbb{R}^{h_{\text{in}} \times w_{\text{in}}}$ with $k \in G$ and two output annotations $y_c \in \mathbb{R}^{h_{\text{out}} \times w_{\text{out}}}$ as follows:

$$p_j^i = \{x_k, y_c\}\ \forall k \in G,\ \forall c \in \{1, \ldots, C\}$$

All patches are of the same size, that is, $h_{\text{in}} = w_{\text{in}} = 572$ and $h_{\text{out}} = w_{\text{out}} = 388$.

The complete bone-marrow dataset S (Supplementary Table 1) was prepared using the following five markers: DAPI, which stains DNA in all nuclei; endomucin and endoglin, which have both been reported to have a high specificity for bone-marrow sinusoids; collagen, which mostly stains vessel walls (including sinusoids and arteries) as well as extracellular matrices; and GFP (in the CXCL12-GFP mouse), which is a genetically encoded marker that stains

reticular mesenchymal stromal cells. We use $C=2$ annotated classes: sinusoids and arteries (the latter accounting for both central larger arteries and endosteally located smaller arterioles). The immunostaining and imaging protocols are detailed extensively in ref. [24]. Note that such a dataset with a fixed combination of markers and classes is challenging to obtain in practice. We artificially ablate parts of the data as described in the different experiments to simulate different realistic scenarios of marker and class combinations. The foetal liver dataset is separately described in the segmentation of foetal liver vasculature subsection.

Marker combination strategies. We herein study two fundamentally different settings. First, we study a relatively simpler scenario in which training data are acquired with a consistent, fixed set of available markers $m_{1,\ldots,K}$; at the test time, segmentation is requested for samples only with a subset of trained markers. In this setting we ablate subsets of markers to generate test samples with missing markers, that is, $m_G \forall\, G \subseteq \{1,\ldots,K\}$. Note that such training data with many markers are scarce as they require extensive workforce, effort and budget—not only for the sample preparation and acquisitions, but also for their annotations.

We next study the more common (and more challenging) scenario in which a training set with a fixed set of markers is unavailable. As it is computationally not feasible to evaluate each possible set of markers during training, we propose to evaluate a number of illustrative cases (Fig. 4). To this end, we predefine a set of marker combinations M^{train} such that specific combinations $m_G^i \in M^{\text{train}}$ are assigned to each s^i in the training set. The test set is constructed with all possible combinations of markers as above, that is, $m_G \forall\, G \subseteq \{1,\ldots,K\}$. We use a validation set with samples that contain all the markers to avoid overfitting to a specific marker combination, that is, $m_{1,\ldots,K}$. Furthermore, M^{train} is shuffled for each cross-validation step so that different m_G^i are applied to different samples.

The resulting segmentation performance understandably varies largely across different m_G. Such variation in any metric would be difficult to interpret and to use to compare the models. We therefore report relative changes, that is, the score Q achieved by a model ϕ on m_G relative to a reference model ϕ_{ref} as $Q_{\phi,\phi_{\text{ref}}}(m_G) = Q_\phi(m_G) - Q_{\phi_{\text{ref}}}(m_G)$. Consequently, a resulting vector encodes metric differences of a method with respect to a reference, on a given m_G. To put such results in context, we also report results referenced to an ideal upper bound (UB) model as $Q_{\phi,\phi_{\text{UB}}}(m_G) = Q_\phi(m_G) - Q_{\phi_{\text{UB}}^{m_G}}$, where $\phi_{\text{UB}}^{m_G}$ is a UNet model trained exclusively on m_G. Note that such upper bound reference results $Q_{\phi_{\text{UB}}^{m_G}}$ can only be computed for marker combinations explicitly available in at least a sample, that is, for marker combinations

$$m_G \in M^{\text{UB}}, \quad \text{where} \quad M^{\text{UB}} = \left\{ m_G \mid G \subseteq m^i,\ \forall m^i \in M^{\text{train}} \right\}.$$

By contrast, Q_ϕ with the proposed models can be computed on any marker combination, that is

$$m_G \in M^{\text{total}}, \quad \text{where} \quad M^{\text{total}} = \left\{ m_G \mid G \subseteq \bigcup_{m^i \in M^{\text{train}}} m^i \right\}$$

In other words, the proposed methods can produce results even for m_G that never occur in any training sample, as long as the individual markers are present in at least one sample. The ratio $|M^{\text{UB}}|/|M^{\text{total}}|$ is thus also reported to indicate the maximum possible marker combinations achievable by conventional networks. Lower ratios then emphasize that—aside from requiring comprehensive training sets—standard neural network models for segmentation can only be utilized for a few marker combinations, whereas our proposed model can accommodate all combinations.

Network architectures. The network architectures employed in this work are described below, first the two baseline models adapted from previous work, followed by our proposed models:

- UNet[41] is one of the most commonly used segmentation models of biomedical images. It is a CNN based on an encoder–decoder architecture with skip connections, and targeted to extract features at different resolution levels. We used its standard settings, except for decreasing the number of parameters in each layer by a factor of two, having empirically found that this produces superior results on our dataset.
- HeMIS[14] is the current state-of-the-art for image segmentation with missing modalities. Each marker is processed by separate models (backends) that are subsequently combined with mean and s.d. (abstraction layer). Such feature aggregation using statistical moments allows to seamlessly apply back-propagation regardless of the missing modalities. The merged output is then processed with additional convolutional layers (frontend) to obtain the final segmentation.
- MZ is a UNet where missing markers are padded with zeros, that is, for any m_G, an input image

$$x_k \leftarrow \begin{cases} x_k & \text{if } k \in G \\ 0 & \text{otherwise} \end{cases}.$$

