MTNeuro: A Benchmark for Evaluating Representations of Brain Structure Across Multiple Levels of Abstraction

Jorge Quesada1, Lakshmi Sathidevi1, Ran Liu1, Nauman Ahad1, Joy M. Jackson1, Mehdi Azabou1, Jingyun Xiao1, Christopher Liding1, Carolina Urzay1, William Gray-Roncal2, Erik C. Johnson2,*, Eva L. Dyer1,*
1- Georgia Institute of Technology
2 - Johns Hopkins Applied Physics Laboratory

Abstract

There are multiple scales of abstraction from which we can describe the same image, depending on whether we are focusing on fine-grained details or a more global attribute of the image. In brain mapping, learning to automatically parse images to build representations of both small-scale features (e.g., the presence of cells or blood vessels) and global properties of an image (e.g., source brain region) is a crucial and open challenge. However, most existing datasets and benchmarks for neuroanatomy consider only a single downstream task at a time. We introduce a new dataset, annotations, and multiple downstream tasks that provide diverse ways to readout information about brain structure and architecture from the same image. Our multi-task neuroimaging benchmark (MTNeuro) is built on volumetric, micrometer-resolution X-ray microtomography imaging of a large thalamo-cortical section of mouse brain, encompassing multiple cortical and subcortical regions, that reveals dense reconstructions of the underlying microstructure (i.e., cell bodies, vasculature, and axons). We generated a number of different prediction challenges and evaluated several supervised and self-supervised models for brain-region prediction and pixel-level semantic segmentation of microstructures. Our experiments not only highlight the rich heterogeneity of this dataset, but also provide insights into how self-supervised approaches can be used to learn representations that capture multiple attributes of a single image and perform well on a variety of downstream tasks. Datasets, code, and pre-trained baseline models are provided at: https://mtneuro.github.io/.

1 Introduction

Our understanding of our natural surroundings entails multiple levels of perceptual processing: we can recognize a macroscopic object (e.g., a tree), while also identifying finer-grain structures within (e.g., leaves and branches), and context-relevant features (e.g., leafiness, height, or season). This multi-level perception scheme also translates to the medical image domain: the process of interrogating medical images (either by a human expert or an algorithm) involves combining macrostructural insights (such as a region of interest) with context-relevant microstructure information and human-interpretable features (e.g., the density of a given cell type in a microscopy image) in order to derive a diagnosis or characterize a target sample.

*Contact authors: (ELD) evadyer@gatech.edu; (ECJ) erik.c.johnson@jhuapl.edu; (JQ) jpa-cora3@gatech.edu; Both senior authors contributed equally.

Preprint. Under review.
In particular, the ongoing effort to understand the connections and dynamics of the brain involves analyzing both macroscopic level properties such as region-level structures (1) as well as detailed microstructures like the size or shape of a given cell type (2). While significant advances have been achieved in unveiling the properties and structures within the brain through several imaging modalities at different scales (3; 4; 5), most existing neuroimaging benchmarks are designed for evaluation at a single-scale level, or geared towards a particular downstream task. This can be attributed to several causes, including the prohibitive cost of manual annotation for data spanning multiple scale levels (6; 7), the associated computational cost of processing multi-scale data, and the fact that neuroimaging technologies have only recently progressed towards pipelines that can capture multi-area and even whole-brain volumes at high resolutions.

To fill this gap, we present the **MTNeuro** benchmark: a multi-task, multi-scale benchmark based on a large 3D X-ray microtomography image dataset spanning multiple areas in a single mouse brain. We host our dataset in the Brain Observatory Storage Service and Database (BossDB, a specialized interactive tool) and provide an integrated dataloader to facilitate transfer experiments and analysis. This benchmark provides a unified framework that allows for evaluating models and representations arising in three distinct tasks, which we summarize below.

**Task 1 - Image-level classification of brain area:** prediction of the brain region (somatosensory cortex, striatum, thalamus, zona incerta) to which a given image or volume belongs.

**Task 2 - Pixel-level segmentation of microstructures:** prediction of neural microstructures (blood vessels, axons, cell bodies, background) from pixel-level annotations within the four core brain regions contained in the dataset.

