scientific reports

OPEN

An open-source tool for automated human-level circling behavior detection

O. R. Stanley[1], A. Swaminathan[1,4], E. Wojahn[1,4], C. Bao[1], Z. M. Ahmed[2] & K. E. Cullen[1,3]

Quantitatively relating behavior to underlying biology is crucial in life science. Although progress in keypoint tracking tools has reduced barriers to recording postural data, identifying specific behaviors from this data remains challenging. Manual behavior coding is labor-intensive and inconsistent, while automatic methods struggle to explicitly define complex behaviors, even when they seem obvious to the human eye. Here, we demonstrate an effective technique for detecting circling in mice, a form of locomotion characterized by stereotyped spinning. Despite circling's extensive history as a behavioral marker, there currently exists no standard automated detection method. We developed a circling detection technique using simple postprocessing of keypoint data obtained from videos of freely-exploring (*Cib2*$^{-/-}$;*Cib3*$^{-/-}$) mutant mice, a strain previously found to exhibit circling behavior. Our technique achieves statistical parity with independent human observers in matching occurrence times based on human consensus, and it accurately distinguishes between videos of wild type mice and mutants. Our pipeline provides a convenient, noninvasive, quantitative tool for analyzing circling mouse models without the need for software engineering experience. Additionally, as the concepts underlying our approach are agnostic to the behavior being analyzed, and indeed to the modality of the recorded data, our results support the feasibility of algorithmically detecting specific research-relevant behaviors using readily-interpretable parameters tuned on the basis of human consensus.

Observable actions serve as noninvasive readouts of underlying biological facts—e.g. injury, disease, gene expression, or neural function. This recognition sits at the core of behavioral sciences. Behavioral analysis has historically relied, and largely still relies, on labor-intensive manual behavior coding of real-time or videotaped behavior as its gold standard. Unfortunately, especially when analyzing long sessions or large numbers of sessions, manual coding can suffer due to rater variability, fatigue, or quirks in precise definitions[1–3]. Though recent advances in computer vision[4] and miniaturized sensors[5,6] have made quantitative data increasingly available, classifying kinematic data into specific behaviors remains a central challenge. Critically, behaviors which seem clear-cut to a human observer can in reality be noisy and subjective. As a result, while algorithmic behavior detection holds the potential for rapid, objective quantification, automated methods face difficulty in explicitly defining complex behaviors.

As one example, rodent behavioral neuroscience studies frequently report a repetitive spinning behavior known as 'circling'. The utility of circling as a behavioral marker, and the resultant need for objective quantification of the behavior, has been recognized for more than 50 years[7]. Examples include studies of basal ganglia damage[7–12] as well as genetically engineered models of Alzheimer's Disease[13,14] and autism[15–17]. Additionally, a large number of mutant mouse strains displaying circling behavior display dysfunction of the vestibular system, which in healthy animals contributes to maintaining balance, steadying gaze, and keeping track of the body within the environment. These include mutants that exhibit loss of vestibular hair cells[18], disrupted development of stereocilia[19,20], or disrupted structural development of the inner ear[21,22].

However, despite its long history as a behavioral marker, the existing literature lacks a standard, quantitative definition of circling. Rather, studies often report the simple presence or absence of circling[5,23–25]. Studies which quantify frequency of occurrence rely on manual coding and use disparate definitions such as complete rotations[26,27], sequences of complete rotations[13], or 270-degree turns during which the body travels a minimum distance[28]. Older studies which deployed video analysis relied on tracking the center of mass of a mouse against a

[1]Department of Biomedical Engineering, Johns Hopkins University, 720 Rutland Ave, Traylor 504, Baltimore, MD 21205-2109, USA. [2]Departments of Otorhinolaryngology-Head and Neck Surgery, Biochemistry and Molecular Biology, Ophthalmology, University of Maryland School of Medicine, Baltimore, MD, USA. [3]Departments of Neuroscience, Otolaryngology-Head and Neck Surgery, Johns Hopkins University, Baltimore, MD, USA. [4]These authors contributed equally: A. Swaminathan and E. Wojahn. ✉email: kathleen.cullen@jhu.edu

high-contrast background[9,10] and could thus apply limited analysis, whereas more recent work has incorporated commercial and open source tracking solutions but focused on total amount of rotation[29,30] rather than circling per se. Accurate and objective quantification of behavioral parameters such as the frequency of occurrence, duration of bouts of circling behavior, and velocity of movements would facilitate comparison between specific etiologies. This inconsistency reduces the utility of circling as a tool for comparisons across models and setups. Thus, there is a need for a broadly-accessible, quantitative, automated tool for the detection of circling behavior.

Here, we present a technique for detecting circling behavior by tuning algorithmic parameters based on consensus occurrence times among human observers. Specifically, we assessed (*Cib2*$^{-/-}$;*Cib3*$^{-/-}$) dual knockout mice, a mouse strain we and others have previously reported exhibiting circling[31–33]. We track the snout and base of the tail in mice using the open-source software package DeepLabCut[4] ("DLC"), then compare the performance of several behavioral detection algorithms which analyze characteristics of the animals' paths. By identifying the behavioral parameters that result in labels most closely matching human behavioral coding, our technique achieves statistical equivalence to individual observers' independent labels at matching consensus times. Our methodology thus provides a simple, inexpensive process for recording and quantifying mouse circling behavior. Furthermore, the success of our technique suggests its applicability for use in comparing against other etiologies and in detecting and quantifying other research-relevant behaviors.

Results

The goal of the present work was to build and validate a tool to automatically identify mice exhibiting circling behavior during free exploration. Figure 1 illustrates an overview of our data collection and analysis pipeline. We first recorded a set of videos of five wild-type (C57BL/6) mice and five (*Cib2*$^{-/-}$;*Cib3*$^{-/-}$) dual knockout mice in six different recording conditions (enumerated in Methods: "Data generation"). To establish a standard against which to compare the quality of automatic circling detection, three human observers independently marked times at which circling behavior occurred in two subsets of these videos: (1) a *test set* containing all videos of one mutant mouse and one wild type mouse and (2) an equally-sized *training set* randomly selected from the remaining videos, evenly split between mutant and wild type. These behavior labels were compared to identify consensus instances in which all observers marked circles sufficiently close to one another (0.1 s or less), which subsequently served as our gold standard (see Methods: "Gold standard development").

All videos not used in the test set were used to train and evaluate a computer vision model for tracking two keypoints on the mouse body, the tip of the snout and base of the tail. Once trained, the computer vision model was run on the manually-screened training set videos to generate tracking information. After minimal preprocessing of the resulting keypoint position data, we applied three progressively more sophisticated algorithms intended to identify instances of circling behavior. Parameters for each algorithm were optimized to match consensus times as closely as possible as measured by F1 score on the human-scored training set (see Methods: "Algorithm development and parameter search"). After this optimization, the trained computer vision model and finalized detection methods were applied to the manually-screened test set to establish whether these methods generalized to completely novel videos. Finally, to establish whether this large amount of data was required to achieve human-level performance at detecting circling behavior, we trained computer vision models on varied subsets of our overall dataset and applied our most performant algorithm to the test set.

Comparison and consensus among human observers

To obtain a standard against which to measure our behavioral labeling algorithms, we examined the degree of consensus among human labels of circling behavior. To this end, first we selected one mutant mouse and one wild-type mouse at random to form a held-out test set. All videos of these two animals (4 videos in each of 6 conditions; 24 videos each animal) manually screened for instances of circling independently by three observers, who were instructed to mark times at which they noted complete rotations during bouts of circling behavior but not during normal exploratory behavior shared by wild-type mice (e.g. turning around after reaching the edge of the arena). This process was repeated for 24 randomly selected mutant videos and 24 randomly selected wild-type videos from among the 188 not included in the test set. Videos were not labeled with information regarding mouse genotype.

We hypothesized that human observers marking occurrences of circling behavior would show consistent agreement. To assess this, we calculated F1 scores for pairs of observers, treating one as ground truth for another (Fig. 2A). Contrary to our expectation, we found that although pairwise F1 scores were similar on average (0.53, 0.52, and 0.49), the distribution of those scores varied widely enough that one pair differed significantly from both others (pair CA, p = 3.5E−2 and 1.4E−4 versus pairs AB and BC).

We next examined whether human performance differed between our training and test sets as well as how well independent labels matched ultimate consensus circling occurrence times, i.e., cases in which all three observers indicated a circle sufficiently close to one another. Independent observers scored against each other produced an average F1 score of 0.51 (95% CI 0.47–0.55) on the training set and 0.53 (0.41–0.62) on the test set. (not significantly different, p = 0.65; Fig. 2B, Independent vs Independent) When all observers were scored against a group consensus gold standard, average F1 scores were similar to pairwise scoring: 0.51 (0.44–0.57) and 0.53 (0.38–0.65) on the training and test sets, respectively. (p = 0.75; Fig. 2B, Independent vs Consensus) We observed no significant difference between these alternative gold standards (p = 0.20). Consensus times were used for subsequent tuning and evaluation of automated behavior detection methods.

