Image recognition based on deep learning in *Haemonchus contortus* motility assays

Martin Žofka, Linh Thuy Nguyen, Eva Mašátová, Petra Matoušková*

Department of Biochemical Sciences, Faculty of Pharmacy, Charles University, Heyrovského 1203, 500 05 Hradec Králové, Czech Republic.

Correspondence:

*Corresponding Author

matousp7@faf.cuni.cz

Abstract

Poor efficacy of some anthelmintics and rising concerns about the widespread drug resistance have highlighted the need for new drug discovery. The parasitic nematode *Haemonchus contortus* is an important model organism widely used for studies of drug resistance and drug screening with the current gold standard being the motility assay. We applied a deep learning approach Mask R-CNN for analysing motility videos containing varying rates of motile worms and compared it to other commonly used algorithms with different levels of complexity, namely the Wiggle Index and the Wide Field-of-View Nematode Tracking Platform. Mask R-CNN consistently outperformed the other algorithms in terms of the detection of worms as well as the precision of motility forecasts, having a mean absolute percentage error of 7.6% and a mean absolute error of 5.6% for the detection and motility forecasts, respectively. Using Mask R-CNN for motility assays confirmed the common problem with algorithms that use non-maximum suppression in detecting overlapping objects, which negatively impacts the overall precision. The use of intersect over union as a measure of the classification of motile / non-

motile instances had an overall accuracy of 89%, indicating that it is a viable alternative to previously used methods based on movement characteristics, such as body bends. In comparison to the existing methods evaluated here, Mask R-CNN performed better and we anticipate that this method will broaden the number of possible approaches to video analysis of worm motility.

Keywords: parasite; nematode; instance segmentation; object detection; Mask R-CNN

1. Introduction

Parasitic nematode infections present a substantial problem to human and veterinary medicine. Anthelmintics are often used as a weapon of choice to control or eliminate parasitic worms. With increasing reports of drug resistance, there is an ongoing effort to develop new anthelmintics [1, 2]. In view of this, drug screening and drug testing methods are clearly necessary. A number of important human and animal nematodes were already studied in large chemical library-scale screenings [3, 4], although nearly all of the studies were aimed at well-established non-parasitic *Caenorhabditis elegans* because of the scientific interest and medical applications as potential human disease models [5-9].

Nowadays, *in vitro* assays can measure the effects of compounds on development, growth, behaviour, and motility. Several approaches have been pursued in attempts to deliver robust, automated assays on the nematode phenotype [10-14]. The current gold standard for measuring drug effectiveness is *in vitro* assessment of worm motility, as measured via microscopy. The automated phenotyping from short video recordings offers a useful assay for high-throughput whole-organism phenotypic screening and avoids manual scoring. In assays with parasitic worms under experimental conditions, several challenges need to be overcome to enhance the performance of motility measurements. These challenges are associated with the

50 different worm lengths, their tendency to clump, and also the range of diverse, complex
51 movement patterns different from that seen in *C. elegans*.

52 Focusing on veterinary importance, the parasitic nematode *Haemonchus contortus*,
53 known as the barber's pole worm, is a model organism often used for drug screening and
54 studying drug resistance; due to its economic importance, known genome and transcriptome
55 [15, 16]. *H. contortus* infects predominantly small ruminants. It has a direct life cycle that starts
56 with eggs being shed onto a pasture in the faeces of the infected animals. The cycle continues
57 through three free-living larval stages to the parasitic fourth larval stage and lastly adults, which
58 reside in the abomasum where they feed on blood. The ability to maintain *H. contortus* infection
59 in a natural host animal and produce different larval stages *in vitro* provides experimental
60 advantage. The infective third-stage larvae (L3s) can be stored, and researchers commonly use
61 L3s or *in vitro* fourth larval stage for motility studies. Many automated or semi-automated
62 approaches to measure motility have already been described such as via electrical impedance
63 [17] or infrared light beam-interference [7]. Nevertheless, in our study of analysing the video
64 recordings in *H. contortus* larvae to assess the motility, we have compared various
65 (representative) algorithms with different levels of complexity in terms of model preparation as
66 well as computational cost. We were interested to assess if the additional complexity of the
67 methods leads to a sufficient increase in precision, that would justify the time-performance
68 trade-off of these methods.

69 The first one mentioned is an algorithm which calculates a motility index value based
70 on measuring the standard deviation of the pixel's light intensity averaged for a number of
71 frames [18]. This Wiggle Index (WI) does not allow for the detection of the phenotype of
72 individual worms; still, this assay led to many "hit compounds" in screening of a large variety
73 of compound libraries [19, 20]. The second approach called Wide Field-of-View Nematode
74 Tracking Platform (WF-NTP) is built on the use of the Gaussian adaptive threshold of the image

75 frames to identify worms and was applied in model free-living nematode *C. elegans* [21]. It
76 was initially developed for drug screening purposes; nevertheless, it has also found applications
77 in other studies such as detailed genetic or behavioural studies. Nonetheless, WF-NTP has a
78 major limitation for images with overlapping objects as these objects get discarded by the
79 algorithm based on an upper pixel size threshold per object, resulting in underestimation of the
80 the total number of detected objects in an image [21]. In our study, we customized the associated
81 WF-NTP software written in Python and compared it to the last approach which involves the
82 application of deep learning.

83 The advent of machine learning has revolutionized the surveying and classifying of
84 biological data including image recognition, enabling the automation of many tasks. The most
85 important aspect of large-scale computerization is the possibility to automate *in vitro* assays
86 and scale them, allowing researchers to avoid tedious repetitive tasks and focus on activities
87 where they can use their knowledge to provide added value, leaving the analysis and
88 interpretation of the data as the only rate-limiting step. In the field of computer vision, deep
89 learning algorithms with a convolutional neural network (CNN) architecture have made rapid
90 advancements on a variety of image classification tasks in cell cultures [22], and the model
91 organism *C. elegans* [23, 24], even parasitic nematode in plants [25], yet none of them have
92 been applied to parasitic gastro-intestinal nematodes of livestock, including *H. contortus*.
93 Object-detection-based deep learning methods have also been adapted for instance
94 segmentation. Here we used a state-of-the-art region-based CNN called Mask R-CNN [26],
95 which can be trained based on a set of annotated images (Fig 1a) to detect the objects of interest,
96 in our case the larval stages of *H. contortus*. In the first step, the algorithm uses a regional
97 proposal network (RPN), i.e., a binary classifier that identifies rectangular subsets of the image
98 that are of interest for the given task and are called regions of interest (ROI). In our case that
99 would be parts of the image that potentially contain worms. These ROIs are then refined to

100 cover the whole object (Fig 1b) and filtered using non-maximum suppression (NMS). NMS is
101 a process of removing highly overlapping ROIs, to avoid detecting the same object multiple
102 times before passing the filtered ROIs as proposals for classification. The classifier performs a
103 refinement of the ROI and outputs a final rectangle surrounding the detected object called a
104 bounding box, the class of the predicted instance along with the detection confidence and the
105 object mask, which are the pixels that make up the object (Fig 1c).

106 The aim of the study was to train a Mask R-CNN model for image recognition of
107 different life cycle stages of *H. contortus* and evaluate the application of this deep learning
108 algorithm to the field of drug discovery of anthelmintics. Thereby, we compared the Mask R-
109 CNN detections and motility predictions to the WI and WF-NTP methods.. Additionally for the
110 motility, we tested a metric called intersection over union (IoU) instead of body bends from
111 WF-NTP readout. IoU is used regularly in image recognition to quantify the extent of overlap
112 between two objects, the intersection of the objects is divided by the union of the objects, which
113 results in a numerical value between 0 and 1 (see Graphical abstract). Furthermore, we
114 discussed the possible application of automated Mask R-CNN in other assays related to the
115 veterinary practice.