- MS is a UNet prepended with a sampling layer that randomly deletes markers at training time with probability $r_{\text{drop}} \in (0,1)$, that is, $x_k \leftarrow \text{Bern}(r_{\text{drop}})x_k$, where $\text{Bern}(r)$ denotes random sampling from a Bernoulli distribution with probability r; $r_{\text{drop}}=0$ would be equivalent to MZ, whereas $r_{\text{drop}}=1$ is excluded as that would create an input image without markers. Note that no such sampling occurs at inference, but it is implicit to the application strategy in that samples with different markers are expected. We set $r_{\text{drop}}=0.5$ in all our experiments, hence all marker combinations are sampled with equal probability. In-depth analysis revealed that this choice leads to the best overall accuracy (Extended Data Fig. 4a). Note that in this study we consider the scenario where all marker combinations are equally likely to occur at inference, which justifies the choice of setting $r_{\text{drop}}=0.5$; however, if combinations of specific number of markers are expected to occur more frequently at inference, r_{drop} can be adjusted accordingly (Extended Data Fig. 4b), potentially also incorporating marker-specific preferences.
- MS-DR is similar to MS, where the sampling layer is replaced with a dropout layer, that is, $x_k \leftarrow r_{\text{drop}}^{-1}\text{Bern}(r_{\text{drop}})x_k$ only at training time.
- MS-VR is a modification of MS where the sampling layer normalizes the output according to the number of sampled markers, that is, $x_k \leftarrow p^{-1}\text{Bern}(r_{\text{drop}})x_k$ at training time, with p being the ratio of markers chosen by the sampling layer. In this case, the normalization p is also applied at inference time (that is, $x_k \leftarrow p^{-1}x_k$), as it it is a function of the number of utilized markers, instead of an expected probability as in MS-DR.
- HeMIS-MS is a HeMIS network that is prepended with our MS sampling layer. By contrast to zeroing out missing markers as in the UNet counterpart, we herein ablate any backend corresponding to a missing marker, to avoid statistical moment calculations from empty feature maps.
- MS–SE is an MS model with a squeeze and excitation module (as proposed in ref. [19]) in each UNet block (see Fig. 3b). These modules each learn a weighting function $\delta : \mathbb{R}^{1\times F} \rightarrow \mathbb{R}^{1\times F}$ of a feature map X with F activations, where each activation layer $Xf \in \{1,\ldots,F\}$ is weighted as a function of its layer-wise mean value $\langle X_f \rangle$, that is, $X_f \leftarrow \delta_f(\langle X_f \rangle)X_f$; δ is parametrized using two fully connected network layers, respectively, with $F/2$ and F nodes and no biases. The first layer is followed by a ReLU and the second one by a sigmoid function.
- MS–ME instead weights the activations using our proposed ME module (illustrated in Fig. 3a), which learns a weighting from a binary vector V that encodes which of the K markers are the K markers presented to the network as input. Our module thus learns a function $\omega : \{0,1\}^K \rightarrow \mathbb{R}^{1\times F}$ to weight the activations as $X_f \leftarrow \omega_f(V)X_f$. Similar to squeeze and excitation, we parametrize ω using two fully connected layers. The first layer has as many nodes as possible marker combinations (2^K-1) and we employ biases in both layers, which increases the F_1 score (Extended Data Fig. 5). Similarly to MS–SE, we add ME modules after each UNet block (Fig. 3b).

We do not apply batch normalization in any of the models as we empirically found this to decrease the F_1 score, presumably due to the relatively small batch sizes. Supplementary Table 4 shows that the HeMIS architecture has less parameters than UNet, but a higher memory footprint and lower speed. Meanwhile, our MS and ME methods have little to no additional burden, neither in memory space nor in speed.

Training and evaluation details. We implement all CNN models in TensorFlow 2.1 (ref. [42]) and train them on an NVIDIA GeForce GTX TITAN X GPU with 12 GB of VRAM, with three Intel Xeon E5-2640 v3 CPU cores and 40 GB of host RAM. Due to large image sizes and GPU memory limitations, we use a rather small batch size of two. Using the Adam optimizer[43], we minimize the following weighted cross-entropy loss:

$$L = \sum_{c\in C} W_c \sum_{y_j^c \in Y^c} y_j^c \log\left(\overline{y_j^c}\right),$$

where y_j^c is the ground truth annotation for class c, $\overline{y_j^c}$ is its corresponding network prediction, and W_c is a class-specific weight to account for class imbalance. We found that using a weight that is linear in the class cardinality leads to training instabilities, especially with large class imbalance. We therefore instead use a logarithmic weighting as follows:

$$W_c = \log\left(\epsilon + \frac{1}{C|Y^c|}\sum_{i\in\{1,\ldots,C\}} |Y^i|\right),$$

where $|Y^c|$ is the total number of annotated pixels for class c and ϵ is a small number for stability in logarithm calculation.

The results are evaluated individually for each of the classes using F_1 score, as follows:

$$F_1 = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$$

where TP are true positives, FP are false positives, and FN are false negatives.

All networks are trained for 200 epochs, and we choose the one that yields the highest mean F_1 score across all classes on the validation set for evaluation. The evolution of the loss function and the F_1 score is shown in Extended Data Fig. 6. Note that there exists a considerable class imbalance (see Supplementary Table 1), which can explain the inferior F_1 score when classes are targeted simultaneously (Fig. 1b,f) as compared with targeting them individually, which is why we chose the latter approach to conduct our experiments. There has been substantial research in class imbalance[44], but this is not within the scope of this paper, where we focus on effects of markers, that is input image modalities.

The 230 image patches were split according to which of the eight fluorescence-based microscopy samples in Supplementary Table 1 they belonged to, with five samples for training, one for validation and two for testing. For all experiments we employ fourfold cross-validation, by ensuring that all samples are once in test set and that the same sample folds are used for different methods in each experimental setting such that their results can be directly compared.

Tiling strategy for handling large datasets. Wide-tissue fluorescence-based microscopy images typically have very large pixel sizes, exceeding memory capacity of typical GPUs used for deep learning. Furthemore, fluorescence-based microscopy samples are often acquired with different resolutions. To address these problems, we herein adopt a pipeline similar to those included in recent UNet variants[41,45] by decomposing each sample s^i in J patches p_j^i, $j \in \{1, \ldots, J\}$ with constant size and resolution to segment them individually and to subsequently stitch them together. These patches are constructed as follows (illustrated in Extended Data Fig. 7):

1. We resample the resolution of all fluorescence-based microscopy samples to 1 µm pixel size.
2. As convolution operations decrease the spatial extent of the image, we first zero-pad the complete sample with a margin of 92 pixels (half of the difference between input and output patch sizes) required to preserve the original size when stitching the output patches.
3. The input patches are taken with an overlap of 92 pixels (same as the padding) between them so that there is no overlap in the output patches. In this way, we limit padding artefacts to the border of the samples instead of introducing them for each individual patch used in the CNNs.
4. To normalize appearance differences across samples, for each patch we apply Gaussian standardization (zero mean, unit variance) using the mean and s.d. of the respective sample that the patch comes from.
5. Output patches are neither overlapping nor padded. As the sample size does not have to be divisible by the patch size, the last patch of each dimension was taken with as much overlap with the previous one as needed to cover the whole sample. At inference, the result for the overlapping region is averaged for the involved patches. Finally, the slices reconstructed by two-dimensional tiling are inserted in their corresponding axial position to form three-dimensional images.

