# Automated skull-stripping and segmentation with Multi-Task U-Net in large mouse brain MRI databases

**Riccardo De Feo[1,2,3,*], Artem Shatillo[4], Alejandra Sierra[3], Juan Miguel Valverde[3], Olli Gröhn[3], Federico Giove[2,5], and Jussi Tohka[3]**

[*] ***riccardo.defeo@uniroma1.it***
[1] ***Sapienza Università di Roma, 00184 Rome Italy***
[2] ***Centro Fermi–Museo Storico della Fisica e Centro Studi e Ricerche Enrico Fermi, 00184 Rome Italy***
[3] ***A.I. Virtanen Institute for Molecular Sciences, University of Eastern Finland, 70210 Kuopio, Finland***
[4] ***Charles River Discovery Services, Kuopio, Finland***
[5] ***Fondazione Santa Lucia IRCCS, 00179 Rome Italy***

## ABSTRACT

Skull-stripping and region segmentation are fundamental steps in preclinical magnetic resonance imaging (MRI) studies, and these common procedures are usually performed manually. We present Multi-task U-Net (MU-Net), a convolutional neural network designed to accomplish both tasks simultaneously. MU-Net achieved higher segmentation accuracy than state-of-the-art multi-atlas segmentation methods with an inference time of 0.35 seconds and no pre-processing requirements. We evaluated the performance of our network in the presence of skip connections and recently proposed framing connections, finding the simplest network to be the most effective. We tested MU-Net with an unusually large dataset combining several independent studies consisting of 1,782 mouse brain MRI volumes of both healthy and Huntington animals, and measured average Dice scores of 0.906 (striati), 0.937 (cortex), and 0.978 (brain mask). These high evaluation scores demonstrate that MU-Net is a powerful tool for segmentation and skull-stripping, decreasing inter and intra-rater variability of manual segmentation. The MU-Net code and the trained model are publicly available at https://github.com/Hierakonpolis/MU-Net.

## 1 Introduction

Preclinical imaging studies serve a fundamental role in biological and medical research, relating research results at the molecular level to clinical application in diagnosis and therapy. Magnetic Resonance Imaging (MRI) represents about 23% of all small-animal imaging studies providing the opportunity to monitor the development of pathological condition and response to treatment in a non-invasive way[1]. Its unique qualities also include the availability of different imaging contrasts, making MRI especially useful in the context of preclinical neuroscience, from drug development[2] to structural, parametric and functional studies[3].

Skull-stripping and region segmentation represent an integral part of processing pipelines in both human and murine MR imaging. Skull-stripping refers to the identification of the brain within the MRI volume and region segmentation refers to the labeling of specific anatomical regions of interest (ROIs) within the brain. Especially in preclinical imaging, these tasks are often performed manually. While manual segmentation represents the gold standard and is employed as the ground truth when evaluating automated segmentation algorithms, it is a slow process that depends on the expertise of the individual expert. Furthermore, manual segmentation suffers from both intra- and inter-rater variability. Dice scores are typically used to quantify segmentation agreement[4]. Inter- and intra-rater Dice scores reported in literature range between 0.80 to 0.96 across different segmentation tasks, and can be as low as 0.75 in the presence of pathological states[5–7].

In preclinical MRI, state-of-the-art automated region segmentation pipelines are based on atlas registration: individual MRI volumes are aligned with a labeled template (atlas) to propagate the labels to the individual volumes[8–10]. The accuracy of registration-based segmentation depends on the accuracy of the template and the registration itself. Its accuracy can be further improved by registering multiple atlases to the same volume and combining the so-resulting maps. Bai et al.[11] compared different single and multi-atlas methods for atlas-based segmentation of the mouse brain, finding that the combination of a diffeomorphic registration algorithm and multi-atlas segmentation provided the most accurate results. While multi-atlas segmentation accounts for individual variability more effectively than single-atlas segmentation, it also requires multiple labeled atlases and multiple registration steps, significantly increasing the segmentation time.

Registration-based methods also lose effectiveness when the target volumes present large anatomical alterations or deformations, for example, as a consequence of neurodegeneration or injury. Small registration errors can still result in a marked decrease in accuracy depending on the shape of the regions considered, strongly affecting small, thin or elongated

regions. For example, while Ma et al.[12] achieved Dice scores of about 0.90 in several gray matter ROIs, for the ventricles and several white matter structures such as the fimbria, the external capsule, the internal capsule and the anterior commissure the Dice overlap remained around 0.70.

Instead of directly employing one or more manually segmented atlases, deep neural networks (DNNs)[13] can use these as training data to learn a mapping function, from the image data to the segmentation maps. In this way, the anatomical information is not explicitly represented in a set of maps but implicitly encoded in the trained network. DNNs have been successfully applied in a large number of computer vision tasks, including biomedical imaging[14]. Roy et al.[15] trained a Convolutional Neural Network (CNN) algorithm based on Google Inception[16] for the skull stripping in humans and mice after traumatic brain injury, achieving better performance than other state-of-the-art methods, in particular 3D Pulse Coupled Neural Networks (3D-PCNN)[17] and Rapid Automatic Tissue Segmentation (RATS)[18]. Wachinger et al.[19] developed a region segmentation CNN significantly outperforming state-of-the-art, registration based methods for the healthy human brain, both in terms of inference time and accuracy. Roy et al.[20] further improved on both aspects with a network based on the U-Net architecture[21], with a reported segmentation time of 20 s.

A specific type of CNN architecture, U-Net, has proved to be valuable in biomedical image segmentation. U-Net is based on the encoder/decoder structure, adding skip connections between the encoder branch and the decoder branch, allowing it to easily integrate multi-scale information and better propagate the gradient during training. This architecture has been shown to generalize even from a limited amount of annotated data[22], and as such is well suited for medical imaging, where datasets as large as the ones commonly used for CNNs are uncommon. Valverde et al.[23] recently demonstrated the effectiveness of U-Net-like architectures in preclinical research, designing the first DNN for the segmentation of ischemic lesions in rodents and achieving overlap scores comparable with human inter-rater Dice overlaps.

In this work, we introduce multi-task U-Net (MU-Net) to simultaneously perform skull-stripping and region segmentation of the mouse brain, based on the U-Net architecture. Our training data consisted of 128 coronal $T_2$ MRI volumes from 32 mice at 4 different ages as well as five manually annotated regions (cortex, hippocampi, ventricles, striati and brain mask) from these images. We tested MU-Net in 1782 heterogeneous volumes acquired over the course of four years from both wild type (WT) and Huntington (HT) C57BL/6J mice, allowing us to evaluate MU-Net in a variety of experimental conditions. Our network achieves a higher accuracy than state-of-the-art multi-atlas segmentation methods with an inference time of approximately 0.35 s, achieving Dice overlaps comparable with the highest examples of inter-rater Dice scores in literature.

# 2 Results

## 2.1 MU-Net

MU-Net is a U-Net based network combining region segmentation and skull stripping to build a shared representation for both tasks, providing at the same time a more efficient feature encoding and a regularizing effect. We present an overview of MU-Net's architecture in Figure 1. Briefly, MU-Net presents an encoder-decoder U-Net-like architecture, with each branch articulated in four dense blocks. Unlike U-Net, the final block of the decoder branch further branches into two different output maps representing our two tasks, sharing the same feature representation. MU-Net was trained using the Adam optimizer[24] to jointly minimize the Dice loss for both skull-stripping and region segmentation[25]. For further details refer to section 4.3.