116

117 **Fig 1**: Mask R-CNN training and detection process **(a)** annotated images used for training the
118 algorithm, **(b)** regions of interest (ROI) detected in the first step by the regional proposal
119 network (solid boxes) and the refined regions (dashed boxes), **(c)** Mask R-CNN predictions
120 containing the object boundaries, predicted object classes (i.e., L3 – third-stage larva) with an
121 associated confidence level and the mask of the object (coloured contour).

122

123 ## 2. Materials and methods

124 ### *2.1.* Procurement of L3s of *H. contortus*

125 Six-month-old sheep were orally infected with 8,000 infective L3s of MHco3 strain
126 (Inbred Susceptible Edinburgh, ISE) [27]. Four weeks post-infection faecal samples were
127 collected. L3s were produced from eggs by incubating faeces in a plastic box covered with foil
128 at 27 °C for 7 days. Then, the faeces were rinsed twice in tap water which was poured into
129 1,000mL conical measuring cups, in which the larvae sank to bottom. To remove dirt or dead
130 individuals, the pellet of larvae was pipetted out and filtered through a 20µm sieve submerged
131 in water (27 °C). Clean L3s at a concentration of approximately 4,000 L3 per mL were stored
132 in culture flasks in water at 10 °C for several months.

133 During the storage, some larvae naturally die. Prior to the experiment, we filtered larvae
134 through a 20µm sieve (for 12 h), this time collecting both live larvae from the bottom of the
135 sieve and dead individuals remaining on the sieve.

136 All experimental procedures were examined and approved by the Ethics Committee of
137 the Ministry of Education, Youth and Sports (Protocol MSMT-25908/2014-9).

138

139 ### 2.2. Video processing and manual counting

140 Videos from a microscope camera were recorded for a duration of 10 seconds with a
141 framerate of 20 frames per second (fps) with a pixel quality of 2560 by 2160 (see
142 Supplementary Video). The videos (.avi with MJPEG compression, ~ 0.45GB) were acquired
143 using NIS-Elements Imaging Software (version 4.20) using a Nikon Eclipse Ti microscope (4x
144 magnification) with the camera (Andor Zyla 5.5 sCMOS, 12bit: 2560 x 2160 pixels). The
145 videos were recorded for different ratios of live and dead larvae (100:0, 75:25, 50:50, 25:75,
146 0:100; denoted as motility group 100, 75, 50, 25 and 0, respectively) in 9 replicates and they
147 contained on average 60 worms. The live larvae were kept at 37 °C to maintain their motility.

148 The video recordings were processed so that each worm was labelled with a numeric
149 identifier. A trained human with prior experience of manual motility assays evaluation assessed
150 each worm from each video and information about motility and the dead phenotype was
151 recorded. The discrepancies were dealt with as follows (i) a dead worm which was moving due
152 to the interference of a motile worm was counted as motile; (ii) a live coil non-motile worm
153 was counted as non-motile; (iii) the worms which were not labelled by an algorithm were given
154 an identifier manually; (iv) worms were excluded from the counts when they appeared in less
155 than 20 frames; (v) dead phenotype was recorded when overt microscopical pathologies, such
156 as the presence of large vacuoles in the tegument and straight phenotype, were observed.
157 This manual counting was taken as the referential method for comparison of accuracy
158 among the three models.
159
160 **2.3. Wiggle Index analysis**
161 We used an ImageJ implementation of the WI for the analysis [28]. We did not apply
162 any scaling of the videos, the Gaussian image blur was set to 2, the number of frames to be
163 averaged was set to 50 and the motility output was based on the standard deviation. Since the
164 values from the WI are relative, we used motility group 100 as the control group to be able to
165 convert these relative values to percentages. We calculated the mean value for motility group
166 100 and divided all values by the mean to obtain normalized motility values.
167
168 **2.4. WF-NTP analysis**
169 WF-NTP presents an automated evaluation of motility in *C. elegans*. For its successful
170 implementation, we adopted the method for the *H. contortus* motility assay. The algorithm was
171 based on three steps;in the first instance all worms in each frame were detected, in the second
172 step, these detections were linked based on a neighbourhood search, and in the final step, each

173 worm was classified as motile or non-motile based on bends per second and the speed of the
174 worm. However, as these parasites differ in their sizes as well as their movement, some
175 modifications to the original configuration and algorithm were necessary. These modifications
176 belonged predominantly to two categories, the first was a recalibration of the available input
177 parameters of the WF-NTP method. This included aspects such as the image size, the size of
178 the worms that we want to detect and other configurations (see Data availability). The other
179 modifications were changes in the evaluation of the conditions for worm motility and a limit
180 on the minimum number of frames that a worm needed to be detected in for it to be included in
181 the evaluation.

182 The original method relied on bends per second and the speed of individual worms to
183 classify them as either motile or non-motile [21]. When a worm was below a certain threshold
184 of bends per second and below a certain speed threshold, it was classified as non-motile. In our
185 settings, the speed threshold was obtained from the motility group 0. From the initial test for
186 the motility group 50, the result was that out of 169 worms classified as motile, 23% exceeded
187 the speed threshold, but no bends were recorded. This failure to measure bends per second for
188 motile worms, was most likely caused by the more complex movement of *H. contortus.*, This
189 led us to replace the original approach with an evaluation metric, IoU, that is regularly used in
190 image recognition [29]. The use of IoU allowed us to use the capabilities of WF-NTP in terms
191 of the detection and tracking of individual worm instances in an image and then apply the IoU
192 metric to determine the worm motility. The calculation of IoU was done for the same worm
193 instance in neighbouring frames to detect changes on a frame-by-frame basis. Afterwards, the
194 mean was calculated and worms above a certain IoU threshold were labelled as non-motile. To
195 ensure that the comparison of WF-NTP and Mask R-CNN was aimed at the quality of the
196 detection of individual worm instances rather than the approach to determine the motility, we

used the same tracking for detections across frames and the same IoU conditions for both methods.

Another modification made to the original method was the introduction of a threshold for the minimum number of frames that a worm needed to be detected in, in order to be included in the analysis. The reason for this restriction was that as the individual worms move, they have a tendency to change shape or overlap with other worms, therefore in terms of the detection the individual worm can be lost for a certain number of frames and reappear later, however this can lead to counting some instances multiple times. To limit these potential duplicates (Supplementary Fig S1) as much as possible and not bias the total number of worms detected in the video, we selected the highest number of worms detected in a given video across all frames and added a buffer of 10% in order to account for situations where some worms were not detected, for example due to their overlaps. This maximum number then served as a limit for selecting the number of worms to be evaluated as motile or non-motile. The instances were selected based on the number of frames in a decreasing order, therefore ensuring that we had data from enough frames to classify the worm correctly. Again, this restriction was applied to detections from Mask R-CNN to enable a comparison under the same conditions.

2.5. Mask R-CNN analysis

For Mask R-CNN we used the Matterport implementation (https://github.com/matterport/Mask_RCNN) of the algorithm, including a trained model using a backbone architecture of Resnet101. The pre-trained model weights were obtained from training the model on the MS COCO dataset [30]. The process of using a pre-trained model is called Transfer Learning, which is an approach where the knowledge gained from solving one problem is applied to a related problem. The motivation for selecting a pre-trained model is that in general deep neural networks require a lot of annotated data for training and using an existing

222 model and re-training the top layer of the network allows us to have a smaller dataset while
223 maintaining good performance. We manually annotated a total of 95 images and divided them
224 into training (69) and validation (26) using the VGG (Visual Geometry Group) Image
225 Annotator [31], these images contained three categories of objects; eggs, first-stage larvae (L1),
226 and L3, to fully utilise the capabilities of Mask R-CNN which is able to do instance
227 segmentation and classification. The total number of annotated objects was around 10,400
228 divided into training (6,800 objects) and validation (3,600 objects). Then we ran the training on
229 a GPU instance until the loss function showed no signs of improvement. To avoid overfitting,
230 we used data augmentation and applied one or multiple of the following augmentations:
231 flipping, contrast normalization, additive gaussian noise, multiplying the pixels by a number
232 within a given interval to make the whole image lighter or darker. The final model had a mean
233 average precision at 0.5 IoU of 64.1% on the validation set.