Bone-marrow quantification pipeline. For the results reported on the characterization of bone-marrow vasculature, we employed additional unlabelled samples from the dataset described in ref. [24]. In that earlier work, several samples had to be discarded as their image quality was not sufficient for the simplistic image-processing tools (MIP) employed therein.

The MS–ME model proposed herein allows for processing of samples with diverse marker combinations; however, as shown in Fig. 1c–e,g–i, not all markers are beneficial for achieving a precise segmentation. To provide an accurate quantification without sacrificing many samples, we employ in our analysis only samples that are stained with the marker sets in M_2^s for sinusoids and M_2^a for arteries, as shown in Fig. 1c,g. Quantified on the samples stained with the best marker combination, the above marker sets guarantee an F_1 score of 75% or higher than that achievable given the best marker combination. With this criterion, 47 samples are employed for the quantification of sinusoids and 29 for arteries.

All employed samples were visually inspected to qualitatively confirm that the segmentation was satisfactory (examples in Extended Data Fig. 2). The F_1 score reported for MS–ME in Supplementary Table 2 is calculated from the results reported in Fig. 3e aggregated over the markers in M_2^s for sinusoids and M_2^a for arteries. The MIP method was evaluated on the same annotations as the other methods proposed in this work (Supplementary Table 1), although not all of these samples were employed for quantification in ref. [24].

Fluorescence-based microscopy images often contain out of tissue regions, which must be discarded for analysis purposes. To this end, in this analysis we account for an extra class denoted as tissue. In several works, segmentation of this class is easily achieved by thresholding of m_1 (DAPI) and some simple morphological image processing[25]. Given the simplicity of the task, we train a separate UNet model that uses only m_1 (available in all of the samples) and achieves an F_1 score of 88.7 ± 6.8 with the same evaluation strategy applied for other models in this work. The reported fraction of bone-marrow volume for sinusoids and arteries is calculated as the ratio of their respective foreground pixels within the newly defined tissue class to the total number of foreground tissue pixels.

Segmentation of foetal liver vasculature. For the foetal liver dataset presented in the "marker sampling and excite for the quantification of foetal liver vasculature" section, we employed the same strategies previously described for the earlier bone-marrow dataset. We herein report the dataset description and technical details that differ from the bone-marrow images.

We denote the five available markers with letters instead of numbers (that is, the complete marker set is denoted as m_{abcde}: DAPI (a), lyve1 (b), Hlf (c), Evi1 or α-catulin (d) and smooth muscle actin (e). The class to segment in this dataset is denoted as vessels. The number of samples, patches and percentage of annotated pixels can be seen in Supplementary Table 3.

The vasculature quantification for embryonic stages E13.5 and E17.5 (displayed in Fig. 6g) is performed on fifteen unlabelled samples: eight samples of E13.5 and seven samples of E17.5. A manually annotated three-dimensional tissue mask is included in these images, which defines the total tissue volume and is employed to disregard out-of-tissue regions in the analysis of vasculature occupation. Note that that the annotated foetal liver vasculature data employed for training and evaluation of segmentation CNNs is a subset of these samples.

Statistical tests. Unless otherwise specified, the two-sided Wilcoxon signed-rank test was employed to assess differences between paired test results; the test is non-parametric to avoid any assumption of normal distribution. As the aim of the models was to perform best across different classes, the employed tests aggregated the data for both sinusoids and arteries in a paired-wise manner. When the data under evaluation was unpaired, we used a two-sided Mann–Whitney U test, which is also non-parametric. Unless otherwise stated, measurements were always taken from distinct samples. Significance was established with a P-value ≤ 0.05. Box plots employed across the different figures consist of a box representig the data quartiles, and whiskers indicating the extent of data points within 1.5 times the interquartile range.

Animal studies. Mice C57BL/6J were purchased from Charlesriver. Mice were analysed at 12–20 weeks of age for the bone-marrow studies. For foetal liver analyses, embryos were extracted from the previously euthanized pregnant dams, at the developmental stages indicated, through a small abdominal incision and further dissected with forceps under a stereo microscope, where single lobes were separated using a surgical scalpel. The lobes were fixed in 2% paraformaldehyde (diluted in PBS) (6 h, 4 °C), washed twice in ice-cold PBS for 5 min and subsequently dehydrated in 30% sucrose in PBS (24–48 h, 4 °C). The liver lobes were placed into base moulds (15 × 15 × 5 mm) and completely covered in OCT medium, snap-frozen using liquid nitrogen and stored at −80 °C until use. Slices were stained following the same protocols described in ref. [24] for marrow tissues. Experimental animals were not randomized and experiments were performed in a non-blinded fashion. Animals were maintained at the animal facility of the University Hospital Zürich and treated in accordance with the guidelines of the Swiss Federal Veterinary Office. Experiments and procedures were approved by the Veterinäramt des Kantons Zürich, Switzerland.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The labelled dataset employed for training and evaluation of the models described is included as a single HDF5 file within a CodeOcean capsule in https://codeocean.com/capsule/8424915/tree/v1 (ref. [46]).

Code availability

The code employed for training the models described in this paper is publicly available on the CodeOcean platform as https://codeocean.com/capsule/8424915/tree/v1 (ref. [46]). This capsule also includes the trained models employed for the different presented figures. MS–ME is also implemented within MiNTiF, our Fiji plugin for ImageJ for user-friendly training and deployment of CNNs by non-experts of deep learning: https://github.com/CAiM-lab/MiNTiF.

Received: 27 August 2020; Accepted: 2 July 2021;
Published online: 9 August 2021

References

1. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
2. Litjens, G. et al. A survey on deep learning in medical image analysis. *Med. Image Anal.* **42**, 60–88 (2017).
3. Ronneberger, O., Fischer, P. & Brox, T. U-net: convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* 234–241 (Springer, 2015).
4. Zhang, Y. & Yang, Q. An overview of multi-task learning. *Natl Sci. Rev.* **5**, 30–43 (2018).
5. Wang, M. & Deng, W. Deep visual domain adaptation: a survey. *Neurocomputing* **312**, 135–153 (2018).

