scientific reports

Check for updates

OPEN

Machine learning-based analysis of microfluidic device immobilized *C. elegans* for automated developmental toxicity testing

Andrew DuPlissis[1], Abhishri Medewar[1], Evan Hegarty[1], Adam Laing[1], Amber Shen[1], Sebastian Gomez[1], Sudip Mondal[1]✉ & Adela Ben-Yakar[1,2,3]✉

Developmental toxicity (DevTox) tests evaluate the adverse effects of chemical exposures on an organism's development. Although current testing primarily relies on large mammalian models, the emergence of new approach methodologies (NAMs) is encouraging industries and regulatory agencies to evaluate novel assays. *C. elegans* have emerged as NAMs for rapid toxicity testing because of its biological relevance and suitability to high throughput studies. However, current low-resolution and labor-intensive methodologies prohibit its application for sub-lethal DevTox studies at high throughputs. With the recent advent of the large-scale microfluidic device, vivoChip, we can now rapidly collect 3D high-resolution images of ~ 1000 *C. elegans* from 24 different populations. While data collection is rapid, analyzing thousands of images remains time-consuming. To address this challenge, we developed a machine-learning (ML)-based image analysis platform using a 2.5D U-Net architecture (vivoBodySeg) that accurately segments *C. elegans* in images obtained from vivoChip devices, achieving a Dice score of 97.80%. vivoBodySeg processes 36 GB data per device, phenotyping multiple body parameters within 35 min on a desktop PC. This analysis is ~ 140 × faster than the manual analysis. This ML approach delivers highly reproducible DevTox parameters (4–8% CV) to assess the toxicity of chemicals with high statistical power.

Keywords U-Net, Few-shot learning, *C. elegans*, Developmental toxicity, Microfluidics, High-throughput screening

Traditional developmental toxicity (DevTox) studies have relied on mammalian models such as mice, rats, and rabbits to study adverse effects on their development when exposed to chemicals. Scientific and technological advances have led to the development of new approach methodologies (NAMs), such as in silico, in vitro, or small model organisms (*C. elegans*, *Daphnia*, zebrafish embryos), reducing the use of vertebrates. Among model organisms, *C. elegans*, is the only model that combines several advantages, including small body size, ease of culture, homologous genes to humans, conserved xenobiotic pathways, several organ systems, etc., making it suitable for high-throughput toxicology screening platforms[1–6]. Developmental parameters from *C. elegans* models have demonstrated high concordance with mammalian toxicity endpoints and outperformed other small model organism species in some situations[7–15].

Although *C. elegans* have been used as a model organism in major discoveries in the fields of neurodegeneration, aging, and toxicology assessments[11,15–24], these studies either had low throughput or used gross phenotypes. Specifically, the automated studies were performed by either taking low-resolution images of *C. elegans* from well plates[25–28] or 1D scattering signal using a flow cytometer to extract developmental parameters[8,29–31]. Both technologies rely on using anesthetics to reduce worm movement, which is known to have adverse effects on worms, causing their bodies to shrink or curl, eventually introducing errors in the body length measurements[28,31]. Flow cytometers can provide such data from a large number of worms, however, with high variability (coefficient of variance, CV > 20%)[32]. Plate-based assays also demonstrate poor statistical power due to the small number of worms used per well to avoid overlap and thus simplify image analysis[28,32–34].

[1]vivoVerse, LLC, Austin, TX 78731, USA. [2]Department of Mechanical Engineering, The University of Texas at Austin, Austin, TX 78712, USA. [3]Department of Biomedical Engineering, The University of Texas at Austin, Austin, TX 78712, USA. ✉email: sudip.mondal@vivoverse.com; ben-yakar@mail.utexas.edu

We previously developed a microfluidic-based immobilization method to collect high-resolution images of many *C. elegans* nematodes without using anesthetics, while also addressing the challenge of random orientation[35–38]. This microfluidic device, known as the vivoChip, facilitates the rapid immobilization of thousands of *C. elegans* from 24 different populations in parallel microchannels. Specifically, the vivoChip enables the collection of high-resolution images from 40 microfluidic channels for each of the 24 populations, with each worm body spanning over thousands of pixels (5056 × 354 pixels per channel). By imaging along the length of the microchannels, we can capture the entire body of 8 *C. elegans* in a single field of view (FOV), eliminating the need for image stitching. The vivoChip can immobilize worms across a wide range of body sizes, which is particularly valuable given the adverse effects on the development and health caused by the exposure to high concentrations of toxic chemicals. The variation of body sizes results in differing contrast levels across images, introducing additional challenges for image analysis.

DevTox analysis of such high-resolution data generated by vivoChip devices is highly time-consuming and would greatly benefit from automation, enabling consistent and efficient processing of large datasets in a shorter timeframe. Several machine learning (ML)-based vision systems have been developed for the segmentation of *C. elegans* bodies in images obtained by plate readers, facilitating the analysis of growth and behavioral parameters[39,40]. However, these systems, face difficulties in multi-object identification in dense settings where overlapping specimens are common[41]. The vivoChip overcomes this concern by immobilizing up to 40 worms side by side within 40 parallel microchannels per well, efficiently eliminating overlap. Automated segmentation of *C. elegans* bodies in vivoChip images requires processing 3D image stacks to accurately detect each worm and precisely delineate its boundaries. This task is complicated by the need to extract the necessary features from high-frequency content that spans the entire volume.

In this paper, we present vivoBodySeg, an ML-based model to automatically segment *C. elegans* bodies immobilized inside the vivoChip devices and, thus, streamline accurate and multiparametric DevTox studies. To create extensive and balanced ground truth data, we developed a user-friendly toolbox that enables the visualization, classification, and segmentation of vivoChip-generated *C. elegans* images. vivoBodySeg utilizes a 2.5D U-Net architecture with an attention mechanism at the bottleneck that is trained for classification and semantic segmentation of the *C. elegans* body. The base model achieved highly accurate segmentation with a Dice score of 97.8% across a population of adult *C. elegans* exposed to different concentrations of toxic chemicals. The predicted segmentations are indistinguishable from humans while taking ~ 140 × less time. Further, we demonstrate that by performing few-shot learning (FSL) with only a small number of samples with severe developmental delays and four hours of training, we can improve the Dice score for this population from 94.9% to 96.9%. This automated image analysis pipeline reduces human error, eliminates user bias, and achieves repeatable high-accuracy analysis of DevTox parameters for methylmercury treated worms.

Methods

C. elegans culture and chemical treatment

C. elegans culture and chemical treatment are performed using N2 strain (*Caenorhabditis* Natural Diversity Resource—CaeNDR) according to previously published protocols[35,36,42]. Briefly, eggs are collected from gravid *C. elegans* adults by sodium hypochlorite treatment. The eggs are hatched and allowed to develop into synchronized larvae 1 (L1) stage worms overnight. Approximately 100 L1s are placed in each well (population) of a 24-well plate with *E. coli* (HB101 strain) food in S medium. The L1 larvae are treated with methylmercury (methylmercury (II) hydroxide, CAS# 1184-57-2, Sigma) at 18 concentrations (0.5–9 µM), a known developmental toxicant[23,43–46], in conventional 24-well plastic plates. The concentrations of methylmercury were selected based on our preliminary experiments, which were inspired by published results on lethal concentrations[23,46]. The remaining 6 wells among the 24 wells were treated with 0.2% dimethyl sulfoxide (DMSO, solvent controls). The plates are incubated at 20 °C for 72 h until the worms in the control wells reach the day 1 (D1) stage of adulthood. To test technical and biological variabilities, we repeated experiments with the same batch of L1s (used on 2 experiments on the same day) and different batches of L1s (used on 3 multiple days). In total we repeated the same experiment five times using five vivoChip-24x devices on three different days.

High-resolution imaging of *C. elegans*

To acquire high-resolution images of worms, we load 24 worm populations (~ 100 per population) growing in a standard 24-well plates into a 24-well microfluidic device (vivoChip-24x, vivoVerse). Underneath each well of the vivoChip-24x device, there are 40 parallel, gently tapering 3 mm long microfluidic trapping channels to trap *C. elegans* (Fig. 1a–c). The vivoChip-24x device contains a total of 960 trapping channels.

To enable immobilization of worms of varying sizes, we designed two types of vivoChip-24 devices: a 3-layer (3L) and 4-layer (4L) microfluidic devices with distinct channel geometries (Supplementary Fig. 1a, b). The entrance of the 3-layer vivoChip-24x device (vivoChip-24x-3L) has a cross-section of 98 µm × 102 µm (width x height), tapering down to 24 µm × 40 µm (width x height) (Supplementary Fig. 1c). The device is suitable for immobilizing adult to young adult worms with a minimum 2–6 embryos in their bodies. However, younger worms tend to slip through the channels of the 3L chip.

To capture smaller worms, particularly during range-finding experiments with a wide range of chemical concentrations, we designed a second chip with 4 layers (vivoChip-24x-4L). The smaller cross-section of 4L-chip (14 µm × 17 µm) at the end of the immobilization channels facilitates the trapping worms as small as young larval 4 (L4) stage (Supplementary Fig. 1d). This capability is crucial in characterizing developmental parameters in severely intoxicated worms exposed to high chemical concentrations. Such data from severely retarded worms are essential in fitting the Hill-function and determining LOAEL, EC_{10}, and EC_{50} values with a low confidence interval.