234 The trained model was then used on the individual frames of the video to detect the
235 instances of parasites and we calculated the centroid of the detected object in terms of the
236 position on the x and y axis using the scikit-image regionProps functionality [32]. These
237 detected instances were then linked together using the trackpy library [33], which detected the
238 trajectories of objects based on certain conditions. In our case we restricted the distance that the
239 centroid could move a maximum of 100 pixels in each dimension. As a result of detecting the
240 trajectories, we were able to identify the same parasite instance across multiple frames and
241 assign it with a unique identifier. In cases where the tracking did not meet the above conditions,
242 a new identifier was assigned, and therefore a single worm could be detected multiple times,
243 which occurred for complex cases with extensive movements and a high number of overlaps.
244 Subsequently, we calculated the IoU and the threshold for the number of parasites to select for
245 the evaluation as described in the WF-NTP section. We observed a decreasing mean percentage

246 error (MPE) with an increase in the length of the video, and therefore we decided on using the
247 full length of the videos for the motility assay (Supplementary Fig S2).
248
249 **2.6. Error metrics**
250 For the evaluation of the performance of the individual algorithms, we used standard
251 metrics for the quantification of the forecast precision. These metrics work with the error term,
252 which is defined as the difference between the actual and the forecasted value.In our case, the
253 actual value was obtained by manually processing the videos. Two aspects were measured; the
254 first one was the bias of the forecasts to determine if the algorithm is systematically over or
255 under-estimating, for that purpose, we used mean error (ME) or mean percentage error (MPE),
256 which both work with the mean of the error term. The second aspect was the size of the error
257 term irrespective of the error term being positive or negative to compare the overall
258 performance of the algorithms. For that purpose, we used mean absolute error (MAE) or mean
259 absolute percentage error (MAPE), which both work with the mean absolute value of the error
260 term.
261
262 **2.7. Monte Carlo drug screening simulation**
263 To assess the impact of the differences in detection accuracy we ran a Monte Carlo drug
264 screening simulation [34] for all three algorithms. To remove the effect of individual motility
265 groups we worked with the differences between the predicted motility rates and the manually
266 obtained motility rates. Using these differences we were able to obtain the correlations between
267 the algorithms as well as the means and standard deviations. A multivariate normal distribution
268 was used as the data generation process for the simulation to ensure that the correlation among
269 the algorithms was taken into account. The SciPy package (version 1.3.3) implementation of
270 the multivariate normal distribution was used.

For the simulation, we generated 1 million random motility rates from a continuous uniform distribution within the range of zero to one. Similarly, we generated 1 million observations of the multivariate normal distribution. These observations were the detection errors of each of the algorithms. By adding this error to the randomly generated motility rate, which served as our true observation, we received simulated motility predictions from the three algorithms. A "hit" in our simulation was defined as having a motility rate lower than 50%.

3. Results

3.1. Detection of motility

The observed typical cases for detection scenarios in the motility videos were depicted in Fig 2, where the results of Mask R-CNN were used for illustrating these examples. The most common situation, which did not pose a problem to the algorithms, were worms with few or no overlaps (Fig 2a). Other common occurrences were collisions between worms that caused a non-motile worm to be moved and therefore being perceived by the algorithms as motile (Fig 2b). Lastly, the most challenging scenarios contained a high number of worms with significant overlaps with mostly motile worms, which resulted in problematic detections and negatively impacted the ability to track worms across multiple frames (Fig 2c).

Fig 2: Sample frames of common detection scenarios in motility videos. The blue outlines are the masks detected by Mask R-CNN. Each detected worm has an associated identification number displayed in white based on the tracking across the frames. **(a) c**ommon scenario with low or no overlaps of individual worms. **(b) e**xample of a motile worm (id 32) colliding with a non-motile worm (id 26) causing the later to be perceived as motile. **(c) c**omplex detection scenario with a high number of worms with significant overlaps

296 **3.2. Motility forecast errors**

297 In the manual video processing, we observed the motility and the live rate. To measure
298 the performance of the algorithms, we used the motility rate rather than the live rate even though
299 the stillness alone does not denote a dead worm. The reason for this choice was that the
300 algorithms detect movement and are not able to differentiate if the movement was caused by
301 the object itself or by an interaction with another object. The general manual motility rate was
302 always higher than the manual live rate (Supplementary Table S1). We measured the
303 performance of the individual algorithms by evaluating the errors, defined as the differences
304 between the manual processing of the file and the respective algorithm. Out of the three
305 methods, Mask R-CNN had the lowest MAE with a value of 5.6%, followed by WF-NTP
306 (8.76%) and the WI (14.2%). In terms of the bias of the predictions, the WI had the lowest ME
307 of -0.71%, therefore slightly overestimating the motility rate on average, while both WF-NTP
308 and Mask R-CNN underestimated the motility rate with values of 6.52% and 1.95%
309 respectively (Table 1).

310 Shown in Fig 3, the WI had an increasing volatility with an increase in the ratio of alive
311 worms in the group. Mask R-CNN overestimated the motility rate for the motility group 0;
312 while four videos had a zero motility rate, showing that the overestimation is not systematic,
313 two videos had a motility rate error below 3.6%, and the remaining three videos had high
314 motility rates without any apparent reason. WF-NTP and Mask R-CNN had a tendency to
315 systematically underestimate for the motility group 100 (Fig 3). The mean and standard
316 deviation of motility rates per motility group and algorithm is shown also in Supplementary
317 Table S2 and the same statistics are reported as differences between the manual count and the
318 respective algorithm in Supplementary Table S3.

319

320

321 **Table 1:** Motility error metrics for algorithms

Algorithm	MAE, %	ME, %
Wiggle Index	14.2	-0.71
WF-NTP	8.76	6.52
Mask R-CNN	5.61	1.95

322

323 The differences between the manual processing and the respective algorithm were expressed as
324 the mean absolute error (MAE) and mean error (ME).

325

326 **Fig 3:** Motility rates comparisons for individual motility groups per algorithm. The number of
327 the motility group indicates the percentage of live larvae. The motility rate is visualized as box
328 plots.