6. Guo, S.-M. et al. Revealing architectural order with quantitative label-free imaging and deep learning. *eLife* **9**, e55502 (2020).
7. Ounkomol, C., Seshamani, S., Maleckar, M. M., Collman, F. & Johnson, G. R. Label-free prediction of three-dimensional fluorescence images from transmitted-light microscopy. *Nat. Methods* **15**, 917–920 (2018).
8. Christiansen, E. M. et al. In silico labeling: predicting fluorescent labels in unlabeled images. *Cell* **173**, 792–803 (2018).
9. Ouyang, W., Aristov, A., Lelek, M., Hao, X. & Zimmer, C. Deep learning massively accelerates super-resolution localization microscopy. *Nat. Biotechnol.* **36**, 460–468 (2018).
10. Li, R. et al. Deep learning based imaging data completion for improved brain disease diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* 305–312 (Springer, 2014).
11. Iglesias, J. E. et al. Is synthesizing MRI contrast useful for inter-modality analysis? In *International Conference on Medical Image Computing and Computer-Assisted Intervention* 631–638 (Springer, 2013).
12. Chartsias, A., Joyce, T., Giuffrida, M. V. & Tsaftaris, S. A. Multimodal MR synthesis via modality-invariant latent representation. *IEEE Trans. Med. Imaging* **37**, 803–814 (2017).
13. Lee, D., Kim, J., Moon, W.-J. & Ye, J. C. Collagan: collaborative gan for missing image data imputation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 2487–2496 (IEEE, 2019).
14. Havaei, M., Guizard, N., Chapados, N. & Bengio, Y. Hemis: hetero-modal image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* 469–477 (Springer, 2016).
15. Dorent, R., Joutard, S., Modat, M., Ourselin, S. & Vercauteren, T. Hetero-modal variational encoder–decoder for joint modality completion and segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* 74–82 (Springer, 2019).
16. Varsavsky, T., Eaton-Rosen, Z., Sudre, C. H., Nachev, P. & Cardoso, M. J. Pimms: permutation invariant multi-modal segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* 201–209 (Springer, 2018).
17. Jaderberg, M. et al. Spatial transformer networks. In *Advances in Neural Information Processing Systems* 2017–2025 (NIPS, 2015).
18. Wang, F. et al. Residual attention network for image classification. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 3156–3164 (2017).
19. Hu, J., Shen, L. & Sun, G. Squeeze-and-excitation networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 7132–7141 (IEEE, 2018).
20. Roy, A. G., Siddiqui, S., Pölsterl, S., Navab, N. & Wachinger, C. 'Squeeze & excite' guided few-shot segmentation of volumetric images. *Med. Image Anal.* **59**, 101587 (2020).
21. Roy, A. G., Navab, N. & Wachinger, C. Recalibrating fully convolutional networks with spatial and channel 'squeeze and excitation' blocks. *IEEE Trans. Medical Imaging* **38**, 540–549 (2018).
22. Rickmann, A.-M., Roy, A. G., Sarasua, I., Navab, N. & Wachinger, C. 'Project & excite' modules for segmentation of volumetric medical scans. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* 39–47 (Springer, 2019).
23. Wang, X., Cai, Z., Gao, D. & Vasconcelos, N. Towards universal object detection by domain attention. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 7289–7298 (IEEE, 2019).
24. Gomariz, A. et al. Quantitative spatial analysis of haematopoiesis-regulating stromal cells in the bone marrow microenvironment by 3D microscopy. *Nat. Commun.* **9**, 1–15 (2018).
25. Gomariz, A., Isringhausen, S., Helbling, P. M. & Nombela-Arrieta, C. Imaging and spatial analysis of hematopoietic stem cell niches. *Ann. N. Y. Acad. Sci.* **1466**, 5–16 (2019).
26. Acar, M. et al. Deep imaging of bone marrow shows non-dividing stem cells are mainly perisinusoidal. *Nature* **526**, 126–130 (2015).
27. Coutu, D. L., Kokkaliaris, K. D., Kunz, L. & Schroeder, T. Three-dimensional map of nonhematopoietic bone and bone-marrow cells and molecules. *Nat. Biotechnol.* **35**, 1202–1210 (2017).
28. Coutu, D. L., Kokkaliaris, K. D., Kunz, L. & Schroeder, T. Multicolor quantitative confocal imaging cytometry. *Nat. Methods* **15**, 39–46 (2018).
29. Christodoulou, C. et al. Live-animal imaging of native haematopoietic stem and progenitor cells. *Nature* **578**, 278–283 (2020).
30. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).
31. Kunisaki, Y. et al. Arteriolar niches maintain haematopoietic stem cell quiescence. *Nature* **502**, 637–643 (2013).
32. Gal, Y. & Ghahramani, Z. *Bayesian Convolutional Neural Networks with Bernoulli Approximate Variational Inference* (ICLR, 2016).
33. Kendall, A. & Gal, Y. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems 30* 5574–5584 (NIPS, 2017).
34. Raghu, M., Zhang, C., Kleinberg, J. & Bengio, S. Transfusion: understanding transfer learning for medical imaging. In *Advances in Neural Information Processing Systems* Vol. 32, 3347–3357 (NIPS, 2019).
35. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25* (eds. Pereira, F., Burges, C. J. C., Bottou, L. & Weinberger, K. Q.) 1097–1105 (NIPS, 2012).
36. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2016).
37. Xie, Y. et al. Efficient and robust cell detection: a structured regression approach. *Med. Image Anal.* **44**, 245–254 (2018).
38. Ren, S., He, K., Girshick, R. & Sun, J. Faster r-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intelligence* **39**, 1137–1149 (2017).
39. He, K., Gkioxari, G., Dollar, P. & Girshick, R. Mask r-CNN. In *2017 IEEE International Conference on Computer Vision (ICCV)* (IEEE, 2017).
40. Yang, L. et al. NuSeT: a deep learning tool for reliably separating and analyzing crowded cells. *PLoS Comput. Biol.* **16**, e1008193 (2020).
41. Falk, T. et al. U-net: deep learning for cell counting, detection, and morphometry. *Nat. Methods* **16**, 67–70 (2019).
42. Abadi, M. et al. TensorFlow: large-scale machine learning on heterogeneous systems (2015); https://www.tensorflow.org/
43. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. In *International Conference on Learning Representations* (ICLR, 2015).
44. Li, Z., Kamnitsas, K. & Glocker, B. Overfitting of neural nets under class imbalance: analysis and improvements for segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* 402–410 (Springer, 2019).
45. Gómez-de Mariscal, E. et al. DeepImageJ: a user-friendly plugin to run deep learning models in ImageJ. Preprint at https://www.biorxiv.org/content/10.1101/799270v2 (2019).
46. Gomariz, A. et al. *Marker Sampling and Excite: Deep Learning with Heterogeneous Marker Combinations in Fluorescence Microscopy* (CodeOcean, 2021); https://codeocean.com/capsule/8424915/tree/v1