**Task 3 - Multi-task decoding from fixed representations:** estimation of human-interpretable semantic features (such as the average cell size or axon density) from the representation of a given image, obtained after “freezing” or fixing the weights of a trained encoder.

To understand how current models perform on these different tasks, we evaluate a family of different supervised and self-supervised model baselines. Our results in Tasks 1 and 3 highlight a significant generalization gap between self-supervised over supervised approaches, which opens up interesting opportunities for further evaluation and development of self-supervised methods for these tasks. Through testing across a family of different models across a variety of tasks, our proposed benchmark provides both an exciting platform for evaluating self-supervised learning (SSL) methods for the machine learning community, and a rich tool in the effort to extract fundamental insights at different levels of the brain for the neuroscience community.

## 2 Background and Related Work

### 2.1 The need for a benchmark in brain mapping and connectomics

Over the past decade, there have been major advances in our ability to resolve fine-scale neuroanatomical structures in the brain. With these advances, we have generated large amounts of brain data that span many spatial scales, and can reveal different features of brain organization.

At the nanoscale, electron microscopy has provided detailed wiring diagrams of small portions of cortex (8). At micron scale, microscopy techniques have provided detailed pictures of cytoarchitecture - or how neurons and cells are organized (4). Efforts at even larger scales to capture many brain areas simultaneously, like connectivity atlas and X-ray microtomographic datasets (9), have provided information about the interplay between long-range connections across brain areas and microstructures such as cell body densities and other morphological features of brain structure.

Accompanying these new tools for data generation have been major advances in machine learning and computational approaches for modeling and analyzing these datasets, for problems such as object detection, segmentation, and classification. While the information provided by these methods is incredibly rich and has a great deal of structure at many scales, any given method is typically tested on an individual challenge problem at a particular scale. The expense of annotating and proofreading can be considerable, and significant neuroanatomy knowledge is typically required of annotators (7). Moreover, many efforts to provide high quality data, such as (4), have not been focused on building

---

2[^Note2]: [https://mtneuro.github.io/](https://mtneuro.github.io/)
specific benchmarks for the ML community, but rather on providing references and resources for
the neuroscience community.

Using machine learning tools to understand these emerging brain datasets at different spatial scales is
both a challenge and an increasingly critical need. Large tera- and peta-scale connectomics datasets
are being collected using electron microscopy and X-ray microtomography, including data from the
entire brain of *Drosophila* (10; 11), large portions of the mouse brain (12; 2) and even a cubic mi-
limeter of human cortex (5). Advances in imaging technologies promise to continually increase the
spatial extent, number of species, and number of imaged individuals. These datasets have the micro-
or nanoscale resolution and large spatial extents required to resolve sub-cellular structures (e.g., mi-
 tochondria and synapses), microstructures (e.g., glia, neurons, and vasculature), and macrostructure
(e.g., brain regions, cortical layer structure, and long-range white matter projections). This rich,
multi-scale structure combined with the large size of these datasets requires new ML tools, which
drives the need for datasets and benchmarks for high-resolution neuroimaging data that can leverage
the many scales at which neural structure exists.

### 2.2 Existing neuroimaging datasets and benchmarks

Due to the large variety of spatial scales, neuroanatomical structure, and imaging modalities, a wide
range of segmentation problems have been formulated for neuroimaging data. At the macroscale,
there has been a long history of developing benchmarks for different datasets in MRI and related
modalities like DTI and fMRI. For example, the BraTS dataset (13) focuses on the MRI of brain
tumor and motivated many segmentation works (14; 15; 16; 17). The ADNI (18) and MIRIAD (19)
datasets provide MRI-based Alzheimer’s disease imaging that focuses on tracking disease progres-
sion (20; 21; 22; 23; 24; 25; 26). UK Biobank (27) offers a huge collection of Brain MRI data of
5,00,000 human participants of ages 40-69 across several years. The low spatial resolution of MRI
and related methods, however, limit the ability to observe microscale structures at the cellular or
subcellular level.