Developing and testing algorithmic circling detection methods

Accurately tracking an animal's position, or even the position of many body parts, is insufficient on its own to establish the behavior an animal is exhibiting. Rather, this raw data must be processed once collected. To this

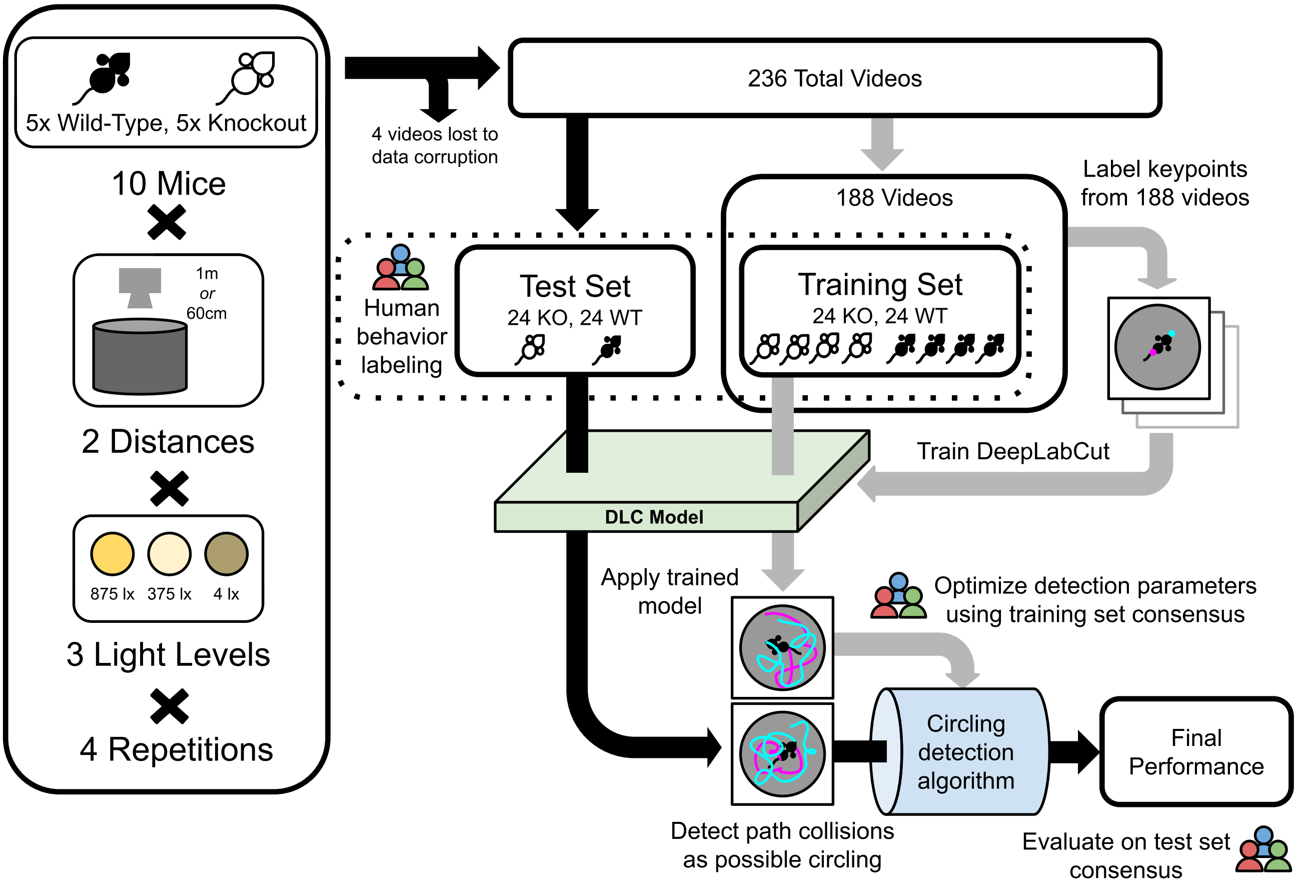


Fig. 1. Data collection conditions and analysis pipeline. We collected videos of five wild-type and five (*Cib2*$^{-/-}$;*Cib3*$^{-/-}$) dual knockout mice exploring a 30 cm-diameter cylindrical arena. Each of 6 combinations of light and distance conditions was repeated 4 times for each mouse, resulting in a total of 236 videos as 4 became corrupted. After behavior videos were recorded, all videos of one mutant mouse and one wild-type mouse were set aside for human behavioral labeling as a test set. For each of these held-out videos, three observers independently marked occurrences of circling behavior. These behavioral labels were compared to produce a set of consensus labels on which all observers agreed. A separate training set of human behavior labels was constructed by randomly selecting 24 mutant and 24 wild-type videos from among the remaining 188 videos. Additionally, positions of the snout and tailbase were manually labeled in 20 random frames from each of these 188 videos. Manually-labeled bodypart locations were used to train a computer vision model using DeepLabCut. This trained model was then used to track animals in the human-scored videos, and the resulting paths were analyzed by three candidate circling detection algorithms. After the parameters of these algorithms were optimized for F1 score on the training set, they were applied to the test set for evaluation.

end, we applied several candidate algorithms for detecting circling behavior using keypoint tracking data after minimal preprocessing. In our case, this preprocessing consisted of setting a threshold on frame-by-frame keypoint label jumps equal to a velocity of 40× the median length of the labeled tail-snout vector per second, a value selected to be high enough that no plausible physical process could produce it. Datapoints exceeding this limit were replaced with simple linear interpolation.

Each behavior detection algorithm first detects instances in which the path of the mouse's snout intersects itself (see example timelapse, Fig. 3A) to analyze as candidate circling occurrences. Importantly, not all intersections are the result of circling—many will be produced by normal exploratory head movements made by both mutant and wild-type animals. Each algorithm therefore attempts to exclude false positives (i.e., instances incorrectly marked as circling) using one or more features of the animal's path between the points of collision.

We observed that the underlying distribution of putative circling parameters (duration, rotation, and size) could be well described by an initial decay for parameter values near zero and a roughly normal peak. We thus sought to model the distribution of each of these parameters as the sum of an exponential and a Gaussian distribution (see Methods: "Algorithm development and parameter search"). Figure 3B illustrates these distributions and fits for two example videos. Bounds outside of which collisions were excluded as circles were parameterized by considering how far, and in what direction, a candidate circle falls from the center of the fitted Gaussian in terms of standard deviations. For example, a snout path collision produced by an exploratory head movement during which a mouse rotated its body only 10 degrees would fall many standard deviations short of the average and thus be correctly excluded. Values of these features, as measured by F1 score against human consensus labels

