

## APPENDIX

### A IMPLEMENTATION DETAILS

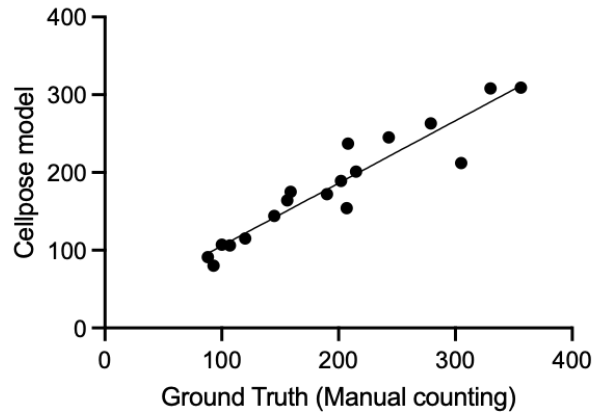
**Dataset.** We use an in-house dataset of digital microscopy images obtained from the PD mice models. This dataset consists of 1500 images among which a small fraction of 108 images have been annotated with the segmentation masks for dopamine neurons. The images’ resolutions are in the range [3000, 6000]. We use all of the images for self-supervised learning, and then fine-tune the self-supervised pre-trained models with the labeled images (supervised learning). For supervised learning, we randomly divided the dataset into training (70%), validation (10%), and testing (20%).

**Network Architecture.** For target segmentation task, we use a U-Net network which consists of encoder ( $f_\theta$ ) and decoder ( $g_\theta$ ) parts. The encoder is a standard ResNet-50, which is initialized with the self-supervised pre-trained encoder.

**Tile sampling and augmentation.** For target segmentation task, we divide images into non-overlapping patches of size  $512 \times 512$  to ensure we sample from every part of the image. In all experiments, the raw image intensities per channel are normalized to the  $[0,1]$ . Data augmentation is essential for biological and medical image analysis due to the typically limited amount of available annotated data. We use different data augmentation techniques to enforce the model to capture more robust and generalizable representations. In particular, we use Flip, Rotation, RGBShift, Blur, GaussianNoise, and RandomResizedCrop to teach the expected appearance and color variation to the deep model.

**Fine-tuning protocol.** We initialize the encoder of the target model (i.e. U-Net) with the pre-trained models and fine-tune all target model parameters. We train the target models using the Adam optimizer with a learning rate of  $1e-3$  and  $(\beta_1, \beta_2) = (0.9, 0.999)$ . We use ReduceLROnPlateau learning rate decay scheduler. We use batch size of 32 and train all models for 200 epochs. We employ early-stop mechanism using the validation data to avoid over-fitting. We use Dice coefficient loss function for training the target task. Dice coefficient is used for evaluating the accuracy of the target segmentation task. We run each method ten times on downstream task and report the average and standard deviation performance over all runs.

**Cell counting.** The automatic cell counting is a challenging task due to the overlapping cells which share boundary; distinguishing overlapping cells requires certain post-processing to enhance the counting accuracy. In particular, we first calculate the minimum and average cell size using the ground truth for the training data. Then, we take the models predictions (segmentation masks) and extract the connected components within the prediction masks; each connected component represents one or more cells (in the case of overlapping cells). We then filter out components that are smaller than the minimum cell size. For the remaining components, we count cells by dividing the cell size by the average cell size.



	Ground Truth (Manual counting) vs. Cellpose
Pearson r	
r	0.9443
95% confidence interval	0.8538 to 0.9794
R squared	0.8917
P value	
P (two-tailed)	<0.0001
P value summary	****
Significant? (alpha = 0.05)	Yes
Number of XY Pairs	18

Figure 1: Correlation plot depicting the number of DA neurons counted by a biologist vs the number of DA neurons counted by Cellpose model (Robitaille et al., 2022). Blind data set was used to count the neurons that was previously used to analyze model efficiency in Figure 4.