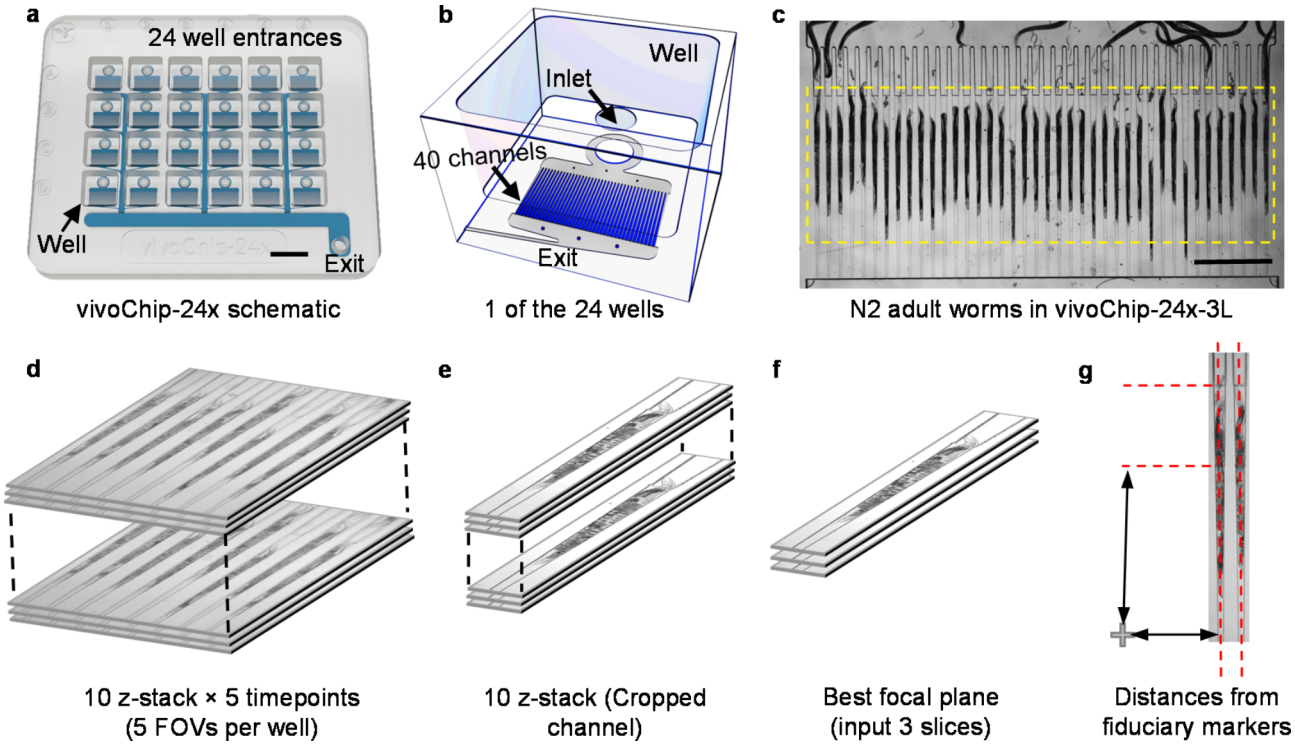

Fig. 1. High-resolution *C. elegans* body images are acquired using a vivoChip-24x platform. (**a**) Schematic of vivoChip-24x technology designed to immobilize *C. elegans* and capture high-resolution images from 24 different populations with up to 40 worms per population. Scale bar = 9 mm. (**b**) Detailed schematic of a single well from the 24-well device. Each well features an opening at the top and an inlet on the bottom, leading to 40 parallel immobilization channels beneath it. Worms are loaded into the well, flow through the inlet, and are immobilized within the 40 channels by the pressure of the fluid flow directed toward the exit. (**c**) Brightfield image of 40 adult *C. elegans* immobilized inside the microfluidic channels of the vivoChip-24x-3L device. Scale bar = 1 mm. (**d**) Graphic demonstrating a z-stack of 10 images collected for each FOV. Each FOV covers 8 channels using a 10×, 0.4 NA objective. For each FOV, we collect a total of 50 images over 5 time points, with 10 z-slices (6 µm z-step size) captured at each time point (1-s interval). (**e**) We crop each channel for the first time point. (**f**) We identify the best focal plane for analysis and use 3 slices, comprising the focal plane, one slice above, and one slice below to train the network. (**g**) The location of the worm within the channel is used to determine the channel height by referencing its position with respect to the fiduciary cross mark present within each well of the vivoChip-24x.

We redesigned our previous gasket system to robustly seal the vivoChip-24x devices and provide fluidic connections to all wells[35,36]. A single input in the gasket applies fluidic pressure to push the worms into individual channels (one *C. elegans* per channel) using intermittent ON/OFF fluidic pressure cycles. An automated fluid control system (vivoCube+, vivoVerse) is used to apply the cyclic air pressure (0 and 3.5 PSI) to a liquid reservoir containing M9 buffer. The pressure in the reservoir drives the buffer through the gasket assembly, which in turn forces the worms to flow from the well into the microchannels. Once all channels fill up with worms as they are immobilized inside the parallel, narrowing channels, a constant fluid pressure holds them still for performing blur-free imaging.

Automated high-resolution imaging is then performed on all 960 channels to collect volumetric (z-stack) time-lapse brightfield and z-stack fluorescence images. This process takes 30 min to acquire 7,200 images (216 GB of data) using a customized automated microscope (IX73, Evident) equipped with a high-quantum efficiency, fast, and large-area camera (Iris 15, Teledyne). All 40 channels underneath each well are imaged in five FOVs using 10×, 0.4NA objective (UPLXAPO10X, Evident). The 10× objective enables imaging a 2.15 × 1.26 mm^2 FOV, capturing the full lengths of adult worms in 8 channels a pixel resolution of 0.425 µm. This high-resolution is necessary to precisely analyze the body dimensions and autofluorescence intensity distribution.

The entire worm volume is captured in 10 z-slices at 6 µm steps. Based on the Nyquist theorem, ideal imaging for 0.4NA objective would require z-slice intervals of 3.25 µm (considering an axial resolution of 6.5 µm). However, two practical constraints influenced the selection of 6 µm step size: the maximum data readout speed (frames per second) and the minimum volumetric imaging speed (volumes per second) required to capture dynamic phenotypes. In particular, the twitching frequency of embryos inside eggshells necessitated high temporal resolution to ensure accurate capture of these rapid movements. The z-slices are centered around the best focal plane of a fiduciary marker. To balance between temporal imaging accuracy and data size, we selected

to acquire five time-lapse 3D hyperstack images at 1-s intervals (Fig. 1d). Following the time-lapse brightfield imaging, a single z-stack of autofluorescence images is also acquired using a GFP filter and the same objective.

Pre-processing of images

Images are automatically uploaded to a local server for processing and analysis. Each channel within the FOV capturing 8 channels is then cropped into individual hyperstacks by clipping the full hyperstack into eight 150 µm-wide sections (Fig. 1e). The cropping is centered along the centerline of each channel, which is determined using their predefined locations relative to a fiduciary marker (Fig. 1g).

Manual segmentation for ground truth data

Manual annotation of each volumetric image of individual *C. elegans* is done using an in-house graphical user interface (GUI) toolbox, vivoSegmenter (Supplementary Figs. 2 and 3). This GUI allows users to scroll over multiple z-plane images and time points for each cropped channel. Since the morphology of a worm's body inside the channel does not change substantially between time points, users segment only the first time point. In the GUI, users first classify each channel by selecting one of three classes: no worm (empty channel), partial worm (a partially visible worm is present inside the channel), or full worm (a full worm body is present in the channel). For channels classified as full worms, users perform segmentation by clicking multiple points along the worm body in different z-slices in the cropped channel images. The GUI registers the coordinates (x, y, and z) of each point clicked by the user. Once all the points are entered and connected to form a closed shape, a polygon is created with these points as vertices. Most body segmentations require users to examine 3–5 z-slices for accurate delineation. Although depth information is essential for accurate boundary delineation, wide-field microscopy lacks the optical sectioning needed for effective body segmentation over depth. We, therefore, collapse all points in a segmentation polygon, namely our ground truth, into a single 2D binary image. Finally, each channel is assigned a class label (full, partial, or empty) and a single polygon is generated only for channels labeled as full worms.

Data pre-processing for ML analysis

Our network is designed to work with either a full z-stack image volume or a subset of the data volume. Specifically, while we collect 10 focal planes per time point during imaging, we have found that successfully solving most vision-related problems is optimally done with a subset of a z-stack volume centered around a relevant focal plane (Fig. 1f). Finding this central focal plane of a cropped channel is done by applying the Laplace transform to each z-slice and selecting the one with the highest frequency components[47,48]. We note that the mode (most frequently occurring value) of the z-values from all the vertices clicked by users during worm segmentation matches the best focal plane identified by the Laplace transform. After identifying the central focal plane, N planes on either side of the central plane are selected and stacked along the channel dimension ($H \times W$). The image stacking forms a 2.5D tensor, together represented as a $(2N + 1) \times H \times W$ tensor. For each cropped channel, the image height (H) is fixed to 5,056 pixels, while the width (W) varies between 340 and 360 pixels, depending on the cropping process. To standardize the size of all 960 channels and make the size suitable for our network, images are padded on both sides to obtain a fixed width of 384 pixels.

vivoBodySeg architecture for *C. elegans* body analysis

We propose a 2.5D U-Net for the vivoBodySeg model, incorporating an attention mechanism at the bottleneck to enable for the classification and semantic segmentation of *C. elegans*[49]. The proposed architecture consists of three sub-networks: a fully convolutional encoder, a bottleneck layer utilizing a small vision transformer (ViT), and a fully convolutional decoder (Fig. 2a). Additionally, a classifier connected to the bottleneck predicts the class label for the input image, assigning it to one of the three classes: full worm, partial worm, or no worm.

Our model is trained using supervised learning on a large, diverse dataset. During training, the model iteratively improves its ability to accurately segment and classify new images of *C. elegans* by minimizing the loss function. Once trained, the encoder arm extracts features from input images, which are then processed by the bottleneck layer to generate image-wise classification and by the decoder to generate pixel-wise segmentation masks. This dual-output design allows the network to effectively perform both segmentation and classification tasks, leveraging learned features for precise and efficient analysis.

A traditional 2D U-Net processes individual image slices, making it suitable for datasets where each slice can be analyzed independently[50]. However, it does not account for volumetric or spatial information across slices, limiting its ability to leverage depth information. On the other hand, a 3D U-Net extends convolution and pooling operations in all three dimensions, allowing the model to analyze entire 3D blocks simultaneously and learn from features present in 3D. While depth information is important for accurately delineating boundaries, wide-field microscopy lacks the optical sectioning required to produce fine-grained segmentation over depth. Additionally, 3D convolution operations are more computationally expensive and require significantly more memory, making them less practical for large datasets or resource-constrained environments.