Acknowledgements

We thank T. Nagasawa for the Cxcl12-GFP reporter mice, T. Yokomizo for the Hlf-tdTomato mouse strain and M. Kurokawa for the Evi1-GFP mouse strain. This work was enabled by funding from the Hasler Foundation (A.G. and O.G.). We also acknowledge the support from the Swiss National Science Foundation (SNSF) 179116 (O.G.) and 310030_185171 (C.N.-A.), the Swiss Cancer Research Foundation KFS-3986-08-2016 (C.N.-A.), the Clinical Research Priority Program 'ImmunoCure' of the University of Zurich to (C.N.-A.). We extend our thanks to NVIDIA for their GPU support.

Author contributions

A.G. designed and performed the experiments and drafted the manuscript. T.P. and O.G. provided key methodological ideas and input. A.G., P.M.H., S.I. and U.S. conducted data acquisition, manual labelling of the images, and helped interpret bone marrow analysis in the biological context. T.P., O.G. and C.N.-A. critically revised the text. A.G., C.N.-A. and O.G. conceived this project. This work was jointly directed by C.N.-A. for the biological and data acquisition aspects and O.G. for computer-scientific and methodological aspects. All authors discussed the results and provided feedback to the writing.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s42256-021-00379-y.

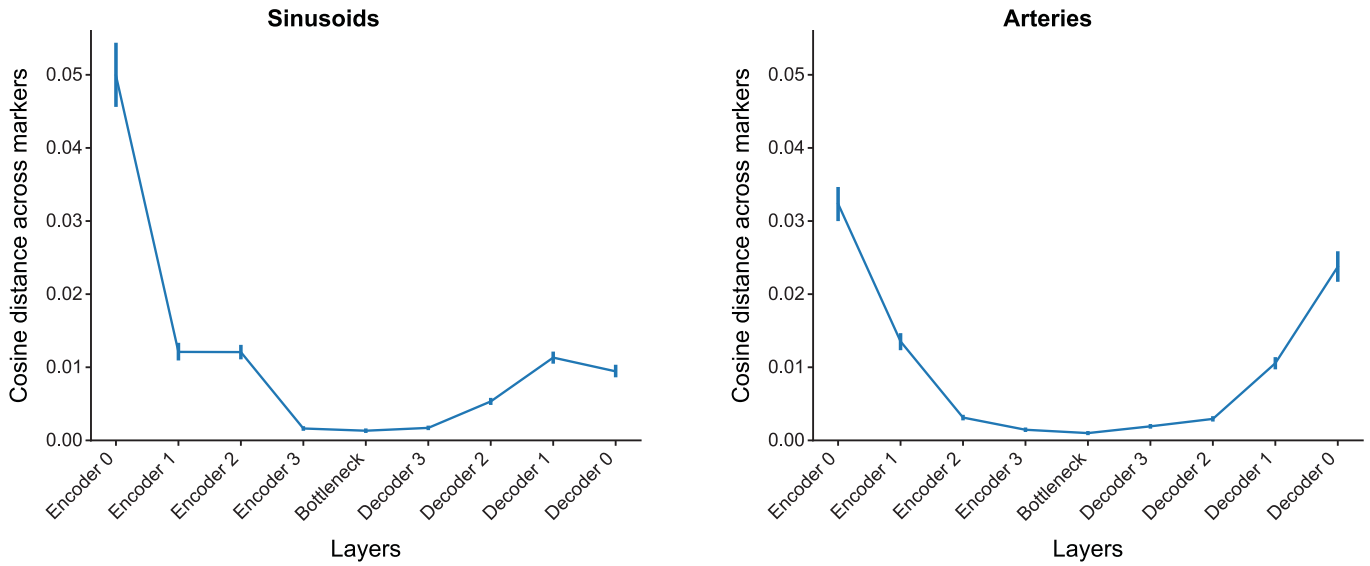
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-021-00379-y.

Correspondence and requests for materials should be addressed to A.G.

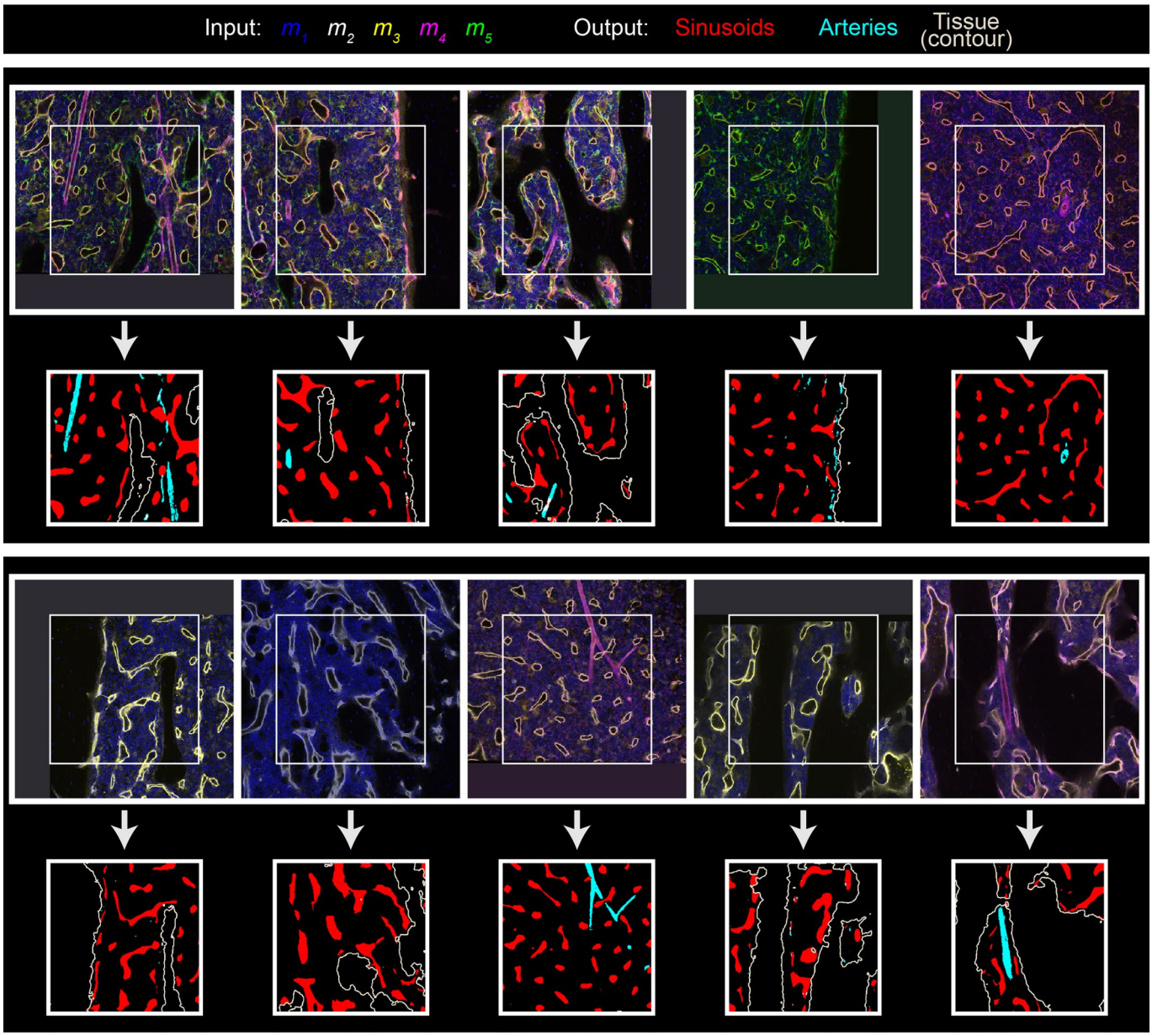
Peer review information *Nature Machine Intelligence* thanks Shalin Mehta and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