MU-Net was trained and validated on 128 mouse MRI volumes acquired from 32 animals at four different ages (5, 12, 16, and 32 weeks). The training set was annotated for the brain mask and four more regions: cortex, striati, hippocampi and ventricles, each category including both left and right regions. For validation, we used 5-folds cross validation and measured overlaps using the Dice score[4]. The hypothesis tests between the Dice coefficients of different segmentation method were performed using permutation tests and all the reported p-values are obtained from a permutation test. Validation and metrics are further expanded upon in section 4.2.

Using the training and validation dataset, we compared the performance of different network architectures based on the presence of skip connections, framing connections and 2D or 3D filters. Furthermore, we compared MU-Net with multi-atlas segmentation and evaluated the impact of mouse age on the accuracy of our segmentation maps. Finally, we tested MU-Net on a large dataset including 1782 MRI volumes from 817 mice. These volumes were acquired as parts of 10 studies on Huntington's disease and included both WT and HT mice with ages ranging from 4 to 60 weeks. This heterogeneous dataset included three segmented regions, for which we measured average Dice scores of 0.906 (striati), 0.937 (cortex), and 0.978 (brain mask) between MU-Net and manual segmentations.

## 2.2 Architecture comparison

We compared the performance of different networks trained with and without dense connections and dual framing connections, in both 2D and 3D implementations. By a dense connection (also termed residual connection) we refer to skip connections inside each dense block of the MU-Net architecture. Dual framing connections refer to the skip connections originally introduced by Han and Ye[26] for an image reconstruction task.

**Table 1.** Convolutional neural network and Multi-STEPS accuracy across all examined regions, for each examined architecture.

| Dim | SC | FC | Brain mask | Cortex | Hippocampi | Ventricles | Striati | ROI mean |
|---|---|---|---|---|---|---|---|---|
| 2D | | | **0.984±0.005** | **0.966±0.009** | **0.925±0.017** | **0.907±0.020** | **0.939±0.010** | **0.935±0.026** |
| 2D | x | x | **0.984±0.006** | 0.963±0.010 | **0.924±0.016** | 0.905±0.022 | 0.937±0.009 | 0.932±0.026 |
| 2D | | x | **0.984±0.006** | 0.963±0.011 | **0.924±0.017** | 0.905±0.022 | 0.938±0.009 | 0.932±0.026 |
| 2D | x | | **0.984±0.005** | 0.964±0.011 | 0.923±0.018 | 0.905±0.024 | 0.937±0.010 | 0.932±0.027 |
| 3D | x | x | 0.982±0.007 | 0.956±0.016 | 0.914±0.033 | 0.900±0.025 | 0.926±0.045 | 0.924±0.038 |
| 3D | | x | 0.982±0.007 | 0.958±0.016 | 0.916±0.032 | 0.900±0.025 | 0.928±0.029 | 0.925±0.034 |
| 3D | x | | 0.982±0.006 | 0.957±0.016 | 0.913±0.041 | 0.899±0.028 | 0.926±0.042 | 0.924±0.040 |
| 3D | | | 0.982±0.007 | 0.957±0.013 | 0.916±0.033 | 0.899±0.026 | 0.926±0.039 | 0.924±0.036 |
| 3DConv 2DPool | | | 0.983±0.006 | 0.961±0.010 | 0.919±0.026 | 0.902±0.026 | 0.934±0.014 | 0.929±0.030 |
| 2D SLP | | | **0.984±0.005** | **0.965±0.009** | **0.924±0.016** | **0.907±0.021** | **0.939±0.010** | **0.934±0.026** |
| 2D + N3 | | | **0.984±0.005** | **0.965±0.009** | **0.924±0.020** | **0.907±0.020** | **0.939±0.009** | **0.934±0.026** |
| STEPS (affine) | | | \ | 0.920±0.058 | 0.827±0.079 | 0.761±0.090 | 0.873±0.062 | 0.845±0.093 |
| STEPS (diffeo) | | | \ | 0.948±0.036 | 0.844±0.048 | 0.812±0.090 | 0.871±0.045 | 0.869±0.070 |
| STEPS* (affine) | | | \ | $0.936 \pm 0.013$ | $0.831 \pm 0.029$ | $0.781 \pm 0.049$ | $0.887 \pm 0.019$ | $0.859 \pm 0.066$ |
| STEPS* (diffeo) | | | \ | $0.954 \pm 0.009$ | $0.845 \pm 0.048$ | $0.826 \pm 0.039$ | $0.885 \pm 0.016$ | $0.879 \pm 0.055$ |
| Majority Voting | | | \ | 0.889±0.179 | 0.780±0.232 | 0.677±0.208 | 0.816±0.245 | 0.791±0.230 |

Listed values are the average Dice scores between automatic and manual segmentation in 5-fold CV $\pm$ standard deviations of these Dice scores. ROI mean column refers to the mean Dice coefficient of the cortex, the hippocampi, the ventricles and the striati. SC and FC mark the presence of skip connection and framing connections. MU-Net results are displayed in the first row. STEPS refers to Multi-STEPS using randomly selected templates; STEPS* refers to Multi-STEPS runs using randomly selecting mice of the same age only; affine indicates that only affine registration was used, whereas diffeo indicates this was followed by a diffeomorphic registration step; Majority voting refers to the selection of the most occuring label after diffeomorphic registration; 3DConv 2DPool: network featuring no in-block skip connections or framing connections, with 3D 5x5x5 filtering and 2D pooling in the coronal plane only; 2D SLP: 2D network with in-block skip connections and a limited number of parameters; 2D +N3: 2D network trained on data bias-corrected using the N3 algorithm. Boldface characters indicate the best performing network, achieving significantly higher Dice scores than all other networks for that ROI

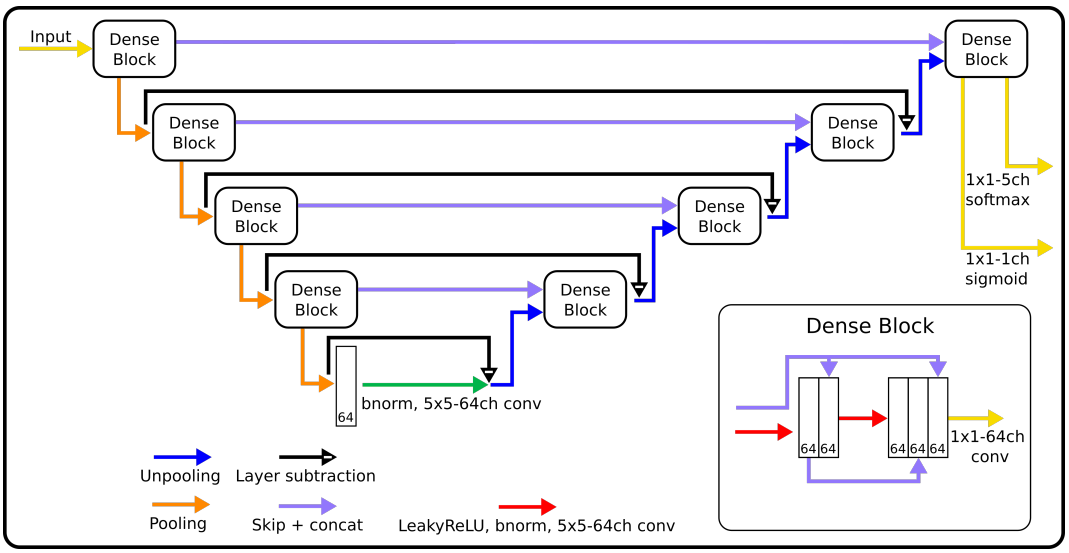


**Figure 1.** General outline of the architectural features implemented and compared in the networks discussed, varying according to the presence or absence of the in-block dense connections (purple arrows in the dense block box), presence or absence of the layer subtraction connections (black), and the use of 2D or 3D filters.