Given these considerations, we selected a 2.5D U-Net model architecture, inspired by previous work on N.5D CNN[51]. An N.5D CNN refers to a convolution neural network (CNN) that processes N + 1 dimensions but only N of them processed through in convolutional operations. The "+1" dimension is stacked over the channel/feature dimension similar to how RGB information is treated in standard images. In our context, the + 1 dimension refers to the 'z' dimension (z slices) of the collected 3D volumes along the channel height. Each layer of our vivoBodySeg network employs the same residual convolutional layer design in both the encoder and decoder components. This layer structure follows the design principles outlined by He *et. al.* in their ResNet architecture (Fig. 2b)[52]. Following the standard practices for semantic segmentation, the final decoder layer

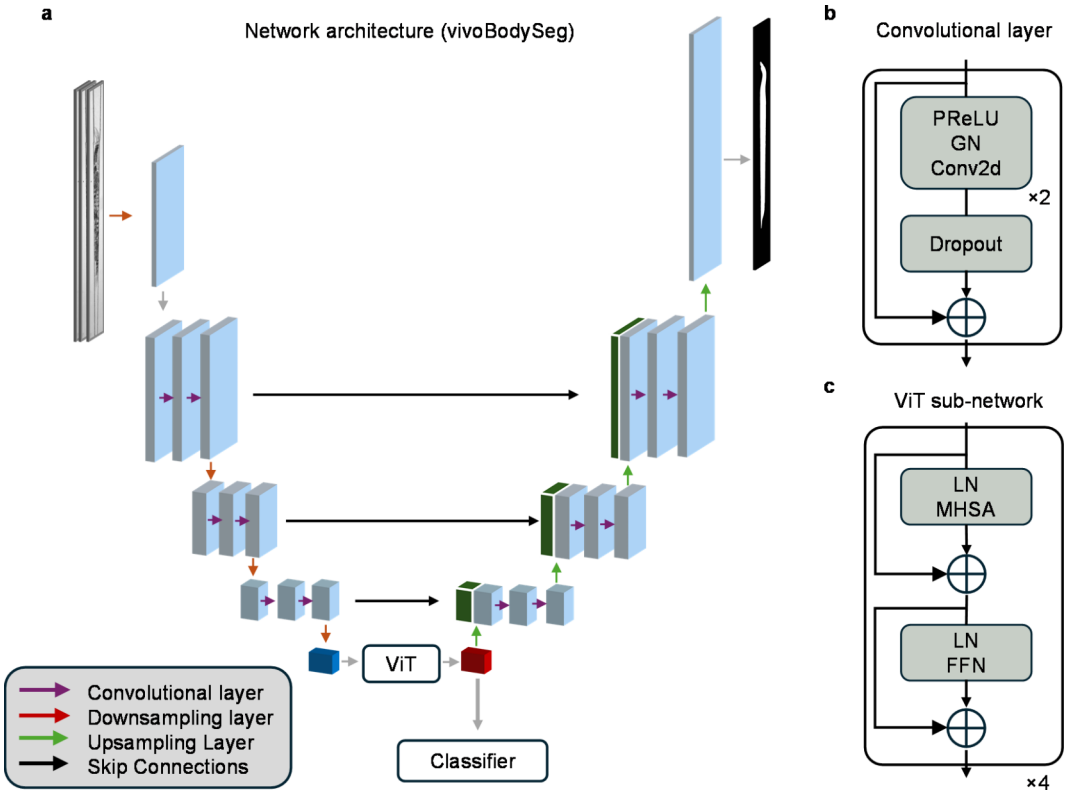

Fig. 2. The U-Net architecture of vivoBodySeg for automated analysis of *C. elegans* body. (**a**) Schematic of our 4-layer U-Net architecture. (**b**) Overview of our basic convolutional layer used in both the encoder and decoder components. (**c**) Overview of the vision transformer (ViT) integrated into the bottleneck layer.

is followed by a series of linear layers and a softmax function to produce a soft segmentation output, enabling accurate pixel-wise segmentation across the input image.

To augment the standard convolutional approach, we introduce a ViT at the bottleneck to enable efficient, long-range communication between embedded voxels (Fig. 2c). While standard images collected by non-scientific cameras may span hundreds of pixels, our images with Iris 15 camera span over 5,056 pixels along the worm length. By replacing the standard convolutional bottleneck with a ViT, our goal is to ease the classification task as relevant image patches span thousands of pixels. The output of our ViT is routed to two separate sub-networks: the convolutional decoder and the classification subnetwork. The classifier utilizes a pooled attention mechanism, introduced by Lee *et. al.* involving a single seed vector followed by a series of linear layers[53]. This design produces a vector with elements that serve as inputs for our classification task. This approach enables more robust handling of large-scale image features and improves performance of the classification task.

Unlike a standard ViT that generates a tokenized version of images through a linear embedding of voxels[54] or by utilizing a secondary generative model such as a VQ-VAE[55], the encoder of our U-Net serves as our embedding mechanism. At the bottleneck, the 4D tensor is rearranged into a tokenized format,

$$X_e = Flatten\left(F_{enc}\left(X\right)\right) \in \mathbb{R}^{B,\frac{H}{S}\times\frac{W}{S},C} \tag{1}$$

where s is equal to 2^{Layers} and $c = 256$. After this reshaping, the tensor is combined with learnable positional encodings initialized as a series of sinusoidal functions. The tokenized data is then processed as a sequence by a small 4-layer ViT. Adhering to the standard design introduced by Vaswani *et. al.* for natural language processing[56], the data first undergoes normalization before being routed through a multi-head self-attention (MHSA) block. Following a residual connection, the normalized data is passed to a feed-forward network (FFN), where we maintain the standard feature expansion with respect to the FFN hidden dimension e.g., $|C_{FFN}| = 4 * |C_{MHSA}|$. The input and output of the self-attention block are also connected by a residual connection, ensuring stability and efficient gradient flow. Our code and network configuration files for the vivoBodySeg framework are available for academic use upon request.

Network training, validation, and testing for *C. elegans* images

The training dataset includes 3,637 channels acquired using the vivoChip-24x-3L devices from experiments conducted under different treatment conditions. Among these, 81% contain channels where the entire body of *C. elegans* is fully visible (full worm), 14% contain channels with only partially visible *C. elegans* bodies within

(partial worm), and 5% contain empty channels (no worm). We split the data into training, validation, and testing sets in an 8:1:1 ratio, with consistent class distribution percentages across all subsets.

We utilize horizontal and vertical flips, small rotations, and contrast adjustments to enhance the diversity of the training dataset. Each training step randomly selects a mini batch of 32 images and processes it in a single forward pass. We use an AdamW optimizer with an initial learning rate of 2×10^{-4} and weight decay of 1×10^{-2}[57]. During training, the learning rate is scheduled according to cosine annealing with warm restarts ($W_0 = 10, F = 2$) over 1,200 total epochs[58]. Network weights are updated according to focal loss functions, equally weighted for segmentation and image classification losses[59]. Only channels labeled as "full worms" are sent to the decoder for segmentation training with the corresponding weight updates applied. All vivoBodySeg networks are trained on a computer with 128 GB of RAM and an NVIDIA A6000 GPU with 48 GB of VRAM.

Post-processing and inference to find *C. elegans* body parameters

During inference or testing, we only consider those channels with full worms and apply a streamlined post-processing procedure to clean the data and extract relevant endpoints. During post-processing, a threshold of 0.50 is applied to the model's output to form a binary mask representing the *C. elegans* body. After this step, the connected component analysis identifies large binary objects corresponding to the worm bodies (detects L4 to adult stage worms) and removes smaller objects outside this binary mask, such as laid eggs, small larvae, debris, or other artifacts within the channel. The binary mask is then used to estimate three body parameters: length, area, and volume. The body length is retrieved from the longest spanning tree in the skeleton of the binary mask. The binary object provides the area of the *C. elegans* body. The total volume is estimated by taking into account the known height of each pixel inside the predicted *C. elegans* mask.

Evaluation metrics for model performance

When evaluating our network, we report several metrics to quantify the overall segmentation quality. To understand the network performance, we use the Dice score to track validation progress and quantify how the model performs in a test setting. The Dice score is calculated as.

$$Dice\ score(\%) = (2 \times |A \cap B|)\,/(|A| + |B|) \tag{2}$$

where A represents the set of predicted pixels and B represents the set of ground truth pixels. The Dice score ranges from 0% (no overlap) to 100% (perfect overlap), providing a measure of the similarity between the predicted segmentation and the ground truth. This metric allows us to assess how well the model delineates the object of interest in comparison to human annotations. To statistically compare the performance of different models, we use the Wilcoxon signed rank sum test to determine if there are significant differences in the average Dice score ranks between different models. We use our post-processed data to further elucidate the model accuracy by reporting the ratio of the predicted skeleton length over the ground truth skeleton length. We also estimate the volume ratio (predicted volume to the ground truth volume) and classification accuracy using the weighted F1 score. To compare the model performance with human scorers, we calculate inter-individual Dice scores between multiple scientists and compare these with the Dice score estimated from different models.

Autofluorescence analysis

We measure the autofluorescence signal within the predicted body mask using the z-stack fluorescence images captured with a GFP filter set. First, we create a maximum-intensity projection image from all 10 z-stacks. Using the control wells with 0.2% DMSO treatment, we determine a threshold intensity above which the brightest 5% of the pixels lie. These 5% pixels correspond to the granules of lysosomes in the worm gut, which are major contributors to the increase in autofluorescence under stressors. We calculate the average pixel intensity considering the pixels with an intensity above this threshold and within the predicted body mask for all worms. We use the average autofluorescence value per unit body length and per unit body area to identify concentration-dependent responses to a chemical treatment.

Statistical analysis of concentration-dependent body parameters

For DevTox testing, body parameters and autofluorescence signals are calculated for each full worm across all 40 channels per population (well). Measurements of body lengths and autofluorescence signals that significantly deviate from the median of each well are identified using Tukey fences (1.5 × the interquartile range). Worms with values outside these fences are excluded from further analysis. After filtering, the remaining worms are used to calculate the well mean (μ) and standard deviation (σ) for body length, area, and volume. The data is presented as the mean ± standard error of the mean (μ ± SEM) from multiple replicates. Additionally, the coefficients of variance ($CV = \sigma/\mu$) are calculated from all control wells. To characterize assay variability, we measure the CV from solvent controls (0.2% DMSO) across five vivoChip experiments with methylmercury.