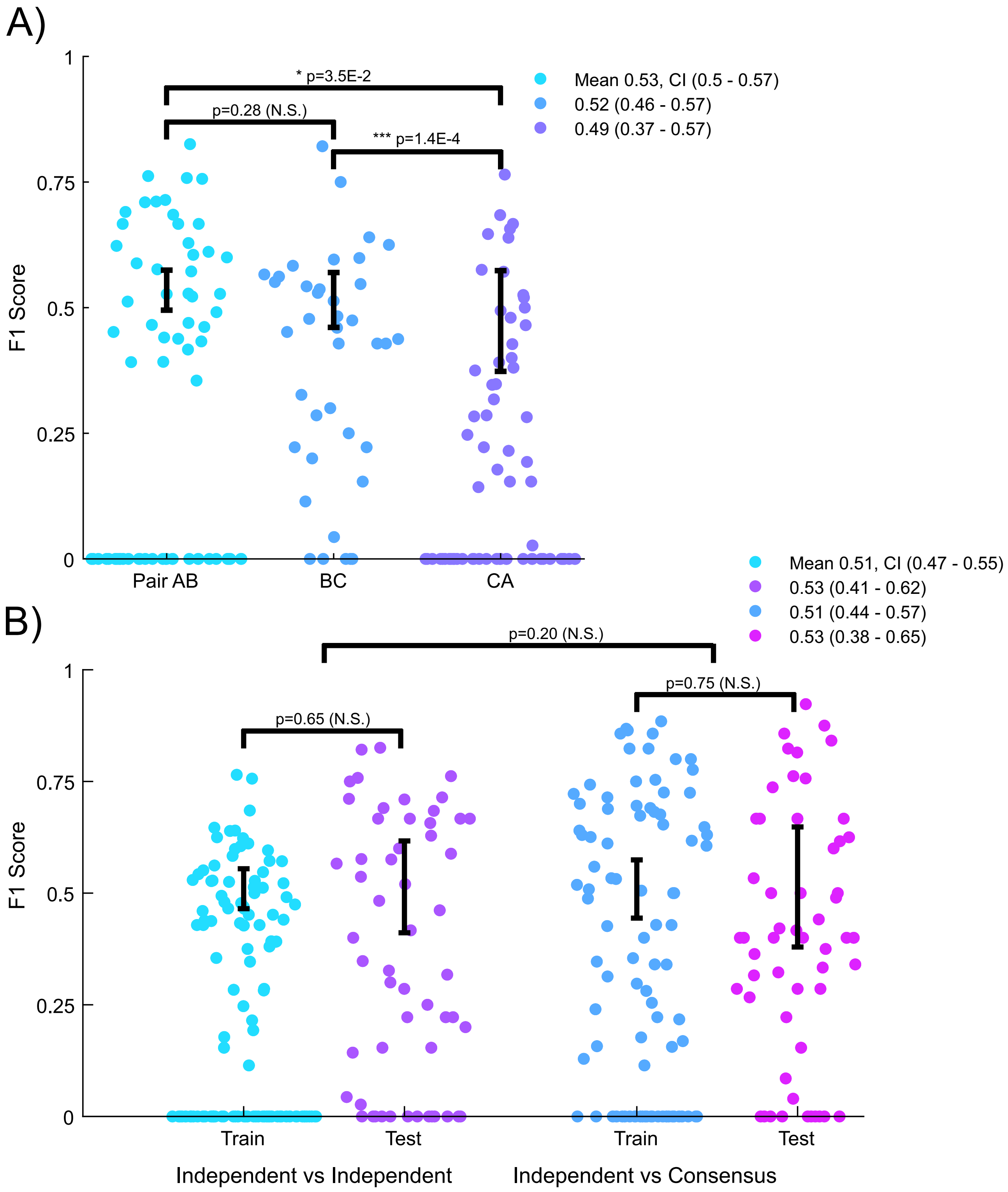
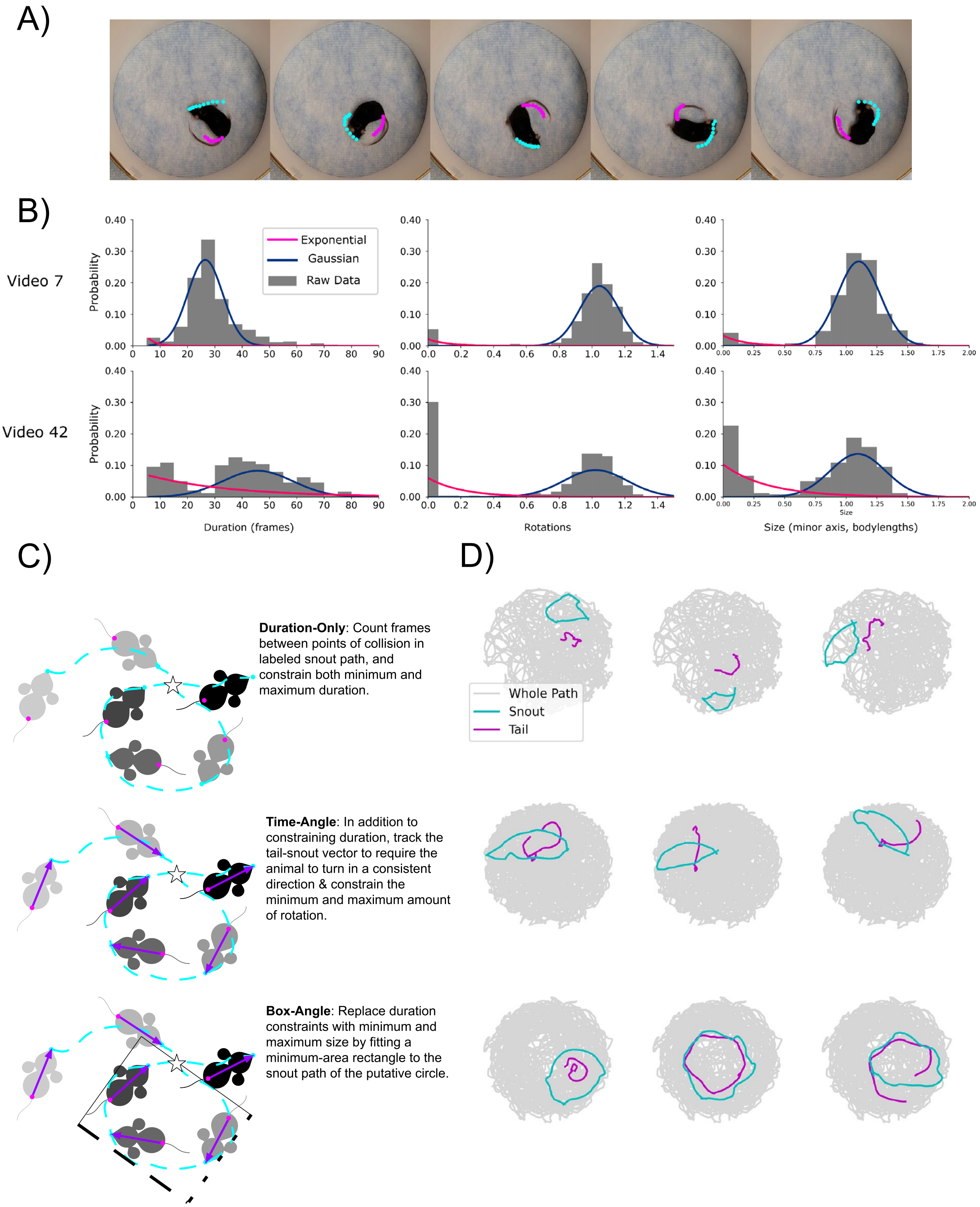


Fig. 2. Human F1 scores. (**A**) Treating one independent observer as the gold standard for another reveals that humans show substantial variability in labeling circling behavior. In particular, although average F1 scores for each pair (AB, BC, CA) are similar (0.53, 0.52, 0.49), the distributions of scores across videos differ significantly between one pair and the other two (pair CA, p = 3.5E−2 and 1.4E−4 versus pairs AB, BC respectively) while the other pair did not differ significantly (AB versus BC, p = 0.28). (**B**) Scoring of independent observers' labels against another observer (left columns) or against consensus labels (agreement among 3 observers, right columns) produce similar results (p = 0.2), as does comparing between our two human data subsets (train versus test subset, p = 0.65 and 0.75). Pooled pairwise F1 scores averaged 0.51 (95% CI 0.47–0.55) in the training set and 0.53 (0.41–0.62) in the testing set. Scoring against consensus occurrences, in which all observers mark a complete circle within 0.1 s of one another, produced similar scores of 0.51 (0.44–0.57) in the training set and 0.53 (0.38–0.65). Each point in a column represents a single video. Labeler-video combinations for which F1 score is undefined (i.e., both scorer and ground truth marked no circling instances), are not displayed for either paired or consensus scoring but were included in bootstrapping for purposes of calculating confidence intervals.



Duration-Only: Count frames between points of collision in labeled snout path, and constrain both minimum and maximum duration.

Time-Angle: In addition to constraining duration, track the tail-snout vector to require the animal to turn in a consistent direction & constrain the minimum and maximum amount of rotation.

Box-Angle: Replace duration constraints with minimum and maximum size by fitting a minimum-area rectangle to the snout path of the putative circle.

Fig. 3. Method parameters and performance levels. (**A**) Timelapse of keypoint-labeled frames of a mouse engaged in circling behavior. (**B**) Parameter distributions and associated exponential and Gaussian fits from two sample videos. To accommodate the substantial variability observed across videos, we relied on a two-step process of Gaussian kernel estimation followed by fitting to a weighted sum of an exponential and normal distribution. This allowed the same technique to account for differences in e.g. average duration (left column, compare blue Gaussian fits) or greater numbers of small collisions likely to be false positives (right column, compare pink exponential fits). (**C**) Illustration of circle detection using each of the described methods. Duration-Only considers only time taken to complete the putative circle, Time-Angle additionally calculates the angle of the tail-to-snout vector for each frame and considers its total net change, and Box-Angle removes duration requirements and instead constraints the geometry of the circle based on the axes of a rectangle bounding the candidate circling instance. (**D**) Examples of false-positive detections using each method. There are clear features which indicate an instance should be filtered out for the Duration-Only (minimal head movement relative to the tail) and Time-Angle (oblong or missized snout path geometry) methods.

in our training set, were optimized via an iterative Bayesian search over the range of ± 3 standard deviations from the mean of the Gaussian fit. Table 1 lists the parameters and final values for each method; parameter ranges were searched independently for each method.

Duration-only method

To establish a lower bound on automatic behavioral detection performance, we first assessed a simple detection algorithm which considers only how long a candidate circling instance takes. The 'Duration-Only' method locates points at which the path of the mouse's snout crosses over itself and excludes those which are either too short or too long (Fig. 3C, Row 1). Our parameter search produced an F1 score of approximately 0.21 in the training set. The same parameters applied to the manually-screened test set scored 0.10 (95% CI 0.02–0.17). As expected given its simplicity, this fell well below human performance and was not sufficient to filter out many false positives which appear obvious upon review. Specifically, it incorrectly labeled many cases of head-only exploratory movements as circling, examples of which are illustrated in Fig. 3D, Row 1. To filter these out more effectively and thereby improve behavioral detection, we next explored incorporating the labeled tailbase position.

Time-angle method

For a mouse that is spinning rather than exploring with its snout alone, the angle of its snout relative to its tail should change noticeably and in a consistent direction. Accordingly, we next considered what we term the 'Time-Angle' method for excluding false-positive instances of circling. As illustrated in Fig. 3C, Row 2, this method calculates the angle of the mouse's body in each frame using the vector from labeled tailbase position to labeled snout position. It then screens candidate circles using bounds on duration as well as minimum and maximum total rotation.