For microscale structure, there are fewer benchmarking datasets due to the scale and complexity of
the annotation and processing. Example problems include segmentation of synapses in the CREMI
challenge (28), which provides high-resolution imaging of synapses in 5 cubic microns of non-
isotropic EM with nanometer resolution. Other problem formulations include 2D image segmenta-
tion of cell membranes (ISBI 2012 Challenge), with data from (29), and 3D segmentation of cells
(30; 31). Specific benchmarks have also been developed for axon instance segmentation (32) and mi-
 tochondria segmentation (33). Different forms of microscopy with micrometer resolution, including
calcium fluorescence microscopy, are suitable for segmenting cell bodies and extracting functional
time-series data, but lack other microstructure information. Benchmarking datasets have also been
set up for estimation of functional traces from two-photon calcium fluorescence microscopy, in-
cluding spikefinder (34). These benchmarks have been instrumental in driving progress on specific
problems at specific spatial scales. In general, however, these microscale dataset benchmarks focus
on relatively small spatial extents, and lack the multi-scale, macrostructure also found in the brain.
For example, to the best of our knowledge, there is no existing publicly available microtomogra-
phy dataset of brain structure with dense microstructure annotations and macroscale labels currently
available.

Encompassing larger spatial extents, projects such as the BigBrain atlas provides high-resolution
sections from the mouse brain with Nissl contrast (35), but the resolution only allows resolution
of cells around 20 micrometers. Large-scale EM datasets are also being used to benchmark per-
formance and segment neurons, augmented with iterative human proofreading (36; 37; 5), which
are leading to large segmented datasets with increasingly complex annotation. These data, how-
ever, are not suitable for large-scale benchmarking and algorithm development in the general ma-
achine learning community due to their size and ongoing refinement. To accelerate progress towards
machine learning tools which can operate in multiple level of spatial abstraction within the same
high-resolution dataset, benchmarks are needed which encompass large spatial extents at high res-
olution. This will enable a broader community of researchers to apply state-of-the-art methods to
these important application areas.
3 Dataset and Tasks

3.1 Overview of dataset

We build our benchmark on a large open access high-resolution (1.17µm isotropic) 3D X-ray microtomography imaging dataset that encompasses multiple brain areas (12) of a single mouse subject. The size of the intact brain volume, along with the high-resolution of the volumetric data, results in a uniquely rich set of both macroscale (region of interest) and microscale (cells, blood vessels, axons) structures that can be interrogated throughout the dataset (38; 39). As such, this benchmark addresses the aforementioned need for datasets that encompass multiple levels of abstraction, allowing us to extract insights into the performance of the evaluated baselines on distinct ML challenges that represent different axes of interest. On top of the existing dataset and sparse annotations, we add additional dense annotations (continuous volume of annotation) at both structural levels, as well as three independent multi-scale tasks which are described in Section 3.3.

The full volumetric dataset provides micron resolution of an intact brain sample totalling 5805 × 1420 × 720 pixels. From this volumetric dataset, we extracted four three-dimensional subvolumes, each of size 256 × 256 × 360 (see Appendix 1 for more details) in each of the four regions of interest: either somatosensory cortex (CTX), striatum (STR), ventral posterior region of thalamus (VP), or the zona incerta (ZI) (see Figure 1 A-B).

For each subvolume, we generated ground truth annotations for pixel-level segmentation models (Task 2) by training a Unet model on the sparse 2D annotations (cell, blood vessel, axon, background) provided in the original data resource provided in Prasad et al.(12); this allowed us to create dense 3D reconstructions which we had a manual annotator proofread and correct. These densely-annotated volumetric cutouts contain pixel-level microstructural labels (see Figure 1B), identifying each point as either part of an axon, cell, blood vessel, or background. (For Task 3) We then leveraged these pixel-level labels to compute a number of semantic features from these reconstructions: the density of blood vessels and axons, the number and size of cells, and the average inter-cell distance in each slice. These semantic labels provide information into different features of the cytoarchitecture that can be used to interpret the embeddings learned by models tested on this dataset.