The average values for each body parameter and autofluorescence signal are plotted for different concentrations. The effective concentration (EC_{10}) values, representing a 10% change in the parameter, are determined by fitting the dataset to a 4-parameter, variable slope Hill function using the "Find ECanything" nonlinear fit function of GraphPad Prism. For body length, area, and volume, the bottom (floor) of the nonlinear function is constrained to zero. For autofluorescence signal, the top (ceiling) of the nonlinear function is left unconstrained. The EC_{10} values are reported alongside their ± 95% confidence interval (CI) for each parameter.

To calculate the lowest observable adverse effect level (LOAEL) for each phenotype, we perform a normality test (Shapiro–Wilk test) and identify the lowest concentration at which the phenotype significantly deviates (p-value < 0.05) from the baseline value of the control population. This is assessed using Welch ANOVA with post hoc Dunnett's T3 multiple comparison tests.

Model name	Dice score (%)	Length ratio	Volume ratio	Weighted F1 score
vivoBodySeg-2D (no attention)	96.62 ± 0.12	0.987 ± 0.002	1.026 ± 0.003	0.986
vivoBodySeg-2D, Att (with attention)	97.24 ± 0.10	0.988 ± 0.001	0.999 ± 0.003	1.000
vivoBodySeg-2.5D, Att (2.5D, with attention)	97.80 ± 0.08	0.991 ± 0.001	1.008 ± 0.002	0.995

Table 1. Test results of the proposed U-Net models using 362 images.

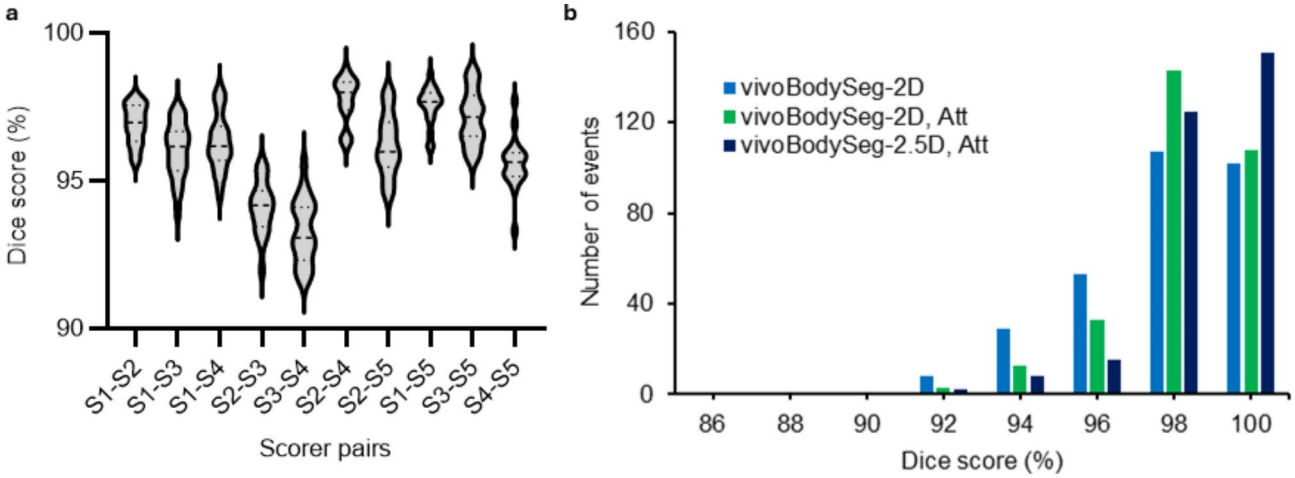

Fig. 3. Comparison of Dice scores for body segmentation by multiple scorers and different vivoBodySeg models. (**a**) Dice scores for all possible pairs of five individual scorers for a subset of 20 channels. (**b**) Bar plot showing Dice scores ranging from 86 to 100% ($n = 301$ *C. elegans* samples) for three U-Net models.

Results

A 2.5D U-Net vivoBodySeg model segments *C. elegans* body

Making use of our training corpus, we trained a base model (U-Net with no attention, vivoBodySeg-2D) alongside an improved U-Net model making use of the attention bottleneck (vivoBodySeg-2D, Att) and the 2.5D U-Net with an attention bottleneck (vivoBodySeg-2.5D, Att). All three models were trained and validated using our in-house computational power (Supplementary Fig. 4). We tested all three models on an unseen $M = 362$ test samples. We calculated segmentation accuracy using the Dice score, length ratio, volume ratio, and classification accuracy for all three models (Table 1). Using additional ablation experiments, we found that using $N = 1$(3 z-stack images) is the optimal configuration for this problem (Table 1). Expanding beyond 3 z-slices did not improve performance (Supplementary Table 1), making this configuration both efficient and effective. We note the length and volume ratios are highly correlated with predicted body segmentation. We found vivoBodySeg-2.5D, Att to be highly performant, with an average Dice score of 97.80% ± 0.08%, a length ratio of 0.991 ± 0.001, and a volume ratio of 1.008 ± 0.002. The vivoBodySeg-2.5D, Att model classified all 362 test images into three groups (301 full worms, 48 partial worms, and 13 no worms) with a weighted F1 score of 0.995. The model could detect *C. elegans* bodies (in channels with full worms) with high accuracy and ignore foreign particles with high confidence (Supplementary Fig. 5). The output from the vivoBodySeg-2.5D, Att model was noticeably better than the vivoBodySeg-2D output due to improved detection of worm bodies, especially the tip region of the head and the tail (Supplementary Fig. 6).

To evaluate our model's performance relative to manual segmentations, we compared the models' results with those of human scorers. First, we assessed inter-individual variability by comparing segmentation results from 5 scientists (scorers), each of whom segmented the same 20 channels of *C. elegans*. We then calculated the Dice scores for all the possible pairs of scorers, resulting in 200 total comparisons. The average Dice scores among 5 scorers was 96.10% ± 0.12% ($n = 200$) (Fig. 3a).

To evaluate the performance of different models, we then calculated the Dice scores for all 3 models using 301 body segmentations (under the full worm classification) from two scorers representing the test sample set. vivoBodySeg-2.5D, Att had 9% of the samples with a Dice score below the average value of 96.10% compared to 31% for vivoBodySeg-2D (Fig. 3b and Supplementary Fig. 7a–b).

Few-shot learning to improve the body detection for smaller worms immobilized in vivoChip-24x-4L device

New chemicals are often tested across a wide range of concentrations in range-finding experiments to identify their lethal concentration. In such assays, several worm populations may become developmentally arrested or severely retarded, resulting in much smaller worms. In addition, highly potent chemical conditions can create heterogeneous worm populations with variable body sizes. To study such conditions, we treated *C. elegans*

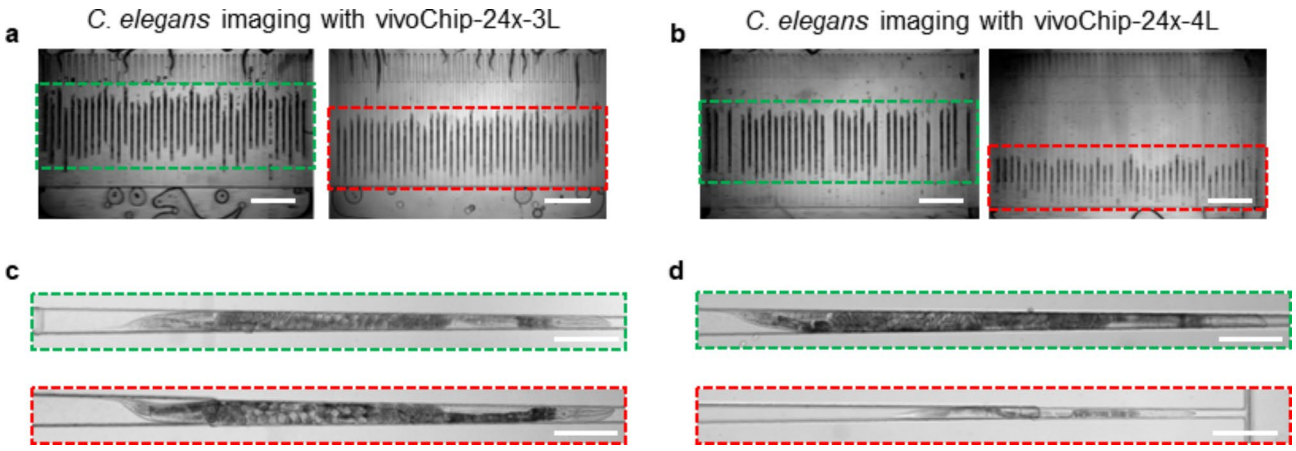

Fig. 4. Images of *C. elegans* with developmental defects captured using two vivoChip-24x device designs. (**a**) Images of 40 trapping channels from the 3-layer chip (vivoChip-24x-3L) with immobilized D1 adult worms from a control population boxed in green (left panel) and a young adult worm population that was treated with 4 µM methylmercury in 0.2% DMSO boxed in red (right panel). (**b**) Images of 40 trapping channels from a 4-layer chip (vivoChip-24x-4L) with immobilized D1 adult worms from a control population boxed in green (left panel) and a young L4 stage worm population that was treated with 8 µM methylmercury in 0.2% DMSO (right panel). (**c**, **d**) High-magnification images of single control worms (green boxes) and methylmercury-treated worms (red boxes) from vivoChip-24x-3L (**c**) and vivoChip-24x-4L devices (**d**). Scale bars = 1 mm (a, b) and 200 µm (**c**, **d**).

	vivoChip-24x-3L device		vivoChip-24x-4L device	
Number of images (%)	Dice score (%)	Weighted F1 score	Dice score (%)	Weighted F1 score
0 (0.0%)	97.80 ± 0.08	0.995	94.90 ± 0.47	0.860
32 (6.3%)	96.57 ± 0.14	0.938	95.90 ± 0.31	0.986
64 (12.5%)	96.95 ± 0.12	0.971	95.93 ± 0.30	0.986
128 (25.0%)	97.28 ± 0.11	0.992	96.52 ± 0.32	0.986
256 (50.0%)	97.32 ± 0.10	0.995	96.90 ± 0.25	0.986
512 (100.0%)	97.39 ± 0.11	0.995	96.91 ± 0.19	0.986

Table 2. Dice score for the few-shot learning of vivoBodySeg-2.5D, Att model with different amounts of training data (6.3%–100% of 512 images).

populations with high concentrations of the reference toxicant methylmercury[23,46]. Under these conditions, *C. elegans* are expected to show delayed development and thus significantly smaller body sizes and lower image contrasts.

