Deep learning for robust and flexible tracking in behavioral studies for *C. elegans*

Kathleen Bates[1,2,+], Kim Le[3,+], and Hang Lu*[,1,2,3]

[1] Interdisciplinary Program in Bioengineering, Georgia Institute of Technology, Atlanta, Georgia, USA

[2] School of Chemical & Biomolecular Engineering, Georgia Institute of Technology, Atlanta, Georgia, USA

[3] Wallace H. Coulter Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, Georgia, USA

*Author for correspondence: hang.lu@gatech.edu

[+] These authors contributed equally

ABSTRACT

Robust and accurate behavioral tracking is essential for ethological studies. Common methods for tracking and extracting behavior rely on user adjusted heuristics that can significantly vary across different individuals, environments, and experimental conditions. As a result, they are difficult to implement in large-scale behavioral studies with complex, heterogenous environmental conditions. Recently developed deep-learning methods for object recognition such as Faster R-CNN have advantages in their speed, accuracy, and robustness. Here, we show that Faster R-CNN can be employed for identification and detection of *Caenorhabditis elegans* in a variety of life stages in complex environments. We applied the algorithm to track animal speeds during development, fecundity rates and spatial distribution in reproductive adults, and behavioral decline in aging populations. By doing so, we demonstrate the flexibility, speed, and scalability of Faster R-CNN across a variety of experimental conditions, illustrating its generalized use for future large-scale behavioral studies.

INTRODUCTION

Ethology has been crucial in the fields of neuroscience, genetics, and aging (Anderson and Perona 2014; Krakauer et al. 2017; Brown and De Bivort 2018; Niepoth and Bendesky 2020) . This rings true even in the simplified *C. elegans* model, which has been used to probe a variety of ethological questions (Hedgecock and Russell 1975; Chalfie et al. 1985; Gray et al. 2005; Stephens et al. 2008; Brown et al. 2013; Brown et al. 2016) . In these experiments, it is extremely valuable to robustly and accurately track and measure the behavior of *C. elegans* on a large scale. The ability to collect large-scale behavioral data has significantly improved throughout the years. While many behavioral assays consist of manually recording small populations on agar plates under a stereomicroscope, recent automated methods have drastically increased the variety of biological questions researchers can explore. Ranging from multi-camera systems, to time-shared imaging systems, to low-cost imaging systems, advances in hardware have allowed users to more easily obtain large amounts of raw behavior video (Stroustrup et al. 2013; Churgin et al. 2017; Maia Chagas et al. 2017; Stern et al. 2017; Le et al. 2020) . In addition to improvements in data acquisition, there have been advances in culture methods, which enable the exploration of more complex environmental conditions. These range from individual "arenas", which allow the tracking of populations with individual level resolution, to microfluidic devices, which allow for precise spatiotemporal environmental control (Chung et al. 2011; Zhang et al. 2016; Churgin et al. 2017) . While these technological advancements have enabled the ability to explore complex behavior relevant to neuroscience and aging, this increase in behavioral recordings and data shifts the bottleneck to the analysis of large-scale image datasets. This is especially crucial for images taken in heterogeneous environments, such as those in more complex, naturalistic conditions.

One of the major challenges in analyzing behavioral data is the detection and identification of the object of interest, especially under a variety of imaging and environmental conditions. While there are many existing image processing tools that are currently used to detect, identify, and subsequently analyze the behavior of worms (Swierczek et al. 2011; Javer et al. 2018), there are unmet needs. These tools often use traditional image processing methods, such as background subtraction, thresholding based on the color or intensity of the object, or the use of morphological features (such as size), to detect and identify the object of interest. For example, in the popular worm tracker Tierpsy Tracker (Javer et al. 2018), users manually optimize parameters based on experimental conditions and are subsequently able to extract behavioral data from their dataset. With these segmentation and tracking tools, users can extract a variety of informative behavioral phenotypes, such as size, speed of movement, and the posture of individuals. However, with the advent of more complex experimental setups that introduce more heterogeneous experimental or environmental conditions, it is not straightforward to adapt these traditional methods to robustly and accurately detect objects of interest. For example, in conditions with low or uneven imaging contrast, basic thresholding based on intensity values may not be accurate. If animals or the objects of interest move only subtly, background subtraction cannot be used to easily differentiate between the object of interest and the background of the

image. Additionally, if there is a wide range in morphological properties, such as the dramatic size change of animals during development, it is difficult to rely on traditional morphological features such as size as a method of identifying objects of interest. Thus, coupled with the increased scale of behavioral datasets, there is a need for a robust, flexible, and facile method to detect and identify worms that would be able to work across a variety of different experimental conditions, with minimal user input.

To address this problem, we turn towards deep learning, which has emerged as a powerful data-driven tool for object detection. While there are many deep learning object detection methods, the Faster R-CNN architecture is a widely-used method that uses region proposal networks (RPN) coupled with convolutional neural networks (CNN) to extract the location (in the form of bounding boxes) and estimated likelihood for each detected object (Ren et al. 2015) . It is one of the top performing object detection methods, as measured by the mean average precision (mAP) of detections on the standardized COCO dataset (Huang et al. 2017). Compared to other CNN methods with equivalent or higher mAP, the Faster R-CNN architecture is less computationally costly and thus advantageous for large volumes of data. Further, the Faster R-CNN architecture has been tested in a wide range of applications, ranging from vehicle and pedestrian detection to malarial detection via cell classification (Ren et al. 2015; Hung et al. 2018) .

Here, we implement Faster R-CNN to identify and detect worms across a variety of different conditions without extensive user input. We find that after the initial training, the deep learning model quickly and accurately detects objects of interest. We demonstrate its flexibility across a variety of different recording platforms and imaging modalities. We also demonstrate its ability to detect worms across a variety of different ages (from L2 to death), showing its flexibility across different body sizes and movement levels, and illustrate how it can be used to extract useful behavioral metrics to give insights into biological questions, such as egg laying, development, and behavioral decline in aging. Finally, we provide a web-based pipeline (https://github.com/lu-lab/frcnn-all-in-one) for testing our trained models with novel data and to enable other researchers to annotate and train their own object detection models with novel data and classes.

RESULTS

To illustrate the difficulty in identifying and tracking objects in complex conditions, we examined three common experimental set-ups. The first tracks an individual worm from the L2 larval stage to Day 1 of adulthood on an agar plate seeded with food (**Figure 1A**). During this period, the worm was free to roam throughout the field of view, and animals were imaged through early adulthood. This type of low-magnification imaging setup is common for long-term and high-throughput imaging (Swierczek et al. 2011; Mathew et al. 2012; Zheng et al. 2012; Buckingham et al. 2014; Perni et al. 2018), as well as lifespan imaging (Stroustrup et al. 2013) . For small L2 animals, a major challenge in tracking is the small size of the individual (starting at around 360 µm in length) and differentiating it from the background despite the low contrast and low

magnification of the image. Tuning heuristics-based image processing tools to optimize for the small size and low contrast of young animals leads to further challenges as the worm grows (**Figure S1 A-E**). The contrast from background improves as the worm develops; however, other subtle changes in background (such as eggs or tracks formed on lawns) may be identified as animals when using heuristics tuned for young animals (**Figure S1 C-E**). These heuristics are also highly dependent on environmental and imaging conditions. In the case that changes in illumination or environment are an integral part of the experiment, this leads to an inability to process data without further tuning (**Figure S1 E**). Together, these challenges make processing developmental behavior data a labor-intensive task.

The second experimental system measures worm fecundity over time. An adult worm was allowed to roam across a seeded agar plate and freely lay eggs (**Figure 1C)**. Due to the small size of eggs (~50 µm) (Riddle DL, Blumenthal T, Meyer BJ 1997), low magnification and contrast, and the tendency for eggs to be laid in clusters, it can be both time and labor intensive to manually count the number of eggs over time and mark their spatial location. Further, some studies may involve egg-laying behavior in different environments (e.g. on or off bacterial lawns) where imaging conditions and contrast pose significant problems in identifying these objects and distinguishing them from other objects in the field of view. In contrast to animals, the immobility of eggs also prevents the use of background subtraction as a useful tool. These practical constraints make it difficult to track fecundity and other egg-laying phenotypes at a large scale.

The third example is tracking the behavior and movement of individuals during the aging process. Worms were cultured within a microfluidic chamber array (**Figure 1D, top)** from the L4 stage to their death. Individuals were longitudinally monitored and their behavior was recorded intermittently throughout their lifespan under a variety of different food concentrations (**Figure 1D, bottom rows)**. While the size of the worm and the contrast are better than those for young animals during development, there are two inherent challenges. First, as before, the environment is heterogeneous - within the chamber there are often moving objects aside from the worm (such as debris or eggs laid by the individual), making it difficult to accurately identify and detect the location of the worm even in instances with high levels of movement. Second, there are low levels of movement as the worm ages and eventually dies, making it difficult to identify the worm through traditional image processing techniques that rely on movement. These challenges are cumbersome to address when using traditional image processing tools that require tuned user parameters. For instance, when the parameters are chosen for an individual video **(Figure S2 A-B, top row)**, it fails to accurately identify and detect the location of worms of the same age and under the same environmental and imaging conditions (**Figure S2 A-B, bottom rows**). When the worm is in its reproductive period, the presence of lain eggs that cluster together in the chamber can cause misidentification and inaccurate segmentation of the worm (**Figure S2 A**). When the worm is aged and only performs small, subtle movements the algorithm often truncates or misidentifies the location of the worm entirely (**Figure S2 B).**

For both the on-plate and on-chip conditions, traditional detection and tracking methods are unable to robustly identify the location of the worm. While classic segmentation methods based on heuristics can provide posture information that CNN object detection methods cannot, the need to specifically tune parameters for a wide range of videos makes it unrealistic to deploy these methods at large scales in each of the demonstrated experiments. Thus, there is a need for a quick and generalizable method of identifying and tracking objects of interest in challenging imaging conditions such as these.