As shown in Table 1, all MU-Nets achieved Dice overlaps with the ground truth comparable to or higher than the typical intra-rater variability of manual segmentation (Dice scores from 0.80 to 0.96). The skull-stripping task achieved an excellent Dice score of 0.984. The ventricles were characterized by the lowest segmentation performance (average Dice score 0.907), while the cortex displayed the highest overlap with the ground truth (average Dice score 0.966). Dice scores for each animal in all ROIs are provided as a supplementary csv-file.

The network displaying the highest average Dice overlaps was, in fact, the simplest one, including no in-block skip connections nor framing connections, and using 2D convolutions. The accuracy of this network was significantly higher than the accuracy of other all other 2D networks ($p < 0.00003$). Because of its high performance and simplicity this network is the natural choice for the MU-Net architecture.

The choice between 2D and 3D architectures was the most important factor in increasing performance, resulting in a marked increase in mean Dice scores for both tasks ($p < 0.00001$) between all 2D networks compared to the 3D ones. We further compared MU-Net with one featuring less channels per filter (49, 49, 50, 50, from the shallowest to the deepest dense block) to match the number of parameters to the number of parameters of the simplest 2D network. We registered a slightly (but not significantly, $p = 0.077$, permutation test) lower accuracy compared to MU-Net, indicated as 2D SLP in Table 1.

To test whether the increased performance of 2D architectures compared to the 3D implementation depended on the reduced number of parameters or on an excessive loss of information when pooling in the anterior-posterior direction, we trained a network using 3D filters while only pooling in the coronal plane. This network achieved intermediate performance between the 3D and 2D implementations (Table 1), suggesting that both above mentioned aspects were relevant in increasing the algorithm's performance.

We trained MU-Net on the N3 bias-corrected data[27]. The validation accuracy achieved in this way was indistinguishable from the one of MU-Net trained on the non-corrected data; applying N3 bias correction did not lead to any noticeable difference.

## 2.3 Comparison with multi-atlas segmentation

We compared MU-Net with multi-atlas segmentation, applying the state-of-the-art Multi-STEPS[28,29] label fusion method to combine the labels obtained from the registration of multiple labeled volumes. Multi-STEPS takes into account local and global image matching, combining an expectation-maximization approach with Markov Random Fields[30] to improve on the segmentation based on the quality of the registration itself. We repeated this procedure using both diffeomorphic and affine registration methods, and both by using randomly selected templates and by restricting the random selection of templates to mice of the same age only. The brain mask segmentation was not evaluated as the manually drawn mask was used during the diffeomorphic registration procedure.

MU-Net achieved higher accuracy than all Multi-STEPS implementations ($p < 0.00001$). There was a marked qualitative difference between Multi-STEPS segmentation and MU-Net (Fig. 2), the latter achieving results visually indistinguishable

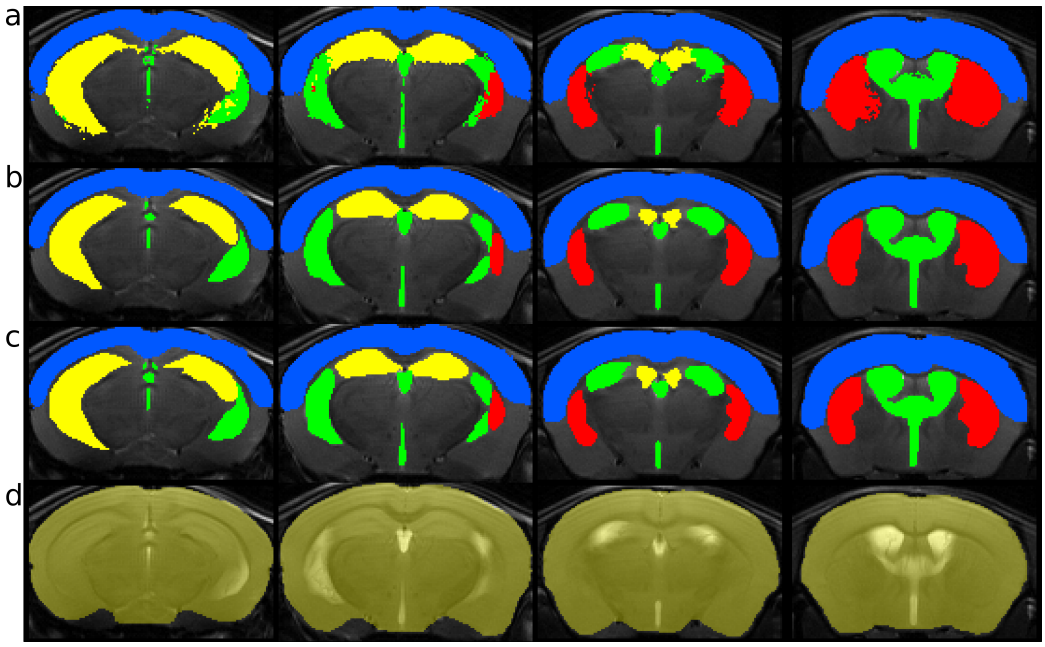


**Figure 2.** Segmentation comparison in four slices from a single animal: (a) Multi-STEPS, (b) MU-Net, and (c) manual annotation. In (a)-(c), the regions highlighted are the cortex (blue), ventricles (green), striati (red), and hippocampi (yellow). Panel (d) shows the inferred brain mask by MU-Net.

**Table 2.** Mean and standard deviation of average Dice overlaps evaluating the accuracy of MU-Net trained on volumes segmented via Multi-STEPS.

| Training Set | Mask | Cortex | Hippocampus | Ventricles | Striatum | ROI mean |
|---|---|---|---|---|---|---|
| STEPS* | / | 0.954±0.011 | 0.867±0.027 | 0.866±0.035 | 0.898±0.017 | 0.896±0.043 |
| STEPS | / | 0.953±0.009 | 0.872±0.022 | 0.849±0.041 | 0.885±0.016 | 0.890±0.046 |

from manual segmentation.

MU-Net had an inference time of about 0.35 s (one forward pass) and a training time of 12 hours. Multi-STEPS segmentation procedure required total inference time of 117 minutes for each labeled volume (on average 440 s for each pairwise diffeomorphic registration and 7.85 s for label fusion). Implementing Multi-STEPS segmentation using only templates of the same age led to a small but significant ($p < 0.0007$) performance improvement over randomly choosing templates of any age. The employment of diffeomorphic registration was the most important factor affecting the performance of Multi-STEPS, as displayed in Table 1. A simple majority voting strategy led to significantly lower performance in all ROIs compared to all other label fusion strategies ($p < 0.003$).