Notably, the images obtained using the vivoChip-24x-4L devices are significantly more distinct than those collected with vivoChip-24x-3L devices as the 4L experiments involve phenotypically different worm populations. Beyond the slightly different microfluidic environments, the visual differences between larvae and adult *C. elegans* are particularly striking. Developmentally retarded worms (young L4) are transparent, lack eggs, and have fewer gut granules compared to normally grown adult *C. elegans* (Fig. 4a–d, Supplementary Fig. 8). Using vivoBodySeg-2.5D, Att, we measured a zero-shot Dice score of 94.90% ± 0.47% on images obtained with the vivoChip-24x-4L device (Table 2). Since vivoBodySeg-2.5D, Att performed exceptionally well on images acquired with 3L devices (Dice score of 97.80% ± 0.08%), we wanted to understand few-shot learning (FSL) performance on phenotypically different worm populations imaged using the 4L devices. FSL refers to a machine learning approach that enables a model to learn effectively with only a small amount of labeled data[60,61]. This is particularly useful in scenarios where acquiring large amounts of labeled data is challenging, such as in biological image segmentation tasks where different phenotypes may require extensive data annotation.

We segmented additional channels with worms captured using the vivoChip-24x-4L devices, producing an ideal training corpus with masks representing worms ranging from early L4 up to D1 adult worms (Supplementary Fig. 9). This dataset provided sufficiently diverse range of *C. elegans* phenotypes, including variations in body dimensions and number of embryos *in-utero*. For the FSL test, we used 512 and 107 channel images from vivoChip-24x-4L for training and testing, respectively. To avoid catastrophic forgetting, an equal number of worms from the previous 3L dataset was mixed with the new 4L data.

To assess the impact of dataset size on the fine-tuning process, we fine-tuned the network in five steps with varying amounts of training data. Specifically, we performed the training with 32 (6.3%), 64 (12.5%), 128 (25.0%),

256 (50.0%), and 512 (100.0%) samples from the vivoChip-24x-4L devices. Each split was trained with a batch size of 32 over 250 epochs using a cosine annealed one-cycle learning rate scheduler, with a maximum learning rate of 10^{-6}. All trained models were tested with $M = 362$ images from 3L chips and $M = 107$ images from 4L chips. To understand model improvement, we calculated the Dice scores and classification accuracies using weighted F1 scores for images acquired with each device independently (Table 2 and Supplementary Figs. 10a–b). Although the classification saturated rapidly (upon training with 32 samples), the segmentation accuracy converged slowly as we increased the number of training data.

While phenotypic differences between the two populations reduced the performance of our baseline model (vivoBodySeg-2.5D, Att), we successfully fine-tuned its performance in detecting worm bodies across a wide range of worm populations with just four hours of additional training (Fig. 5a–f). Ideally, the training corpus would represent the full distribution of images associated with any test set. However, practical limitations pertaining to ground truth creation make this challenging. We could overcome these limitations using our fine-tuning strategy that allowed us to effectively process new and highly disparate data classes using hundreds rather than thousands of examples. The final vivoBodySeg-2.5D, Att model, trained with 512 images, detected the adult worms from vivoChip-24x-3L devices with slightly lower accuracy (Dice score 97.39% ± 0.11% compared to 97.80% ± 0.08%). However, its performance on the new class of data improved (Dice score 96.91% ± 0.19% compared to 94.90% ± 0.47% from the baseline model). The final vivoBodySeg-2.5D, Att (100.0%) model detected adult worms in vivoChip-24x-3L images with accuracy comparable to the baseline model before fine-tuning (Supplementary Fig. 11). Additionally, the final model achieved a weighted F1 score of 0.986, a significant improvement over the baseline model's score of 0.860 prior to implementing the FSL approach. Additional examples of full worms from both 3L and 4L devices are shown in Supplementary Figs. 12 and 13. We manually inspected specific channels that were classified as no worm or partial worm, but no masks were predicted (Supplementary Fig. 14).

Studying DevTox in *C. elegans* with methylmercury exposure

We exposed age-synchronized *C. elegans* larvae to methylmercury in well plates and imaged them using the vivoChip-24x-4L device (Fig. 6a). We analyzed all 4,800 microfluidic channels from 5 independent experiments using the final vivoBodySeg model (fine-tuned with few-shot learning on 512 images). This model was used to automatically identify channels with *C. elegans* and estimate the body parameters (length, area, and volume). The inference code with this model could analyze a full chip with 1,200 brightfield images in 35 min.

The assay results for all body parameters across the five chip experiments were highly consistent. The coefficient of variance (CV) for DMSO controls across the replicates was 3.7% for the body length, 7.9% for

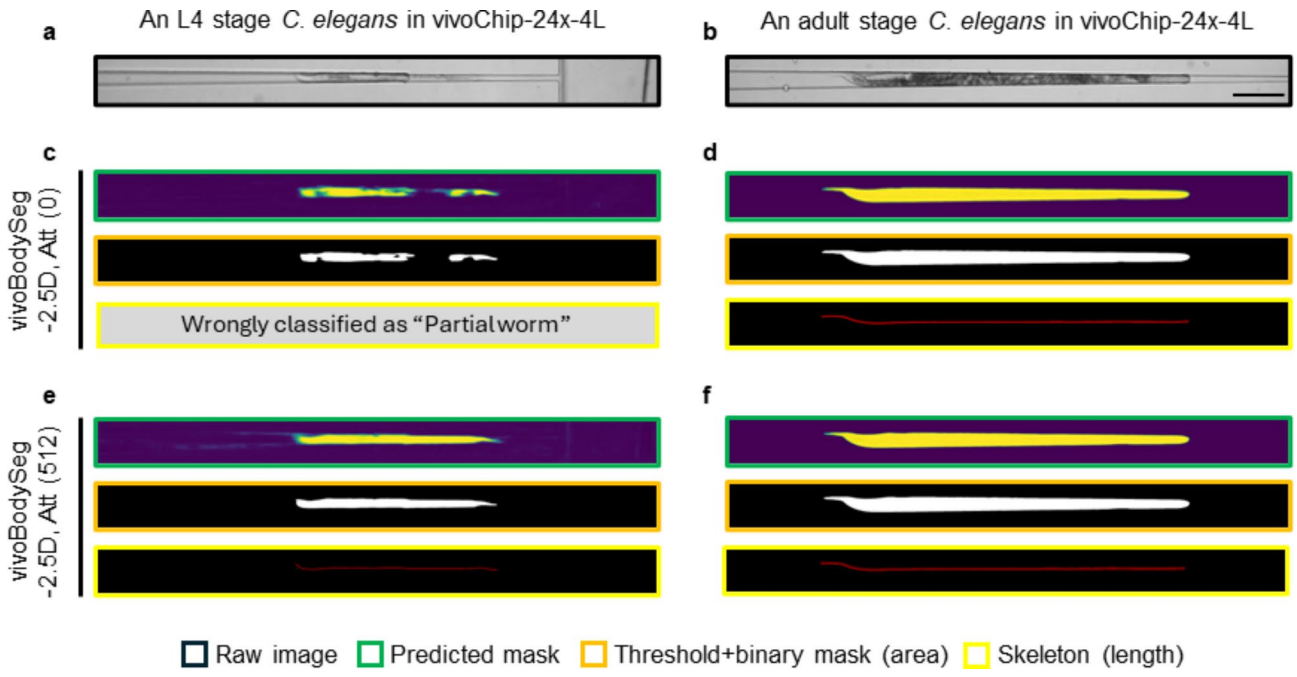

Fig. 5. Improved worm detection using few-shot learning with vivoBodySeg model. Images of a young L4 (**a**) and an adult stage (**b**) *C. elegans*, imaged in the vivoChip-24x-4L device. (**c**, **d**) Results from the baseline vivoBodySeg-2.5D, Att model with zero-shot learning. The model wrongly classified the channel as having a partial worm and thus discarded it from further analysis. The color in the predicted mask (green box) represents the pixel-wise probability of being a part of the worm mask. (**e**, **f**) Results from the vivoBodySeg-2.5D, Att model after few-shot learning using 512 images. The model correctly predicted the worm mask and accurately classified the channel images. Raw image, predicted mask, binary mask, and skeleton are compared. The scale bar = 200 μm.