329

330 **3.3. Instance detection performance**

331 The motility rates are an indicator of the degree of correctly classified worms as motile
332 or non-motile based on IoU. However, they do not provide the information of the number of
333 detected worm instances in the image. In order to ensure that the motility rates were estimated
334 reliably and not only based on a small subset of detected worms, the detection performance was
335 validated. To achieve this, we compared Mask R-CNN and WF-NTP for the total worm count
336 (WI was excluded as the algorithm does not detect individual instances of worms, therefore this
337 information was not available). The precision was measured based upon the error term between
338 the manual counts and the respective algorithm. In terms of the size of the error, the MAPE was
339 used and we observed that the error was significantly lower ($P < 0.0001$, Wilcoxon signed-rank
340 test) for Mask R-CNN (7.6%) than for WF-NTP (40.23%). Both algorithms had a tendency to
341 underestimate the number of worms on average, therefore not detecting all the worm instances

342 in a video, mainly due to the challenges of detecting overlapping or clustered worms. This was
343 confirmed by the MPE with values of 5.61% for Mask R-CNN and 40.23% for WF-NTP (Table
344 2). The reported MAPE and MPE for WF-NTP had the same value because all the detections
345 underestimated the total number of worms. WF-NTP underestimated the worm count for all
346 motility groups, while Mask R-CNN had similar box plots in comparison to the manual counts
347 apart from underestimating the motility group 0, as shown in Fig 4. The mean error and standard
348 deviation of the differences between the manual count and the respective motility algorithm is
349 shown in Supplementary Table S4.
350
351 **Table 2:** Worm count error metrics for algorithms

Algorithm / Metric	MAPE, %	MPE, %
WF-NTP	40.23	40.23
Mask R-CNN	7.6	5.61

352
353 The differences between the manual counts and the respective algorithm were expressed as the
354 mean absolute percentage error (MAPE) and mean percentage error (MPE).
355
356 **Fig 4:** Worm count comparisons for individual motility groups per algorithm. The number of
357 the motility group indicates the percentage of live larvae. The number of worms is visualized
358 in box plots.
359
360 **3.4. Mask R-CNN detection and classification precision**
361 The precision of detecting worm instances in a video was on average 98.6%. This was
362 calculated based on all the tracked worms across all the motility groups without setting a
363 minimum number of frames for which a worm needed to be tracked. The detections were

manually validated and the detection was considered correct despite only a part of the worm being detected in some frames. The majority of cases where the detection was incorrect was due to debris that had a similar size and shape as a worm. For the detected worms across all the motility groups, there was an overall accuracy of 89% to correctly label them as either motile or non-motile based on their mean IoU values. In Fig 5, the distribution for different IoU values obtained from all the worms across all the motility groups clearly showed that most of misclassified cases were located around the 0.8 threshold, while the classification accuracy increased for values further away from the threshold. The precision and recall values (Table 3) confirm that the prevalent case was the misclassification of motile worms as non-motile.

The errors defined as differences between Mask R-CNN and the manually evaluated motility rates were tested for normality using the D'Agostino-Pearson test as well as Shapiro-Wilk test. In both cases we could not reject the null hypothesis that the data is from a normal distribution with p-values 0.85 and 0.35 respectively (Table 4).

Fig 5: The distribution for different Intersection over Union (IoU). Both **(a) a** histogram of IoU and the associated number of worms per bin classified into motile / non-motile and **(b)** a histogram of IoU and the associated probability per class (motile / non-motile) show that the majority of misclassified cases were located around the 0.8 IoU.

Table 3: Classification performance metrics

Classification	**Precision, %**	**Recall, %**
Motile	94	88
Non-Motile	84	92

Precision and recall metrics for classification performance

386 **Table 4:** Normality tests for the motility rate error terms between Mask R-CNN and the manual
387 processing

Normality test	Statistic	p-value
D'Agostino-Pearson	0.32	0.85
Shapiro-Wilk	0.97	0.35

388

389 For both normality tests we cannot reject the null hypothesis that the data is from a normal
390 distribution.

391 **3.5. Monte Carlo drug screening simulation**

392 The Monte Carlo simulation was based on the results from the motility forecasts and
393 the obtained forecast errors. While WF-NTP and Mask R-CNN had a positive correlation of
394 forecast errors (0.65), the forecast errors for WI were not correlated with the other algorithms
395 (Table 5).

396

397 **Table 5**: Correlation matrix for the prediction errors of the individual algorithms

	Wiggle Index	WF-NTP	Mask R-CNN
Wiggle Index	1		
WF-NTP	-0.04	1	
Mask R-CNN	0.08	0.65	1

398

399 Mask R-CNN achieved the highest accuracy in the simulation with a value of 94%
400 followed by WF-NTP 90% and the WI 85% (full classification performance in Supplementary
401 Table S5). The simulation estimated that an improved detection performance of 3.15% (MAE)
402 between Mask R-CNN and WF-NTP led to a 3.5% improvement of drug screening accuracy,
403 while an improvement of 8.59% (difference of MAE between Mask R-CNN and WI) led to a
404 9% improvement of drug screening accuracy.

405

406 ## 4. Discussion

407 The motility phenotype is often used for drug screening and drug testing of
408 anthelmintics. In our study, we deployed and compared three algorithms to measure motility in
409 the parasitic nematode *H. contortus*. Apart from evaluating motility for both WF-NTP and
410 Mask R-CNN, we were also interested in the detection precision of worms by these algorithms.
411 The reason for that was the fact that for both algorithms, the detection was the first step followed
412 by the tracking across the frames and the classification into motile and non-motile. Therefore,
413 accurate detection was the pre-requisite for reliable motility quantification.

414 The results showed that the performance increased with an increased complexity of
415 algorithm. The WI, despite requiring minimal configuration and being the fastest algorithm,
416 had some limitations in comparison to the other methods; it was by design not able to detect
417 individual instances of worms and a control group with only motile worms needed to be created
418 in order to be able to translate the WI into a percentage for the motility rate for each video,
419 otherwise it was hard to interpret the numerical value of the WI. Because of these limitations
420 and the lack of precision, it would be advisable to use WI only for initial studies to quickly test
421 experimental results.

422 For WF-NTP, overlapping parasites had the highest negative impact on the detection
423 ability, as the algorithm filters out objects that are above or below a certain pixel size. This
424 resulted in detecting overlapping worms as a single object and these objects were then excluded
425 based on the size threshold. As a consequence, the total number of detected instances in an
426 image was significantly lower than the actual number obtained by manually processing the
427 videos. Although WF-NTP struggled with detecting all instances, the classification of motile /
428 non-motile performed well once an instance was detected.

429 Mask R-CNN had the lowest MAE. However, similarly to WF-NTP, it struggled with
430 overlapping objects, primarily due to the fact that the algorithm uses NMS. As a result, some

instances were not detected, although in contrast to WF-NTP, the worm clusters were not discarded as a whole. The problem of detecting overlapping objects is currently an area of active research in object detection, particularly for biological data. For instance, Böhm et al. [35] are working on algorithms that can handle overlapping objects and have had success with applying their algorithm to *C. elegans*. Motile worms had a higher chance of being detected, because they had higher variability across the frames due to their movement. In terms of precision, the algorithm was slightly biased towards underestimating the motility rate; however, the errors were normally distributed enabling us to increase the number of acquired videos in order to get to a desired level of confidence and therefore successfully replace manual processing.

The results confirmed that IoU could be used as a metric for determining the motility of larvae, although some modifications might be required in order to deal with slow-moving worms, such as including the total distance travelled during the video to complement the frame-by-frame information. Furthermore, the IoU threshold should be considered a parameter to be optimized as it can vary based on the age of the larvae, the temperature of the environment and other factors. The downside of using Mask R-CNN is that it requires a quite big amount of annotated data (provided in the Data availability), which is time consuming to create and the training runs for 1-2 days on a GPU instance (an Amazon Web Services p2.xlarge compute instance with 1 GPU, 4 vCPUs and 61GB RAM was used).

The drug screening simulation showed that the improved detection performance resulted in increased accuracy for drug screening of roughly the same magnitude as the MAE improvement, illustrating that the improvement is not negligible for practical purposes even though some of the forecast errors of the algorithms were correlated. For applications with multiple observations for a given category / group, the impact would be most likely lower, because with an increase in the number of observations per category / group, the sample mean of the forecast errors would converge to the expected value of the forecast errors. In this

456 situation, the bias of the estimate measured in our case by ME would then have a higher impact
457 than the variance which would make WF-NTP, with the highest bias, the least accurate
458 algorithm.