In addition to these microstructural annotations, we have expanded the original region-level annotations for the original dataset to include interpolations of the region labels across all 720 slices of size...
From these interpolated sections, we extracted 12 new subvolumes (three for each of the four regions of interest) for examining the generalization of models in Task 1.

3.2 Data access

The dataset and all corresponding labels are stored in BossDB (40), the Brain Observatory Storage Service and Database. The dataset project page can be found at http://bossdb.org/project/prasad2020. BossDB is a specialized spatial database for Electron Microscopy and X-Ray Microtomography Datasets, with seamless visualization through Neuroglancer, which enables interactive visualization of large-scale 3D annotated volumes and annotations. All data are available publicly, using public log on credentials (no account creation required). The project page documents project metadata, citation instructions, and the data creators and curators.

For benchmarking, data are accessed through the Python intern API (41). This API allows a remote connection to the BossDB system, including downloads of arbitrary, on-demand 3D cutouts of data, including raw images and annotations, without the need to download the entire dataset to disk. To facilitate the use of this API, we provide a Pytorch DataLoader 3 for rapid algorithm development and testing. We also provide sample Jupyter notebooks to allow downloading of task cutouts, saved as Numpy files, for development in other frameworks. The tasks are defined with task-specific JSON files specifying the metadata for each task. The use of this dataloader is illustrated in Fig. 1C, where we detail how the dataset can be efficiently accessed through BossDB.

The data are structured into different channels of raw images and annotations, and different channels are used for each task as described below.

**Raw images:** All tasks utilize the same raw images, consisting of single color channel with 8-bit unsigned integer values. The total raw data volume is $720 \times 1420 \times 5805$ voxels at a resolution of $1.17 \mu m$. The data are available at https://api.bossdb.io/v1/mgmt/resources/prasad/prasad2020/image.

**Brain region annotations:** Tasks 1 and 3 utilize dense pixel-level annotations of different brain regions (macrostructure) spanning x. The labels are 0: CTX; 1: STR; 2: VP; 3: ZI. At this scale/level, there equal number of samples of each class and hence the classes are balanced. The image data are represented as 64-bit unsigned integers and stored in https://api.bossdb.io/v1/mgmt/resources/prasad/prasad_analysis/roi_labels.

**Microstructure Annotations:** Task 2 and Task 3 utilize pixel level annotations of brain microstructure. Volumes of $256 \times 256 \times 360$ are densely annotated with microstructure labels. The labels are 0: no label (background); 1: blood vessels; 2: cells; 3: mylenated axons. These are stored as 64-bit unsigned integers data in the channel https://api.bossdb.io/v1/mgmt/resources/prasad/prasad_analysis/pixel_labels.

A key aspect of our approach is portability of data access and training infrastructure across neuroimaging datasets in BossDB. This allows for easy extension of code and baselines to new volumetric imaging datasets stored in BossDB. These include data from new species, with new microstructure labels (synapses, membranes, mitochondria), and with new macrostructure labels (brain regions, experimental state). By modifying the dataloader JSON, developers can specify different BossDB datasets, spatial regions, and annotation (label) sources. This allows for the flexibility required to support the different tasks in this dataset, and will enable further training and deployment of these baseline models to new datasets. This is an important contribution of this work towards developing machine learning tools for emerging large-scale neuroimaging datasets.

3.3 Tasks

3.3.1 Task 1: Image-level classification of brain area

When analyzing imaging data that spans many different brain regions, one important question is how much global image features correlate with the brain region from which the sample is drawn. Thus, we can pose this as a classification problem, where we pull a small patch (or small region) from the data and estimate which of the 4 brain regions of interest the sample was drawn from. This task can be further sub-divided into three sub-tasks, which we outline below and detail in Figure 2.

3https://github.com/MTNeuro/MTNeuro
Tasks considered in the benchmark. Task 1: image classification to predict brain area; Task 2: microstructure segmentation at the pixel-level; Task 3: prediction of microstructure features from latent representations obtained in Task 1.