To address these challenges, we implemented Faster R-CNN, a deep learning network with high precision in object detection, including small objects. From an existing Faster R-CNN model pre-trained on the COCO image dataset, we tuned the model using our respective behavioral data sets (Lin et al. 2014; Abadi et al. 2015) . For each of the three different experimental conditions, we trained a separate Faster R-CNN model. For the condition with developing worms cultured on an agar plate we created the worms-on-plate (WoP) model. For the WoP model, we annotated worms with bounding boxes in 1,122 randomly chosen and representative images from a much larger dataset and trained the model with 1,008 of these, holding out the remainder as a test set to evaluate the model. To measure worm fecundity, we created the egg-finder model. We annotated eggs and worms in 127 images and used 114 of these to train the model, with the remainder used to for evaluation of model performance. Lastly, for the more specialized condition of worms cultured within the microfluidic chambers, we created the worms-in-chamber (WiCh) model. For the WiCh model, we annotated eggs and worms in 5,176 images, with 4,658 of these used to train and the remainder to evaluate the model.

When we qualitatively examined the bounding box output of these trained models, we found that many of the failure cases using traditional methods were resolved (**Figure 1, Figures S1-2**). In the WoP dataset, worms were identified accurately in both very low-contrast images when worms were very small as well as in much higher-contrast images later in life (**Figure 1A**). We also found that this model could detect worms in very different imaging conditions with which it was not trained (**Figure 1B, File S3**). This is highly significant as it indicates the model is much more generalizable compared to heuristic techniques and thus more widely usable in real applications. In the egg-finder dataset, we found that eggs were identified well despite their small size and tendency to cluster together (**Figure 1C**). Notably, when we applied the egg-finder model to publicly available videos from the Open Worm Movement Database, we were able to detect both worms and eggs (**File S1**). Finally, in the more specific WiCh dataset, worms were identified accurately, even in the most food-dense, low-contrast settings and at later ages where worm movement is reduced (**Figure 1D**). In addition, clusters of eggs could be accurately identified, making it possible to differentiate active worm movement from passive movement of the egg cluster (**Figure 1D**).

Next, we quantitatively evaluated our models to ensure their accuracy (**Table 1**). A common metric used in object detection is the average precision (AP), which uses the overlap between actual bounding boxes and those predicted by the model at varying confidence thresholds to

evaluate the model performance. An AP equal to unity would indicate perfect predictions. For the WoP model, we were able to obtain an average precision of 0.969. AP for the youngest animals (66 test images) was 1.0 with our test set, compared to the oldest animals (22 test images), for which false positives reduced the AP slightly to 0.876 (**Figure S3**), making the model robust across age without further tuning. For our egg-finder model, the worm AP was 0.932 and the egg AP was 0.74 (**Figure S3**). While the average precision for eggs is not as high as for the other objects we detected, we found that it was able to identify 79% of the eggs in our test set and that the sensitivity of the model for our test set was 0.84 (**Figure S3**). Conservative identification of eggs by the model likely stems from the size of eggs making the overlap threshold (intersection over union) of detections and ground truths particularly sensitive, as well as the intensity of the eggs being similar to other image features such as the tracks created by animal movement, and the occlusion of eggs by each other. In practice, we found this model worked well to identify egg-laying phenotypes in further experiments (see below, **Figure 3**), and was able to detect eggs in data collected by other labs (**File S1**). It also may be possible to improve the average precision by using more training data. For the WiCh model, we obtained an AP of 0.998. This model is also robust across different conditions within the dataset, ranging from different ages, contrast levels, and objects of interest (**Figure S3**), with the models detecting not just worms but also the lain eggs within the field of view (AP of 0.932).

In addition to the models being flexible across conditions without the need for additional parameter tuning, the inference time for each image is short (~131 ms/frame on our equipment). This is significant because in practice, these generalizable strategies can reduce the time and effort it takes to quantify new data. For instance, in cases with large sets of data under varied conditions, traditional hand annotation or the implementation of user-tuned parameters would require excessive amounts of time and manual labor. In contrast, deep learning enables users to analyze large behavioral datasets in a more efficient manner.

We next put the algorithm to a real use-case - monitoring egg-laying phenotypes of *C. elegans.* Egg-laying rate is indicative of health (Trent et al. 1983), evolutionary fitness (Duveau and Félix 2012; Zhao et al. 2018; Anholt et al. 2020), and is also important in understanding the regulatory mechanisms of the reproductive circuit (Trent et al. 1983; Mendel et al. 1995; Ségalat et al. 1995; Chen et al. 2020). However, the small size (approximately 50 µm) and large number of eggs (about 300 per adult hermaphrodite) makes measuring fecundity a challenging task. Manually counting eggs is time-intensive, but it is often still the method of choice because existing automated methods for egg detection are very sensitive to imaging conditions, requiring high image uniformity and often high magnification (Geng et al. 2005; Jung et al. 2014; Chen et al. 2020). Likely due to the challenging nature of the task, several popular software packages for worm behavior quantification do not include methods to track egg-laying (Swierczek et al. 2011; Wählby et al. 2012; Javer et al. 2018). Other egg-counting methods rely on specialized cytometry to count eggs as they are flushed from liquid culture (Sofela et al. 2018; Atakan et al. 2020). Another factor that makes counting eggs difficult is the burst-like timing of egg-laying events (Waggoner et al. 1998; Schafer 2005; Collins et al. 2016). Many egg-laying events in close

temporal proximity causes eggs to cluster, making identification of individual eggs challenging. Further, the bacterial lawn that adults feed on becomes highly textured as adults crawl on it, which makes identifying eggs and animals significantly more challenging for image thresholding methods. To determine whether the egg-finder model would resolve these issues, we applied it to count and locate eggs (**Figure 2**). We collected images of individual day 1 adult worms and the eggs they laid at two time points, 2 and 5 hours after transferring animals onto plates for individual culture. While we qualitatively found that our model did not successfully identify every egg, the agreement between manual egg counts and those from our model was significant (**Figure 2A & B**). Even when illumination was uneven across the field of view, when eggs were laid on the lawn, and when eggs clustered together, the model produced accurate results (as exemplified in **Figure 2A**). This robustness in the face of highly variable conditions indicates that this method is a much faster alternative to manual counting of eggs. This method is especially well suited for large datasets and movies where manual annotation of every frame would be prohibitively difficult or when imaging conditions like textured bacterial lawns prevent thresholding techniques from performing well.

Next, we showed that we can apply the algorithm to accurately estimate two egg-laying rates for each animal in early adulthood from images of individual animals' brood at two time points (**Figure 2C**). These results match previously reported egg-laying rates of about 4-10 eggs/hr (Waggoner et al. 1998). The distribution of egg-laying rates for the manual and Faster R-CNN model egg counts was statistically indistinguishable at both timepoints, whereas comparing the Faster R-CNN model egg count distributions and manual egg count distributions between timepoints showed a significant difference. We noted that egg-laying rates decreased over time, which we suspect is due to transferring animals from a crowded growth plate to individual plates. This change in the experienced levels of oxygen, carbon dioxide, and food may promote a higher egg-laying rate until the individuals habituate to the new environment and begin to deplete food (Trent et al. 1983; Schafer 2005; Fenk and De Bono 2015). The ability to detect this difference using the egg-finder model demonstrates that Faster R-CNN can be used to identify biologically relevant phenotypes in a less time-intensive way than manually counting eggs.

In addition to identifying a rate phenotype, we used the egg-finder model to identify a spatial phenotype using the food-sensing mutant *tph-1*. These animals are known to spend a greater fraction of time in a roaming state and are also slower to pause upon encountering food (Flavell et al. 2013; Entchev et al. 2015; Stern et al. 2017). Based on this and anecdotal evidence (Dhaval Patel, personal communication), we expected that *tph-1* animals would lay more eggs off of bacterial lawns compared to wild type animals. To examine whether this was the case, we defined an egg location preference score such that a greater number of eggs laid on the bacterial lawn would result in a positive preference score, while a negative preference score would indicate that a greater number of eggs were laid off the bacterial lawn. We found that the distribution of preference scores for WT and *tph-1* animals was consistent with *tph-1* animals having a lower preference for laying eggs on the bacterial lawn, and that the two distributions

were significantly different both when counted manually and when counted using the Faster R-CNN model (**Figure 2D**). This example demonstrates that the model can distinguish eggs on or off the lawn regardless of lighting, contrast, and despite the width of adult animal's tracks being almost the same size as an egg. Further, we applied the egg-finder Faster R-CNN model to movies from the publicly available Open Worm Movement Database and found that we were able to successfully detect both eggs and worms in these movies without additional training (**File S1**). Overall, the egg-finder model performed well across both datasets in different imaging conditions and was able to uncover the same egg-related phenotypes as a human annotator in our own dataset again suggesting that Faster R-CNN models can replace manual labor particularly for large-scale datasets and movies with complex imaging conditions.

In addition to quantifying aspects of behavior through endpoint snapshots, we reasoned that Faster R-CNN could also be used to track animals over time. Estimating animal linear and angular velocity is a useful indicator of the animals' behavior state (e.g. dwelling/ roaming) as well as potentially an indicator of health when observed over sufficiently long timescales (Hahm et al. 2015; Zhang et al. 2016). Using the dynamic location of worm bounding boxes detected using our WoP model, we evaluated whether it was possible to obtain accurate movement measurements. We compared the centroids of detected bounding boxes obtained from our model to the centroids of hand-annotated postures at 5 series of time points over about 2 days of worm development (**Figure 3A-B**). Throughout the ~ 2 days of observation, the object detection model was able to accurately identify worms, with the smallest animals an average of 124 pixels (~0.12 mm^2) in area (**Figure 3A**, left column). This timescale includes development from late L2 stage through to adulthood. We found that the bounding box centroids detected were typically close (0.212 mm +/- 0.197 mean+/- standard deviation) to the centroids of the hand-annotated bounds of the worm, indicating that this method can provide accurate worm locations at discrete time points (**Figure 3A**, right column, **Figure S6 B-D**). To test whether we could also measure motion accurately with this method, we used time points a minute apart and calculated motion between the bounding box centroids and hand-annotated worm shape centroids at these time points. The difference between these two motion measurements was an average of 0.126mm +/- 0.083mm (mean +/- std) and motion trends were clearly replicated between the hand-annotated shape centroids and the bounding box centroids (**Figure 3B**, **Figure S6 C**). We further compared peak and mean velocity in our hand-annotated dataset to peak and mean velocity computed using the WoP Faster R-CNN model centroids. We found that both peak and mean velocities were very similar, with differences on the order of pixels in 5MP images (**Table S1**). Similarly, we found that we were able to very accurately recapitulate the linear and angular velocity and peak and mean velocities of a publicly available dataset using our WoP Faster R-CNN model (Stern et al. 2017) (**Figures S4 and S5, File S3**). Even without knowledge of the precise pixels that comprise the worm, the extents of the worm can be used as a rough measurement of movement and speed.