Furthermore, we trained MU-Net on the outputs of the implemented Multi-STEPS procedures and measured the Dice overlaps of each network's output with the ground truth (Table 2). As evidenced in Tables 1 and 2, and Figure 3, MU-Net trained on Multi-STEPS segmentations achieved higher Dice overlap with the ground truth than the same Multi-STEPS segmentations constituting their training sets ($p < 0.00001$). With the exception of the network trained on 5 weeks old mice exclusively, these networks were still under-performing compared to training on manually segmented data ($p < 0.00001$).

## 2.4 Age stratified training sets

We evaluated the performance of MU-Net when restricting the training set to mice of a specific age only. Networks trained on data from mice of 12, 16 and 32 weeks achieved higher accuracy, both on their respective validation set and the overall ground truth, compared to the networks trained on 5 weeks mice ($p < 0.00001$). As shown in Fig. 4, while all networks trained on one specific age displayed a statistically significant ($p < 0.05$, unpaired) decrease in mean accuracy when validated on animals of a

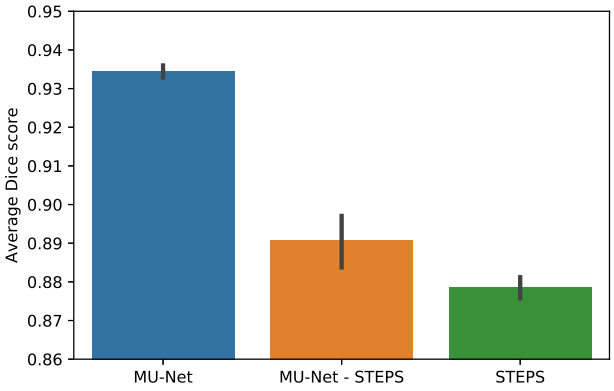


**Figure 3.** Dice score comparison between different segmentation methods. MU-Net: MU-Net trained on the manually segmented data;MU-Net - STEPS: MU-Net trained on volumes segmented using Multi-STEPS; STEPS: Multi-STEPS segmentation. The error bar represents standard deviation.

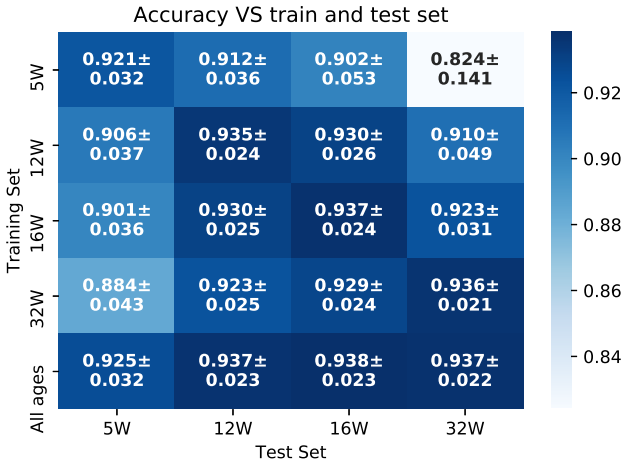


**Figure 4.** Mean accuracy $\pm$ standard deviation for the average accuracy of MU-Net trained and evaluated on different datasets according to mouse age. Networks exclusively trained on older animals achieved lower accuracy when attempting to generalize to the youngest animals, and vice-versa.

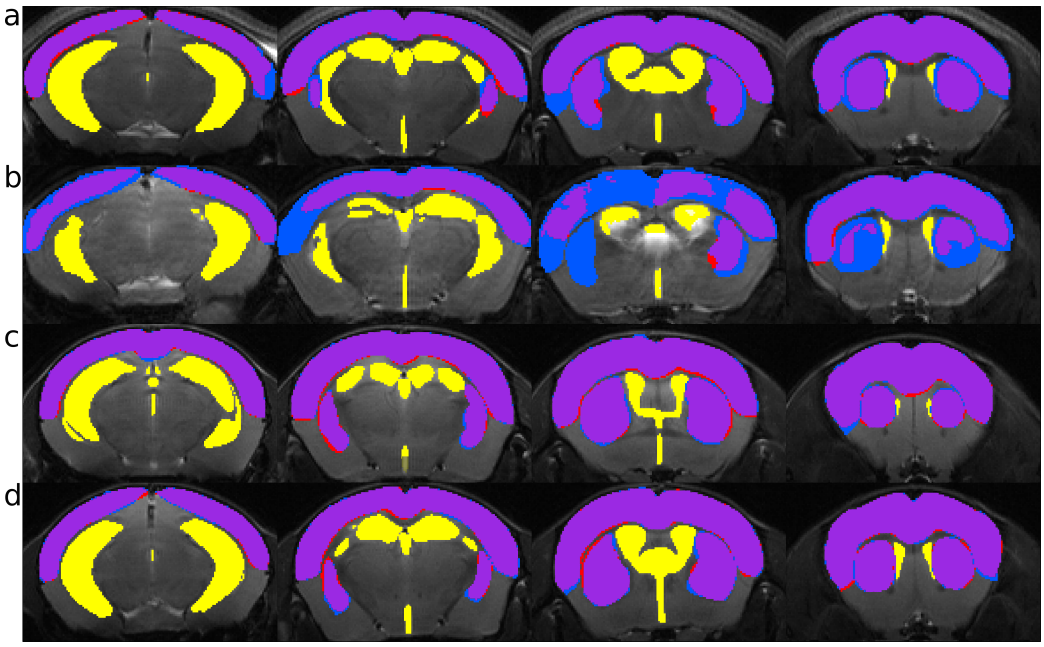


**Figure 5.** MU-Net segmentation compared to the manual segmentation in four slices of four volumes of the test set. Blue and red indicate, respectively, ground truth and inferred segmentation, purple their overlap (striati and cortex); yellow ROIs (ventricles and hippocampi) are inferred ROIs for which manual annotations were not available. Rows indicate (a) the highest performing volume (mean Dice 0.964, 8 weeks old R6/2 mouse); (b) the lowest performing volume (mean Dice 0.685, 12 weeks old R6/2 mouse); (c) the volume displaying performance closest to the mean performance on the entire test set (Dice 0.923, 12 weeks old Q175DN mouse); (d) one randomly selected volume (Dice 0.919, 8 weeks old Q175DN mouse)

different age, this difference was highest between the 5 weeks data and the other datasets.

Limiting the training data to one specific age implies that these networks were trained only on a quarter of the data used to train the networks in section 2.2. Irrespective of that, these networks still achieved average Dice overlaps on the mixed-age validation dataset comparable with the accuracy of manual segmentation. The worst performing was the network trained on 5 weeks old mice. Training on the 12, 16 and 32 weeks data and validating on mice of the same age, we observed Dice scores comparable with the overall performance of MU-Net trained on the entire dataset ($p > 0.15$, unpaired). However, we measured a lower overall performance when including mice of all ages in the validation data ($p < 0.00001$), slightly overfitting for each specific age.

## 2.5 Evaluation with a large test dataset

The MU-Net model trained on a train and validation dataset was tested on a large test set of 1782 MRI volumes, acquired from 817 mice with ages ranging from 4 to 60 weeks, and including both WT and HT mice. As the 5-fold cross-validation experiment produced five different MU-Net models, the segmentation maps for the test set were obtained by averaging the prediction maps produced by each of the five models.