**ROI-C1.** For this sub-task, we use only the 4 densely-annotated cubes (shown in blue in Figure 2, Task 1), each corresponding to one of the 4 regions of interest (CTX, STR, VP and ZI). We divide each of the four cubes in $256 \times 256 \times 300$ sub-volumes used for training, and $256 \times 256 \times 50$ sub-volumes used for testing, leaving 10-frames between the splits to avoid any structural overlap. The resulting overall sample size for this sub-task is thus 1400 images, with 1200 images for the train set and 200 for the test set.

**ROI-C2.** In this sub-task, we wanted to evaluate how well our models generalize when tested on new areas within the larger 3D context. We generated three additional $256 \times 256 \times 360$ cubes per class (shown in yellow in Figure 2, Task 1), and tested our models from (ROI-C1) on the other three held-out subvolumes (yellow). The overall dataset used is thus expanded to 5600 images, using the same 1200 images as ROI-C1 for training, and 4400 images for testing.

**ROI-C3.** In this sub-task, we evaluate the performance of the models when allowed to learn from a larger set of data. Given the three additional sets of cubes generated for sub-task C2, we train the evaluated models on the original (blue) annotated volumes plus one of the new (yellow) generated volumes. We then evaluate a model’s capability to predict the class of the two remaining sets of (yellow) cubes. This task employs 5600 images, with 2800 in the train set, and 2800 in the test set.

### 3.3.2 Task 2: Pixel-level segmentation of microstructures

Another important requirement for a holistic understanding of brain data is correctly identifying the structural features of small scale elements such as cells and blood vessels. In this task, we provide the pixel-level labels corresponding to the neural microarchitecture of the volumetric cutouts, and evaluate different baselines in terms of their prediction accuracy for classifying each pixel in a separate test set of volumes in the four considered brain regions. The size of the dataset depends on whether the processing happens at the 2D level (1400 image samples) or 3D level (348 volume samples considering 4-slice volumes). Given the nature of the microstructure annotations, we consider two different segmentation modes, detailed in Figure 2, Task 2:

**3-class segmentation task.** We first consider the pixel-level segmentation of images into one of three classes: either cell bodies, blood vessels, or other (background and axons). We use the same train and test split as in ROI-C1 in Task 1 (300 images for train, 50 for test, with a gap of 10 slices between datasets) on the main subvolumes that are densely annotated at the pixel level.

**4-class segmentation task.** In this task, we consider the pixel-level segmentation of images into one of four classes: either cell bodies, blood vessels, background or axons. When we consider dense axonal segmentation, we remove the ZI region from our training and testing set due to the inability to reliably segment axons in this subvolume through human-level annotations.

### 3.3.3 Task 3: Multi-task decoding from fixed representations

A unique feature of a dataset with both semantic and region-level information is that the representations obtained by models trained for a region classification task can be interrogated in terms of semantic feature encoding. Moreover, models that are not trained with a specific downstream task
Table 1: Results on image classification accuracy for brain area prediction (Task 1) across the 4 interest regions. We consider three different testing and training conditions. In (C1), we train and test in one subvolume, splitting across sections. In (C2), we report the test accuracy of our model on new volumes unseen during train. In (C3), we train on two and test on the remaining 2 heldout subvolumes.

<table>
<thead>
<tr>
<th></th>
<th>ROI - C1</th>
<th>ROI - C2</th>
<th>ROI - C3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised</td>
<td>0.88 ± 0.03</td>
<td>0.77 ± 0.04</td>
<td>0.88 ± 0.02</td>
</tr>
<tr>
<td>Sup w/ Mixup</td>
<td>0.90 ± 0.04</td>
<td>0.78 ± 0.04</td>
<td>0.90 ± 0.02</td>
</tr>
<tr>
<td>BYOL</td>
<td>0.88 ± 0.02</td>
<td>0.75 ± 0.03</td>
<td>0.97 ± 0.01</td>
</tr>
<tr>
<td>MYOW</td>
<td>0.90 ± 0.02</td>
<td>0.77 ± 0.06</td>
<td>0.98 ± 0.01</td>
</tr>
<tr>
<td>MYOW-m</td>
<td>0.94 ± 0.02</td>
<td>0.76 ± 0.04</td>
<td>0.98 ± 0.01</td>
</tr>
<tr>
<td>PCA</td>
<td>0.59</td>
<td>0.24</td>
<td>0.07</td>
</tr>
<tr>
<td>NMF</td>
<td>0.62</td>
<td>0.25</td>
<td>0.50</td>
</tr>
</tbody>
</table>