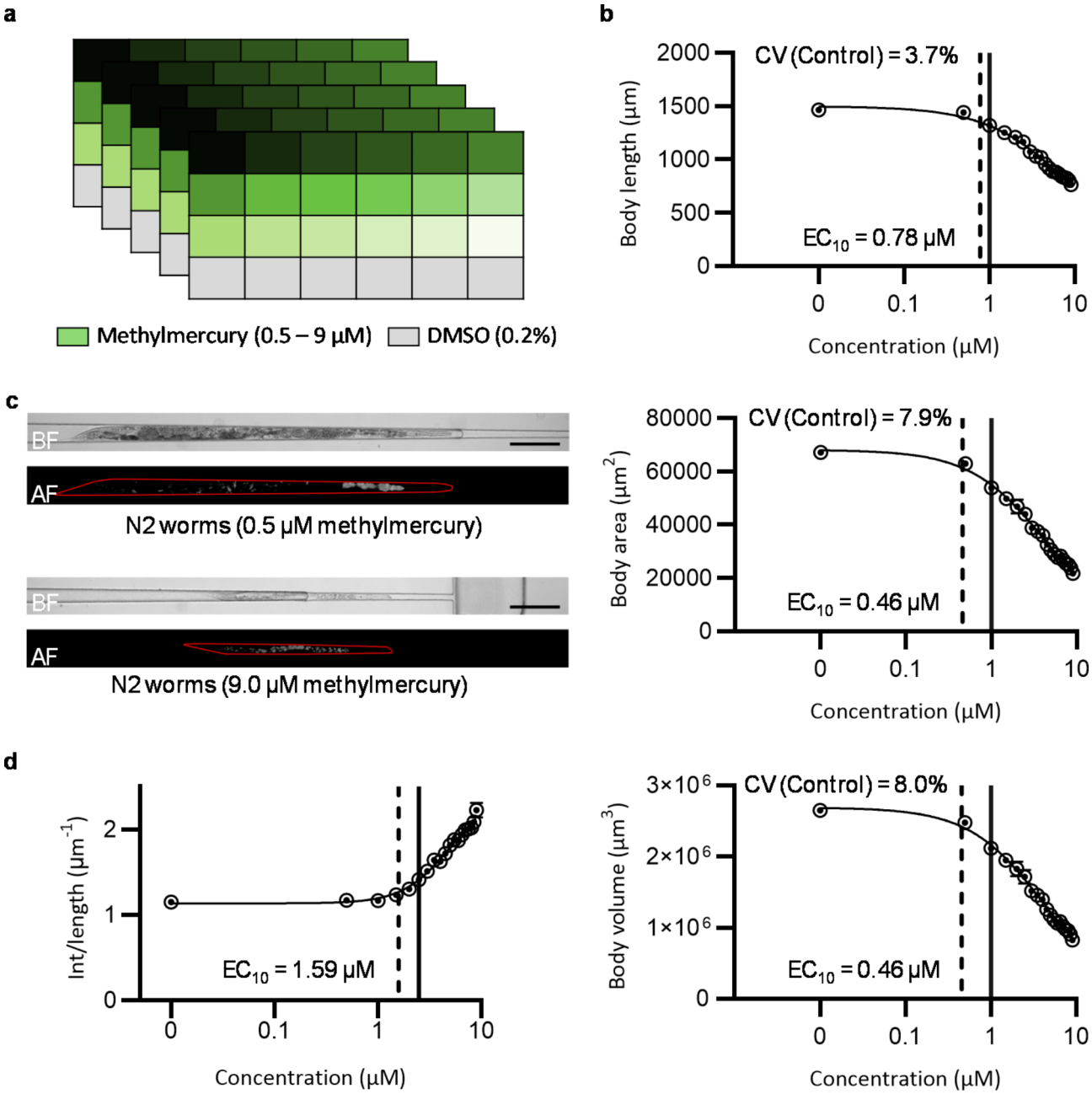

Fig. 6. Automated body parameter and autofluorescence analysis using the vivoBodySeg model. (**a**) Map showing exposure of 24 *C. elegans* populations to methylmercury (0.5 – 9.0 µM) and 0.2% DMSO. The experiment was repeated with five vivoChip-24x-4L devices. (**b**) The data from all five experiments were used to quantify average body length, body area, and body volume. The results are presented as mean ± SEM ($n = 5$ repeats). The solid lines represent the Hill-function fits to the experimental data, while the vertical dotted lines represent the EC_{10} values for each parameter, while the vertical solid lines represent the LOAEL value of 1.0 µM for all 3 parameters. Coefficients of variation (CV) are provided for the control wells. (**c**) Example images of worms from 0.5 µM and 9.0 µM methylmercury. Brightfield (BF) and autofluorescence (AF) images obtained using the GFP filter set, are shown. The scale bars = 200 µm. (**d**) Average autofluorescence intensity per unit body length for methylmercury and DMSO control populations. Data are presented as mean ± SEM ($n \geq 4$ repeats). The vertical dotted line represents the EC_{10} value, while the vertical solid line represents the LOAEL value of 2.5 µM for mean intensity per length.

body area, and 8.0% for body volume. All three CV values were far below the 30% threshold specified in OECD test guidelines for similar species[62]. These results indicate that our automated DevTox assay using *C. elegans* is highly robust and capable of detecting small changes in body parameters with high confidence.

To demonstrate the utility of our DevTox assay, we performed a concentration–response study on the methylmercury, a reference toxicant previously tested in *C. elegans* and other species, and found to be toxic in

humans[23,43,44,63–67]. The average body parameters from five experiments consistently decreased with increasing methylmercury concentrations (Fig. 6b). Toxicity effects were observed at slightly lower concentrations for body volume (EC_{10} = 0.46, 0.32–0.63 µM, ± 95% CI) and body area (EC_{10} = 0.46, 0.32–0.64 µM) compared to body length (EC_{10} = 0.78, 0.56–1.05 µM). The LOAEL value for all three body parameters was estimated to be 1.0 µM. Worms exposed to the highest concentration of the toxicant (9.0 µM) only developed to the early L4 stage. These smallest worms were, thus, trapped in the fourth layer of the microchannels, towards the exit with the smallest channel height, matching our body parameters calculated using the model (Fig. 6c).

We utilized predicted body masks to analyze the autofluorescence signals within each *C. elegans* body and quantified the changes in the average autofluorescence signal (Fig. 6c, Supplementary Fig. 15). The average autofluorescence signal per unit body length for worms exposed to 9.0 µM methylmercury (2.23 ± 0.09) was ~ 2 × higher than that for 0.5 µM methylmercury (1.17 ± 0.06), a statistically significant difference (p-value < 0.001). In contrast, the average autofluorescence signal per unit body length for 0.5 µM methylmercury was comparable to the baseline value for DMSO control (1.15 ± 0.01, p-value = 0.63, Fig. 6d). Similarly, the average autofluorescence signal per unit body area for worms treated with 9.0 µM methylmercury (0.081 ± 0.008) increased by ~ 3 × than that for 0.5 µM methylmercury (0.027 ± 0.001, p-value = 0.002, Supplementary Fig. 16). The EC_{10} value for the average autofluorescence signal per unit body length was1.59 µM (95% CI is 1.27–1.93 µM), which was lower than the EC_{10} value for per unit body area (2.98 µM, 95% CI is 2.48–8.44 µM). The LOAEL values for autofluorescence per unit length and per unit area were 2.5 µM and 1.5 µM, respectively. Interestingly, the increase in autofluorescence signal was delayed compared to the developmental features, suggesting it may be a less sensitive parameter for toxicology assessment of methylmercury. Additionally, *C. elegans* exposed to high methylmercury concentrations displayed reduced motility when observed in their culture plates, consistent with previously published correlations between high autofluorescence and impaired locomotion[68].

Discussion

This study presents an ML-based image analysis platform to perform DevTox studies using *C. elegans* as a NAMs model. This platform enables rapid, high-content analysis of thousands of *C. elegans* images, collected while they are immobilized in parallel channels within the large-scale microfluidic platform, vivoChip-24x. DevTox assessment is one of the mandatory tests needed for evaluating human and environmental risk of new chemicals before they are approved for commercial use. Given the large number of chemicals currently in commercial use (85,000) and the addition of more than 2,500 new chemicals annually[69], industries are seeking innovative solutions to assess chemicals in a high-throughput manner with high predictive power.

C. elegans has been widely used as one of the whole organism models for early drug discovery and toxicology screens. To improve assay sensitivity, capture multiple phenotypes, and facilitate high-content image-based analyses, we developed an automated imaging and image-analysis pipeline to quantify multiple phenotypes in *C. elegans* models. The vivoChip technology captures brightfield, fluorescence, z-stack, and time-lapse images from 960 *C. elegans* across 24 populations in each device. We designed two device variants: a 3-layer device (vivoChip-24x-3L) for adult worms and a 4-layer device (vivoChip-24x-4L) for young L4 stage worms. The 4L device is optimal for range-finding experiments with wide concentration ranges, while the 3L device best suited for experiments targeting narrow concentration range to capture sub-lethal phenotypes. The 3L device ensures sufficient data points for sub-lethal embryo phenotypes to fit the Hill function and precisely determine the LOAEL and effective concentrations (EC_{10} and EC_{50}). The vivoChips are fabricated using plastic instead of PDMS to minimize chemical absorption[70]. The microfluidic channels are sealed with a thin substrate to allow the capturing of high-resolution images with sufficient contrast from transparent worm body parts such as the tail tip or young larvae with few gut granules.

The ML-based vivoBodySeg, trained on images obtained using vivoChip-24x devices, performs exceptionally well in experiments with both 3L and 4L devices. By including a ViT at the bottleneck layer of our encoder, the model easily communicates features spanning the entire image length in a memory-efficient manner. By leveraging information from multiple z-stack images, the network generates volumetric features effectively without relying on higher dimensional convolutions. Ablation experiments guided our design decisions, including incorporating an attention mechanism at the bottleneck (Table 1), choosing 3 z-slices (Table 1, Supplementary Table 1), selecting a 6 µm z-step size (Supplementary Fig. 17), and increasing U-Net depth to 4 layers (Supplementary Table 1), leading to significant performance improvements. The vivoBodySeg model achieves segmentation accuracy indistinguishable from our highly trained scientists. The model processes data from one vivoChip-24x experiment (960 *C. elegans*, 9,600 single-channel images, 36 GB data) in just 35 min using a single desktop PC (with 128 GB of memory and an A6000 GPU with 48 GB of VRAM), compared to ~ 5,000 min for uninterrupted manual expert analysis.

Since the vivoBodySeg model was optimized for brightfield images, its performance may not be generalized well to other wide-field imaging modalities, such as dark-field, phase contrast, or differential interference contrast (DIC). To address this issue, we propose a few-shot learning (FSL) approach for rapid adaptation of the model to new imaging modalities with minimal labeled data.

Our *C. elegans* DevTox analysis quantifies key developmental endpoints, including worm length, body area, and volume, as they prove to be highly useful developmental endpoints with varying levels of sensitivity. Existing automated methods for analyzing *C. elegans* developmental parameters include length estimations using the COPAS Biosorter via 1D signal analysis or plate readers via low-resolution image analysis[27–29,33]. While several groups have used these tools, they exhibit high variability and limited statistical power[32]. Microfluidic technologies have enabled on-chip *C. elegans* imaging, but they typically produce low-resolution data resulting in significant variability in body measurements[71,72]. In contrast, our vivoChip microfluidic platform captures high-resolution images, which are precisely analyzed by the ML pipeline described in this study. This integrated platform achieves high-precision measurements with > 80% statistical power, allowing detection of small

changes, such as a 4% change in length and area or a 7% change in volume, using three experimental replicates. In repeat experiments with methylmercury, both technical and biological replicates showed a strong correlation (~ 97% across 5 samples from 3 different days, Supplementary Table 2). The fully automated ML-based approach eliminates user bias, does not suffer from user fatigue, helps reduce assay costs, and achieves high throughput to screen many chemicals in a low-resource setting.