Out of the entire test set, segmentation failed completely on two volumes, where no brain mask was detected at all. The remaining 1780 volumes were successfully segmented with an average Dice overlap of $0.978 \pm 0.012$ for the brain mask, $0.906 \pm 0.041$ for the striati, and $0.937 \pm 0.035$ for the cortex, distributed as illustrated in Fig. 6. There was no significant difference between the segmentation accuracy of male and female animals ($p > 0.1$, unpaired). However, there was a significant difference in accuracy between HT and WT mice ($p < 0.00001$, unpaired) for all ROIs. Dice scores of WT animals were 0.4% higher for the brain mask, 1.7% higher for the cortex, and 1.9% higher for the striati. Applying N3 bias correction on all volumes before segmentation did not result in a significant Dice score difference. A detailed list of the Dice scores for each animal and each ROI is available as an supplementary csv-file.

A visual inspection of the segmentation maps (Fig. 5) revealed that ROIs were qualitatively similar to those obtained on the validation set and displayed in Fig. 2. We observed, however, a visible decrease in performance in the presence of strong ringing artifacts (Fig. 5.b)

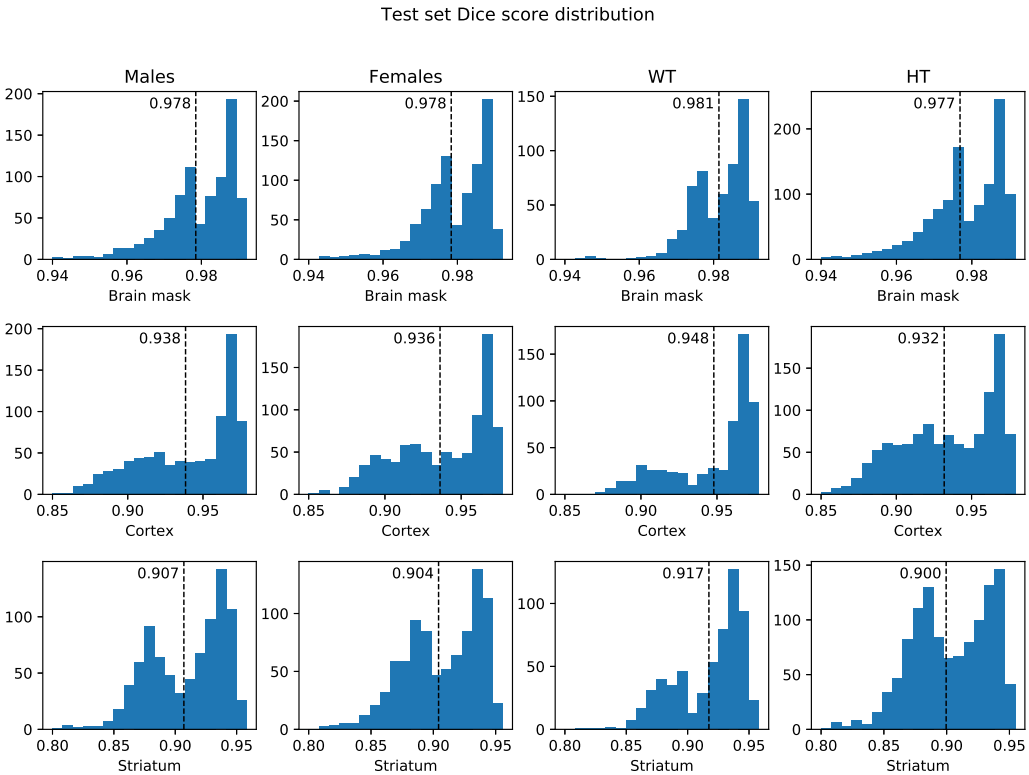


**Figure 6.** Test set Dice score distribution for the brain mask, cortex and striati ROIs. Males and Females include all mice of each gender, both WT and TG. Likewise, WT and TG include both males and females.

# 3 Discussion

We presented a multi task deep neural network, MU-Net, for the simultaneous skull-stripping and segmentation of mouse brain MRI. We selected the best performing network between a number of proposed architectures and found it to achieve higher segmentation accuracy on the validation set compared to state-of-the-art multi-atlas segmentation procedures, with a much lower inference time. We then evaluated the performance of MU-Net on a large and heterogeneous test set test set of 1,782 mice coming from 10 different studies of Huntington disease, with varying ages and genetic backgrounds (WT as well as HT Q175 and R6/2 variants). In this test set we measured Dice overlaps of 0.978, 0.906 and 0.937 for the brain mask, striati and cortex, rivaling human-level performance. The employment of deep neural networks for the segmentation of animal MRI is a promising strategy for the reduction of both rater bias and segmentation time.

While Snehashis et al.[15] have proposed a CNN for skull-stripping for mouse MRI, to our knowledge this work represents the first CNN performing both region segmentation and skull-stripping in mouse brain MRI. The advantages of CNNs to the atlas based region segmentation[8,11] are clear. First, compared to atlas-based segmentation MU-Net is much faster and produces accurate results without pre-processing. Second, we found MU-Net to be significantly more accurate than Multi-STEPS region segmentation. Regarding label fusion strategies, in contrast to Bai et al.[11], we found that majority voting was the least effective method, while in agreement with Bai et al.[11] the usage of diffeomorphic registration was the most important factor in increasing the accuracy of registration-based segmentation.

Interestingly, the networks trained on Multi-STEPS segmentations achieved higher overlap with the ground truth than Multi-STEPS, highlighting the generalization ability of MU-Net. This supports the use of atlas based segmentation methods to augment MRI segmentation datasets in analogy with Roy et al.[20], leveraging unlabeled data. The results obtained by training on Multi-STEPS segmentations alone remain, however, of insufficient quality to eliminate the need for manual annotations in the training data, as the CNN attempts to replicate any form of systematic error present in the atlas-based labeling procedure.

The Dice score of MU-Net for skull-stripping was higher than the reported Dice score from the skull-stripping DNN implemented by Snehashis et al.[15] (our Dice score was 0.984 vs 0.949 in Snehashis et al.[15]. However, a direct comparison between the two segmentation accuracies is unreliable as the networks are trained and validated on different datasets. The increased performance might also depend on the higher in-plane resolution of the coronal slices in our dataset and the larger quantity of data available for both training and testing MU-Net.

In literature both 3D and 2D implementations of CNNs are available for different segmentation tasks[20,31,32], and other architectural variants have been proposed: Roy et al.[20] added dense connections[33] in the convolution blocks of U-Net while keeping the number of output channels constant; Han and Ye[26] proposed two variants based on signal processing arguments for the reduction of artifacts in a sparse image reconstruction task. We, however, found that a more complex model did not improve and in fact lowered the accuracy of our results, perhaps given the simplicity of the task. Thus in agreement with Isensee et al.[34] we found that a 2D approach was preferable to 3D approach in the presence of anisotropic voxels. We also found the Dice loss to be sufficient to effectively train our model without the addition of a cross-entropy loss.

The highest Dice overlap was measured in the cortex, while the lowest in the ventricles. This might appear counter-intuitive, as $T_2$ MRI offers much higher contrast for the ventricles compared to our other ROIs, thanks to the longer $T_2$ relaxation time. However, ventricles are more subject to partial volume effects and present more elongated shapes, increasing the effect of errors on the region boundary. This result might also depend on the manual rater's variability when assigning a class to voxels presenting partial volume effects.