459 In general, for quick low-precision experiments, such as primary screening of large
460 compound libraries, the best option is to use the WI as it can be used out-of-the-box with slight
461 modifications to a few parameters. If the detection of individual instances is desired, Mask R-
462 CNN is the best choice, although it is more time-intensive to prepare the data and train the
463 model. Moreover, the ability to precisely detect larvae in an image then allows the researchers
464 to perform multiple types of analyses that involve microscopy of the parasites and are not only
465 limited to motility assays, but also include other common *in vitro* assays for the diagnosis of
466 drug resistance, e.g. larval development tests and egg hatch tests [36]. In the aspect of
467 potentially repurposing the algorithms for other assays, Mask R-CNN is the most flexible of
468 the models, because it is able to detect multiple classes of objects. On contrary, WF-NTP
469 decides solely on the size of the objects, making a distinction between classes of similar sizes
470 impossible.

471

472 **5. Conclusions**

473 Our study confirmed that the state-of-the-art machine learning algorithm Mask R-CNN was
474 able to outperform less computationally complex algorithms that are based on pixel intensity
475 or Gaussian threshold. The algorithm performed better in the detection of individual worm
476 instances as well as the subsequent motility classification. This was consistent with how the
477 algorithm is designed as the tracking and motility classification depend on the detection
478 accuracy. For the motility classification, we used IoU as a new metric, that successfully
479 replaced previous ones based on locomotion specific to the studied organisms. Whilst IoU is
480 very flexible, as it only requires a single parameter, it would need further extension to be able

481 to deal with cases of heterogeneous groups of viable worms that have a high variance in their
482 speed of movement.

483 The gain in precision of Mask R-CNN came at the cost of requiring an extensive annotated
484 dataset as well as computational resources for training the model. Therefore, the time-precision
485 trade-off needs to be evaluated by the researchers on a case-by-case basis, unless the study
486 allows reapplying an already existing trained model. The output of the present study provides
487 an annotated dataset containing various developmental stages of *H. contortus* that can be
488 utilised by researchers for future machine learning applications.

489

490 Data availability

491 All underlying data for training the model, the trained model itself as well as the input and
492 output of the motility analysis for all three algorithms is available at
493 https://doi.org/10.5281/zenodo.5734143. The source code for Mask R-CNN is available in a
494 repository on GitHub https://github.com/zofkam/mask_rcnn_motility. We have also used
495 Zenodo to assign a DOI to the repository https://doi.org/10.5281/zenodo.5735068.

496

497 Additional files

498 **Supplementary Table S1**: Manual counts of motility and live rates per motility group

499 **Supplementary Table S2**: Motility rates per motility group and algorithm

500 **Supplementary Table S3:** Mean error between the manual count and the respective motility
501 algorithm

502 **Supplementary Table S4**: Mean of differences in worm counts between the manual processing
503 and the respective algorithm

504 **Supplementary Table S5:** Drug screening simulation classification performance

Supplementary Fig S1: Average number of worms in the videos per motility group for WF-NTP and Mask R-CNN.

Supplementary Fig S2: The impact of the length of the video in frames on the percentage error of the detected worms for Mask R-CNN.

Supplementary Video: Mask R-CNN sample detection video of motility group 50.

Abbreviations

CNN: convolutional neural network; fps: frames per second; GPU: graphics processing unit; IoU: intersection over union; L3: third-stage larva; MAE: mean absolute error; mAP: mean average precision; MAPE: mean absolute percentage error; Mask R-CNN: region based convolutional neural network; ME: mean error; MPE: mean percentage error; NMS: non-maximum suppression; ROI: region of interest; RPN: regional proposal network; WF-NTP: Wide Field-of-View Nematode Tracking Platform; WI: Wiggle Index

Competing Interests

The authors declare that they have no competing interests.

Funding

This work was supported by Charles University UNCE/18/SCI/012 and SVV 260 550 as well as by the project EFSA-CDN [CZ.02.1.01/0.0/0.0/16_019/0000841], co-funded by ERDF.

Authors' contributions

M.Ž. and L.T.N. designed this study and wrote the manuscript. M.Ž. implemented the deep learning model, wrote the code, ran the statistical analyses and generated the figures. L.T.N. created the graphical abstract. L.T.N., E.M. acquired and annotated the images used to train the

deep learning model. M.Ž. and L.T.N. conducted the experiments and analyzed the results. P.M. managed the project. All authors contributed to literature review and critical revision of the manuscript.

Acknowledgements

We thank MVDr. Tomáš Kaňka for assisting with the sheep infection. The authors thank prof. Lenka Skálová for critical discussion and comments. We would like to express our gratitude to Jon Nolan for proofreading and revisions.

References

[1] Jiao Y, Preston S, Hofmann A, Taki A, Baell J, Chang BCH, et al. A perspective on the discovery of selected compounds with anthelmintic activity against the barber's pole worm-Where to from here? Adv Parasitol. 2020;108:1-45. doi: 10.1016/bs.apar.2019.12.003.

[2] Geary TG, Sakanari JA, Caffrey CR. Anthelmintic Drug Discovery: Into the Future. Journal of Parasitology. 2015;101(2):125-33. doi: 10.1645/14-703.1.

[3] Partridge FA, Murphy EA, Willis NJ, Bataille CJ, Forman R, Heyer-Chauhan N, et al. Dihydrobenz[e][1,4]oxazepin-2(3H)-ones, a new anthelmintic chemotype immobilising whipworm and reducing infectivity in vivo. PLoS Negl Trop Dis. 2017;11(2):e0005359. doi: 10.1371/journal.pntd.0005359.

[4] Abriola L, Hoyer D, Caffrey CR, Williams DL, Yoshino TP, Vermeire JJ. Development and optimization of a high-throughput screening method utilizing *Ancylostoma ceylanicum* egg hatching to identify novel anthelmintics. PLoS One. 2019;14(6):e0217019. doi: 10.1371/journal.pone.0217019.

[5] Buckingham SD, Partridge FA, Sattelle DB. Automated, high-throughput, motility analysis in *Caenorhabditis elegans* and parasitic nematodes: Applications in the search for new anthelmintics. Int J Parasitol Drugs Drug Resist. 2014;4(3):226-32. doi: 10.1016/j.ijpddr.2014.10.004.

[6] Burns AR, Luciani GM, Musso G, Bagg R, Yeo M, Zhang Y, et al. *Caenorhabditis elegan*s is a useful model for anthelmintic discovery. Nat Commun. 2015;6. doi: 10.1038/ncomms8485.

[7] Taki AC, Byrne JJ, Wang T, Sleebs BE, Nguyen N, Hall RS, et al. High-Throughput Phenotypic Assay to Screen for Anthelmintic Activity on *Haemonchus contortus*. Pharmaceuticals (Basel). 2021;14(7). doi: 10.3390/ph14070616.

[8] Elfawal MA, Savinov SN, Aroian RV. Drug Screening for Discovery of Broad-spectrum Agents for Soil-transmitted Nematodes. Scientific Reports. 2019;9. doi: 10.1038/s41598-019-48720-1.

[9] Pulak R. Techniques for analysis, sorting, and dispensing of *C. elegans* on the COPAS flow-sorting system. Methods Mol Biol. 2006;351:275-86. doi: 10.1385/1-59745-151-7:275.

[10] Ramot D, Johnson BE, Berry TL, Carnell L, Goodman MB. The Parallel Worm Tracker: A Platform for Measuring Average Speed and Drug-Induced Paralysis in Nematodes. PLoS One. 2008;3(5). doi: 10.1371/journal.pone.0002208.

[11] Javer A, Ripoll-Sanchez L, Brown AEX. Powerful and interpretable behavioural features for quantitative phenotyping of *Caenorhabditis elegans*. Philos Trans R Soc Lond B Biol Sci. 2018;373(1758). doi: 10.1098/rstb.2017.0375.