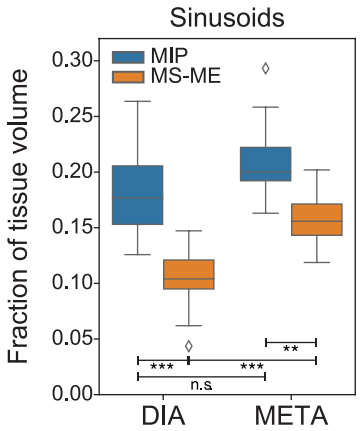
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



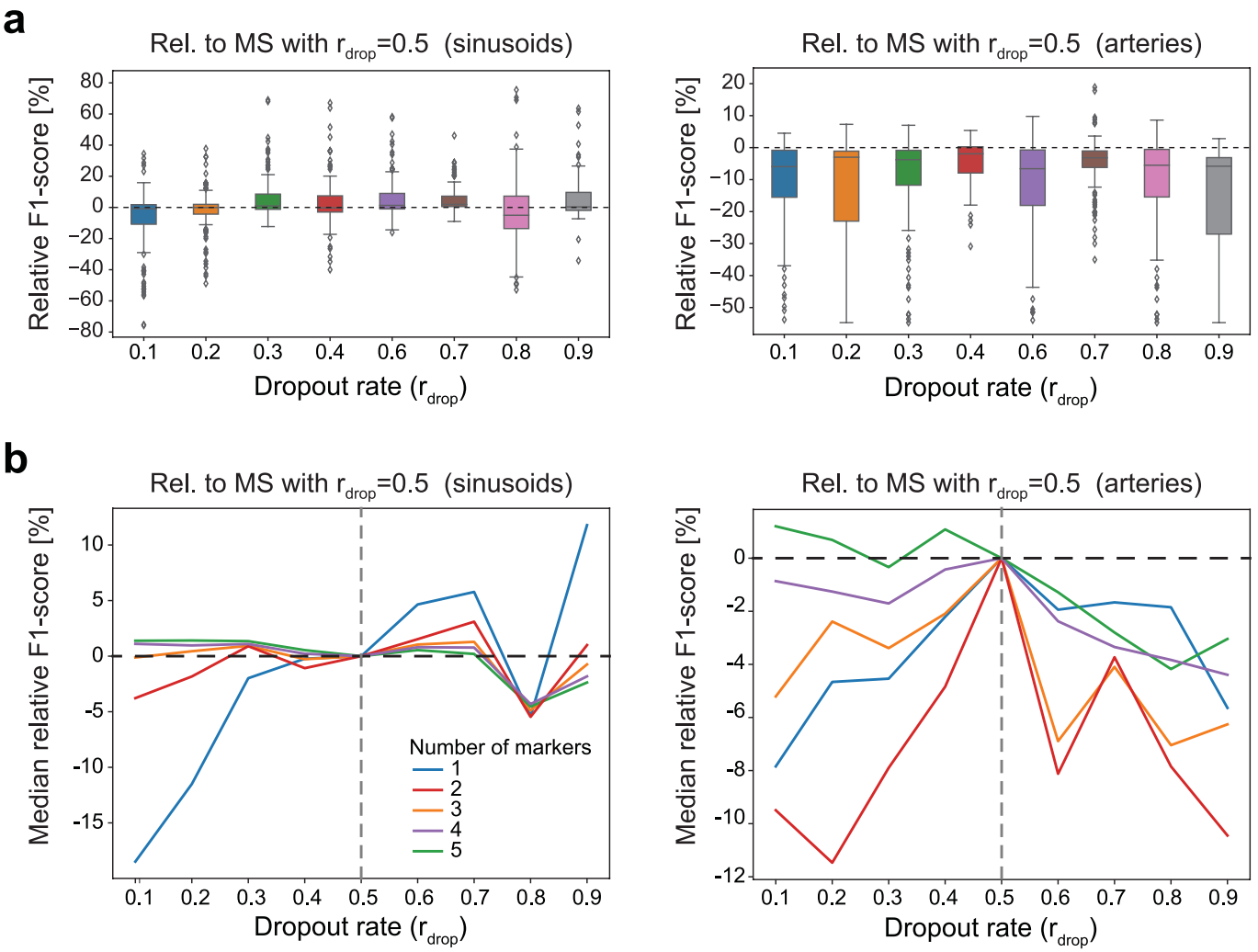
Extended Data Fig. 1 | Analysis of attention parameters in *ME* modules for sinusoids (left) and arteries (right). We estimate recalibration strength by calculating cosine distances between the *ME* attention subnetwork outputs obtained for each of the possible input marker combinations. Results are represented as the mean of all such pair-wise distances between all possible marker combinations, at a given layer where *ME* is placed, with the bars depicting the standard deviations of these distances. Using the colored network layers shown in Fig. 3b, *Encoder* layers correspond to the network layers in blue, *Decoder* to the layers in green, and *Bottleneck* to the yellow layers. The numbers next to each layer indicate the resolution level, where 0 corresponds to the highest (original resolution) and 3 to the lowest (that is, right before and after the bottleneck, for the encoder and decoder, respectively). It can be seen in this representation that recalibration strength is higher in layers with higher resolution, especially near the input of the network. This observation may indicate that high resolution layers of the network focus on effectively combining features from available markers, and in this way create shared abstract features that are common across markers for subsequent processing in lower resolution layers.