We next examined how this approach could be used to track behavior continuously over development. Behavior during development is individualistic and can affect long-term behavior

and neuropeptide signaling in *C. elegans* (Stern et al. 2017). The dual challenges of the small size and low contrast of young animals have previously imposed stringent hardware requirements to ensure high-quality images (e.g. at higher magnification), therefore limiting the scalability of long-term developmental experiments. In contrast, the WoP model can extract worm position despite low image quality and the extreme variation in the size and contrast of developing worms. We collected time-lapse images at 1-minute intervals of 10 animals over a 2-day period from late L2 stage to adulthood, and detected worm bounding boxes for each image (**Figure 3C, File S2**). The magnitude of motion increased over time, and likewise, we found that the size of the detected bounding box grew approximately 5-fold over time as the animals developed (**Figure 3C, Figure S6A**). We next examined whether our centroid data could be used to identify roaming and dwelling states, which are an indicator of satiety and which are influenced by neuromodulators. *C. elegans* spends the greater portion of its time moving at slower speeds while eating (dwelling) and a small portion of its time searching for other food sources (roaming) (Waggoner et al. 1998; Ben Arous et al. 2009; Flavell et al. 2013). We calculated the linear and angular velocity for each animal and found that while there was little separation in linear velocity between the two states, there was a striking split in angular velocity that was consistent with the roaming and dwelling state separation in other datasets at the same 1-minute sampling rate (**Figure 3D, Figure S7**). We found that by classifying the behavior state using an angular velocity threshold of 90 degrees we were able to accurately predict roaming and dwelling in a ground truth data set with the same sampling rate. We then used the same 90 degree threshold to identify roaming and dwelling states throughout animal development in our own data (**File S4**). Thus, this technique can enable researchers to infer high-level information about the animal's behavioral state from the limited information provided by bounding box identification.

Next we tested how well the model can be used to track motion in a realistic biological discovery context: examining behavioral decline in the aging process. Behavior and movement are common methods to gauge the health and physiological age of an individual (Golden et al. 2008; Hahm et al. 2015; Zhang et al. 2016; Churgin et al. 2017; Le et al. 2020). For these experiments, it can be challenging to accurately measure how the movement of individuals changes throughout their entire lifespan due to complex environmental conditions and the large scale of the data. As a specific example, we examined wild-type individuals cultured in a microfluidic device, allowing us to identify and track individuals throughout the entirety of their adult lifespan. Due to the size of each chamber (~1.5 mm in diameter), as the worm grows, the extent of movement becomes limited making common metrics (such as tracking the distance traveled by the centroid of the segmented worm) unable to clearly provide insights into the decline of movement over time. As a result, to gauge movement we examined the normalized sum of the difference in pixels across frames for segmented individuals (**Figure 4A**). As the worm ages and its movement declines, the difference across frames decreases as well for the overall population (**Figure 4B**). However, although this method provides useful insight into the behavioral decline of the individual, it can be difficult to accurately obtain the properly segmented worm in large datasets. This is due to issues with background contrast, the presence of eggs in the chamber, and the low mobility of

older worms. Furthermore, the process typically requires large amounts of manual parameter tuning (to account for the different sizes of worms as they age, changing levels of movement with age, and variation across the individuals) and substantial computational time to segment and extract features of interest. The need for manual tuning and intensive computational resources makes it difficult to scale this method for large sets of behavioral information.

Faster R-CNN can serve as a quick and accurate alternative to gauge behavioral decline with aging across a population. By tracking the bounding box locations of the worm detected by the WiCh model and measuring the IoU (intersection over union) of the detection bounding boxes across the video, we can get a rough metric of movement. Young, highly active worms have little to no bounding box intersection across frames, while older, slower moving worms have increasing levels of intersection across frames (**Figure 4C).** To examine how movement changed with time we examined a movement score (1 – IoU), and observed individual decline in movement, as well as a population-level behavioral declines with age (**Figure 4D**). Not only were we able to view similar patterns of movement decline to that observed with segmented frames, we were able to do so on a larger scale with minimal processing time (~131 ms/frame).

Further, we wanted to examine whether this movement score could discern how perturbations influence decline in movement with age. Dietary restriction (DR) is an evolutionary conserved perturbation that has been shown to modulate aging (Mair et al. 2003; Colman et al. 2009; Greer and Brunet 2009). We examined the movement score of worms cultured under constant DR (OD_{600}2.5) starting at Day 2 of adulthood and were able to demonstrate that the worms under lower food levels had a statistically significant difference in behavioral decline compared to worms cultured at higher food levels (OD_{600}10), trends observed in a prior study (**Figure 4E**) (Le et al. 2020). In addition, to verify the performance of the model on the dataset, we also validated that the detected bounding boxes from the WiCh model were comparable to the bounding box of the hand annotated, segmented worms (**Figure S8**). This exercise demonstrates that motion quantitatively estimated by Faster R-CNN can be used as a quick metric to track and examine behavioral decline within an aging population.

Discussion

Processing big sets of behavior data remains a major challenge currently facing large-scale ethological studies. As a model organism, *C. elegans* is well-poised as a subject for large-scale investigation, but typical computer vision analysis pipelines may still be insufficient in complex imaging conditions, where the animal itself may change size, contrast with background, or where inhomogeneities in the environment lead to failure of heuristic models. Here we have shown that applying Faster R-CNN object detection models to identify, count, and track behaviors in challenging environments is a fast and flexible alternative to more traditional analysis methods. We first demonstrated this method's applicability in identifying eggs and determining egg-laying rates and spatial distribution of eggs. We were also able to show that this method was effective

in tracking movement of animals ranging from L2 stage through the end of life, providing high-level behavior state information as well as information relevant to animal health. This effectively includes most applications in behavioral and aging studies using *C. elegans*.

Compared to other conventional methods, we found this deep learning method to be significantly more generalizable across a variety of experimental conditions. It requires no specialized hardware or dedicated imaging set-up and, once trained, we found the Faster R-CNN models worked very well, even under conditions for which they had not been trained (**File S1, S3**). In comparison, other methods (including other machine-learning image classification tools like Ilastik (Berg et al. 2019)) needed additional manual tuning for even slightly different lighting conditions, differently sized animals, or changing contrast levels (**Figure S9**). While for small datasets, the time investment for annotating images and training the model may be high, for large datasets with imaging variability the high detection accuracy and elimination of video-by-video or frame-by-frame hand-tuning represents a significant gain. For datasets that require extensive human correction, the accuracy of manual annotation degrades unpredictably over time as attention wanes; in comparison, machine learning models are predictably biased based on the content of the training set. These advantages, combined with the high speed of processing with a GPU (on average 7.6 fps), makes the deep learning approach a very good alternative to more conventional methods.

For cases where the existing trained models (WoP, egg-finder, or WiCh) are unable to robustly detect objects of interest, we have also developed an easy-to-use web-based tool that enables users to annotate, retrain, and evaluate their dataset (accessible from https://github.com/lu-lab/frcnn-all-in-one). There is a low barrier-to-entry for use, with no dependence on paid, commercial software (such as MATLAB) and no requirement for downloading external programs. While we found that the Faster R-CNN architecture suited our accuracy and speed requirements, this pipeline can also be used to train models using other object detections architectures that achieve faster inferencing speeds with somewhat lower accuracy. Using the Faster R-CNN architecture with our pipeline, we found that training on as few as 10-20 annotated images for several hours provided very high-quality tracking results (**File S6**). This time investment will likely pay off for large-scale datasets and datasets where imaging conditions create challenges for classic thresholding methods.

We also showed that even without segmenting images as many traditional methods do, we were still able to extract information about animal and egg location, size, linear and angular velocity, and animal behavior state that could be used to identify biologically meaningful phenotypes. These simple methods have a lower computational cost than that of segmentation, and in cases where precise knowledge of posture is not necessary, this method provides a fast and environmentally robust estimation of relevant metrics. Additionally, there are related deep learning approaches that provide semantic segmentation, such as a Mask R-CNN model, which would enable extraction of posture information. The application of this model is a natural next step, although the annotation necessary to train such a model is significantly more intensive.

In addition to the ability to accurately identify and track animals quickly, the success of this deep learning method in extreme imaging conditions suggests that this method can be used to push the current limitations in quantifying animal behavior in ethologically relevant environments. Researchers face a trade-off between performing assays in naturalistic environments and the ability to extract more information from more uniform and controlled environments. Deep learning methods such as this one may provide us with a greater ability to extract the necessary information from richer environments with greater ethological relevance.

Model	**Category**	**Average Precision (AP) @ threshold 0.5**	**Average Precision (AP) @ threshold 0.1**	**Average Precision (AP) @ confidence threshold 0.01, IoU threshold 0.3**
Development	Worms (all ages)	0.969	0.969	0.969
Egg counting	Eggs	0.398	0.430	0.740
Aging	Worms (all ages)	0.998	1.00	1.00

Table 1. DETECTION RESULTS ACROSS DIFFERENT CONFIDENCE THRESHOLDS ON THE DEVELOPMENT, EGG LAYING, AND AGING DETECTION MODELS USING FASTER R-CNN.

Author Contributions

KB, KL, and HL conceptualized the work, KB and KL collected and analyzed data and developed models and generated figures, KB, KL, and HL wrote the paper, and HL supervised the work.

Acknowledgments

The authors are grateful to Shay Stern, Cori Bargmann, Yuehui Zhao, and Patrick McGrath for generously providing video data, to Dhaval Patel for advice regarding *tph-1* animals and to QueeLim Ch'ng for strains.

Funding

This study was funded by US NSF (1764406) and US NIH (R01AG056436, R01GM088333) grants to HL, US NIH F31 fellowship to KB (1F31GM123662) and US NSF GRF to KL (DGE-1650044).

Conflict of Interest

The authors declare that no conflicts of interest exist.

Supplemental Material

Supplemental figures and media can be found on figshare (https://doi.org/10.6084/m9.figshare.13681675.v3). Pre-trained models can also be found on figshare (https://doi.org/10.6084/m9.figshare.13678705.v1).