[12] Partridge FA, Brown AE, Buckingham SD, Willis NJ, Wynne GM, Forman R, et al. An automated high-throughput system for phenotypic screening of chemical libraries on *C. elegans* and parasitic nematodes. Int J Parasitol-Drug. 2018;8(1):8-21. doi: 10.1016/j.ijpddr.2017.11.004.

[13] Puckering T, Thompson J, Sathyamurthy S, Sukumar S, Shapira T, Ebert P. Automated Wormscan. F1000Res. 2019;6:192. doi: 10.12688/f1000research.10767.3.

[14] McDermott-Rouse A, Minga E, Barlow I, Feriani L, Harlow PH, Flemming AJ, et al. Behavioral fingerprints predict insecticide and anthelmintic mode of action. Mol Syst Biol. 2021;17(5):e10267. doi: 10.15252/msb.202110267.

[15] Gilleard JS. *Haemonchus contortus* as a paradigm and model to study anthelmintic drug resistance. Parasitology. 2013;140(12):1506-22. doi: 10.1017/S0031182013001145.

[16] Preston S, Jabbar A, Gasser RB. A perspective on genomic-guided anthelmintic discovery and repurposing using *Haemonchus contortus*. Infect Genet Evol. 2016;40:368-73. doi: 10.1016/j.meegid.2015.06.029.

[17] Smout MJ, Kotze AC, McCarthy JS, Loukas A. A Novel High Throughput Assay for Anthelmintic Drug Screening and Resistance Diagnosis by Real-Time Monitoring of Parasite Motility. Plos Neglected Tropical Diseases. 2010;4(11). doi: ARTN e885 10.1371/journal.pntd.0000885.

[18] Preston S, Jabbar A, Nowell C, Joachim A, Ruttkowski B, Baell J, et al. Low cost whole-organism screening of compounds for anthelmintic activity. Int J Parasitol. 2015;45(5):333-43. doi: 10.1016/j.ijpara.2015.01.007.

[19] Jiao Y, Preston S, Koehler AV, Stroehlein AJ, Chang BCH, Simpson KJ, et al. Screening of the 'Stasis Box' identifies two kinase inhibitors under pharmaceutical development with activity against *Haemonchus contortus*. Parasit Vectors. 2017;10(1):323. doi: 10.1186/s13071-017-2246-x.

[20] Preston S, Jiao Y, Jabbar A, McGee SL, Laleu B, Willis P, et al. Screening of the 'Pathogen Box' identifies an approved pesticide with major anthelmintic activity against the barber's pole worm. Int J Parasitol Drugs Drug Resist. 2016;6(3):329-34. doi: 10.1016/j.ijpddr.2016.07.004.

[21] Perni M, Challa PK, Kirkegaard JB, Limbocker R, Koopman M, Hardenberg MC, et al. Massively parallel *C. elegans* tracking provides multi-dimensional fingerprints for phenotypic discovery. J Neurosci Methods. 2018;306:57-67. doi: 10.1016/j.jneumeth.2018.02.005.

[22] Moen E, Bannon D, Kudo T, Graf W, Covert M, Van Valen D. Deep learning for cellular image analysis. Nat Methods. 2019;16(12):1233-46. doi: 10.1038/s41592-019-0403-1.

[23] Bates K, Le K, Lu H. Deep learning for robust and flexible tracking in behavioral studies for *C. elegans*. bioRxiv. 2021:2021.02.08.430359. doi: 10.1101/2021.02.08.430359.

[24] Fudickar S, Nustede EJ, Dreyer E, Bornhorst J. Mask R-CNN Based *C. Elegans* Detection with a DIY Microscope. Biosensors-Basel. 2021;11(8). doi: ARTN 257 10.3390/bios11080257.

[25] Divya V Naik PS. Leaf disease detection using Mask R-CNN. International journal of current engineering and scientific research. 2020;7(7):52-7.

[26] He K, Gkioxari G, Dollar P, Girshick R. Mask R-CNN. IEEE Trans Pattern Anal Mach Intell. 2020;42(2):386-97. doi: 10.1109/TPAMI.2018.2844175.

[27] Roos MH, Otsen M, Hoekstra R, Veenstra JG, Lenstra JA. Genetic analysis of inbreeding of two strains of the parasitic nematode *Haemonchus contortus*. Int J Parasitol. 2004;34(1):109-15. doi: 10.1016/j.ijpara.2003.10.002.

[28] Preston S, Jabbar A, Nowell C, Joachim A, Ruttkowski B, Cardno T, et al. Practical and low cost whole-organism motility assay: A step-by-step protocol. Mol Cell Probes. 2016;30(1):13-7. doi: 10.1016/j.mcp.2015.08.005.

[29] Rezatofighi H, Tsoi N, Gwak J, Sadeghian A, Reid I, Savarese S. Generalized Intersection over Union: A Metric and A Loss for Bounding Box Regression. 2019 Ieee/Cvf Conference on Computer Vision and Pattern Recognition (Cvpr 2019). 2019:658-66. doi: 10.1109/Cvpr.2019.00075.

[30] Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft COCO: Common Objects in Context. Computer Vision – ECCV 2014. 2014;8693:740-55. doi: 10.1007/978-3-319-10602-1_48.

[31] Dutta A, Zisserman A. The VIA annotation software for images, audio and video. Proceedings of the 27th ACM international conference on multimedia. 2019:2276-9. doi: 10.1145/3343031.3350535.

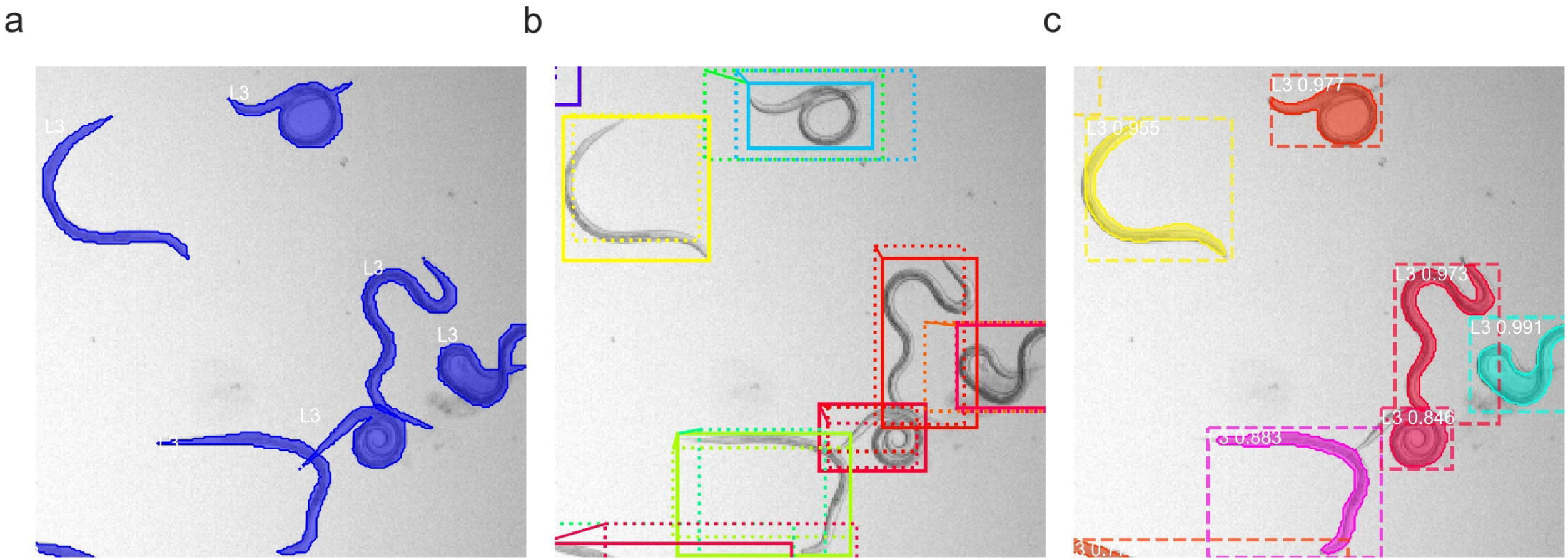
[32] van der Walt S, Schönberger JL, Nunez-Iglesias J, Boulogne F, Warner JD, Yager N, et al. scikit-image: Image processing in Python. PeerJ. 2014;2:e453. doi: 10.7717/peerj.453.