4 Results

4.1 Task 1: Image-level classification of brain area

In this task, we consider the classification of different images into a number of candidate brain areas (CTX, STR, VP and ZI). We benchmarked 2 supervised Resnet18 (42) models: the first trained using standard approaches for regularization (dropout and weight decay factor of 0.3) and the other trained using mixup (43). Furthermore, we consider a number of self-supervised learning methods, given the high label efficiency of these approaches: (i) BYOL (44), (ii) MYOW (45), (iii) and a variant of MYOW trained with a single projector and predictor (MYOW-merged, or MYOW-m). For the SSL models, we fix the network weights after training, and then train a linear layer on top of the fixed representations. This tells us how well the SSL loss captures the classes in the data after only a linear transformation (cropping and rescaling). All models have a latent dimensions of 256. As additional baselines, we also extracted 256-dimensional embeddings from our data using Principal Component Analysis (PCA) and Non-Negative Matrix Factorization (NMF), and trained a linear layer on these representations.

In our experiments, we evaluate all methods across 5 training instances with different random seeds, and report the overall mean accuracy and standard deviation. All models are trained for 100 epochs using an SGD optimizer with a learning rate of 0.03 and \( \lambda = 0.2 \). For more details on our experimental setup and models, see Section 4 in the Appendix.

Results in classifying brain region-of-interest (ROI). Results for this small scale setting are shown on the first column of Table 1. We find that many of the SL and SSL models are comparable, with the MYOW-m model achieving the highest accuracy. At the same time, we find that our simple unsupervised baselines (PCA, NMF) do not provide separable classes.

In the second column of Table 1, we test the generalization capabilities of the considered models by evaluating how well these models performed on subvolumes in other parts of the larger dataset (Table 1, ROI-C2) when trained on a single subvolume per class. In this case, we find somewhat similar performance across all the models, with a significant decrease in accuracy that can be attributed to some amount of domain shift or differences in data.

We report our large-scale training task in the third column of Table 1, in which we train models on two subvolumes per area and test on the remaining held out images (the other two subvolumes). In this case, we can observe that SSL methods significantly outperform the SL models, which remain stagnant at around the same accuracy as the original train and test condition. We can thus observe a significant generalizability gap in SSL over SL models, in that exposure to additional data (in this case, the additional cube with respect to task ROI-C1) drastically improves the generalization of the SSL models to unobserved data.

in mind (i.e. SSL models) can also be evaluated to analyze the degree to which they encode certain human interpretable or task-relevant features. In this task, we test all models trained for Task 1 on their capability to provide representations that are able to predict semantic image properties such as blood vessels density, cell count and size, axon density and average distance between cells through a simple linear readout. We use the trained models with fixed weights to compute the representations of each of the images in the original four densely-annotated cubes used in task ROI-C1, and fit a simple linear regression model on these representations to predict the desired semantic features.
4.2 Task 2: Pixel-level segmentation of microstructures

In Task 2, we consider 4 different variants of our segmentation task. Due to the difficulty of segmenting axons in some conditions, we examine the performance of models in a 3-class (combining axons and background into one class) and 4-class setting (with all 4 components). In the 4-class setting we exclude the ZI region in our training and evaluation because the axons in this region are difficult to distinguish accurately even for a human annotator. In both cases, we train 2D and 3D models.

**Experiment setup.** In our experiments, we perform pixel-level segmentation using a selected set of 2D and 3D models. Each model is put through a separate hyper-parameter tuning process for finding an optimal learning rate and batch size by training on the train set and evaluating on the validation split. Each model is trained for 20 epochs with its optimal learning rate and batch size and evaluated across 5 training instances (each with its own random seed). The class-wise F1-score, the class-wise IoU, and the overall mean and standard deviation of both metrics are reported for each model (Table 2). We do not report accuracy because for this particular task of pixel-level segmentation, it does not aptly represent the model performance due to class imbalance (see Appendix Section 5.2.3 for breakdown of different classes in the training and test sets).