Materials and Methods

C. elegans maintenance

C. elegans strains were maintained under standard conditions at 20°C unless otherwise noted (Stiernagle 2006) . Strains used in this work include N2 and QL101[*tph-1(n4622) II*].

Plate assays

To prevent animals from leaving the microscope field of view (FoV), we prepared special plates. Palmitic acid has been demonstrated as an effective barrier for worms in behavior experiments (Fletcher and Kim 2017) . It is typically applied as a solution in ethanol to a standard plate and the ethanol is allowed to evaporate off. However, it is hard to deposit in a controlled way due to the palmitic acid solution wetting the agar. We used an ethanol-sterilized piece of PDMS as a negative to prevent a 10 mg/mL palmitic acid in ethanol solution from wetting the center of a 5cm NGM plate, allowing the ethanol to evaporate for at least 30 minutes before removing the PDMS with tweezers. These plates were subsequently seeded with 10 ul (developmental experiments) or 5

ul (egg-laying experiments) of OP50. Plates used in developmental experiments were incubated at room temperature for about 24 hours to allow a thin lawn to form and were stored at 4°C until an hour before use. For egg-laying experiments, plates were seeded approximately 2 hrs before transferring animals onto plates.

For developmental assays, adult animals were bleached to obtain eggs. Eggs were allowed to hatch and larvae allowed to reach L1 arrest by agitating eggs overnight in M9 buffer. L1s were then pipetted onto an unseeded NGM plate and single animals were pipetted onto the prepared seeded palmitic acid plates. These plates were then parafilmed and incubated at 20°C until animals reached late L2 stage (20 hours after plating), when each plate was placed on a Raspberry Pi-based imaging system. The Raspberry Pi imaging system used a Raspberry Pi v3 Model B (Raspberry Pi Foundation) with official Raspberry Pi touchscreen (Raspberry Pi Foundation) and a Raspberry Pi Camera Module v2 (Raspberry Pi Foundation) with no additional lens to capture images at minute time intervals. Darkfield illumination was provided by an LED Matrix (Adafruit), with a center circle of LEDs dark and the surrounding matrix illuminating animals with red light. Developmental experiments lasted 44 hours, at which point worms have typically reached sexual maturity and plates were removed from imaging systems.

For egg-laying assays, gravid day 1 adults animals were picked onto prepared palmitic acid plates and plates were imaged at 2 and 5 hours at 1.6x on a stereomicroscope (Leica M165 FC) using a 1.3 MP CMOS camera (Thorlabs DCC1645C) with a 0.5x coupler.

C. elegans on-chip culture

Synchronized L4-stage wildtype animals were loaded into a worm chamber array microfluidic device. Microfluidic devices were fabricated from polydimethylsiloxane using standard soft lithography techniques and sterilized by autoclaving.

Worms were cultured at 20°C in *E. coli* (HB101) spiked with Pluronic F-127 (0.005%), carbenicillin (50µg/ml), and kanamycin (50µg/ml) to prevent the risk of bacterial aggregation and contamination during long-term culture. The bacteria was at a concentration of OD_{600} 10 to prevent any harmful effects of dietary restriction on the developmental process. The bacteria also contained 5uM of C22, which interrupts eggshell formation and results in non-viable progeny. At Day 2 of adulthood, worms were then shifted to 25°C and to the desired food level. Individuals shown in Figure 4 were maintained at OD_{600} 10 unless otherwise stated. Individuals in Figure 1 were cultured at OD_{600}10 and OD_{600}2.5. We used an average flow rate of approximately 15µL/min across all conditions. See prior work for more details. (Le et al. 2020)

Training Faster R-CNN network

For the egg detection and aging model, we used TensorFlow GPU (v 1.14) to train the model. For the developmental tracking model, we used TensorFlow CPU (v 1.14). For all models except the model trained using the web-based pipeline we used the pre-trained 'Faster_rcnn_inception_v2' model from the Tensorflow 1 model zoo (https://github.com/tensorflow/models/blob/master/research/object_detection/g3doc/tf1_detection_zoo.md) and fine-tuned it with our data sets of interest. We trained the models and processed the images on a system with an Intel(R) Xeon(R) CPU E5-1620 v4 processor and a NVIDIA Quadro M4000 GPU.

For the egg detection model, images were taken of a mixed population of wild-type worms on a seeded plate at 1.6x magnification on a stereomicroscope (Leica M165 FC) using a 1.3 MP CMOS camera (Thorlabs DCC1645C) with a 0.5x coupler. We manually annotated 127 images of worms and eggs using the labelImg Python package. Images were randomly split into training and testing sets using a rough 90/10 split (114 images for training, 13 test images).

For the developmental tracking model, images were taken as described for plate developmental assays above. 1,122 images were randomly selected from a large set of developmental imaging data (> 10,000 images taken on 8 different imaging setups) and annotated using the labelImg Python package. This annotated image set was divided into 1,008 images for training and 114 images for testing, roughly a 90/10 split. No images used in training or testing overlap with image data evaluated in Figure 4.

For the aging model, videos were taken at an acquisition rate of 14 fps using a 1.3 Megapixel monochrome CMOS camera (Thorlabs DCC1545M camera) coupled with a 10X close focus zoom lens (Edmund #54-363). Each video was 10 seconds in length. Illumination was provided by a set of concentric red LED rings (Super Bright LEDs 60 and 80mm LED Halo Headlight Accent Lights) to reduce the amount of blue light exposed to the worm. Videos were sampled evenly throughout the lifespan of individuals in food levels of $OD_{600}10$ and $OD_{600}2.5$. We manually annotated 5,176 frames of worms and, if present, eggs using the labelImg Python package. Images were randomly split into training and testing sets using a rough 90/10 split (4658 images for training, 518 test images).

For the model trained with our web-based pipeline, we annotated 14 frames total and used 12 frames to train a Tensorflow 2 Faster R-CNN model pre-trained with the COCO image dataset for 2 hours. We used the resulting model to perform the detections in Supplemental File 6.

Faster R-CNN model characterization

For each of our annotated datasets, we evaluated how well our model performed by calculating precision and recall as well as average precision. Precision is a measure of the false positive rate, as calculated by $recision = \frac{TP}{TP+FP}$, where TP is the number of true positives and FP is the number of false positives. Recall is a measure of the false negative rate, as calculated by $ecall =$

$\frac{TP}{TP+FN}$, where FN is the number of false negatives. Average precision is the integral of the precision recall curve for a set of images that have ground truth bounding box annotations as well as model predictions. To determine whether detections by the models were true positives, false positives, or false negatives, we used a measure of the overlap of detections and ground truth known as intersection over union (IoU), calculated as $IoU = \frac{|GT \cap P|}{|GT \cup P|}$, where GT is the bounding box of the ground truth and P is the bounding box of the prediction. An IoU ≥ 0.5 is counted as a true positive and an IoU < 0.5 as a false positive for our worm detections, with the IoU threshold lowered to 0.3 for egg detections in our egg-finder model. A false negative is counted when a ground truth annotation has no overlap with a detection by the model. Once all images with ground truth annotations are evaluated in this way, the maximum precision at each recall level is used to interpolate between points of the precision-recall plot.

The AP, recall, and precision for the egg-finder Faster R-CNN model was evaluated using a score threshold of 0.01. The AP, recall and precision for the WoP Faster R-CNN model and the aging model was evaluated using a score threshold of 0.5. Mask annotation and centroid computation for ground-truth movement comparisons of data in Figure 3 was collected using MATLAB.

Evaluation of egg-laying phenotypes

Images of individually cultured animals were collected at 2 hrs after transfer onto individual plates and again at 5 hrs after transfer. For each image, a human curator manually counted eggs and identified them as being on or off the bacterial lawn. The Faster R-CNN model was also used to detect eggs in each image at a confidence threshold above 0.01, and eggs were manually identified as being on or off the bacterial lawn. We used these detections to overlay bounding boxes on each image and manually classified each detection as on or off the bacterial lawn. As the arena where the worm was able to roam was larger than the microscope field of view, images were tiled to ensure all eggs laid by each individual were captured. In cases where images overlapped with one another, double-counted eggs were subtracted from the overall count. The egg-laying preference score was calculated as follows: $preference\ ratio = \frac{eggs_{on} - eggs_{off}}{eggs_{on} + eggs_{off}}$.

Evaluation of behavioral decline in aging

Images of worms were segmented through hand annotations using Ilastik. To calculate the pixel difference as the worms age we used $\frac{\sum |img1 - img2|}{(img1 + img2)/2}$ where img1 was the initial segmented frame of the video and img2 was the final segmented frame of the video. The IoU of the bounding boxes were calculated using the built-in MATLAB function bboxOverlapRatio. The 1-IoU metric was found by looking at the overlap between the bounding box found in the first frame and the bounding box found in the last frame of the video.

Data Availability

All raw and annotated datasets and trained models are available as supplemental material via figshare (https://doi.org/10.6084/m9.figshare.13678705.v1).

Figure Legends

Figure 1. Deep learning can be used to detect objects in a variety of complex environments

A. Tracking an individual worm throughout its developmental period (from L2 to Day 1 adulthood). *(left)* Images of the plate over time. Scale bar is 5mm. The red box is the Faster R-CNN detection of the tracked worm. *(right)* Zoomed in image of the worm detected by the WoP Faster R-CNN model. Scale bar is 0.5mm.
B. The WoP model applied in a different experimental set-up (adult worms in a microfluidic pillar array chamber). Worms detected by the WoP Faster R-CNN model are marked with red bounding boxes. Scale bar is 3mm.
C. An egg-laying adult on an agar plate, with detected eggs boxed in blue and detected worm boxed in red. All detections made with egg-finder Faster R-CNN model. Scale bar is 1mm. (*right)* Zoomed in image of a cluster of eggs detected by the model.
D. Tracking worms through their adult lifespan. *(top)* Microfluidic chamber array that cultures individual worms within each chamber (white arrows indicate chambers with single worms). White scale bar is 1.5mm. *(bottom)* Zoomed in images of individual worms under different food levels across the entirety of their adult lifespan. Worms detected by the WiCh Faster R-CNN model are boxed in red. Egg clusters detected by the WiCh Faster R-CNN model are boxed in blue. Scale bar is 0.5 mm.