The optimized Time-Angle method reached an F1 score of approximately 0.38 in the training set and 0.22 (95% CI 0.03–0.47) in the test set. Thus, incorporating the additional information of the tail's position resulted in a substantial increase in behavioral detection performance which nevertheless remained well below human level. However, in examining erroneous circles detected by this method (Fig. 3D, Row 2), we identified many cases in which false positives were either clearly too small, too large, or distinctly oblong. To counteract this, we sought to incorporate additional geometric information about candidate circles.

Box-angle method

To filter out cases based on the animal's snout path, we implemented a final method which discards constraints on duration and instead contains an additional step in which a candidate circle is fitted to a minimum-area rectangle to provide additional geometric information (see Fig. 3C, Row 3). The resulting 'Box-Angle' method constrains the side lengths of the resulting fitted rectangle. To avoid relying on specific information about the camera and the geometry of the recording apparatus, the minimum side length is specified relative to the median body length of the mouse as measured by tail-to-snout vector over the course of the video. Notably, during our parameter search we found that the inclusion of temporal information (duration constraints) actually resulted in a slightly worse training set performance (F1 score of 0.39 with vs 0.41 without). The best-performing parameters on the training set produced a test-set F1 score of 0.43 (95% CI 0.21–0.57). We observed that the false positives produced by this more sophisticated method do not suffer the same obvious flaws as those from the less complex techniques (Fig. 3D, Row 3),

Method performance comparison

As illustrated in Fig. 4, the straightforward Duration-Only method performs significantly lower than both human performance (p = 1.1E−11, two-tailed Wilcoxson signed rank test of automatic scores versus human scores) and the automatic Box-Angle methods (p = 2.8E−6, two-tailed Wilcoxson signed rank test between automatic scores from each method). The Time-Angle method, in turn, performs on par with the Duration-Only method (p = 0.2) and underperforms both independent human labeling (p = 4.7E−6) and the automatic Box-Angle methods

Method	Parameter	Final value (SD from mean)
Duration-only	Min duration	− 0.307
	Max duration	+ 0.381
Time-angle	Min duration	− 1.42
	Max duration	+ 0.093
	Min rotations	− 0.524
	Max rotations	+ 1.75
Box-angle	Min rotations	− 1.07
	Max rotations	0.238
	Min minor axis	− 1.23
	Max major axis	− 0.141

Table 1. Method parameters. Optimal values of features calculated from keypoint trajectories were selected for each circling identification algorithm via iterative Bayesian search based on resultant F1 scores.

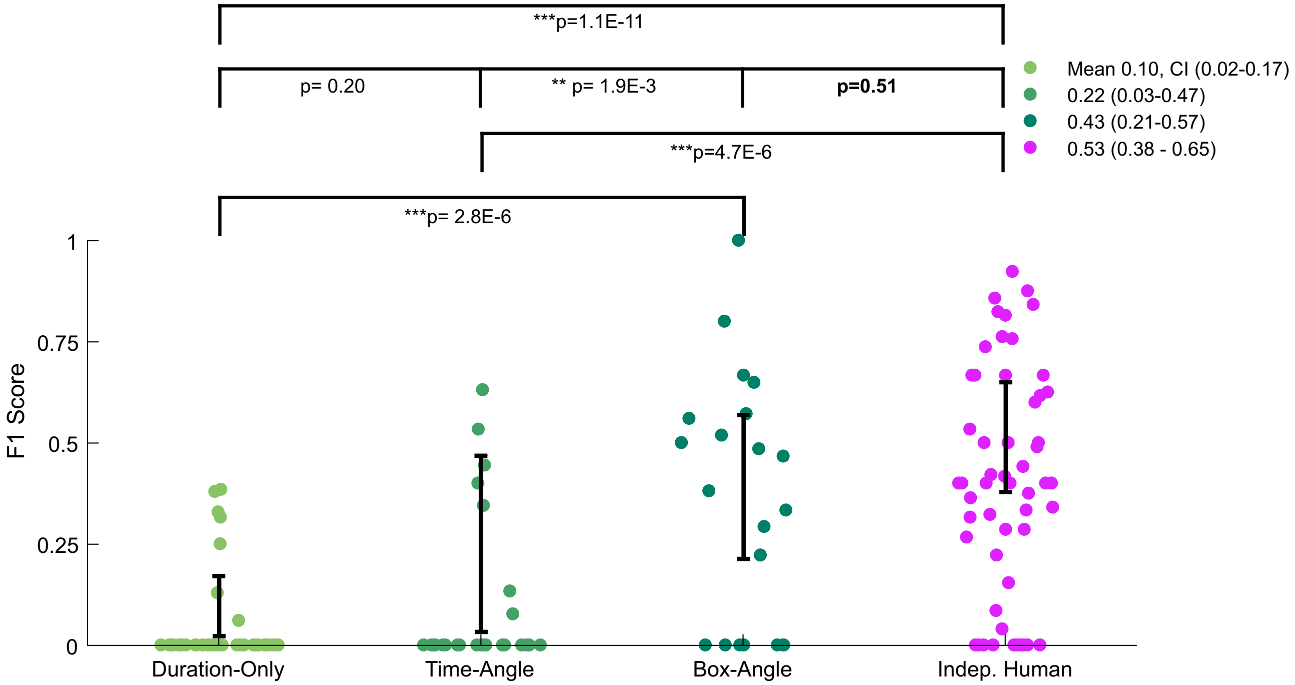


Fig. 4. Method performance comparison. After optimizing behavior detection algorithms on the human-labeled training set, each was scored on the human consensus circling labels of the test set. Each column represents one algorithm, with one dot for each test set video with a defined score. Videos for which F1 score is undefined (i.e., the automated method and human consensus both marked no circling instances) were included in confidence interval calculations but not displayed as individual datapoints. The Duration-Only and Time-Angle methods significantly underperformed independent human observers (mean and 95% CI 0.1 (0.02–0.17) and 0.22 (0.03–0.47), p = 1.1E−11 and 4.7E−6, respectively). Only the Box-Angle method reaches statistical parity (mean F1 0.43 (0.21–0.57), p = 0.51).

(p = 1.9E−3). Finally, the extensive filtering of the Box-Angle method results in a performance distribution not significantly different from human observers (p = 0.51).

Differentiation between mutant and wild-type mice

As the present work was aimed to develop a tool for detecting vestibular-mutant mice, the ultimate determinant of success is whether the presence of an automatically-detected circling instance positively identifies a video of a mutant mouse. The Box-Angle method developed above correctly reports zero circles for 20 of 24 test-set videos of wild-type mice, whereas at least one circle was scored in 16 of 24 mutant mouse videos. Thus, in our test dataset, the presence of at least one automatically-detected circle detects videos of (*Cib2*$^{-/-}$;*Cib3*$^{-/-}$) mutant mice with an F1 score of 0.73.

Impacts of dataset size

All the results reported thus far were produced using a DLC-trained convolutional neural network model, termed the "Full Dataset" model, which used 20 manually-labeled frames from each of 188 mouse behavior videos not used for human labeling. This model represents a substantial investment of experimenter effort, raising the question of whether sufficient behavioral detection performance could be achieved more easily. In particular, the publishers of DeepLabCut observed good labeling performance with substantially smaller datasets than what we use for our “Full Dataset” model[4], but it was unclear a priori what labeling quality would be necessary for successful behavior detection.

In order to investigate the minimum amount of data and thus human labor needed to obtain good automatic behavioral detection performance, we trained several DLC models using different subsets of our manually-labeled frames, detailed in Table 2. Specifically, ten models were trained for each of several training dataset sizes—half, one-quarter, and one-eighth of the full dataset. Figure 5A displays the labeling performance (root-mean-squared error, in pixels) for all 10 networks within each dataset size category. For frames within a model’s training set, these network families reached errors of 9.29 (8.13–10.73), 9.84 (8.53–11.7), and 11.02 (9.11–12.91) pixels (mean & 95% CI). For unseen frames, these errors increased to 19.37 (16.92–22.28), 12.3 (10.51–14.4), and 14.34 (12.66–15.98). Training data for each model was randomly selected to include all 20 labeled frames from the appropriate number of videos, i.e., data were shuffled by video rather than by frame. Subsequently, the ability of each model to accurately label the locations of the snout and tail-base on previously unseen videos was evaluated on a randomly-selected 50% (94 videos) of the full dataset which did not overlap with that network’s training data. Notably, while each family of networks produced larger errors on average than the full dataset model (7.82

Dataset	Full	Half	Quarter	Eighth
# Videos train	188	94	47	24
# Videos test	N/A	94	94	94
# Networks	1	10	10	10
Training RMSE pixels (mean (95%CI))	7.82	9.29 (8.13–10.73)	9.84 (8.53–11.7)	11.02 (9.11–12.91)
Testing RMSE pixels (mean (95%CI))	N/A	19.37 (16.92–22.28)	12.3 (10.51–14.4)	14.34 (12.66–15.98)
F1 score (mean (95%CI))	0.43 (0.21–0.57)	0.39 (0.17–0.54)	0.41 (0.19–0.56)	0.36 (0.14–0.52)
P-value vs full, human	N/A, 0.51	*0.03, ***1.7E−4	*0.03, ***1.4E−4	*0.02, ***3.9E−5

Table 2. Keypoint labeling performance among models. Subsets of our manually-labeled frames were used to train different neural network models using DeepLabCut. All models were initialized using the pretrained ResNet50 model available through DLC and trained for up to 100,000 iterations at a learning rate of 0.001. Performance was assessed using root-mean-squared error, in pixels, between model-assigned and manually-labeled snout and tailbase positions. $^{*}p < 0.05$, $^{***}p < 0.001$.

pixels, dashed horizontal line), reduced dataset size did not monotonically worsen labeling performance on previously unseen videos.