Much like the human eye, MU-Net was not significantly affected by the presence of the bias field, and did not benefit from N3 bias correction. Correcting for the bias field might still be beneficial as it depends on the specific experimental setup, and so N3 bias correction might avoid specializing the network to one particular acquisition procedure. For this reason, we release the trained parameters of the model for MU-Net trained on both the non-corrected and the N3-corrected data.

To ensure the network generalizes to a wide age range, our results indicate that the distinctive features present before adulthood need to be adequately represented in the training data. This is evidenced by the degraded performance observed when testing networks trained on 5-week old mice on the volumes acquired from older ones, and vice-versa. As mice are typically weaned at 3-4 weeks and attain sexual maturity at 8-12 weeks[35], 5-week old mice are not adults. In contrast, training on male mice only did not significantly influence MU-Net's performance on female animals.

An obvious limitation of our approach is its specialization for the specific contrast the algorithm is trained on, in this case $T_2$ MRI. While we would expect a lower accuracy using a registration-based approach, it is possible to attempt an inter-modal registration using an appropriate similarity metric, e.g. mutual information. Expanding our approach to a different contrast would require expanding the training set, image translation or transfer learning. Even considering our limitations, MU-Net successfully generalized to a variety of TG mice in an age range wider than that of the training set.

The employment of CNNs for the segmentation of mouse brain MRI provides a number of benefits for preclinical researchers. Beyond allowing for the employment of large datasets in a time-efficient manner, the ability to generalize and abstract from the training data results in more robust and reproducible predictions. We can thus expect these methods to reduce

the confounding effect of intra- and inter-rater variability inherent in manual segmentation procedures while streamlining animal MRI experimental pipelines.

# 4 Methods

## 4.1 Datasets

### 4.1.1 Animals

A total of 849 C57BL/6J mice (Charles River Laboratories, Germany) were used in this work: 32 mice for our train and validation set and 817 mice for our test set.

Training set animals were scanned at four different ages (5 weeks, 12 weeks, 16 weeks, 32 weeks) resulting in 128 volumes, and did not include any HT genotypes. All training set animals were males.

The test set animals were part of 10 studies scanned at a single or multiple ages from 4 up to 60 weeks, and included both WT and several HT genotypes: R6/2, Q175, Q175DN, Q111, Q50 and Q20 (Supplementary Table S1), for a total of 1,782 MRI scans. The groups included both males and females. These volumes were acquired as part of ten studies of Huntington's disease, kindly provided by the CHDI 'Cure Huntington's Disease Initiative' foundation.

All mice were housed in groups of up to 4 per cage (single sex) in a temperature (22±1°C) and humidity (30-70%) controlled environment with a normal light-dark cycle (7:00-20:00). All animal experiments were carried out according to the United States National Institute of Health (NIH) guidelines for the care and use of laboratory animals, and approved by the National Animal Experiment Board.

### 4.1.2 MRI

Mice were anesthetized using isoflurane (5% for induction, 1.5-2% maintenance) in 70%/30% mix of N2/O2 carrying gas, fixed to a head holder and positioned in the magnet bore in a standard orientation relative to gradient coils. Respiration rate and temperature were monitored using PC-SAMS software and Model 1030 Monitoring & Gating System, Small Animal Instruments, Inc., Stony Brook, NY. The temperature was maintained at $\sim 37\,^{\circ}$C using Small Animal Instruments feedback water heating system.

All acquisitions were performed using a horizontal 11.7 T magnet with a bore size of 160 mm, equipped with a gradient set capable of maximum gradient strength of 750 mT/m and interfaced to a Bruker Avance III console (Bruker Biospin GmbH, Ettlingen, Germany). A volume coil (Bruker Biospin GmbH, Ettlingen, Germany) was used for transmission and a surface phased array coil for receiving (Rapid Biomedical GmbH, Rimpar, Germany). $T_2$ weighted anatomical images were acquired using a TurboRARE sequence with effective TR/TE = 2500/36 ms, 8 echoes, 12 ms inter-echo distance, matrix size 256x256, FOV 20.0x20.0 mm$^2$, 31 0.6 mm thick coronal slices, $-0.15$ mm interslice gap, and 8 averages. Concerning the test data, MRI experimental parameters only differed in acquiring 19 0.7 mm thick contiguous coronal slices.

Volumes within each study were manually segmented by an experienced rater, who had received a training and passed the qualification tests according to SOP (Standard Operating Procedure) for volumetric analysis in mice. Different studies were analyzed by different raters. Each training volume was manually segmented by drawing the brain mask and delineating 4 regions of interest: cortex, hippocampi, striati and ventricles. The brain mask did not include the olfactory bulb or the cerebellum. For the test set, only 3 regions were manually labeled: brain mask, cortex and striati.

## 4.2 Validation and metrics

To assess the overlap between the ground truth and the predicted segmentation masks, we used the Dice coefficient[4]. The Dice coefficient is defined as two times the size of the intersection over the sum of the sizes of the two regions:

$$D = \frac{2\,|X \cap Y|}{|X| + |Y|}$$

This coefficient ranges from 0, meaning no overlap, to 1, indicating a complete overlap between the two regions. A visual rendition for the intuition behind the Dice score is provided in the supplementary Figure S1.

Each experiment reported in sections 2.2 - 2.4 was validated by splitting the training and validation dataset according to a 5-fold cross validation scheme. Unless otherwise specified, training and validation data included animals of all ages. Volumes were distributed in each fold according to the individual identity of each animal, preventing the use of the volumes from the validation animals for training. The animals were randomly assigned to each fold once, and the same animals remained assigned to their respective folds through all experiments. This resulted in a validation set of 6 or 7 animals and a training set of 25 or 26 mice. These correspond to an identical number of MRI volumes when considering each animal at one specific timepoint only, and to four times as many volumes (24-28 validation volumes, 100-104 training volumes) in all other cases.

Unless otherwise specified, we used a paired permutation test to evaluate the significance of differences between the Dice scores obtained by different methods, pairing the results obtained on the same MRI volumes. The unpaired permutation test was used instead when comparing results obtained on different volumes, for example, when comparing the accuracy of a model on volumes from younger mice with that of the same model on older mice. We performed permutation tests using 100,000 iterations, and considered differences in average to be significant when $p < 0.05$.

## 4.3 Neural Networks

We segmented mouse brain MRI volumes into five regions (cortex, hippocampi, ventricles, striati and background) simultaneously generating brain mask binary segmentation via a multi-task learning approach. We refer to the first task as the region segmentation and to the latter task as skull-stripping. We implemented and compared different variants of the U-Net architecture.

### *4.3.1 Architectures*

The MU-Net architecture (Fig. 1) is composed of one encoding path and one decoding path. The encoding path includes 4 stages, each articulated in one dense block followed by a 2x2 max-pooling layer. The last feature map feeds into the bottleneck layer, a 64 channel 5x5 convolutional layer with batch normalization[36] connecting the deepest layer of the encoding path with the decoding path.

The decoding path is composed of 4 more blocks alternating one un-pooling layer[37] and one dense block. Un-pooling operations effectively replace up-convolution layers in U-Net without any learnable parameters, while preserving spatial information. These layers operate by simply placing the elements of the un-pooled layer in the position of the respective maximum activation from the corresponding pooling operation, and setting the rest to zero. Skip connections concatenate the output of each dense layer in the encoding path with the respective un-pooled feature map of the same size before feeding it as input to the decoding dense block.