Figure 2. Automated egg detection using Faster R-CNN

A. Representative image from egg-finder dataset. Blue boxes overlaid on the right inset images indicate egg detections using the egg-finder Faster R-CNN model at the confidence score threshold of 0.01. Scale bar is 1mm.
B. Agreement between manual egg count and egg-finder Faster R-CNN model egg count. Eggs laid by individual animals were counted manually and using the Faster R-CNN model with a confidence score threshold of 0.01. The agreement between these two counts was measured at two time points for n=29 individual animals. Distributions were compared using the Kolmogorov-Smirnov 2-sample test and found to be not significantly different (KS statistic is 0.155, p value is 0.491).
C. Egg-laying rate distribution for 29 animals at 2 time points. Horizontal bars within each distribution represent egg counts for individual animals. Distributions were compared using the Kolmogorov-Smirnov 2-sample test. The egg-laying rate counted manually at timepoint one was significantly different from the egg-laying rate counted manually at timepoint two (KS statistic is 0.448, p value is 0.0053). The egg-laying rate counted using the egg-finder Faster R-CNN model with a confidence threshold of 0.01 at timepoint one was significantly different from the egg-laying rate counted using the egg-finder Faster R-CNN model with the same threshold at timepoint two (KS statistic is 0.448, p value is 0.0053). For each timepoint, the Faster R-CNN egg-laying rate distribution was compared to the manually counted distribution, and for both timepoints, the KS test statistic was determined to be 0.2069 with a p value of 0.5141.

D. Egg-laying preference for N2 (n= 16) and *tph-1* (n=13) animals. Horizontal bars within each distribution represent egg counts for individual animals. Positive scores indicate a higher propensity to lay eggs on the lawn compared to off the lawn. Negative scores indicate a higher propensity to lay eggs off the lawn (see Methods for calculation formula). Distributions were compared using the Kolmogorov-Smirnov 2-sample test. N2 vs. *tph-1* distributions were different at a significant level for both manual counts (KS stat is 0.473, p-value of 0.0018) and egg-finder Faster R-CNN model counts (KS stat is 0.459, p-value is 0.0028).

Figure 3. Tracking behavior in development

A. A single N2 animal tracked throughout development at 5 example time points. At left, the full-size image is overlaid with a red box highlighting the worm location for each timepoint. At right, the boxed portion of the image is overlaid with the manual annotation at the example timepoint (blue) and five minutes later (yellow), as well as the centerpoint of the WoP Faster R-CNN model's detected bounding box at the example timepoint (purple) and five minutes later (orange).
B. Heatmap comparing distance travelled by worms calculated from manual annotations and WoP Faster R-CNN model detections. For each timepoint in A and the subsequent five minutes, the distance travelled by the worm is calculated from manually segmented animals and from the WoP Faster R-CNN detections. The manual distance is calculated from the Euclidean distance travelled at the centroid of the segmented animal for each pair of time points, while the Faster R-CNN distance travelled is calculated from the Euclidean distance between the centroid of WoP Faster R-CNN bounding boxes for each pair of time points.
C. Heatmap of centroid movement measured using the Faster R-CNN model for 10 individuals over the course of development from late L2 stage through adulthood.
D. Histogram of Faster R-CNN derived movement speeds for the example animal in A and B.

Figure 4. Tracking behavioral decline in aging

A. (top) Representative image of movement as the worm ages. The posture at the beginning of the video is shown in blue. The posture at the end of the video is shown in yellow. (bottom) Absolute difference image of the frames shown on the left. Pixel change values are the sum of the absolute difference image normalized by the average worm size in the video.
B. (top) Heatmap of individual behavior decline (via pixel changes) over time (n = 15). (bottom) Average pixel change values over time from L4 to Day 15 of adulthood. Error is plotted as SEM.
C. Representative image of the bounding boxes obtained using Faster R-CNN as the worm ages. The box found at the beginning of the video is shown in blue. The box found at the end of the video is shown in yellow. The IoU of the two boxes is shaded.

D. (top) Heatmap of individual movement (1 – IoU) from L4 to Day 15 of adulthood (n = 31). Individuals are cultured in OD_{600}10. (bottom) Average movement decline over time. Error is plotted as SEM.
E. Average movement decline over time for individuals cultured in high levels of food (OD_{600}10 in blue) and individuals cultured in low levels of food (OD_{600}2.5 in red). Error is plotted as SEM. Movement for OD_{600}10 and OD_{600}2.5 is significantly different via Kolmogorov-Smirnov 2-sample test (p = 0.0.03).

Supplemental Figure Legends

Figure S1. Limitations of traditional image processing techniques in developmental monitoring

A. Detection of animals throughout development using the trained WoP Faster R-CNN model or a common tool that uses traditional segmentation (Tierpsy Tracker). Tierpsy Tracker parameters were manually tuned to detect the animal in the example image in A, and not re-tuned for analyzing the same animal at later timepoints in B-E. Successful segmentation of the worm by Tierpsy Tracker is denoted by a white arrow, with non-worm segmentations marked by red arrows (*middle column*). Worm detections using the WoP Faster R-CNN model are bounded by a red box (*left column*). All worm detections shown reached a confidence threshold of 0.99. The animal in A is detected by both the WoP model and Tierpsy Tracker, but other non-worm objects are identified based on the optimized Tierpsy Tracker segmentation.
B. The same animal as in A at a later timepoint. The animal is identified using the WoP model, but not identified by Tierpsy Tracker.
C. As contrast improves, the same animal is detected by both Tierpsy Tracker and the WoP model, but the segmentation parameters as optimized for small, low-contrast animals also pick up non-worm objects.
D. Once the animal becomes a gravid adult, the animal is identified by both Tierpsy Tracker and the WoP model, but eggs and tracks in the bacterial lawn increase the number of non-worm segmentations by Tierpsy Tracker.
E. Illumination changes increases the number of non-worm segmentations by Tierpsy Tracker, while the WoP model is still able to identify the animal and no other non-worm objects.

Figure S2. Limitations of traditional image processing techniques in aging populations

A. Detection of a young worm using traditional techniques (Tierpsy Tracker) or the trained WiCh Faster R-CNN model *(top row)* Successful detection of a worm via Tierpsy Tracker. Parameters for traditional techniques were manually tuned for this specific video. *(bottom rows)* Detection of worms (under the same age and environmental condition) using the same parameters as before. Detection errors are highlighted by red arrows. Red boxes show detection location via WiCh Faster R-CNN model.

B. Detection of an old, slow moving worm using traditional techniques (Tierpsy Tracker) or the trained WiCh Faster R-CNN model *(top row)* Successful detection of a worm. Parameters for traditional techniques were manually tuned for this specific video. *(bottom rows)* Detection of worms (under the same age and environmental condition) using the same parameters as before. Errors and misidentification are highlighted by red arrows. Red boxes show detection location via WiCh Faster R-CNN model.

Figure S3. Precision-recall curves for the detection models

C. Precision-recall curves for worm detection in the WoP model with confidence threshold of 0.5. Precision recall curve for all worms (left), L2-L3 stage animals (middle), and adult animals (right).
D. Precision-recall curve for the worm detection (left) and egg detection (right) in the egg-finder model with confidence threshold of 0.01. The intersection over union used to determine true positive detections for eggs was 0.3, compared to 0.5 for worms.
E. Precision-recall curve for the overall worm detection (left) and egg detection (right) in the WiCh model with confidence threshold of 0.5 (top row). Precision-recall curves for the worm at varying stages in the lifespan with confidence threshold of 0.5 (middle row), and across different food levels/contrasts with confidence threshold of 0.5 (bottom row).

Figure S4. Accurate centroid tracking in other datasets

A. Comparison of X centroid coordinates from Stern *et al.* to Faster R-CNN WoP model detections of the same data.
B. Comparison of Y centroid coordinates from Stern *et al.* to Faster R-CNN WoP model detections of the same data.

Figure S5. Accurate linear and angular velocity analysis with other datasets

A. Comparison of binned angular velocity over time from Stern *et al.* and Faster R-CNN WoP detections of the same data. Data was binned by first smoothing angular velocities using a 10s moving average window (as in Stern *et al*.), then thresholding the data into low and high angular velocities. The angular velocity values represent the average angular velocity of the low and high angular velocity data for each data set independently.
B. Linear velocity of both datasets vs. time. Linear velocity was calculated in the same way from both datasets, then smoothed with a moving average window of 10s, and finally by removing outliers.
C. Scatterplot comparing Faster R-CNN WoP centroid velocities to Stern *et al.* ground truth velocities.

Figure S6. Accuracy of tracking in development

A. Change of bounding box size over time using WoP Faster R-CNN for animal depicted in 4A. Smoothed using a moving window average over 10 time points (10 minutes). Where no animal is detected, line is not connected.

B. Density histogram of distances between box centroid as detected by WoP Faster R-CNN model and centroid of annotated worm shape for animal in 4A (n=30 time points).
C. Density histogram of difference between movement calculated from annotations and movement calculated from WoP Faster R-CNN model for animal in 4A (n=25 time points).
D. Centroid movement as calculated by the WoP Faster R-CNN model vs. centroid movement as calculated from manually segmented animals for animal in 4A (n=25)

Figure S7. Accurate classification of roaming and dwelling at reduced sampling frequency

A. Linear vs. angular velocity probability plot, calculated as described in Stern *et al*. with centroid data from Stern *et al*. Black dashed line shows split used to classify roaming vs. dwelling states.
B. Linear vs. angular velocity probability plot, calculating angular velocity by using the Stern et al. centroids at the current time as well as the centroid one minute in the past and one minute into the future. Black dashed line shows split at 90 degrees/min angular velocity used to classify roaming/ dwelling states with 79% accuracy based on ground truth classification in (A).

Figure S8. Accurate detection of worms using the WiCh model

Histogram of IoU values for bounding boxes detected by the WiCh Faster R-CNN model compared to bounding boxes of hand annotated, segmented worms of the same frame. (n = 2550 frames).