To demonstrate the implementation of our ML-assisted DevTox assay, we conducted a study with a well-characterized toxicant methylmercury[23,43,45,46,73] at 18 concentrations ranging from 0.5 to 9.0 µM in 0.2% DMSO solvent. We quantified body length, area, and volume observing a concentration-dependent reductions with CVs of 3.7%, 7.9%, and 8.0%, respectively. Our assay produced 2.4× and 6.5× fold improvement in measurement precision compared to conventional image processing method and time-of-flight-based measurements, respectively[32]. Area and volume parameters showed lower EC_{10} values (0.46 µM) compared to length (0.78 µM), indicating they are more sensitive parameters for DevTox. The EC_{50} for body volume (4.14 µM) was 5.5× lower than the previously published LD_{50} after 24 h of exposure[23], emphasizing differences between lethal and sub-lethal endpoints, assay durations, and conditions.

Higher methylmercury concentrations resulted in higher intestinal autofluorescence, possibly indicating an increase in stress levels. Autofluorescence in *C. elegans*, as in mammalian cells, is found in lysosomes granules[74], and linked to metabolism[75], aging[76–79], and toxicant exposure[68,80,81]. The observed autofluorescence increase, previously correlated with reactive oxygen species (ROS) levels, lethality, and behavioral changes[68], underscores its potential as a useful biomarker for assessing chemical toxicity. During autofluorescence analysis, we found a few examples of minor mismatches in body mask predictions (e.g. tip of the tail or the head regions) due to body movement between frames. Immobilization could be improved with a very low concentrations of anesthetics or gels, though care must be taken to avoid adverse effects on worm physiology[82,83].

Our ML pipeline, combined with scalable microfluidic technology, provides a rapid, high-throughput solution for DevTox studies, offering automated, precise quantifications of *C. elegans* volume, which is analogous to body weight assessments in mammalian tests[84]. In the future, we will test a large number of reference chemicals and estimate the accuracy and sensitivity by comparing them with other species[8,11]. Once validated, our DevTox assay can help reduce the use of vertebrate animals in chemical testing while ensuring protection of the environment and human health. Future work will expand DevTox endpoints to include reproductive parameters using *in-utero* embryonic phenotypes, enabling comprehensive developmental and reproductive toxicology (DART) studies. In summary, our DevTox assay can detect small changes in the sub-lethal body parameters with high statistical power from just 40 worms per population, making it suitable for high-throughput toxicity screens. Such high-throughput approaches can generate robust, high-quality toxicology data for large chemical libraries, supporting read-across strategies[85,86]. These strategies are currently being developed by several industries to establish NAMs and design fit-for-purpose toxicology assays.

Data availability

The datasets generated during and/or analyzed during the current study are available from the corresponding authors upon reasonable request.

Received: 24 July 2024; Accepted: 27 December 2024
Published online: 02 January 2025

References

1. Kaletta, T. & Hengartner, M. O. Finding function in novel targets: *C. elegans* as a model organism. *Nat. Rev. Drug. Discov.* **5**, 387–398 (2006).
2. Silverman, G. A. et al. Modeling molecular and cellular aspects of human disease using the nematode *Caenorhabditis elegans*. *Pediatr. Res.* **65**, 10–18 (2009).
3. Hartman, J. H. et al. Xenobiotic metabolism and transport in *Caenorhabditis elegans*. *J. Toxicol. Environ. Health B Crit. Rev.* **24**(2), 51–94 (2021).
4. Leung, M. C. et al. *Caenorhabditis elegans*: An emerging model in biomedical and environmental toxicology. *Toxicol. Sci.* **106**, 5–28 (2008).
5. Corsi, A. K., Wightman, B. & Chalfie, M. A transparent window into biology: A primer on *Caenorhabditis elegans*. *Genetics* **200**(2), 387–407 (2015).
6. Shaye, D. D. & Greenwald, I. OrthoList: A compendium of *C. elegans* genes with human orthologs. *PLoS One* **6**, e20085 (2011).
7. Boyd, W. A. et al. A high-throughput method for assessing chemical toxicity using a *Caenorhabditis elegans* reproduction assay. *Toxicol. Appl. Pharmacol.* **245**(2), 153–159 (2010).
8. Boyd, W. A. et al. Developmental effects of the ToxCast phase I and phase II chemicals in caenorhabditis elegans and corresponding responses in zebrafish, rats, and rabbits. *Environ. Health Perspect.* **124**, 586–593 (2016).
9. Boyd, W. A. et al. Medium- and high-throughput screening of neurotoxicants using *C. elegans*. *Neurotoxicol. Teratol.* **32**, 68–73 (2010).
10. Harlow, P. H. et al. Comparative metabolism of xenobiotic chemicals by cytochrome P450s in the nematode *Caenorhabditis elegans*. *Sci. Rep.* **8**, 13333 (2018).
11. Harlow, P. H. et al. The nematode *Caenorhabditis elegans* as a tool to predict chemical activity on mammalian development and identify mechanisms influencing toxicological outcome. *Sci. Rep.* **6**, 22965 (2016).
12. Hunt, P. R. & The, C. elegans model in toxicity testing. *J. Appl. Toxicol.* **37**, 50–59 (2017).
13. Hunt, P. R. *Building Confidence in the Use of NAMs data for Risk Analysis: C. elegans as a Case Study*, J. Camacho, Editor. (2021).
14. Hunt, P. R., Camacho, J. A. & Sprando, R. L. *Caenorhabditis elegans* for predictive toxicology. *Curr. Opin. Toxicol.* **23–24**, 23–28 (2020).
15. Hunt, P. R. et al. *C. elegans* development and activity test detects mammalian developmental neurotoxins. *Food Chem. Toxicol.* **121**, 583–592 (2018).
16. Morley, J. F. et al. The threshold for polyglutamine-expansion protein aggregation and cellular toxicity is dynamic and influenced by aging in *Caenorhabditis elegans*. *Proc. Natl. Acad. Sci. USA* **99**, 10417–10422 (2002).