The models we used for the 2D segmentation task are the standard 2D U-Net model (46; 47) and selected models from the 'segmentation_models.pytorch' library (48): MA-net (49), FPN (50), U-Net++ (51), PAN (52) and PSPNet (53). The models we used for the 3D segmentation task are the standard 3D U-Net model (46) and selected models from 'MedicalZooPytorch' (54): VNetLight (55; 54) and HighResNet (56). For more details refer Section 5 in the Appendix.

<table>
<thead>
<tr>
<th>Method</th>
<th>Metric</th>
<th>3-Class</th>
<th>4-Class without ZI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bg + Axons</td>
<td>Vessels</td>
<td>Cells</td>
</tr>
<tr>
<td>2D U-Net</td>
<td>F1</td>
<td>0.99</td>
<td>0.76</td>
</tr>
<tr>
<td>2D U-Net</td>
<td>IoU</td>
<td>0.98</td>
<td>0.64</td>
</tr>
<tr>
<td>MA-Net</td>
<td>F1</td>
<td>0.99</td>
<td>0.79</td>
</tr>
<tr>
<td>MA-Net</td>
<td>IoU</td>
<td>0.98</td>
<td>0.68</td>
</tr>
<tr>
<td>FPN</td>
<td>F1</td>
<td>0.99</td>
<td>0.71</td>
</tr>
<tr>
<td>FPN</td>
<td>IoU</td>
<td>0.97</td>
<td>0.59</td>
</tr>
<tr>
<td>U-Net++</td>
<td>F1</td>
<td>0.99</td>
<td>0.79</td>
</tr>
<tr>
<td>U-Net++</td>
<td>IoU</td>
<td>0.98</td>
<td>0.68</td>
</tr>
<tr>
<td>PAN</td>
<td>F1</td>
<td>0.97</td>
<td>0.47</td>
</tr>
<tr>
<td>PAN</td>
<td>IoU</td>
<td>0.94</td>
<td>0.36</td>
</tr>
<tr>
<td>PSPNet</td>
<td>F1</td>
<td>0.97</td>
<td>0.48</td>
</tr>
<tr>
<td>PSPNet</td>
<td>IoU</td>
<td>0.93</td>
<td>0.37</td>
</tr>
</tbody>
</table>

**Pixel-level segmentation in 2D.** The results from the selected models on our 2D pixel-level segmentation task are tabulated in Part I of Table 2 and visualized in Figure 3. The individual slices are the input to the models and they are fed in a batched manner during training. For training and evaluation, we consider both the 3-class and 4-class settings and we use efficientnet-b7 encoder for all the models as it was seen to give the best performance among the 25 encoders that were attempted.

From our results we see that MA-Net performed the best overall among the 2D models with an average IoU of 0.78 followed by U-Net++ with an average IoU of 0.76 in the 4-class setting (Table 2). As can be seen from the class-wise IoU and class-wise F1-scores, for most of the best performing
models, the most challenging components to differentiate are cells and blood vessels, which are also
difficult for human annotators to identify from 2D slices without further 3D context.

**Pixel-level segmentation in 3D.** We provide the same breakdown and accuracy measures as in the
2D case, for this 3D case, in Part II of Table 2. For providing 3D input to the models, we pass in
multiple consecutive slices (8 slices) as a single sub-volume and build a prediction over all slices
jointly in 3D. We input these sub-volumes to the model in a non-overlapping manner, so even though
we have improved 3D context, the model also sees less data. Comparing across the 3D and 2D cases
for the standard U-Net model, we did not see an improvement in the 4-class setting but in the 3-class
setting we see a slight improvement. Within the 3D case, the U-Net model performed best with an
average IoU of 0.65, followed by HighResNet with an average IoU of 0.52 in the 4-class setting.