Figure S9. Limitations with existing machine learning based segmentation tools

A. Representative example frames of issues with segmentation using Ilastik within the same video. Even after training at least 50 frames (including a frame from the same video) the classification predictions and subsequent segmentations truncate the worm. Blue denotes background, yellow marks the worm, and red marks the egg objects.
B. Representative example frames of issues with segmentation using Ilastik across similar videos. All frames were taken under the same imaging condition. (top) Prediction of pixel classification using the trained model. The model was trained with at least 50 images prior. Blue denotes background, yellow marks the worm, and red marks the egg objects. (bottom) Segmentation of objects based on the predictions. Note the truncation of worms and the misclassification of eggs as worms.

Supplemental Files

1. Video from the Open Worm Movement Database (https://www.youtube.com/channel/UCx36wu_Hh0sGvPaCkAMHrMg) with egg and

760 worm detections using the egg-finder Faster R-CNN detection model overlaid in blue
761 (eggs) and red (worms).
762 2. Time-lapse video of a developing N2 animal (the same animal as in Figure 4A-B) imaged
763 at one-minute intervals on an agarose plate. The video speed is such that 1s is
764 equivalent to 30 minutes of images. The left half of the frame consists of the original
765 images, and the right half is the original image with a green box indicating the WoP
766 Faster R-CNN detection overlaid.
767 3. Video from (Stern et al. 2017) with worm detections using WoP Faster R-CNN model
768 overlaid in red (x5 speed).
769 4. Time-lapse video of a developing N2 animal (the same animal as in Figure 4A-B) with
770 behavioral state identified for each timepoint based on the angular velocity. If the
771 location of the worm could not be identified in any set of three consecutive frames, the
772 frame is marked with 'No data'.
773 5. Time-lapse video of an aging worm every other day from L4 to Day 16 within the
774 microfluidic device at food level OD_{600}10. The video is at 2x speed.
775 6. Time-lapse video of a plate with adult animals overlaid with worm detections in red
776 from the Faster R-CNN model trained using our web-based pipeline. The model was
777 trained for ~ 2 hours on 12 annotated images.
778
779
780

References

Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M, et al. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. tensorflow.org.

Anderson DJ, Perona P. 2014. Toward a science of computational ethology. Neuron. 84(1):18–31. doi:10.1016/j.neuron.2014.09.005.

Anholt RRH, O'Grady P, Wolfner MF, Harbison ST. 2020. Evolution of reproductive behavior. Genetics. 214(1):49–73. doi:10.1534/genetics.119.302263. [accessed 2021 Jan 27]. https://doi.org/10.1534/genetics.119.302263.

Ben Arous J, Laffont S, Chatenay D. 2009. Molecular and Sensory Basis of a Food Related Two-State Behavior in C. elegans. Brezina V, editor. PLoS One. 4(10):e7584. doi:10.1371/journal.pone.0007584. [accessed 2021 Jan 12]. https://dx.plos.org/10.1371/journal.pone.0007584.

Atakan HB, Ayhan F, Gijs MAM. 2020. PDMS filter structures for size-dependent larval sorting and on-chip egg extraction of: C. elegans. Lab Chip. 20(1):155–167. doi:10.1039/c9lc00949c. [accessed 2021 Jan 26]. https://pubs.rsc.org/en/content/articlehtml/2020/lc/c9lc00949c.

Berg S, Kutra D, Kroeger T, Straehle CN, Kausler BX, Haubold C, Schiegg M, Ales J, Beier T, Rudy M, et al. 2019. ilastik: interactive machine learning for (bio)image analysis. Nat Methods. 16(12):1226–1232. doi:10.1038/s41592-019-0582-9. [accessed 2021 May 4]. https://doi.org/10.1038/s41592-019-0582-9.

Brown AEX, De Bivort B. 2018. Ethology as a physical science. Nat Phys. 14(7):653–657. doi:10.1038/s41567-018-0093-0. [accessed 2021 Jan 13]. https://doi.org/10.1038/s41567-018-0093-0.

Brown AEX, Gomez-Marin A, Stephens GJ. 2016. Hierarchical compression of Caenorhabditis elegans locomotion reveals phenotypic differences in the organization of behaviour. J R Soc Interface. doi:10.1098/rsif.2016.0466. [accessed 2017 Feb 7]. http://dx.doi.org/10.1098/rsif.2016.0466.

Brown AEX, Yemini EI, Grundy LJ, Jucikas T, Schafer WR. 2013. A dictionary of behavioral motifs reveals clusters of genes affecting Caenorhabditis elegans locomotion. Proc Natl Acad Sci U S A. 110(2):791–6. doi:10.1073/pnas.1211447110. [accessed 2017 Mar 3]. http://www.ncbi.nlm.nih.gov/pubmed/23267063.

Buckingham SD, Partridge FA, Sattelle DB. 2014. Automated, high-throughput, motility analysis in Caenorhabditis elegans and parasitic nematodes: Applications in the search for new anthelmintics. Int J Parasitol Drugs Drug Resist. 4(3):226–232. doi:10.1016/j.ijpddr.2014.10.004.

Chalfie M, Sulston JE, White JG, Southgate E, Thomson JN, And +, Brenners S. 1985. The Neural Circuit for Touch Sensitivity in Caenorhabditis elegans'.

Chen YC, Seyedsayamdost MR, Ringstad N. 2020. A microbial metabolite synergizes with endogenous serotonin to trigger C. elegans reproductive behavior. Proc Natl Acad Sci U S A. 117(48):30589–30598. doi:10.1073/pnas.2017918117. [accessed 2021 Jan 12]. https://www.pnas.org/content/117/48/30589.

Chung K, Zhan M, Srinivasan J, Sternberg PW, Gong E, Schroeder FC, Lu H. 2011. Microfluidic chamber arrays for whole-organism behavior-based chemical screening. Lab Chip. 11(21):3689–3697. doi:10.1039/c1lc20400a.

Churgin MA, Jung S-K, Yu C-C, Chen X, Raizen DM, Fang-Yen C. 2017. Longitudinal imaging of *Caenorhabditis elegans* in a microfabricated device reveals variation in behavioral decline during aging.

820 Elife. 6:e26652. doi:10.7554/eLife.26652. [accessed 2017 Oct 9].
821 http://elifesciences.org/lookup/doi/10.7554/eLife.26652.

822 Collins KM, Bode A, Fernandez RW, Tanis JE, Brewer JC, Creamer MS, Koelle MR. 2016. Activity of the C.
823 elegans egg-laying behavior circuit is controlled by competing activation and feedback inhibition. Elife.
824 5(NOVEMBER2016). doi:10.7554/eLife.21126.

825 Colman RJ, Anderson RM, Johnson SC, Kastman EK, Kosmatka KJ, Beasley TM, Allison DB, Cruzen C,
826 Simmons HA, Kemnitz JW, et al. 2009. Caloric restriction delays disease onset and mortality in rhesus
827 monkeys. Science (80-). 325(5937):201–204. doi:10.1126/science.1173635.

828 Duveau F, Félix MA. 2012. Role of pleiotropy in the evolution of a cryptic developmental variation in
829 Caenorhabditis elegans. PLoS Biol. 10(1). doi:10.1371/journal.pbio.1001230.

830 Entchev E V., Patel DS, Zhan M, Steele AJ, Lu H, Chng QL. 2015. A gene-expression-based neural code for
831 food abundance that modulates lifespan. Elife. 4(MAY). doi:10.7554/eLife.06259.

832 Fenk LA, De Bono M. 2015. Environmental CO2 inhibits Caenorhabditis elegans egg-laying by modulating
833 olfactory neurons and evokes widespread changes in neural activity. Proc Natl Acad Sci U S A.
834 112(27):E3525–E3534. doi:10.1073/pnas.1423808112. [accessed 2021 Jan 26].
835 www.pnas.org/cgi/doi/10.1073/pnas.1423808112.

836 Flavell SW, Pokala N, Macosko EZ, Albrecht DR, Larsch J, Bargmann CI. 2013. Serotonin and the
837 neuropeptide PDF initiate and extend opposing behavioral states in C. Elegans. Cell. 154(5):1023–1035.
838 doi:10.1016/j.cell.2013.08.001.

839 Fletcher M, Kim DH. 2017. Age-Dependent Neuroendocrine Signaling from Sensory Neurons Modulates
840 the Effect of Dietary Restriction on Longevity of Caenorhabditis elegans. Ashrafi K, editor. PLOS Genet.
841 13(1):e1006544. doi:10.1371/journal.pgen.1006544. [accessed 2020 Feb 11].
842 https://dx.plos.org/10.1371/journal.pgen.1006544.

843 Geng W, Cosman P, Palm M, Schafer WR. 2005. Caenorhabditis elegans egg-laying detection and
844 behavior study using image analysis. EURASIP J Appl Signal Processing. 2005(14):2229–2240.
845 doi:10.1155/ASP.2005.2229. [accessed 2021 Jan 26]. https://asp-
846 eurasipjournals.springeropen.com/articles/10.1155/ASP.2005.2229.

847 Golden TR, Hubbard A, Dando C, Herren MA, Melov S. 2008. Age-related behaviors have distinct
848 transcriptional profiles in *Caenorhabditis elegans*. Aging Cell. 7(6):850–865. doi:10.1111/j.1474-
849 9726.2008.00433.x. [accessed 2020 Aug 28]. http://doi.wiley.com/10.1111/j.1474-9726.2008.00433.x.

850 Gray JM, Hill JJ, Bargmann CI. 2005. A circuit for navigation in Caenorhabditis elegans. Proc Natl Acad Sci
851 U S A. 102(9):3184–3191. doi:10.1073/pnas.0409009101. [accessed 2021 May 2].
852 www.pnas.orgcgidoi10.1073pnas.0409009101.

853 Greer EL, Brunet A. 2009. Different dietary restriction regimens extend lifespan by both independent
854 and overlapping genetic pathways in C. elegans. Aging Cell. 8:113–127. doi:10.1111/j.1474-
855 9726.2009.00459.x. [accessed 2018 Jan 5].
856 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2680339/pdf/ace0008-0113.pdf.

857 Hahm J, Kim S, Diloreto R, Shi C, Lee S V, Murphy CT, Nam HG. 2015. C. elegans maximum velocity
858 correlates with healthspan and is maintained in worms with an insulin receptor mutation. Nat Commun.
859 6:1–7. doi:10.1038/ncomms9919. http://dx.doi.org/10.1038/ncomms9919.