The output of the last decoding layer acts as the input of two different classification layers, which share the same feature representation up to this point: a 1x1 single channel convolution with a sigmoid activation function, and a 1x1 5 channels layer followed by a softmax activation function, for the skull-stripping task and the region classification task, respectively.

**Dense block** Each dense block includes 3 convolutional layers preceded by leaky ReLU activation layers and batch normalization. In the models including dense connections[33] the input of each convolution is composed by concatenating the outputs of the previous convolutions within the same block and the block's input. In models where these dense connections are not employed each LeakyReLU / BatchNormalization /convolution sequence simply feeds into the next.

All 3 convolutions are padded and result in 64 output channels, in analogy with Roy et al.[20]. The first and second convolutions employ 5x5 filters, while the third employs a 1x1 filter. This becomes especially relevant in the presence of dense connections, acting as a bottleneck for the 64x3 channels of the concatenated inputs and compressing the size of the feature maps.

**Dual Framing connections** Dual framing connections refer to additional skip connections proposed by Han and Ye[26] in the Dual Frame U-Net model. This architecture was proposed for computed tomography reconstruction from sparse data based on signal processing arguments to reduce artifacts and improve recovery of high frequency edges. We investigated whether it could prove advantageous in our segmentation tasks. Dual framing connections consist in the subtraction of the input of each dense block on the encoding path from the output of the respective dense block of the same size on the decoding path, and as such the implementation of these connections does not increase the number of model parameters.

**3D implementation** A 3D implementation should in principle provide better results by taking into account the features of the adjacent slices, whereas a 2D networks evaluates each coronal slice independently. However, the larger number of parameters also increases the risk of overfitting, and the lower resolution in the anterior-posterior axis compared to the in-plane resolution might constitute confounding factors in the presence of 3D pooling operations.

For these reasons, we confronted 2D and 3D implementations of our network, using 5x5x5 filters and 2x2x2 max-pooling layers, replacing the filters and pooling layers described above. This results in 16008076 and 10286344 parameters for the 3D networks with and without in-block skip connections, respectively. Corresponding 2D networks contain 3297676 and 2087944 parameters, respectively. Thus opting for a 3D architecture increases the number of parameters by factors of 4.85 and 4.93.

### *4.3.2 Loss function*

Recent literature suggests that Dice-based loss functions[20,25,31] would constitute an improvement over cross entropy losses for the segmentation of medical images[38]. We optimize a joint loss function $L$, that is the sum of two Dice loss functions corresponding to the the skull-stripping ($L_{SS}$) and the region classification task ($L_{RS}$). Let $p(i)$ be the predicted probability of voxel $i$ of belonging to the brain mask, and $g(i)$ the ground truth for voxel $i$ ($g(i) = 1$ if the voxel is in the brain mask).

Further, let $p_l(i)$ and $g_l(i)$ be the same quantities for label $l$ ($l = 0,1,2,3,4$ referring to cortex, hippocampi, ventricles, striati, background) encoding the ground truth as a one-hot vector. Then, the loss function can be written as:

$$L = L_{SS} + L_{RS},$$

$$L_{SS} = -\frac{2\sum_i p(i)g(i)}{\sum_i p^2(i) + \sum_i g^2(i)},$$

$$L_{RS} = -\sum_{l=0}^{4} \frac{2\sum_i p_l(i)g_l(i)}{\sum_i p_l^2(i) + \sum_i g_l^2(i)},$$

#### 4.3.3 Training

The networks were implemented using the PyTorch framework and trained with stochastic gradient descent using Adam optimizer[24] with the default parameters (initial learning rate of 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and no weight decay) on an NVIDIA GeForce GTX 1080 GPU. Each training cycle required approximately 12 hours. Training and validation datasets were built according to the 5-fold cross validation scheme described in section 4.2.

Data augmentation was applied on training data. It included random rescaling between 1.01 and 0.95 and random rotation of $\pm 5°$. Smaller sizes were preferred in rescaling to decrease memory requirements. Each transformation was applied with a 50% probability. For the same memory reason, a bounding box was created for each volume using the annotated brain mask as a reference. Each volume was individually normalized to 0 mean and unit variance.

#### 4.3.4 Auxiliary bounding-box network

As MU-Net was trained after cropping the volumes to a bounding box, we trained a lighter 2D network to run a first estimate for the brain mask at inference time from the complete volume. This auxiliary network follows exactly the same architecture of MU-Net, omitting any framing or dense connections, and limiting the number of channels to 4, 8, 16 and 32, from the shallowest to the deepest layer. This results in a network with a total amount of 122454 trained parameters, used to automatically draw a bounding box and run the much heavier MU-Net on the volume thus identified without any manual intervention.

### 4.4 Multi-STEPS multi-atlas segmentation

Before registration, each volume underwent non-parametric N3 bias field correction[27] implemented within the ANTS toolset[39]. Taking each volume as reference, all other volumes were then registered with an affine transformation using FSL FLIRT[40] using a correlation ratio cost function, and then nonlinearly registered via FSL FNIRT[41,42] with the aid of the manually drawn brain mask. Label fusion is achieved with the Multi-STEPS algorithm distributed in the NiftySeg package[28,29].

Multi-STEPS depends on the number of templates employed and the standard deviation of its Gaussian kernel. We performed a grid search to select the optimal parameters, randomly selecting 10 volumes and labeling them using Multi-STEPS. We sampled the standard deviation of the Gaussian kernels between 0.5 and 6 with a stride of 0.5, and the number of templates ranged between 1 and 20 randomly selected volumes. This same process has been performed both using diffeomorphic registration and using affine registration only (supplementary figures 2 and 3), selecting 16 templates and kernel standard deviation of 1.5 for the diffeomorphic case, and 18 templates with kernel standard deviation of 2.5 for the affinely registered volumes. Exploring both grids required in total 287 hours.

Each volume was then segmented using these parameters, randomly selecting an appropriate number of mice as templates for the Multi-STEPS algorithm as emerged from the parameter grid search outlined above. We repeated this procedure randomly selecting the same number of templates from mice of the same age only. The mice randomly selected as reference atlases were selected from the training set associated to each volume according to the same 5-fold cross validation scheme used to train the CNNs as outlined in section 4.2.

### 4.5 Post-processing

The only post-processing steps applied on the segmentation maps were the filling of holes in the resulting 3D volume, the selection of the largest connected component as the brain mask for the skull-stripping task, and assigning all voxels predicted as non-brain to the background class.

# 5 Declarations

## 5.1 Data availability statement

Code and trained model are freely available at https://github.com/Hierakonpolis/MU-Net. The training and validation dataset is property of Charles River Discovery Services, and the test dataset is property of CHDI 'Cure Huntington's Disease Initiative' foundation. All the Dice scores between MU-Net and manual segmentations are available as supplementary files to this manuscript.

## 5.2 Contribution statement

J.T., O.G., and F.G. conceived the study, R.D.F. A.Si, and J.T. designed the study , R.D.F. designed the MU-Net network and wrote the software, R.D.F. and A.Si analyzed the data, J.M.V. contributed to the methods, A.Sh. assembled the datasets, R.D.F. and J.T. wrote the paper with substantial input from all the other authors. All authors commented on the manuscript.