17. Link, C. D. Expression of human beta-amyloid peptide in transgenic *Caenorhabditis elegans*. *Proc. Natl. Acad. Sci. USA* **92**, 9368–9372 (1995).
18. Lakso, M. et al. Dopaminergic neuronal loss and motor deficits in *Caenorhabditis elegans* overexpressing human alpha-synuclein. *J. Neurochem.* **86**(1), 165–172 (2003).
19. Dillin, A. et al. Rates of behavior and aging specified by mitochondrial function during development. *Science* **298**(5602), 2398–2401 (2002).
20. Lee, S. S. et al. DAF-16 target genes that control *C. elegans* life-span and metabolism. *Science* **300**(5619), 644–647 (2003).
21. Xiong, H., Pears, C. & Woollard, A. An enhanced *C. elegans* based platform for toxicity assessment. *Sci. Rep.* **7**, 9839 (2017).
22. Cole, R. D., Anderson, G. L. & Williams, P. L. The nematode *Caenorhabditis elegans* as a model of organophosphate-induced mammalian neurotoxicity. *Toxicol. Appl. Pharmacol.* **194**, 248–256 (2004).
23. Helmcke, K. J. et al. Characterization of the effects of methylmercury on *Caenorhabditis elegans*. *Toxicol. Appl. Pharmacol.* **240**(2), 265–272 (2009).
24. Leung, M. C. et al. *Caenorhabditis elegans* generates biologically relevant levels of genotoxic metabolites from aflatoxin B1 but not benzo[a]pyrene in vivo. *Toxicol. Sci.* **118**(2), 444–453 (2010).
25. Leung, C. K., et al. *High-throughput screening and biosensing with fluorescent C. elegans strains. J. Vis. Exp.* (2011).
26. Leung, C. K. et al. An ultra high-throughput, whole-animal screen for small molecule modulators of a specific genetic pathway in *Caenorhabditis elegans*. *PLoS One* **8**, e62166 (2013).
27. Nyaanga, J. et al. Changes in body shape implicate cuticle stretch in *C. elegans* growth control. *Cells Dev* **170**, 203780 (2022).
28. Gosai, S. J. et al. Automated high-content live animal drug screening using *C. elegans* expressing the aggregation prone serpin alpha1-antitrypsin Z. *PLoS ONE* **5**, e15460 (2010).
29. Pulak, R. Techniques for analysis, sorting, and dispensing of *C. elegans* on the COPAS flow-sorting system. *Methods Mol. Biol.* **351**, 275–286 (2006).
30. Shimko, T. C. & Andersen, E. C. COPASutils: An R package for reading, processing, and visualizing data from COPAS large-particle flow cytometers. *PLoS One* **9**, e111090 (2014).
31. Smith, M. V. et al. A discrete time model for the analysis of medium-throughput *C. elegans* growth data. *PLoS One* **4**(9), e7018 (2009).
32. Moore, B. T., Jordan, J. M. & Baugh, L. R. WormSizer: High-throughput analysis of nematode size and shape. *PLoS One* **8**(2), e57142 (2013).
33. Wahlby, C. et al. An image analysis toolbox for high-throughput *C. elegans* assays. *Nat. Methods* **9**, 714–716 (2012).
34. Koopman, M. et al. Assessing motor-related phenotypes of *Caenorhabditis elegans* with the wide field-of-view nematode tracking platform. *Nat. Protoc.* **15**(6), 2071–2106 (2020).
35. Mondal, S. et al. Large-scale microfluidics providing high-resolution and high-throughput screening of *Caenorhabditis elegans* poly-glutamine aggregation model. *Nat. Commun.* **7**, 13023 (2016).
36. Mondal, S., et al. *High-content microfluidic screening platform used to identify sigma2R/Tmem97 binding ligands that reduce age-dependent neurodegeneration in C. elegans SC_APP model. ACS Chem. Neurosci.* (2018).
37. Ben-Yakar, A. High-content and high-throughput in vivo drug screening platforms using microfluidics. *Assay Drug Dev. Technol.* **17**, 8–13 (2019).
38. Mondal, S. & Ben-Yakar, A. Chapter 11-Caenorhabditis elegans-on-a-chip: Microfluidic platforms for high-resolution imaging and phenotyping. In *Organ-on-a-chip* (eds Hoeng, J. et al.) 363–390 (Academic Press, 2020).
39. Deserno, M. & Bozek, K. WormSwin: Instance segmentation of *C. elegans* using vision transformer. *Sci. Rep.* **13**(1), 11021 (2023).
40. Bates, K., Le, K. N. & Lu, H. Deep learning for robust and flexible tracking in behavioral studies for *C. elegans*. *PLoS Comput. Biol.* **18**(4), e1009942 (2022).
41. Layana Castro, P. E., Garvi, A. G. & Sanchez-Salmeron, A. J. Automatic segmentation of *Caenorhabditis elegans* skeletons in worm aggregations using improved U-Net in low-resolution image sequences. *Heliyon* **9**(4), e14715 (2023).
42. Stiernagle, T. *Maintenance of C. elegans* 1–11 (WormBook, 2006).
43. Ke, T. et al. Genetic factors in methylmercury-induced neurotoxicity: What have we learned from *Caenorhabditis elegans* models?. *Adv. Neurotoxicol.* **9**, 271–290 (2023).
44. Hu, K. et al. Ecotoxicity risk of low-dose methylmercury exposure to *Caenorhabditis elegans*: Multigenerational toxicity and population discrepancy. *Chem. Res. Toxicol.* **34**(4), 1114–1123 (2021).
45. Crawford, N., et al. *Methylmercury-induced metabolic alterations in Caenorhabditis elegans are diet-dependent. Toxics* **9**(11) (2021).
46. Ke, T. & Aschner, M. Bacteria affect *Caenorhabditis elegans* responses to MeHg toxicity. *Neurotoxicology* **75**, 129–135 (2019).
47. Groen, F. C., Young, I. T. & Ligthart, G. A comparison of different focus functions for use in autofocus algorithms. *Cytometry* **6**, 81–91 (1985).
48. Sun, Y., Duthaler, S. & Nelson, B. J. Autofocusing in computer microscopy: Selecting the optimal focus algorithm. *Microsc. Res. Tech.* **65**, 139–149 (2004).
49. Torbunov, D., et al. UVCGAN: UNet vision transformer cycle-consistent GAN for unpaired image-to-image translation. *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (702–712). https://api.semanticscholar.org/CorpusID:247291829 (2023).
50. Ronneberger, O., Fischer, P. & Brox, T. *U-Net: Convolutional networks for biomedical image segmentation*. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*. (Springer, 2015).
51. Avesta, A., et al. *Comparing 3D, 2.5D, and 2D approaches to brain image auto-segmentation. Bioengineering (Basel)* **10**(2) (2023).
52. He, K. et al. Deep residual learning for image recognition. *IEEE Conf. Comput. Vis. Pattern Recogn.* **2016**, 770–778 (2016).
53. Lee, J., et al. Set transformer: A framework for attention-based permutation-invariant neural networks. *Proceedings of the 36th International Conference on Machine Learning* **97**, 3744–3753. https://proceedings.mlr.press/v97/lee19d.html (2019).
54. Dosovitskiy, A., et al. *An image is worth 16x16 words: Transformers for image recognition at scale.* arXiv e-prints. arXiv:2010.11929 (2020).
55. van den Oord, A., Vinyals, O. & Kavukcuoglu, K. Neural discrete representation learning. *Advances in Neural Information Processing Systems* 30. arXiv:1711.00937 (2017).
56. Vaswani, A., et al. Attention is all you need. *Advances in Neural Information Processing Systems*, **30**. arXiv:1706.03762 (2017).
57. Loshchilov, I. & Hutter, F. *Decoupled weight decay regularization.* arXiv e-prints. arXiv:1711.05101 (2017).
58. Loshchilov, I. & Hutter, F. *SGDR: Stochastic gradient descent with warm restarts.* arXiv e-prints. arXiv:1608.03983 (2016).
59. Lin, T.-Y., et al. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **42**(2), 318–327. https://doi.org/10.1109/TPAMI.2018.2858826 (2020).
60. Parnami, A. & Lee, M. *Learning from few examples: A summary of approaches to few-shot learning.* arXiv e-prints. arXiv:2203.04291 (2022).
61. Wang, Y., et al. *Generalizing from a few examples: A survey on few-shot learning.* arXiv e-prints. arXiv:1904.05046 (2019).
62. OECD *Test No. 207: Earthworm, Acute Toxicity Tests* (1984).
63. Tsuchiya, H. et al. Placental transfer of heavy metals in normal pregnant Japanese women. *Arch. Environ. Health* **39**(1), 11–17 (1984).
64. Hong, Y. S., Kim, Y. M. & Lee, K. E. Methylmercury exposure and health effects. *J. Prev. Med. Public Health* **45**(6), 353–363 (2012).
65. Diez, S. Human health effects of methylmercury exposure. *Rev. Environ. Contam. Toxicol.* **198**, 111–132 (2009).

66. Nagashima, K. A review of experimental methylmercury toxicity in rats: Neuropathology and evidence for apoptosis. *Toxicol. Pathol.* **25**(6), 624–631 (1997).
67. Glazer, L. & Brennan, C. H. Developmental exposure to low concentrations of methylmercury causes increase in anxiety-related behaviour and locomotor impairments in zebrafish. *Int. J. Mol. Sci.* **22**(20) (2021).
68. Wu, Q. et al. Chromium exhibits adverse effects at environmental relevant concentrations in chronic toxicity assay system of nematode *Caenorhabditis elegans*. *Chemosphere* **87**(11), 1281–1287 (2012).
69. DTSC. *Chemicals of Emerging Concern*. 2024 [cited 2024 July 17, 2024].
70. Toepke, M. W. & Beebe, D. J. PDMS absorption of small molecules and consequences in microfluidic applications. *Lab. Chip.* **6**, 1484–1486 (2006).
71. Atakan, H. B. et al. Automated high-content phenotyping from the first larval stage till the onset of adulthood of the nematode *Caenorhabditis elegans*. *Lab. Chip.* **19**(1), 120–135 (2018).
72. Rahman, M. et al. NemaLife chip: A micropillar-based microfluidic culture device optimized for aging studies in crawling *C. elegans*. *Sci. Rep.* **10**(1), 16190 (2020).
73. Ke, T. et al. Latent alterations in swimming behavior by developmental methylmercury exposure are modulated by the homolog of tyrosine hydroxylase in *Caenorhabditis elegans*. *Neurotoxicol. Teratol.* **85**, 106963 (2021).
74. Tan, C.H., et al. *Fluorescence dynamics of lysosomal-related organelle flashing in the intestinal cells of Caenorhabditis elegans.* bioRxiv (2023).
75. Schroeder, L. K. et al. Function of the *Caenorhabditis elegans* ABC transporter PGP-2 in the biogenesis of a lysosome-related fat storage organelle. *Mol. Biol. Cell* **18**(3), 995–1008 (2007).
76. Clokey, G. V. & Jacobson, L. A. The autofluorescent "lipofuscin granules" in the intestinal cells of *Caenorhabditis elegans* are secondary lysosomes. *Mech. Ageing Dev.* **35**(1), 79–94 (1986).
77. Davis Jr, B. O., Anderson, G. L. & Dusenbery, D. B. Total luminescence spectroscopy of fluorescence changes during aging in *Caenorhabditis elegans*. *Biochemistry* **21**(17), 4089–4095 (1982).
78. Klass, M. R. Aging in the nematode *Caenorhabditis elegans*: Major biological and environmental factors influencing life span. *Mech. Ageing Dev.* **6**(6), 413–429 (1977).
79. Pincus, Z., Mazer, T. C. & Slack, F. J. Autofluorescence as a measure of senescence in *C. elegans*: Look to red, not blue or green. *Aging* **8**(5), 889–898 (2016).
80. Shen, L. et al. Toxicity evaluation in nematode *Caenorhabditis elegans* after chronic metal exposure. *Environ. Toxicol. Pharmacol.* **28**(1), 125–132 (2009).
81. Hajdu, G. et al. Lysosome-related organelles promote stress and immune responses in *C. elegans*. *Commun. Biol.* **6**(1), 936 (2023).
82. Guo, S. X. et al. Femtosecond laser nanoaxotomy lab-on-a-chip for in vivo nerve regeneration studies. *Nat. Methods* **5**, 531–533 (2008).
83. Mondal, S. et al. Imaging in vivo neuronal transport in genetic model organisms using microfluidic devices. *Traffic* **12**, 372–385 (2011).
84. OECD. *Test No. 421: Reproduction/Developmental Toxicity Screening Test, OECD Guidelines for the Testing of Chemicals*. (OECD, 2016).
85. Serafini, M. M. et al. Recent advances and current challenges of new approach methodologies in developmental and adult neurotoxicity testing. *Arch. Toxicol.* **98**(5), 1271–1295 (2024).
86. Schultz, T. W. et al. A strategy for structuring and reporting a read-across prediction of toxicity. *Regul. Toxicol. Pharmacol.* **72**(3), 586–601 (2015).

Acknowledgements

The authors would like to thank the National Institutes of Health for the Small Business Innovation Research grants from NIEHS (R43ES033579 and R44ES033579) and NIMH (R43MH118841 and R44MH118841). We would like to thank Thao Vu and Laura Murphy for their help with manual segmentations. We also thank Nikko Chavez and Dr. Jacob Moore for the initial version of the image acquisition setup. Some strains were provided by the CGC, which is funded by the NIH Office of Research Infrastructure Programs (P40 OD010440). We thank The Caenorhabditis Natural Diversity Resource (CaeNDR) for providing the N2 strain.

Author contributions

A.D. and A.M. designed, trained, and validated the machine learning models. S.G. developed image acquisition and segmentation software. A.S., A.L., and E.H. ran the experiments. A.D., A.M., A.L., and S.M. analyzed the data, prepared the plots, and performed the statistics. S.M. and A.B. conceptualized the projects, planned the experiments, and finalized the discussion of the results. A.D., S.M., and A.B. prepared the manuscript.

Declarations

Competing interests

E.H., S.M., and A.B. are co-founders of vivoVerse, LLC and its Associates. A.D., A.S., A.L., E.H., S.M., and A.B. are inventors of several approved and pending patents. A.M. and S.G. at the time of their contribution are employed by vivoVerse, LLC and have no conflict of interest to declare.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-84842-x.

Correspondence and requests for materials should be addressed to S.M. or A.B.-Y.