[33] Allan D, Caswell T, Keim N, van der Wel C. trackpy: Trackpy v0.3.2. 2016. doi: 10.5281/zenodo.1213240.

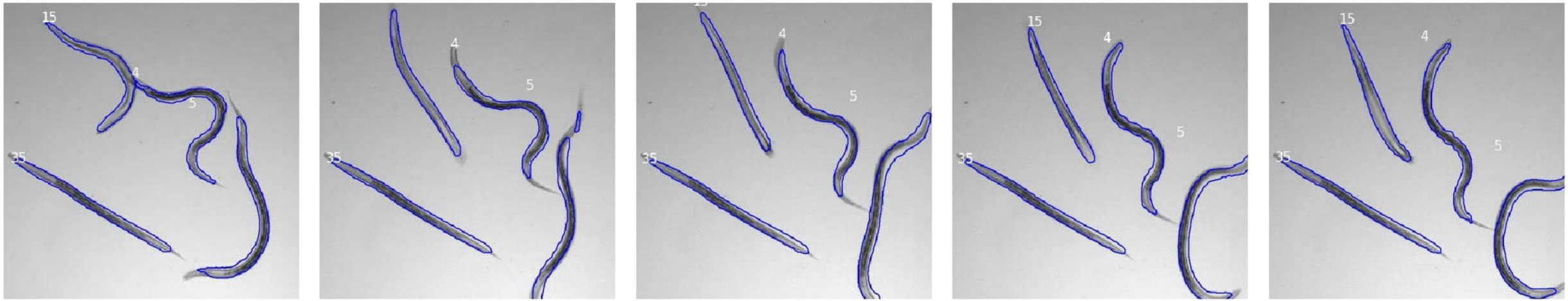
[34] Von Neumann J, Ulam S. Monte carlo method. National Bureau of Standards Applied Mathematics Series. 1951;12(1951):36.

[35] Böhm A, Mayer N, Brox T. Diskmask: Focusing Object Features for Accurate Instance Segmentation of Elongated or Overlapping Objects. 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI). 2020:230-4. doi: 10.1109/ISBI45749.2020.9098435.

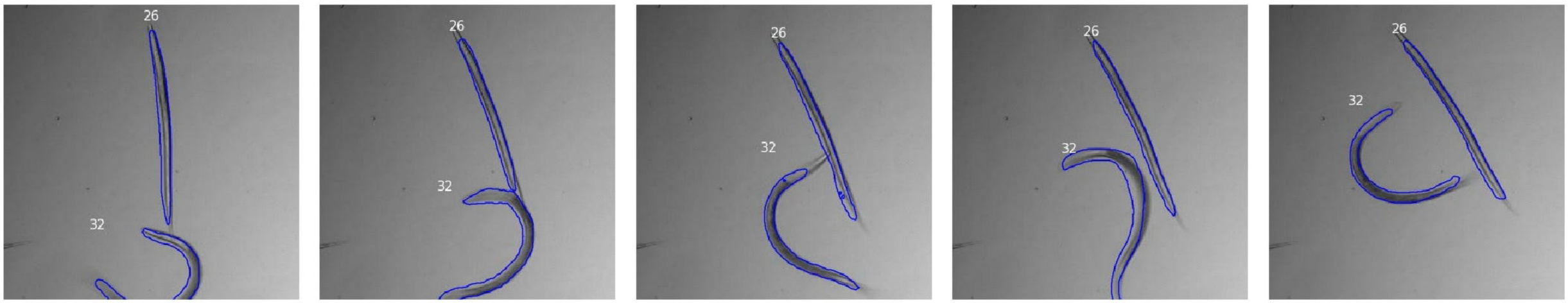
[36] Roeber F, Jex AR, Gasser RB. Advances in the diagnosis of key gastrointestinal nematode infections of livestock, with an emphasis on small ruminants. Biotechnol Adv. 2013;31(8):1135-52. doi: 10.1016/j.biotechadv.2013.01.008.