To determine whether these differences in pixel-wise error impacted the ability of networks to successfully identify circling behavior, we ran F1-score-optimizing parameter searches over the human-scored training videos for each network using the Box-Angle method. As described above, the Full Dataset network achieved an F1 score across the manually-screened videos of approximately 0.43. Model families trained on smaller datasets again showed a non-monotonic relationship between performance and dataset size, with half-, quarter- and eighth-sized datasets producing mean F1 scores across the test video dataset of 0.39, 0.41, and 0.36 respectively (Fig. 5B). These scores fell below the performance of the Full Dataset model (p = 0.03, 0.03, 0.02) and failed to match human performance (p = 1.7E−4, 1.4E−4, and 3.9E−5), suggesting that even small reductions in tracking quality were functionally significant at the behavioral analysis level.

Discussion

In the present study, we developed a technique to automatically detect circling behavior in videos of freely exploring mice using readily-available, consumer-grade hardware and open-source software. Importantly, this method is applicable to analyzing keypoint tracking captured by any method, whether open-source or proprietary, so long as it includes snout and tailbase positions. This makes it a convenient, quantitative tool to screen mice for circling behavior according to specific, objective criteria. More generally, our results suggest that similar procedures to develop consensus behavioral labels among human observers could be straightforwardly applied to enable effective, automatic detection of other behaviors of interest. The development of such quantitative tools with low barriers to entry is essential for the comparative analysis of behavior, as it expands the number of research groups able to produce directly inter-comparable data.

Limitations of prior approaches

Studies which report circling have varied widely in methods of detection and analysis, ranging from qualitatively reporting the presence or absence of circling[5,23–25] to counting rotations per minute using manual or automated video tracking[9,10,29,30]. Manual methods place substantial demands on human time; in our experience, marking circling behavior in a given video took on the order of four to five times as long as the video itself on average, owing to the need to e.g. play videos slowly to avoid missing instances, pause videos to make notes, and replay sections of videos to ensure the timing of instances were precisely noted. Indeed, both the low overall F1 scores and the inter-video variability in those scores observed in comparisons among human observers (Fig. 2) suggest that an automated detection system for circling behavior might be advantageous in that, in addition to not becoming distracted or fatigued[1–3], it would give a consistent output for a given video.

Software to track rodents in videos is commercially available, but such products face issues of both price and opacity; as closed-source software, they limit the ability of researchers to examine how specific results were generated and to customize or otherwise modify those underlying methods[34]. In the present study, we chose to use the open-source markerless feature-tracking toolbox DeepLabCut to track the positions of keypoints on animals. However, the analysis applied to this positional information is agnostic to the tracking method used (see Discussion: "Availability and use"). Existing automated methods using video analysis face limitations in quantifying circling behavior due to false positives arising from grooming or exploratory turns. Our work specifically addresses this issue by incorporating filters for excluding false positives using carefully tuned geometric parameters—by tracking only two key points on the mouse and applying straightforward algorithms, we achieved behavioral labeling performance (F1 score 0.43) statistically similar to that of independent human coders (0.53, p = 0.51, Fig. 5B).

Availability and use

Our Full-Data model, the Python script to detect circles from keypoint positions, and an associated Anaconda environment file can be found at our GitHub (https://github.com/CullenLab/CirclingDetection) along with a step-by-step guide to installing and using the system. Use requires only basic familiarity with the command line or with integrated development environments. To use our code, we recommend installing Anaconda (see anaconda.com) and using it to create a Python environment from the '`circlingmouse.yaml`' file included in the GitHub Repository (i.e., using '`conda env create -n circlingmouse -f circlingmouse.yaml`'). Tracking data in the form of a comma-separated-value (CSV) file, whether from DLC or other sources, should be placed within the downloaded folder. Although we chose to develop our methodology using DLC, the detection algorithm can in principle be employed using any technique to track the snout and tail base. Those utilizing other tracking methods should note that while the '`Circling_BoxAngle.py`' script assumes a particular configuration of columns in the files to be analyzed (described further in the repository documentation), CSVs produced by other tracking methods can be straightforwardly modified to fit this scheme. The code can either be run from the command line ('`conda activate circlingmouse`', then '`python Circling_BoxAngle.py $Your_Tracking_CSV_Here$.csv`' from within the downloaded Repository), or run via one of the development environments Anaconda offers.

In our analysis, we found that human-level performance required accurate keypoint tracking and extensive filtering of false positives. As a result, we encourage the use of our pretrained Full Dataset model by future experimenters wherever possible to make best use of human time and effort. Furthermore, regardless of the tracking method used, we recommend applying the included Box-Angle method code for experiments focused on circling, as it incorporates the most information about the animal's movements and was the only technique to reach parity with human observers. As the analysis is applied directly to the labeled coordinates of the snout and tailbase keypoints, the Box-Angle method will be usable alongside any high-quality tracking method. As noted below, however (see Discussion: "Directions for future work"), our tools may not be appropriate for e.g. mice of varied colors or in complex environments, nor for behaviors whose detection requires tracking additional keypoints.

Conclusions and implications

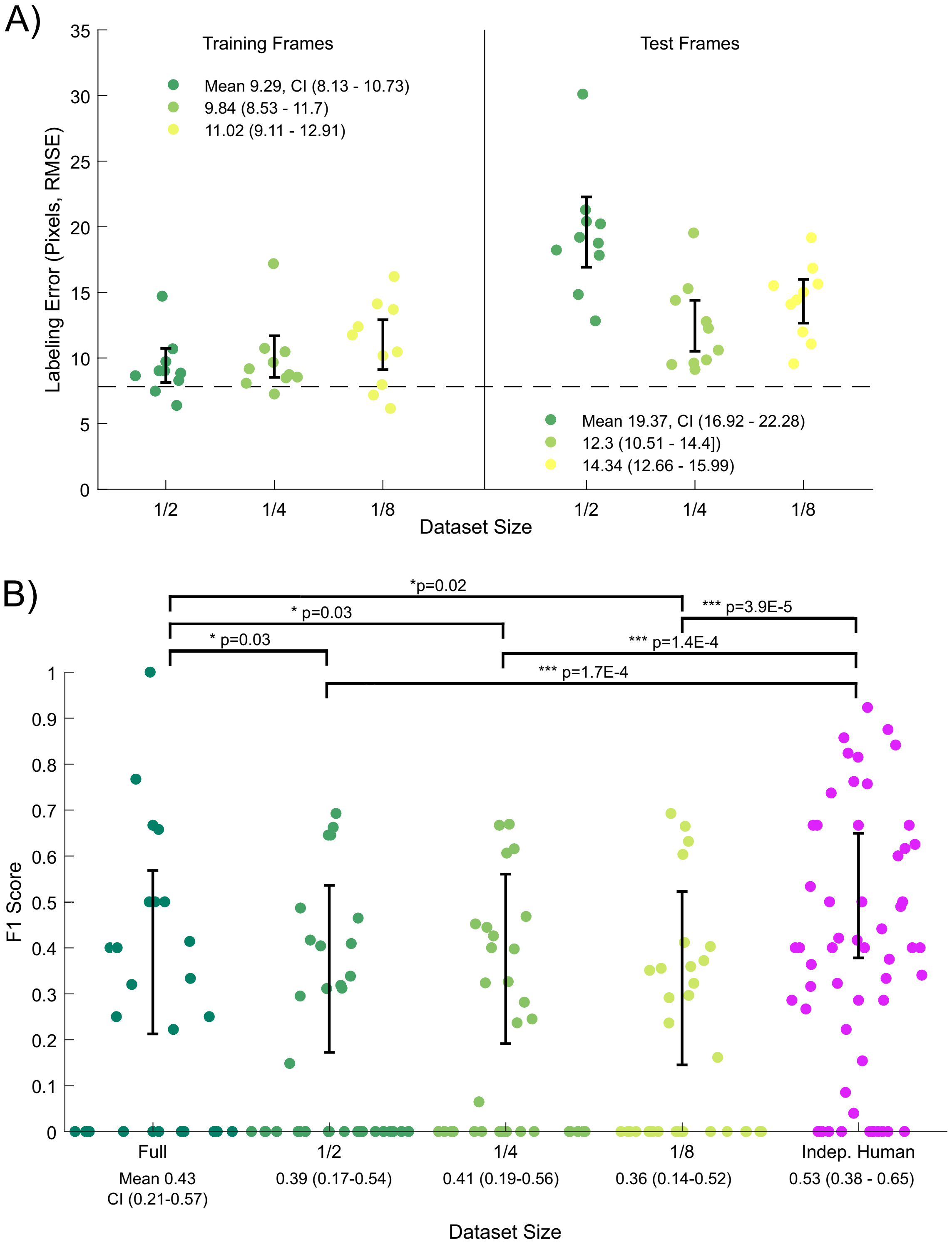
Here, we present an open-source tool which identifies circling, a specific, abnormal locomotor behavior which distinguishes between wild-type mice and many mouse models of interest to biological researchers, in this case a strain of vestibular mutant mice. We first revealed the substantial variability exhibited by human observers in manually marking this behavior, then developed a tool which uses simple, directly interpretable parameters to achieve statistical parity with independent human observers when scored against human consensus.