Hedgecock EM, Russell RL. 1975. Normal and mutant thermotaxis in the nematode Caenorhabditis elegans. Proc Natl Acad Sci U S A. 72(10):4061–4065. doi:10.1073/pnas.72.10.4061. [accessed 2021 May 2]. /pmc/articles/PMC433138/?report=abstract.

Huang J, Rathod V, Sun C, Zhu M, Korattikara A, Fathi A, Fischer I, Wojna Z, Song Y, Guadarrama S, et al. 2017. Speed/accuracy trade-offs for modern convolutional object detectors. In: Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017. Vol. 2017-Janua. Institute of Electrical and Electronics Engineers Inc. p. 3296–3305.

Hung J, Ravel D, Lopes SCP, Rangel G, Nery OA, Malleret B, Nosten F, Lacerda MVG, Ferreira MU, Rénia L, et al. 2018. Applying Faster R-CNN for Object Detection on Malaria Images. http://arxiv.org/abs/1804.09548.

Javer A, Currie M, Lee CW, Hokanson J, Li K, Martineau CN, Yemini E, Grundy LJ, Li C, Ch'ng QL, et al. 2018. An open-source platform for analyzing and sharing worm-behavior data. Nat Methods. 15(9):645–646. doi:10.1038/s41592-018-0112-1.

Jung S-K, Aleman-Meza B, Riepe C, Zhong W. 2014. QuantWorm: A Comprehensive Software Package for Caenorhabditis elegans Phenotypic Assays. Hendricks M, editor. PLoS One. 9(1):e84830. doi:10.1371/journal.pone.0084830. [accessed 2021 Jan 26]. https://dx.plos.org/10.1371/journal.pone.0084830.

Krakauer JW, Ghazanfar AA, Gomez-Marin A, MacIver MA, Poeppel D. 2017. Neuroscience Needs Behavior: Correcting a Reductionist Bias. Neuron. 93(3):480–490. doi:10.1016/j.neuron.2016.12.041. [accessed 2021 Jan 23]. http://dx.doi.org/10.1016/j.neuron.2016.12.041.

Le KN, Zhan M, Cho Y, Wan J, Patel DS, Lu H. 2020. An automated platform to monitor long-term behavior and healthspan in Caenorhabditis elegans under precise environmental control. Commun Biol. 3(1):1–13. doi:10.1038/s42003-020-1013-2. [accessed 2021 Jan 12]. https://www.nature.com/articles/s42003-020-1013-2.

Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL. 2014. Microsoft COCO: Common objects in context. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Vol. 8693 LNCS. Springer Verlag. p. 740–755. [accessed 2021 Jan 12]. https://arxiv.org/abs/1405.0312v3.

Maia Chagas A, Prieto-Godino LL, Arrenberg AB, Baden T, Mathur N, Delfosse K, Doherty T. 2017. The €100 lab: A 3D-printable open-source platform for fluorescence microscopy, optogenetics, and accurate temperature control during behaviour of zebrafish, Drosophila, and Caenorhabditis elegans. PLOS Biol. 15(7):e2002702. doi:10.1371/journal.pbio.2002702. [accessed 2017 Jul 25]. http://dx.plos.org/10.1371/journal.pbio.2002702.

Mair W, Goymer P, Pletcher SD, Partridge L. 2003. Demography of dietary restriction and death in Drosophila. Science (80-). 301(5640):1731–1733. doi:10.1126/science.1086016.

Mathew MD, Mathew ND, Ebert PR. 2012. WormScan: A Technique for High-Throughput Phenotypic Analysis of Caenorhabditis elegans. Lehner B, editor. PLoS One. 7(3):e33483. doi:10.1371/journal.pone.0033483. [accessed 2021 Jan 27]. https://dx.plos.org/10.1371/journal.pone.0033483.

Mendel JE, Korswagen HC, Liu KS, Hajdu-Cronin YM, Simon MI, Plasterk RHA, Sternberg PW. 1995. Participation of the protein Go in multiple aspects of behavior in C. elegans. Science (80-).

267(5204):1652–1655. doi:10.1126/science.7886455. [accessed 2021 Jan 26]. https://pubmed.ncbi.nlm.nih.gov/7886455/.

Niepoth N, Bendesky A. 2020. How Natural Genetic Variation Shapes Behavior. doi:10.1146/annurev-genom-111219. [accessed 2021 Jan 13]. https://doi.org/10.1146/annurev-genom-111219-.

Perni M, Challa PK, Kirkegaard JB, Limbocker R, Koopman M, Hardenberg MC, Sormanni P, Müller T, Saar KL, Roode LWY, et al. 2018. Massively parallel C. elegans tracking provides multi-dimensional fingerprints for phenotypic discovery. J Neurosci Methods. 306:57–67. doi:10.1016/j.jneumeth.2018.02.005. https://doi.org/10.1016/j.jneumeth.2018.02.005.

Ren S, He K, Girshick R, Sun J. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. [accessed 2020 Oct 19]. https://github.com/.

Riddle DL, Blumenthal T, Meyer BJ et al., editor. 1997. Section I, Introduction. Title. In: C. elegans II. 2nd ed. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press. https://www.ncbi.nlm.nih.gov/books/NBK20121/.

Schafer WR. 2005. Egg-laying. WormBook.:1–7. doi:10.1895/wormbook.1.38.1.

Ségalat L, Elkes DA, Kaplan JM. 1995. Modulation of serotonin-controlled behaviors by Go in Caenorhabditis elegans. Science (80-). 267(5204):1648–1651. doi:10.1126/science.7886454. [accessed 2021 Jan 26]. https://pubmed.ncbi.nlm.nih.gov/7886454/.

Sofela S, Sahloul S, Rafeie M, Kwon T, Han J, Warkiani ME, Song YA. 2018. High-throughput sorting of eggs for synchronization of: C. elegans in a microfluidic spiral chip. Lab Chip. 18(4):679–687. doi:10.1039/c7lc00998d. [accessed 2021 Jan 26]. https://pubs.rsc.org/en/content/articlehtml/2018/lc/c7lc00998d.

Stephens GJ, Johnson-Kerner B, Bialek W, Ryu WS, Green D, Swets J, Geng W, Cosman P, Baek J, Berry C, et al. 2008. Dimensionality and Dynamics in the Behavior of C. elegans. Sporns O, editor. PLoS Comput Biol. 4(4):e1000028. doi:10.1371/journal.pcbi.1000028. [accessed 2017 Feb 7]. http://dx.plos.org/10.1371/journal.pcbi.1000028.

Stern S, Kirst C, Bargmann CI. 2017. Neuromodulatory Control of Long-Term Behavioral Patterns and Individuality across Development. Cell. 171(7):1649-1662.e10. doi:10.1016/j.cell.2017.10.041. [accessed 2018 Jan 4]. http://www.ncbi.nlm.nih.gov/pubmed/29198526.

Stiernagle T. 2006. Maintenance of C. elegans. WormBook.:1–11. doi:10.1895/wormbook.1.101.1.

Stroustrup N, Ulmschneider BE, Nash ZM, López-Moyado IF, Apfeld J, Fontana W. 2013. The Caenorhabditis elegans Lifespan Machine. Nat Methods. 10(7):665–70. doi:10.1038/nmeth.2475. http://dx.doi.org/10.1038/nmeth.2475.

Swierczek NA, Giles AC, Rankin CH, Kerr RA. 2011. High-throughput behavioral analysis in C. elegans. Nat Methods. 8(7):592–602. doi:10.1038/nmeth.1625.

Trent C, Tsung N, Horvitz HR. 1983. EGG-LAYING DEFECTIVE MUTANTS OF THE NEMATODE CAENORHABDITIS ELEGANS. Genetics. 104(4).

Waggoner LE, Zhou GT, Schafer RW, Schafer WR. 1998. Control of alternative behavioral states by serotonin in Caenorhabditis elegans. Neuron. 21(1):203–214. doi:10.1016/S0896-6273(00)80527-9. [accessed 2021 Jan 26]. http://www.cell.com/article/S0896627300805279/fulltext.

940 Wählby C, Kamentsky L, Liu Z, Riklin-Raviv T, Conery A. 2012. An image analysis toolbox for high-
941 throughput C. elegans assays. Nat Methods. 9.

942 Zhang WB, Sinha DB, Pittman WE, Hvatum E, Stroustrup N, Pincus Z. 2016. Extended Twilight among
943 Isogenic C. elegans Causes a Disproportionate Scaling between Lifespan and Health. Cell Syst. 3(4):333–
944 345. doi:10.1016/j.cels.2016.09.003. http://dx.doi.org/10.1016/j.cels.2016.09.003.

945 Zhao Y, Long L, Xu W, Campbell RF, Large EE, Greene JS, McGrath PT. 2018. Changes to social feeding
946 behaviors are not sufficient for fitness gains of the Caenorhabditis elegans N2 reference strain. Elife. 7.
947 doi:10.7554/eLife.38675.

948 Zheng M, Gorelenkova O, Yang J, Feng Z. 2012. A liquid phase based C. elegans behavioral analysis
949 system identifies motor activity loss in a nematode Parkinson's disease model. J Neurosci Methods.
950 204(2):234–237. doi:10.1016/j.jneumeth.2011.11.015.