a
b
c
L3
L3
L3
L3
L3
L3
L3 0.977
L3 0.955
L3 0.973
L3 0.991
L3 0.846
L3 0.883

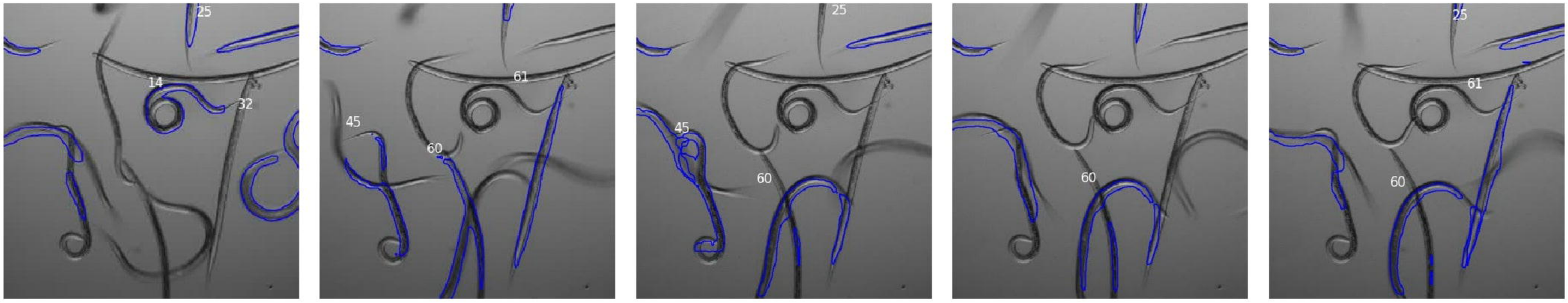
a

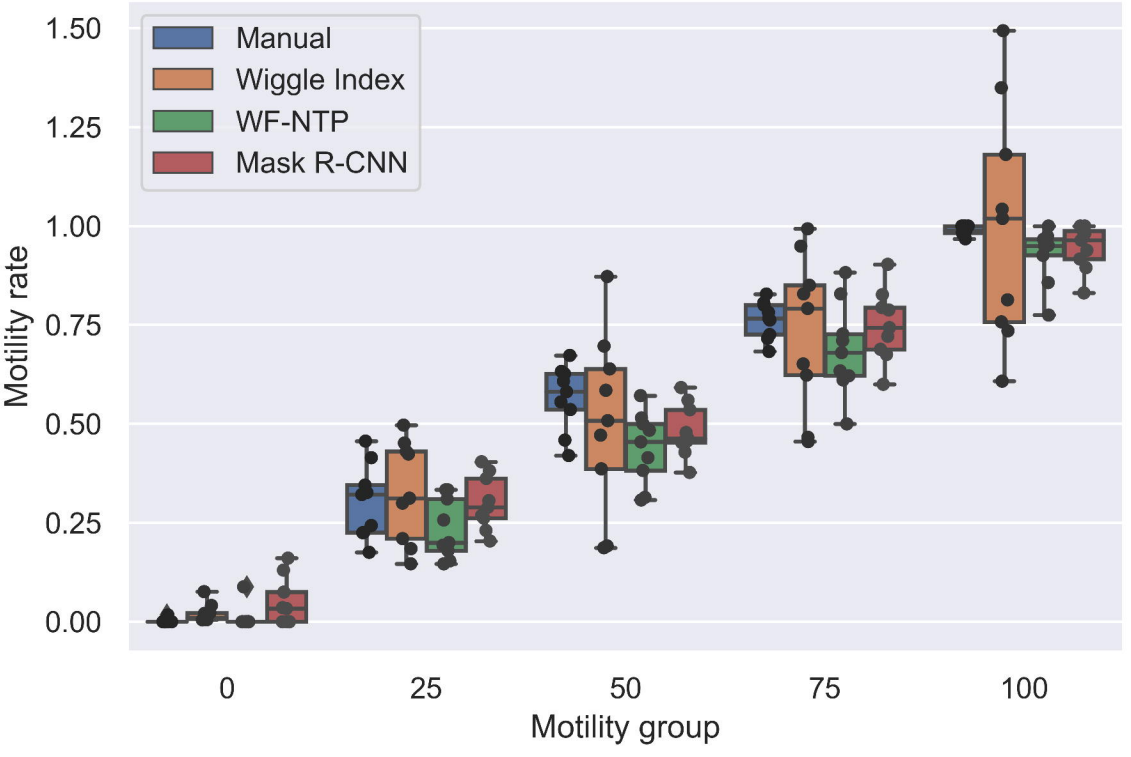


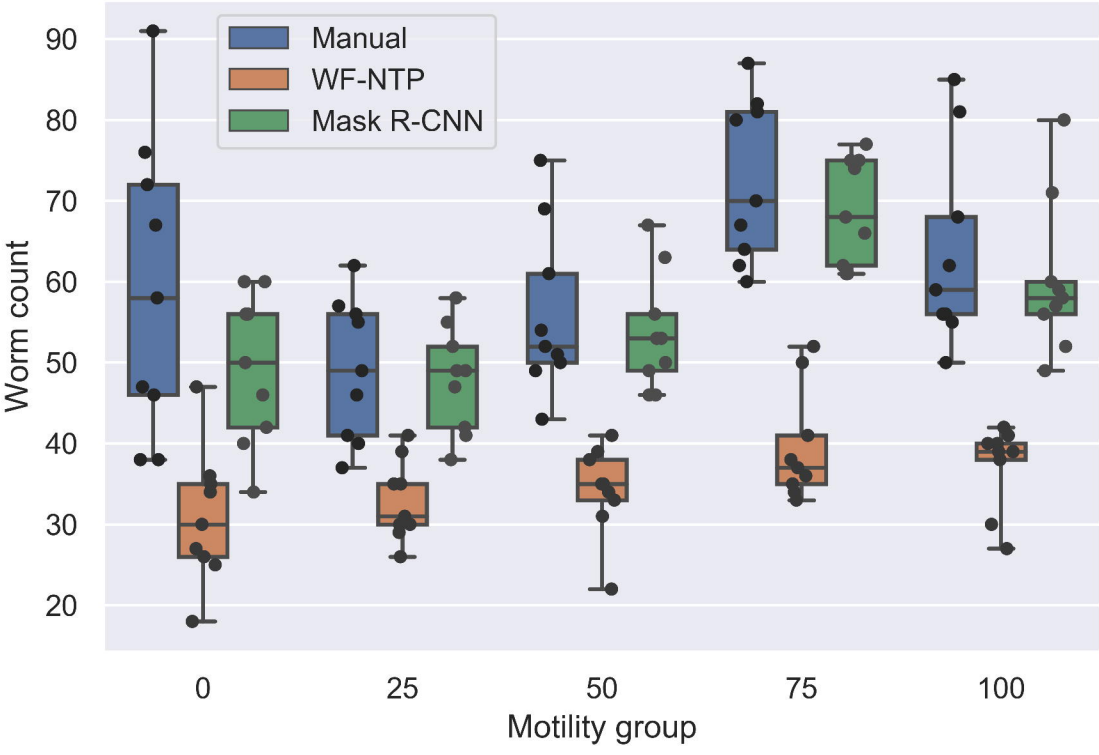
b

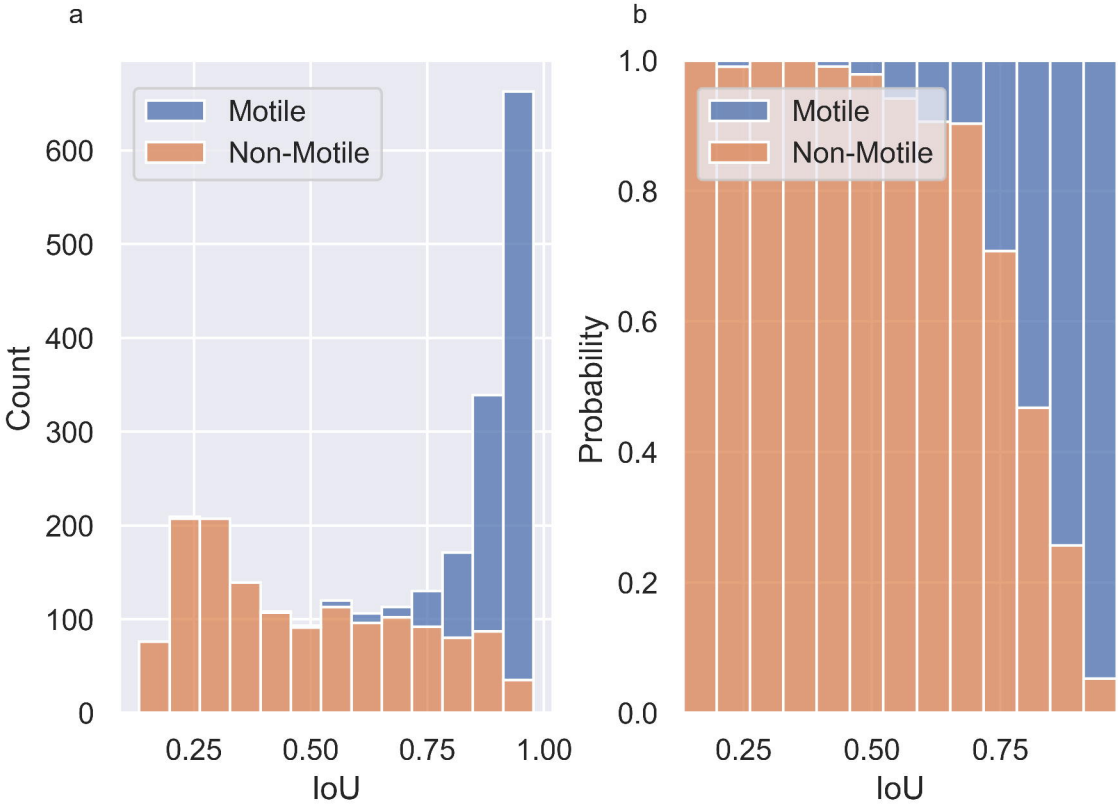


c




Manual
Wiggle Index
WF-NTP
Mask R-CNN
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Motility rate
0
25
50
75
100
Motility group


Manual
WF-NTP
Mask R-CNN
Worm count
90
80
70
60
50
40
30
20
0
25
50
75
100
Motility group


a
Motile
Non-Motile
Count
0
100
200
300
400
500
600
0.25
0.50
0.75
1.00
IoU
b
Motile
Non-Motile
Probability
0.0
0.2
0.4
0.6
0.8
1.0
0.25
0.50
0.75
IoU