## 5.3 Acknowledgments

R.D.F.'s work has received funding from the European Union's Horizon 2020 Framework Programme under the Marie Skłodowska Curie grant agreement No #691110 (MICROBRADAM) and J.M.V.' work was founded from Marie Skłodowska Curie grant agreement No #740264 (GENOMMED). The content is solely the responsibility of the authors and does not necessarily represent the official views of the European commission.

We also extend our thanks to the Academy of Finland, grants (#275453 to A.S. and #298007 to O.G. #316258 to J.T.) and to the CHDI 'Cure Huntington's Disease Initiative' foundation, for kindly providing us with the test data employed in this work. We acknowledge a grant S21770 from European Social Fund to J.T.

# References

1. Cunha, L. *et al.* Preclinical imaging: an essential ally in modern biosciences. *Mol. diagnosis & therapy* **18**, 153–173 (2014).
2. Matthews, P. M. *et al.* Technologies: preclinical imaging for drug development. *Drug Discov. Today: Technol.* **10**, e343–e350 (2013).
3. Febo, M. & Foster, T. C. Preclinical magnetic resonance imaging and spectroscopy studies of memory, aging, and cognitive decline. *Front. aging neuroscience* **8**, 158 (2016).
4. Dice, L. R. Measures of the amount of ecologic association between species. *Ecology* **26**, 297–302 (1945).
5. Ali, A. A., Dale, A. M., Badea, A. & Johnson, G. A. Automated segmentation of neuroanatomical structures in multispectral mr microscopy of the mouse brain. *Neuroimage* **27**, 425–435 (2005).
6. Yushkevich, P. A. *et al.* User-guided 3d active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *Neuroimage* **31**, 1116–1128 (2006).
7. Entis, J. J., Doerga, P., Barrett, L. F. & Dickerson, B. C. A reliable protocol for the manual segmentation of the human amygdala and its subregions using ultra-high resolution mri. *Neuroimage* **60**, 1226–1235 (2012).
8. De Feo, R. & Giove, F. Towards an efficient segmentation of small rodents brain: a short critical review. *J. neuroscience methods* (2019).
9. Pagani, M., Damiano, M., Galbusera, A., Tsaftaris, S. A. & Gozzi, A. Semi-automated registration-based anatomical labelling, voxel based morphometry and cortical thickness mapping of the mouse brain. *J. neuroscience methods* **267**, 62–73 (2016).
10. Schwarz, A. J. *et al.* A stereotaxic mri template set for the rat brain with tissue class distribution maps and co-registered anatomical atlas: application to pharmacological mri. *Neuroimage* **32**, 538–550 (2006).
11. Bai, J., Trinh, T. L. H., Chuang, K.-H. & Qiu, A. Atlas-based automatic mouse brain image segmentation revisited: model complexity vs. image registration. *Magn. resonance imaging* **30**, 789–798 (2012).
12. Ma, D. *et al.* Multi atlas segmentation applied to in vivo mouse brain mri (MICCAI 2012 Workshop on Multi-Atlas Labeling, 2012).
13. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *nature* **521**, 436 (2015).
14. Akkus, Z., Galimzianova, A., Hoogi, A., Rubin, D. L. & Erickson, B. J. Deep learning for brain mri segmentation: state of the art and future directions. *J. digital imaging* **30**, 449–459 (2017).

**15.** Roy, S. *et al.* A deep learning framework for brain extraction in humans and animals with traumatic brain injury. In *Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on*, 687–691 (IEEE, 2018).

**16.** Szegedy, C. *et al.* Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1–9 (2015).

**17.** Chou, N., Wu, J., Bingren, J. B., Qiu, A. & Chuang, K.-H. Robust automatic rodent brain extraction using 3-d pulse-coupled neural networks (pcnn). *IEEE Transactions on Image Process.* **20**, 2554–2564 (2011).

**18.** Oguz, I., Zhang, H., Rumple, A. & Sonka, M. Rats: rapid automatic tissue segmentation in rodent brain mri. *J. neuroscience methods* **221**, 175–182 (2014).

**19.** Wachinger, C., Reuter, M. & Klein, T. Deepnat: Deep convolutional neural network for segmenting neuroanatomy. *NeuroImage* **170**, 434–445 (2018).

**20.** Roy, A. G., Conjeti, S., Navab, N. & Wachinger, C. Quicknat: Segmenting mri neuroanatomy in 20 seconds. *arXiv preprint arXiv:1801.04161* (2018).

**21.** 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).

**22.** Xie, L. *et al.* 4d mri of polycystic kidneys from rapamycin-treated glis3-deficient mice. *NMR Biomed.* **28**, 546–554 (2015).

**23.** Valverde, J. M. *et al.* Automatic rodent brain mri lesion segmentation with fully convolutional networks. *arXiv preprint arXiv:1908.08746* (2019).

**24.** Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

**25.** Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S. & Cardoso, M. J. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, 240–248 (Springer, 2017).

**26.** Han, Y. & Ye, J. C. Framing u-net via deep convolutional framelets: Application to sparse-view ct. *IEEE transactions on medical imaging* **37**, 1418–1429 (2018).

**27.** Sled, J. G., Zijdenbos, A. P. & Evans, A. C. A nonparametric method for automatic correction of intensity nonuniformity in mri data. *IEEE transactions on medical imaging* **17**, 87–97 (1998).

**28.** Cardoso, M. J., Modat, M., Ourselin, S., Keihaninejad, S. & Cash, D. Multi-steps: multi-label similarity and truth estimation for propagated segmentations. In *Mathematical Methods in Biomedical Image Analysis (MMBIA), 2012 IEEE Workshop on*, 153–158 (IEEE, 2012).

**29.** Cardoso, M. J. *et al.* Steps: Similarity and truth estimation for propagated segmentations and its application to hippocampal segmentation and brain parcelation. *Med. image analysis* **17**, 671–684 (2013).

**30.** Bae, M. H., Pan, R., Wu, T. & Badea, A. Automated segmentation of mouse brain images using extended mrf. *Neuroimage* **46**, 717–725 (2009).

**31.** Milletari, F., Navab, N. & Ahmadi, S.-A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, 565–571 (IEEE, 2016).

**32.** Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T. & Ronneberger, O. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, 424–432 (Springer, 2016).

**33.** Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708 (2017).

**34.** Isensee, F. *et al.* nnu-net: Self-adapting framework for u-net-based medical image segmentation. *arXiv preprint arXiv:1809.10486* (2018).

**35.** Dutta, S. & Sengupta, P. Men and mice: relating their ages. *Life sciences* **152**, 244–248 (2016).

**36.** Ioffe, S. & Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167* (2015).

**37.** Noh, H., Hong, S. & Han, B. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, 1520–1528 (2015).

**38.** Karimi, D. & Salcudean, S. E. Reducing the hausdorff distance in medical image segmentation with convolutional neural networks. *arXiv preprint arXiv:1904.10030* (2019).

**39.** Avants, B. B., Tustison, N. & Song, G. Advanced normalization tools (ants). *Insight j* **2**, 1–35 (2009).

**40.** Jenkinson, M. & Smith, S. A global optimisation method for robust affine registration of brain images. *Med. image analysis* **5**, 143–156 (2001).

**41.** Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W. & Smith, S. M. Fsl. *Neuroimage* **62**, 782–790 (2012).

**42.** Andersson, J. L., Jenkinson, M., Smith, S. *et al.* Non-linear registration aka spatial normalisation fmrib technial report tr07ja2. *FMRIB Analysis Group Univ. Oxf.* (2007).