951

Figure 1

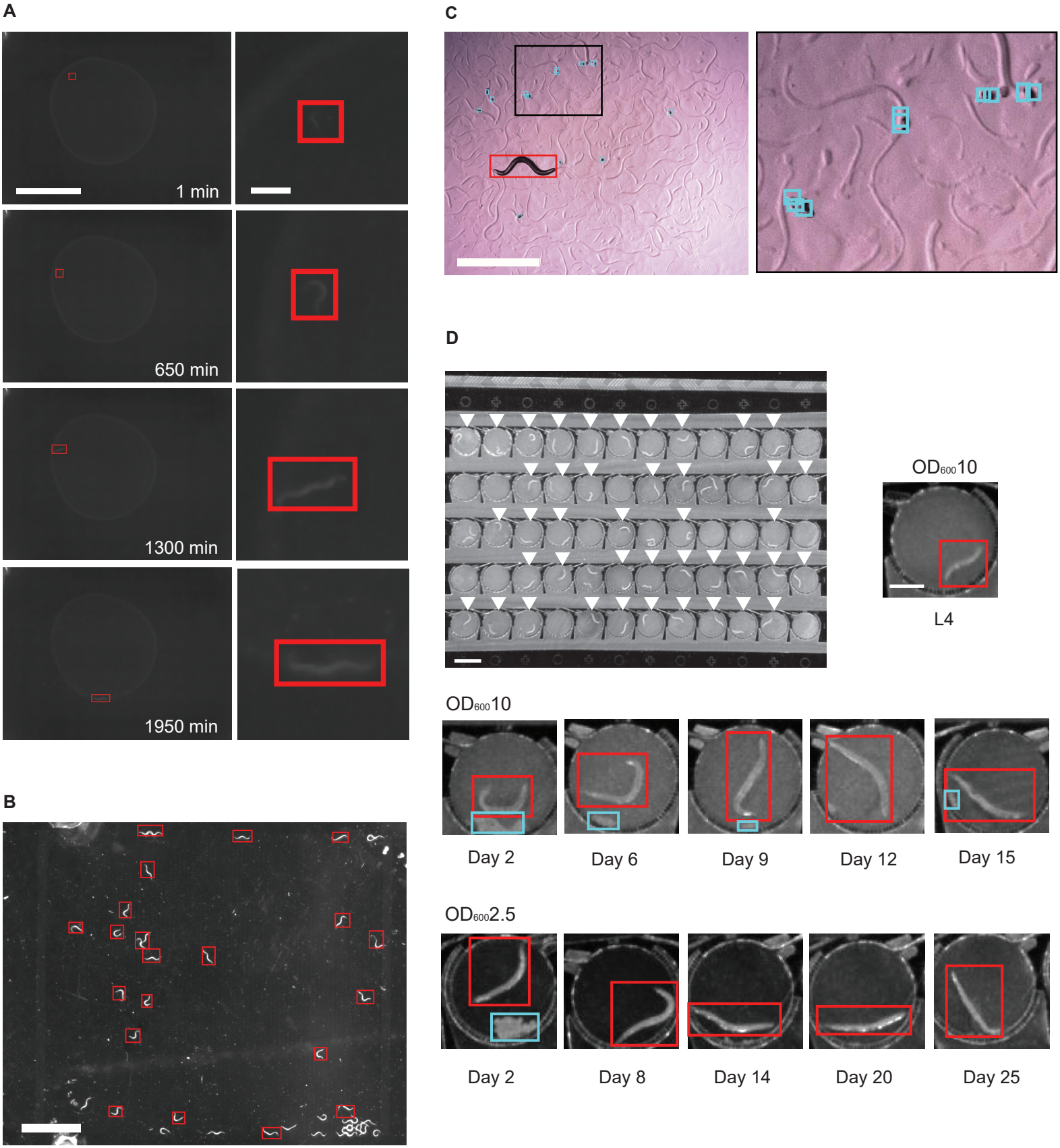

Figure 2

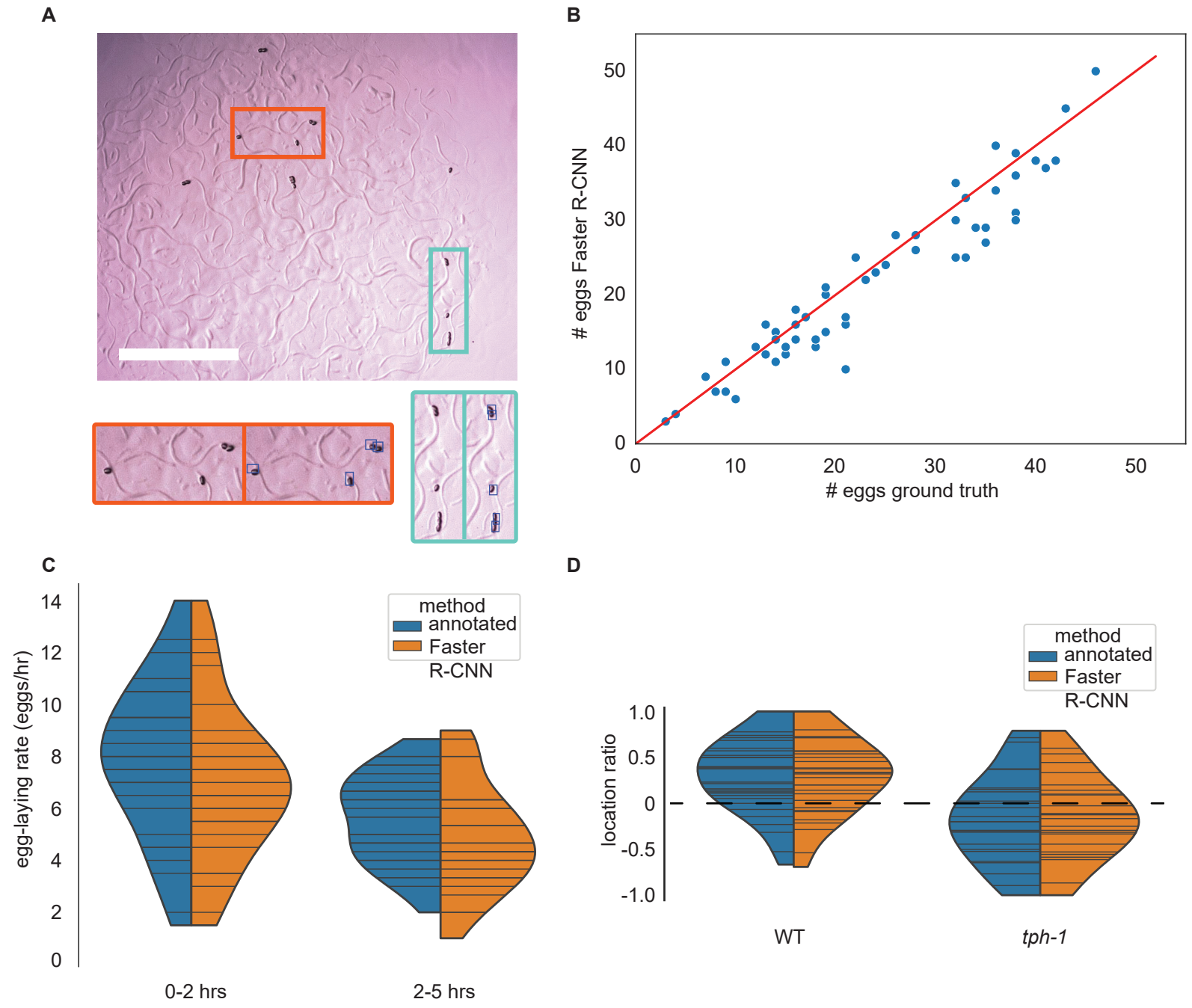

Figure 3

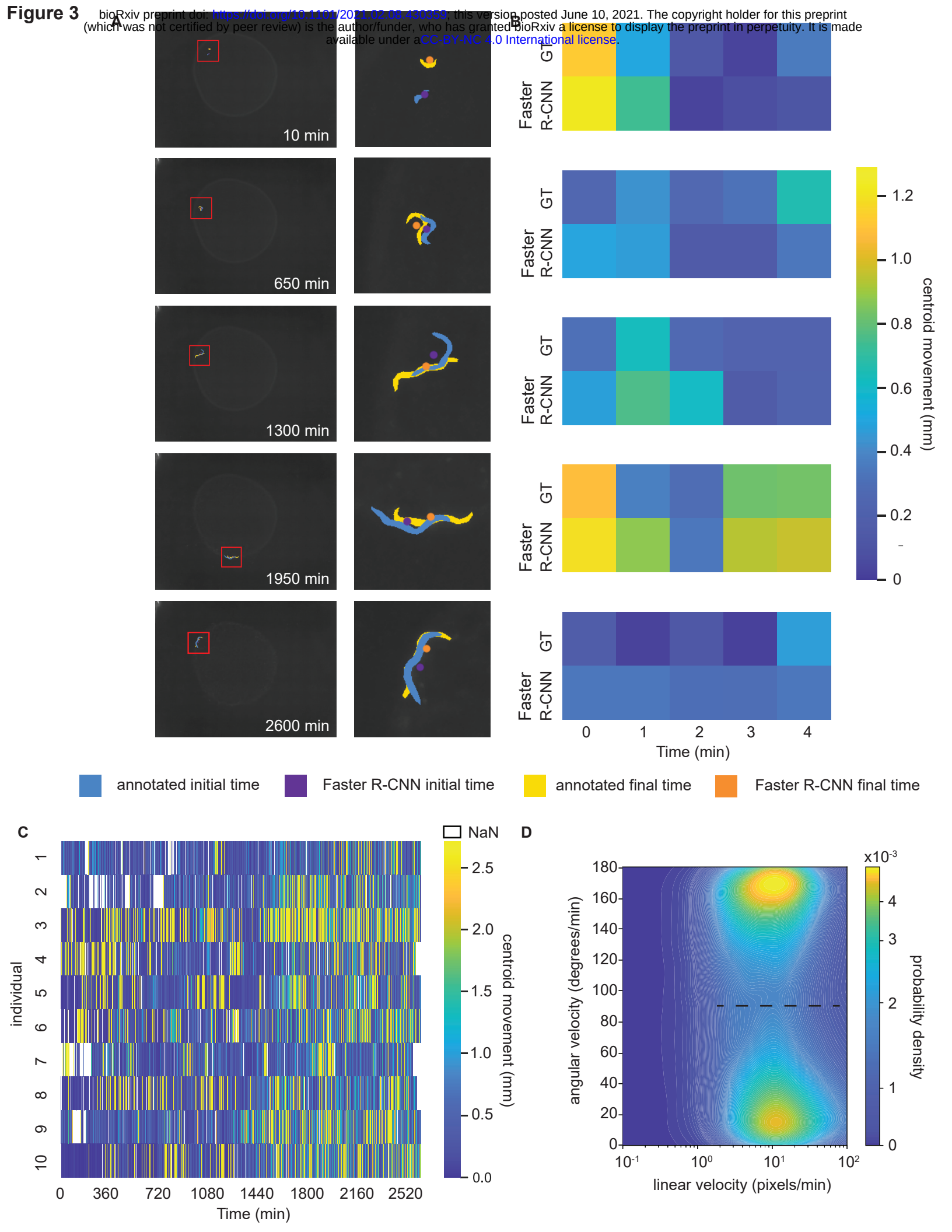

Figure 4