Emerging video-based technologies that facilitate the tracking of key body features are well-suited to the development of accessible methods for objectively quantifying behavior, including the effects of vestibular loss on those behaviors. Such studies are clinically relevant in light of aging populations, as vestibular dysfunction substantially increases fall risk and causes symptoms including dizziness, vertigo, nausea, and blurred vision. In adults over 40, its prevalence has been estimated as ranging from more than one in twenty[35] (using vestibular-specific clinical measures) to more than one in three people over[36] (using broader balance-related symptoms).

Mouse models are advantageous for studies of treatments and causes of vestibular impairment, due in part to the ability of researchers to genetically engineer new strains using increasingly sophisticated tools[37,38]. Recent developments in mouse genetic engineering, involving the creation of transgenic and knockout mutant mice, have provided a novel opportunity to study the relationship between genes and behavior[3]. For example, many mutant mice strains that have been characterized with an underlying impairment of peripheral vestibular function display a circling behavioral phenotype. While there exist a number of non-invasive methods to detect the existence of vestibular dysfunction in mice, such as the rota-rod and balance beam tests[6], the variety of vestibular-loss circling mouse models suggests screening for circling may serve as a convenient screening tool for identifying novel models of vestibular dysfunction. In the present study, we specifically assessed the circling behavior (*Cib2*$^{-/-}$;*Cib3*$^{-/-}$) dual knockout mice[31–33]. CIB2 is found in the stereocilia tips of the receptor cells within the vestibular sensory organs[39] (i.e., vestibular hair cells), suggesting the circling behavior observed in these mice results specifically from deficits in peripheral mechanotransduction. Notably, the strain studied in the present work is just one example of a large number of strains with mutations homologous to subtypes of Usher syndrome[40], including deaf circler[26], waltzer[29], Ames waltzer[41], and Jackson shaker[42] mice. In the present study, comparing those parameters that are shared across the developed methods (i.e., duration and rotation, Table 1) reveals that the parameters which resulted in optimal performance varied depending on what additional information was included; we speculate that this variability arises from the variability of the underlying behavior. Notably, the similar performance observed when applying the optimized Box-Angle method to novel videos (0.41 vs 0.43) suggests we have successfully built a tool which was robust to this variability. Further work automatically and objectively quantifying circling behavior may reveal otherwise undetectable differences in circling parameters (e.g. differences in frequency, rotational velocity, and geometry) between model strains.

An important advantage of video-based approaches to behavioral and especially vestibular neuroscience is that they are non-invasive. This stands in contrast with the emerging use of head-mounted sensors to assess motor impairments in mouse models (see for example[5,6]), which typically require an experimental surgery to securely fix the sensor to the mouse's skull. We speculate that as the spatio-temporal resolution of readily available, consumer-grade hardware continues to improve, the tradeoff in resolution between a completely noninvasive recording technique (video analysis) and a more invasive technique in which inertial sensors are mounted surgically will become less critical.

More broadly, tools for detecting and quantifying behavior which do not rely on specialized experience with software engineering will allow a wider array of research groups to directly compare analyses of neurodevelopmental differences and intervention effects. The use of shared automatic behavioral analysis tools would increase both the speed and consistency of behavioral labeling, especially when analyzing long or numerous videos. For example, the technique presented in this paper could aid in screening novel mutants for vestibular dysfunction.



Common behavior definitions may also provide inter-comparability among studies within and across research groups as well as improving the reproducibility of those studies.

As noted above, the development of this work was motivated in part by substantial variability seen among human observers when labeling behavior independently. The success of the methodology developed here argues for the broader use of tuning behavioral detection algorithms based on consensus among multiple human observers. In particular, by demonstrating the effectiveness of an automated detection system for circling behavior tuned to match group consensus, our results provide a convenient, quantitative screening method for mouse models of vestibular dysfunction, which we hope will serve as an important step toward standardized, automatic measurement of motor dysfunction in mouse models that provide more reliable measurements across studies and laboratories.

◂**Fig. 5.** Dataset size performance comparison. (**A**) Labeling performance (error, in pixels) for each of 10 trained networks on datasets of progressively smaller sizes. All dataset sizes resulted in greater labeling error than the Full Dataset model (dashed horizontal line), particularly for frames not seen during training (test frames). Notably, this trend was not monotonic—the set of quarter-dataset models performed better on test frames, on average, than the set of half-dataset models. Root-mean-squared errors on training set frames were (mean and 95% CI) 9.29 (8.13–10.73), 9.84 (8.53–11.7), and 11.02 (9.11–12.91) pixels respectively. For unseen frames, these errors increased to 19.37 (16.92–22.28), 12.3 (10.51–14.4), and 14.34 (12.66–15.98). Dashed horizontal line represents Full Dataset model training frame error (7.82 pixels). (**B**) To determine whether these changes in labeling quality impacted, we applied the optimized Box-Angle method to the keypoint tracking produced by each network at each dataset size. Within a dataset, the true-positive, false-positive, and false-negative scores for each video were summed to calculate a representative F1 score, plotted here as individual dots in the half-, quarter-, and eighth-sized datasets. The resulting distributions are compared to scores from the Full Dataset network (left column) and to independent human scores (right). As elsewhere, video-net combinations for which F1 score are undefined are included in confidence interval calculations but not displayed as individual datapoints. These smaller datasets underperformed the Full Dataset network (p = 0.03, 0.03, 0.02) as well as human labels (p = 1.7E−4, 1.4E−4, 3.9E−5), indicating that even small reductions in keypoint tracking quality can impact behavioral detection. $^{*}p < 0.05$, $^{***}p < 0.001$.

Directions for future work

The current study focuses on creating as simple a method as possible for detecting a specific behavior. Importantly, however, this work is limited to first-order features of the tracked keypoint paths, i.e. it does not incorporate information regarding velocity or acceleration which human observers have access to when watching behavior videos. Further refinement of automated methods may take advantage of these or other higher-order features. Additionally, it seems clear that incorporating a larger number of keypoint labels will be critical for richer examination of rodent posture and gait[43], and has previously been incorporated into other analyses of circling mice specifically[44]. As our Full Dataset model was trained for only the two keypoints of interest to us, future researchers seeking to apply our findings to more complex behaviors will likely be unable to employ the specific computer vision model used here. Our model is further limited in that its training data consists of black-furred mice on white backgrounds (see example frames, Fig. 3A); we expect this will result in degraded tracking of animals of varied colors or on low-contrast backgrounds.

It is noteworthy that the concepts underlying our approach can be readily applied to other behaviors of interest to researchers. Specifically, by first creating a set of occurrence times coded independently and then constructing a set of consensus occurrence times, we were able to directly quantify human-level performance. In principle, this enables working with a wide range of behaviors which may be difficult to define explicitly ahead of time but which we 'know when we see them'. Indeed, the concept is not limited in application to the analysis of visual information. Rather, in any situation where human observation is currently required to disentangle ambiguity in behavioral data, it is possible to optimize automated analysis to match agreement among multiple independent observers as well as possible to avoid incorporating biases or quirks of any one observer. This is equally true for the identification of freezing behavior[45], which must be distinguished from simply remaining still just as circling must be distinguished from normal exploration; analysis of social interactions among animals[46], the dimensionality of which will rise exponentially with the number of individuals being considered; or attempting to detect a particular category of vocalization[47,48], which may occur in noisy environments or be ambiguous as to whether a call represents e.g. a warning of approaching predators versus a warning to deter conspecific competitors. Although differences will arise in the choice of features being used as input to automated detection methods, which will depend on the behavior and modality of interest (e.g. rotational velocity for circling detection versus linear velocity for freezing detection), in all cases the process of generating consensus from independently-marked behavior timing will be similar.

Methods

Animal care and housing

Five adult wild-type (C57BL/6) mice (Jackson Laboratories) and five adult mutant (*Cib2*$^{-/-}$;*Cib3*$^{-/-}$) mice were used in this study. Generation of mutant mice is described in a separate paper[33]. Animals were group-housed with their littermates on a 12:12 h light: dark cycle at 20 °C with ad libitum access to food and water. All animal procedures complied with the ARRIVE guidelines and were carried out in accordance with National Institutes of Health Guide for the Care and Use of Laboratory Animals (NIH Publications No. 8023, revised 1978) and were approved by the Institutional Animal Care and Use Committees (IACUCs) at University of Maryland (protocol #0420002).

Data generation

We recorded videos of 5 wild-type mice and 5 (*Cib2*$^{-/-}$;*Cib3*$^{-/-}$) dual knockout circling mice during single-animal exploration of a cylindrical arena 30 cm in diameter, similar to ones used in past studies[8,9,15]. For each mouse, we recorded four 2-min videos at 60 fps in each of six recording conditions—low (~ 4 lx), moderate (~ 375 lx), or bright lighting (~ 875 lx), each with the camera either near to (60 cm) or far from (100 cm) the animal. These were selected to span a broad range of potential experimental setups and to provide varied data for neural network training; conditions are illustrated in Fig. 1. Unfortunately, during the course of the study one set of video data (four videos of a mutant mouse in low-light, near-camera conditions) became corrupted; as this mouse was not the one used for human behavioral labeling, we believe the impacts on the study's results were negligible. Our total set of 236 videos thus consisted of 116 videos of wild-type mice and 120 videos of mutant mice.