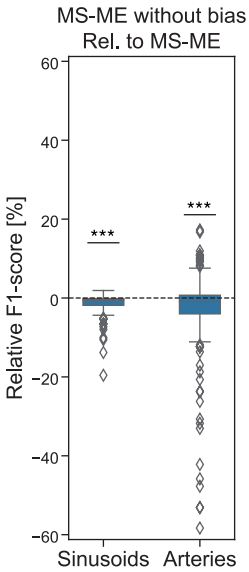
Extended Data Fig. 2 | Example images for the qualitative assessment of the segmentation of bone marrow images employed for the quantification of vasculature. Input images contain different combinations of markers shown as an overlay of different colors. The white rectangle within the input images represents the size of the output image when processing with a neural network. White arrows depict inference with the *MS-ME* model. Different colors are employed in the output images to show the different predicted classes. Since the tissue class overlaps with the other two, its contour is used instead for visual purposes.



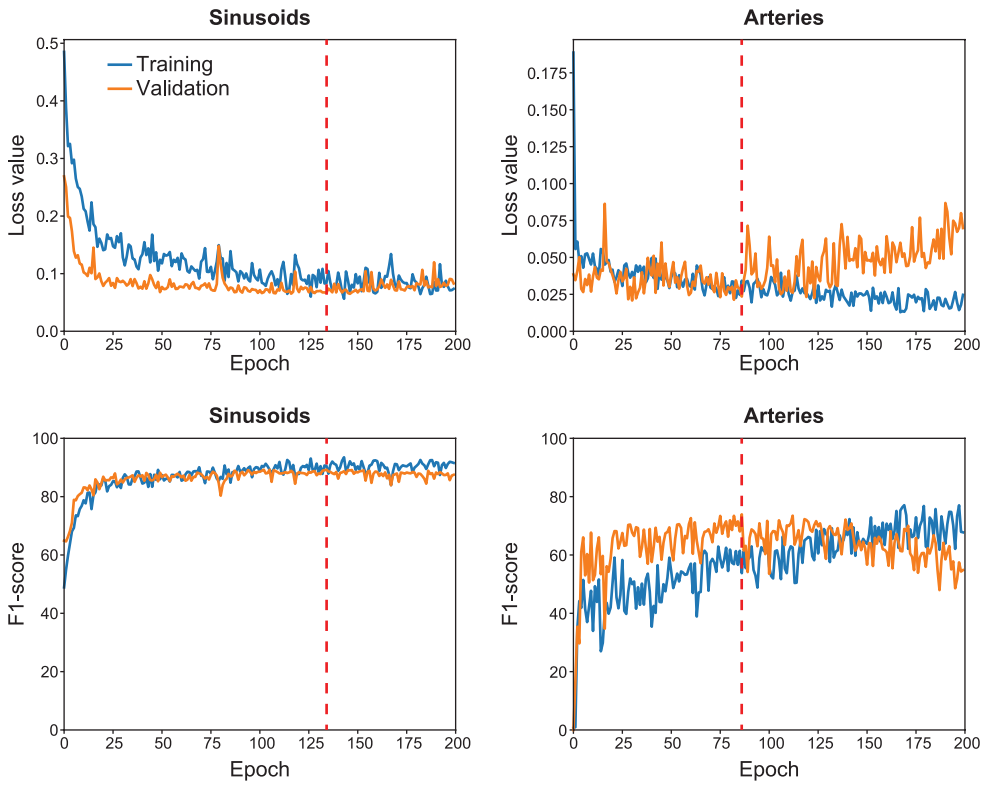
Extended Data Fig. 3 | Bone marrow volume ratio occupied by sinusoids. This volume is compared in both diaphysis (DIA) and metaphysis (META) when segmenting them with the morphological image processing (MIP) algorithm previously proposed (n = 12 for both DIA and META) and with our MS-ME method proposed here (n = 61 for DIA, n = 24 for META).