### 4.3 Task 3: Multi-task decoding from fixed representations

In Task 3, we explore the prediction of different contextual features from the latent space learned
by the models trained in Task 1. We note that the ability to predict fine-scale human-interpretable
features from whole image-level representations is an important step in understanding how relevant
semantic features are for a given task, and can open up interesting studies on how these features
inform neuroanatomy when trained without a specific downstream task in mind (i.e. SSL methods).
To do this, we tapped into the microstructure annotations to extract information about the attributes
each image: the proportion of pixels that are either blood vessels or axons, the cell count and size,
Figure 4: Visualization of the learned representations with different semantic features overlaid. We project the learned MYOW-m (ROI-C1) embeddings using UMAP (57), and overlay different semantic features on top. From left to right, we visualize brain area (class), % blood vessels, % axons, cell count, and cell size.

and average distance between cells and their nearest neighbor. Further details on the experimental setup for this task can be found in the Appendix in Section 6.

We quantify the readout of these different features by fitting a linear regression to the representations obtained from extracting the latent space outputs (readouts) from the trained models and then reporting the $R^2$ values in Table 3 in two conditions. First, we report the scores of reading out the semantic attributes when we examine the representations generated by each of the models trained in ROI-C1 (Table 3 I); then we report the $R^2$ values we obtain on the same test set but using the model trained on two subvolumes per ROI as in ROI-C3 (Table 3 II). In both cases, we find that the SSL models significantly outperform the rest of the baseline models. The gap is more pronounced in the two volume training condition, highlighting the generalization difference between supervised and SSL approaches (58) in providing good representations for a wide range of downstream tasks.

5 Discussion

We introduced a new dataset, annotations, novel tasks, and baseline networks for facilitating the development of models for learning representations of microscopy datasets. Our benchmark considers tasks that can ask for many different attributes of an image, at multiple levels of abstraction.

In addition, we also built general infrastructure for training models from datasets in the BossDB framework, like dataloaders which can be adapted to different datasets. This will expand the use of large-scale volumetric neuroimaging data for machine learning tool development. Ongoing work includes adapting this infrastructure to membrane and synapse segmentation from EM imaging (8), and defining new ways to model experimental conditions (i.e., developmental state).

Limitations and Future Work. In neuroscience, often obtaining high quality labeled data (especially for dense segmentation tasks like those provided in Task 2), is very costly (59). The intensive nature of manual annotation and proofreading data thus limits the amount of labeled and annotated data that we provide to train models on, or the amount of distributional shift that can be assessed.

While this is a limitation of the work, it also is an accurate reflection of the challenges faced in the field, and thus requires more label-efficient approaches for learning like the SSL methods we highlight. Moving forward, we hope to leverage the models tested in this work to generate even more high quality annotated data to further improve model performance and segmentation.

When designing our current benchmark, we focused on building a multi-scale challenge where variability was due to changes in brain structure and not different preparations or imaging parameters. Therefore we focused on a single animal where we have a large intact brain volume that spans many heterogeneous brain regions. While we acknowledge that this limits the generalization of the models to new samples or datasets, this addresses the heterogeneous nature of brain structure that is often overlooked. Our results show that even within a single brain, there is rich heterogeneity in microstructure across different brain regions that makes it difficult for some models to generalize.

This also reveals important generalization gaps between SL and SSL models. In the future, we hope to expand this effort to new multi-scale brain datasets, perhaps using new lightsheet (60) or whole-brain scaling (61; 62) techniques. Also, we anticipate hosting a challenge on this benchmark with community participation.

Broader Impacts. The high heterogeneity of brain data, coupled with the variability in the scale and nature of the tasks presented in this benchmark make it a challenging and useful resource for the broader machine learning community. Furthermore, we hope the provided infrastructure will help accelerate development of machine learning techniques for emerging, high-resolution neuroimaging datasets being collected in the broader community.
Acknowledgements

This project was supported by NSF award IIS-2039741, NIH award 1R01EB029852-01, NIH award R01MH126684, NIH award R24MH114785 in addition to generous gifts from the Alfred Sloan Foundation, the McKnight Foundation, and the CIFAR Azrieli Global Scholars Program.

References