Gold standard development

Assessing the effectiveness of our tool required a ground truth against which we could compare the automatic detection of circling behavior. In order to compare potential automatic methods against the current gold standard of human behavior labeling, we set aside all videos of one randomly selected mutant and wild-type mouse for manual screening (4 videos in each of 6 conditions; 24 videos each). Additionally, within the remaining videos, 24 of mutant mice and 24 of wild-type mice were randomly selected for manual screening. Three observers independently marked times at which circling behavior occurred. Observers were instructed to mark complete rotations during bouts of circling behavior but not during normal exploratory behavior shared by wild-type mice (e.g., turning around after reaching the edge of the arena). Videos were not labeled with information regarding mouse genotype. These behavior labels were compared to produce a set of consensus occurrences.

Consensus behavior labels were established by comparing these independently-marked times. To accommodate variations in the precise timing of marked circling occurrences, times within 6 frames (0.1 s) of one another were counted as the same instance, a timeframe chosen to cover 95% of the observed variation between independent observers. If all three observers agreed by this quantitative definition, the average of these matched times was taken as a consensus instance. Independent labels were not subject to modification during this process. Subsequently, circling instances detected by either human observers or automatic methods were counted as true positives if they fell within 0.1 s of a consensus-labeled occurrence.

Algorithm development and parameter search

We compared three algorithms for detecting circling using labeled locations of two keypoints on the body of freely exploring mice. These methods first search the path of the mouse's snout for cases where it crosses over itself as candidates for instances of circling, then apply thresholds to features of the mouse's path to filter out false positives. The 'Duration-Only' method uses only thresholds on minimum and maximum duration. The 'Time-Angle' method additionally excludes candidate circles based on minimum and maximum angular change of the vector from the tailbase to nose. Finally, the 'Box-Angle' method removes the duration constraints and instead considers the tail-nose vector rotation and the lengths of the major and minor axes of a minimum-bounding rectangle fit to the snout path. As described in Table 1, we thus needed to optimize either 2 or 4 parameters depending on the method being considered. Listed parameter ranges were explored independently for each method rather than the common parameters being frozen as new parameters were added.

The statistics of animal movement can be highly variable. For example, Fig. 3B illustrates the different distributions of path collision parameters observed in two videos in our training dataset. To ensure robustness against behavioral variability between animals or of the same animal at different times, we selected a method for excluding false positives based on the distribution of duration, total rotation, and size within a given video. Specifically, for a given method applied to a specific video n, Gaussian kernel density estimation k for the parameter(s) θ of interest is performed on the set of all M collisions detected in that video's tracking data across possible parameter values x.

$$k_\theta^n(x) = \frac{1}{Mh}\sum_{m=1}^{M}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x-x_m)^2}{2h^2}\right)$$

A combined probability density q consisting of a weighted sum of an exponential with parameter λ and a Gaussian probability density function with mean μ and standard deviation σ, with weights v and w, respectively, is fitted to this density estimate via least-squared-error.

$$\phi_\theta^n = \left\{v_\theta^n,\ \lambda_\theta^n,\ w_\theta^n,\ \mu_\theta^n,\ \sigma_\theta^n\right\}$$

$$q_\theta^n\left(x;\ \phi_\theta^n\right) = v_\theta^n\lambda_\theta^n\exp\left(-\lambda_\theta^n x\right) + w_\theta^n\frac{1}{\sqrt{2\pi(\sigma_\theta^n)^2}}\exp\left(-\frac{(x-\mu_\theta^n)^2}{2(\sigma_\theta^n)^2}\right),$$

$$min_{\phi_n}\left\|q_\theta^n\left(x;\ \phi_\theta^n\right) - k_\theta^n(x)\right\|^2$$

Using the Gaussian component of the fitted distribution, thresholds for rejecting a collision as a circling candidate are then specified in terms of standard deviations above or below the mean.

Lacking an explicit representation of the derivative of F1 score with respect to these threshold parameters, we instead employed constrained Bayesian optimization to identify well-performing parameter combinations[49]. This process first constructs an estimate of the function to be optimized (in our case, F1 score on the training set) based on random sampling of the parameter space, then iteratively explores the location expected the next most exploration-worthy positions in the parameter space and the posterior distributions to find a parameter combination close to the optimal combination. For each parameter combination, we first explore the parameter space via 2000 random samples and then apply Bayesian optimization with 1000 iterations to search for the best parameter combination. Thresholds were constrained to within ± 3 standard deviations. The parameter combination values with the highest train F1 score are reported, evaluated on the test set, and compared with human performance.

Neural network training

We chose to use DLC, an open-source tool for training deep convolutional neural networks to recognize user-labeled image features, to track the locations of mouse-body keypoints due to its accessibility, as it can be straightforwardly used by researchers with little machine learning experience with consumer-grade computing hardware.

As described above, from the 240 free behavior videos originally recorded, 48 were held out as a test set. Four videos unfortunately became corrupted and were not used in the study. From each of the 188 remaining videos not used in manual behavior screening, we labeled 20 random frames with the positions of the mouse's snout and the base of its tail. We utilized data augmentation in the form of the '*imgaug*' dataloader included in DLC, which applies perturbations during network training such as cropping, blurring, and rotating training images. We refer to this as the "Full Dataset" model, in contrast to models trained on subsets of this data.

To investigate the amount of data necessary to reach a plateau in performance, we used different subsets of our 188-video dataset to train several different DLC models. We compared networks trained with one-half, one-quarter, and one-eighth of the full training dataset (97, 48, and 24 videos, respectively). For each such network, the appropriate number of videos were randomly assigned to a training dataset, with 97 of the remaining videos (50% of the training set) then assigned at random to assess the network's ability to label frames from unseen videos. (Table 2) Each model was initialized using a 50-layer pretrained network model (ImageNet-pretrained Resnet50) and trained for 100,000 iterations at a learning rate of 0.001. After training, each DLC network was run on human-scored videos to produce position traces to be analyzed using the Box-Angle method described above.

Statistical analyses

To balance the need to avoid both false positive and false negative errors, we used F1 score to assess three methods of detecting circling behavior in labeled paths, calculated as follows:

$$F_1 = \frac{True\ Positives}{True\ Positives + 0.5 * \left(False\ Positives + False\ Negatives\right)}$$

Notably, in cases where all three of (True Positives, False Positives, False Negatives) are equal to zero, all these metrics F1 score is undefined. When this occurs, the specific set of scores in question is excluded from p-value but not confidence interval calculations. This is the case on many of the manually-screened behavior videos, in which all three of the human observers marked no instances of circling and well-tuned automatic methods do not score any false positives. P-values for differences between F1 score distributions were calculated using a two-tailed Wilcoxson signed rank tests.

As we could not assume normality of performance distributions a priori, confidence intervals were calculated via bootstrap[50], i.e., creating a large number of pseudo-datasets of the same size as the original by repeatedly sampling with replacement from the video scores generated by a given method or human labeler. Specifically, this process involves repeatedly drawing N samples with replacement from among the N scores generated by a particular method, individual, or family of DLC models. The F1 score for a particular draw is then calculated from the sum of true positives, false positives, and false negatives for that draw. In all cases where a confidence interval is reported, we sampled one million such pseudo-datasets. Reported intervals are calculated as the 2.5th and 97.5th percentiles of the resulting population of F1 scores.

Data availability

The datasets used and/or analyzed during the current study available from the corresponding author on reasonable request.

Received: 15 June 2023; Accepted: 29 August 2024
Published online: 08 September 2024