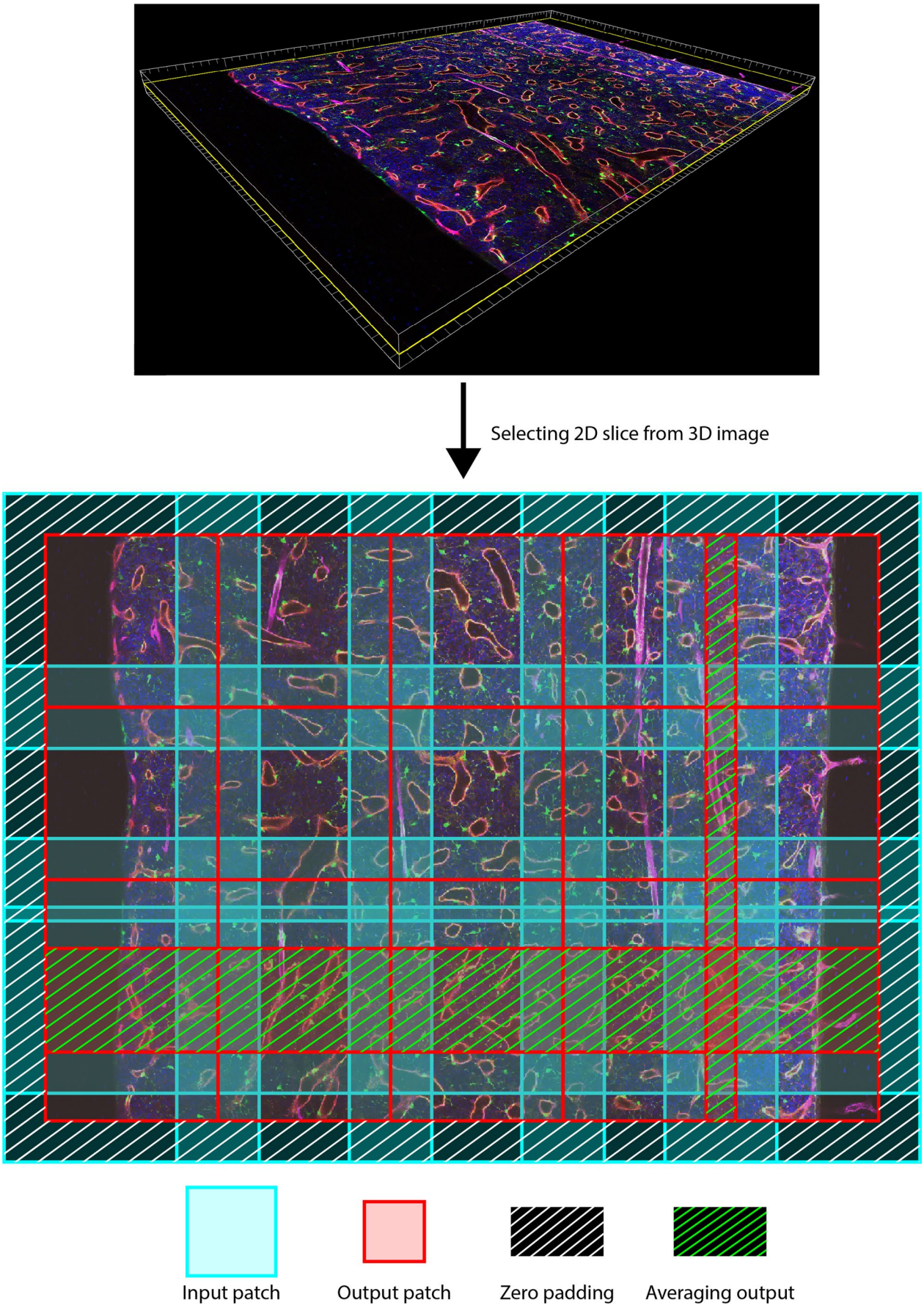
Extended Data Fig. 4 | Effect of marker dropout rate r_{drop} in *MS*. *F*1-score of *MS* models with different r_{drop} evaluated on the sinusoids (left) and arteries (right) relative to the proposed *MS* with $r_{drop} = 0.5$. (**a**) Evaluation for all 31 possible marker combinations (n = 124). Whereas some $r_{drop} \neq 0.5$ produce a slightly superior *F*1-score for sinusoids, $r_{drop} = 0.5$ is the best option for arteries and overall. (**b**) Median relative *F*1-score for models evaluated on combinations of specific numbers of markers, each represented by a different color for the different r_{drop} ($n = \frac{K!}{(K-k)!k!}$, where *K* is the number of markers available, and *k* the number of markers considered for each evaluation). Smaller r_{drop} are shown to be beneficial for combinations of more markers, and vice-versa. However, this trend becomes noisier for $r_{drop} > 0.5$, as illustrated with the gray dashed line. This effect can be due to the decrease in markers observed over time, although it is a question worth of further investigation in future work.



Extended Data Fig. 5 | Effect of bias term on *ME* module. *F*1-score of *MS-ME* model where the bias terms for all ME modules have been removed, relative to the proposed MS-ME model with bias across all marker combinations and cross-validation steps.



Extended Data Fig. 6 | Training evolution with our proposed *MS-ME* model. The weighted cross-entropy loss (top) and the *F*1-score (bottom) are shown across epochs for the training (blue) and validation (orange) sets, both for models trained for segmentation of sinusoids (left) and arteries (right). The red dashed line marks the epoch at which we choose the model, based on the highest validation *F*1-score.



Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Illustration of the image tiling pipeline employed to create suitable patches for CNNs. An example of a slice within the 3D image frame is shown in the upper part using Imaris (Bitplane AG). That slice is decomposed in patches as illustrated in the lower part. Each output patch (red) is smaller than their corresponding input patch (cyan) due to the convolutional operations in CNNs. We position the output patches next to each other without overlap in order to avoid padding artifacts in the application of CNNs. Instead, zero padding is only applied along the borders of the whole slice (area with white stripes). When an overlap between output patches cannot be avoided to fill the slice (area with green stripes), the average of the different patches in that region is used.

nature portfolio

Corresponding author(s): Alvaro Gomariz

Last updated by author(s): Jun 16, 2021

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Described in the cited article where the employed dataset was first published.
Data analysis	Proposed methods: Python 3.7.5, NumPY 1.17.4, tensorflow 2.1, pandas 0.25.3, scikit-image 0.16.2, scikit-learn 0.21.3, scipy 1.4.1. Visualization: Matplotlib 3.1.2, seaborn 0.9.0, Imaris (Bitplane AG) v9.3.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The labeled dataset employed for training and evaluation of the models described is included as a single HDF5 file in a CodeOcean capsule in https://codeocean.com/capsule/8424915/tree/v1.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The sample size was determined based on the available imaging datasets from previous biological studies, some of which had been published. No methodology was employed to calculate and predetermine the sample size.
Data exclusions	For the quantification of vasculature volume within the bone marrow samples, images were excluded when they did not contain the markers necessary for a satisfactory segmentation as described in Section Methods - Bone marrow quantification pipeline.
Replication	No replication was performed in the bone marrow vasculature quantification, since the segmentation and analysis methods employed are deterministic. For evaluation of the methods proposed, we employed cross-validation, which is more trustworthy than replication on the same data.
Randomization	Samples were randomly assigned to different groups for cross-validation of the proposed methods, as described in Section Methods - Training and evaluation details.
Blinding	Blinding was not relevant because the presented results are descriptive.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☐	☒ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	Described in the cited article where the employed dataset was first published. In addition, fetal livers from a-catulin GFP/HlfTdTomato or Evi-GFP/tdTomato mice were fixed and processed as reported for bone marrow tissues and stained with the following primary antibodies: rat anti Lyve-1 (ALY7 Thermo Fisher Scientific), Cy3-anti Smooth Muscle actin, goat anti-tdTomato (SICGEN), rabbit anti-GFP (Takara Bio). Secondary antibodies were donkey anti-rat AF594, Donkey anti-rabbit 680, and donkey anti-goat AF488 (Thermo Fisher Scientific).
Validation	Described in the cited article where the employed dataset was first published.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	Statement included Methods section under Animal studies.
Wild animals	Statement included Methods section under Animal studies.
Field-collected samples	Statement included Methods section under Animal studies.
Ethics oversight	Statement included Methods section under Animal studies.

Note that full information on the approval of the study protocol must also be provided in the manuscript.