References

1. Segalin, C. *et al.* The Mouse Action Recognition System (MARS) software pipeline for automated analysis of social behaviors in mice. *Elife* **10**, e63720. https://doi.org/10.7554/eLife.63720 (2021).
2. van den Boom, B. J. G., Pavlidi, P., Wolf, C. J. H., Mooij, A. H. & Willuhn, I. Automated classification of self-grooming in mice using open-source software. *J. Neurosci. Methods* **289**, 48–56. https://doi.org/10.1016/j.jneumeth.2017.05.026 (2017).
3. von Ziegler, L., Sturman, O. & Bohacek, J. Big behavior: Challenges and opportunities in a new era of deep behavior profiling. *Neuropsychopharmacology* **46**, 33–44. https://doi.org/10.1038/s41386-020-0751-7 (2021).
4. Mathis, A. *et al.* DeepLabCut: Markerless pose estimation of user-defined body parts with deep learning. *Nat. Neurosci.* **21**, 1281–1289 https://doi.org/10.1038/s41593-018-0209-y (2018).
5. Ono, K. *et al.* Retinoic acid degradation shapes zonal development of vestibular organs and sensitivity to transient linear accelerations. *Nat. Commun.* **11**, 63. https://doi.org/10.1038/s41467-019-13710-4 (2020).
6. Chang, H. H. V., Morley, B. J. & Cullen, K. E. Loss of α-9 nicotinic acetylcholine receptor subunit predominantly results in impaired postural stability rather than gaze stability. *Front. Cell. Neurosci.* **15**, 799752. https://doi.org/10.3389/fncel.2021.799752 (2021).
7. Ungerstedt, U. & Arbuthnott, G. W. Quantitative recording of rotational behavior in rats after 6-hydroxy-dopamine lesions of the nigrostriatal dopamine system. *Brain Res.* **24**, 485–493. https://doi.org/10.1016/0006-8993(70)90187-3 (1970).
8. Hefti, F., Melamed, E., Sahakian, B. J. & Wurtman, R. J. Circling behavior in rats with partial, unilateral nigro-striatal lesions: Effect of amphetamine, apomorphine, and DOPA. *Pharmacol. Biochem. Behav.* **12**, 185–188. https://doi.org/10.1016/0091-3057(80)90353-6 (1980).
9. Torello, M. W., Czekajewski, J., Potter, E. A., Kober, K. J. & Fung, Y. K. An automated method for measurement of circling behavior in the mouse. *Pharmacol. Biochem. Behav.* **19**, 13–17. https://doi.org/10.1016/0091-3057(83)90303-9 (1983).
10. Bonatz, A. E., Steiner, H. & Huston, J. P. Video image analysis of behavior by microcomputer: Categorization of turning and locomotion after 6-OHDA injection into the substantia nigra. *J. Neurosci. Methods* **22**, 13–26. https://doi.org/10.1016/0165-0270(87)90084-7 (1987).

11. Ishiguro, A., Inagaki, M. & Kaga, M. Stereotypic circling behavior in mice with vestibular dysfunction: Asymmetrical effects of intrastriatal microinjection of a dopamine agonist. *Int. J. Neurosci.* **117**, 1049–1064. https://doi.org/10.1080/00207450600936874 (2007).
12. Iwamoto, E. T. & Way, E. L. Circling behavior and stereotypy induced by intranigral opiate microinjections. *J. Pharmacol. Exp. Ther.* **203**, 347–359 (1977).
13. Giuliani, F., Vernay, A., Leuba, G. & Schenk, F. Decreased behavioral impairments in an Alzheimer mice model by interfering with TNF-alpha metabolism. *Brain Res. Bull.* **80**, 302–308. https://doi.org/10.1016/j.brainresbull.2009.07.009 (2009).
14. Castillo-Mariqueo, L. & Giménez-Llort, L. Kyphosis and bizarre patterns impair spontaneous gait performance in end-of-life mice with Alzheimer's disease pathology while gait is preserved in normal aging. *Neurosci. Lett.* **767**, 136280. https://doi.org/10.1016/j.neulet.2021.136280 (2022).
15. Portmann, T. *et al.* Behavioral abnormalities and circuit defects in the basal ganglia of a mouse model of 16p11.2 deletion syndrome. *Cell Rep.* **7**, 1077–1092. https://doi.org/10.1016/j.celrep.2014.03.036 (2014).
16. Ryan, B. C., Young, N. B., Crawley, J. N., Bodfish, J. W. & Moy, S. S. Social deficits, stereotypy and early emergence of repetitive behavior in the C58/J inbred mouse strain. *Behav. Brain Res.* **208**, 178–188. https://doi.org/10.1016/j.bbr.2009.11.031 (2010).
17. DeLorey, T. M., Sahbaie, P., Hashemi, E., Homanics, G. E. & Clark, J. D. Gabrb3 gene deficient mice exhibit impaired social and exploratory behaviors, deficits in non-selective attention and hypoplasia of cerebellar vermal lobules: A potential model of autism spectrum disorder. *Behav. Brain Res.* **187**, 207–220. https://doi.org/10.1016/j.bbr.2007.09.009 (2008).
18. Lee, J. W. *et al.* Circling mouse: Possible animal model for deafness. *Comp. Med.* **51**, 550–554 (2001).
19. Mitchem, K. L. *et al.* Mutation of the novel gene Tmie results in sensory cell defects in the inner ear of spinner, a mouse model of human hearing loss DFNB6. *Hum. Mol. Genet.* **11**, 1887–1898. https://doi.org/10.1093/hmg/11.16.1887 (2002).
20. Avni, R. *et al.* Mice with vestibular deficiency display hyperactivity, disorientation, and signs of anxiety. *Behav. Brain Res.* **202**, 210–217. https://doi.org/10.1016/j.bbr.2009.03.033 (2009).
21. Wang, W., Chan, E. K., Baron, S., Van De Water, T. & Lufkin, T. *Hmx2* homeobox gene control of murine vestibular morphogenesis. *Development* **128**, 5017–5029. https://doi.org/10.1242/dev.128.24.5017 (2001).
22. Lv, K. *et al.* A novel auditory ossicles membrane and the development of conductive hearing loss in Dmp1-null mice. *Bone* **103**, 39–46. https://doi.org/10.1016/j.bone.2017.06.007 (2017).
23. Alavizadeh, A. *et al.* The Wheels mutation in the mouse causes vascular, hindbrain, and inner ear defects. *Dev. Biol.* **234**, 244–260. https://doi.org/10.1006/dbio.2001.0241 (2001).
24. Makishima, T. *et al.* Inner ear dysfunction in caspase-3 deficient mice. *BMC Neurosci.* **12**, 102. https://doi.org/10.1186/1471-2202-12-102 (2011).
25. Maskey, D., Pradhan, J., Oh, C. K. & Kim, M. J. Changes in the distribution of calbindin D28-k, parvalbumin, and calretinin in the hippocampus of the circling mouse. *Brain Res.* **1437**, 58–68. https://doi.org/10.1016/j.brainres.2011.12.009 (2012).
26. Pan, B. *et al.* Gene therapy restores auditory and vestibular function in a mouse model of Usher syndrome type 1c. *Nat. Biotechnol.* **35**, 264–272. https://doi.org/10.1038/nbt.3801 (2017).
27. Isgrig, K. *et al.* Gene therapy restores balance and auditory functions in a mouse model of usher syndrome. *Mol. Ther.* **25**, 780–791. https://doi.org/10.1016/j.ymthe.2017.01.007 (2017).
28. Nist-Lund, C. A. *et al.* Improved TMC1 gene therapy restores hearing and balance in mice with genetic inner ear disorders. *Nat. Commun.* **10**, 236. https://doi.org/10.1038/s41467-018-08264-w (2019).
29. Zhao, T. *et al.* Phenotypic differences in the inner ears of CBA/CaJ and C57BL/6J mice carrying missense and single base pair deletion mutations in the Cdh23 gene. *J. Neurosci. Res.* **99**, 2743–2758. https://doi.org/10.1002/jnr.24905 (2021).
30. Rastoldo, G. *et al.* Quantitative evaluation of a new posturo-locomotor phenotype in a rodent model of acute unilateral vestibulopathy. *Front. Neurol.* **11**, 505. https://doi.org/10.3389/fneur.2020.00505 (2020).
31. Holt, J. R. *et al.* Putting the pieces together: The hair cell transduction complex. *J. Assoc. Res. Otolaryngol.* **22**, 601–608. https://doi.org/10.1007/s10162-021-00808-0 (2021).
32. Wang, X. *et al.* CIB2 and CIB3 regulate stereocilia maintenance and mechanoelectrical transduction in mouse vestibular hair cells. *J. Neurosci.* **43**, 3219–3231. https://doi.org/10.1523/jneurosci.1807-22.2023 (2023).
33. Giese, A. P. J. *et al.* Complexes of vertebrate TMC1/2 and CIB2/3 proteins form hair-cell mechanotransduction cation channels. *bioRxiv* https://doi.org/10.1101/2023.05.26.542533 (2023).
34. Sturman, O. *et al.* Deep learning-based behavioral analysis reaches human accuracy and is capable of outperforming commercial solutions. *Neuropsychopharmacology* **45**, 1942–1952. https://doi.org/10.1038/s41386-020-0776-y (2020).
35. Grill, E. *et al.* Prevalence, determinants, and consequences of vestibular hypofunction. Results from the KORA-FF4 survey. *Front. Neurol.* **9**, 1076. https://doi.org/10.3389/fneur.2018.01076 (2018).
36. Agrawal, Y., Carey, J. P., Della Santina, C. C., Schubert, M. C. & Minor, L. B. Disorders of balance and vestibular function in US adults: Data from the National Health and Nutrition Examination Survey, 2001–2004. *Arch. Intern. Med.* **169**, 938–944. https://doi.org/10.1001/archinternmed.2009.66 (2009).
37. Ohlemiller, K. K., Jones, S. M. & Johnson, K. R. Application of mouse models to research in hearing and balance. *J. Assoc. Res. Otolaryngol.* **17**, 493–523. https://doi.org/10.1007/s10162-016-0589-1 (2016).
38. Straka, H., Zwergal, A. & Cullen, K. E. Vestibular animal models: Contributions to understanding physiology and disease. *J. Neurol.* **263**(Suppl 1), S10-23. https://doi.org/10.1007/s00415-015-7909-y (2016).
39. Giese, A. P. J. *et al.* CIB2 interacts with TMC1 and TMC2 and is essential for mechanotransduction in auditory hair cells. *Nat. Commun.* **8**, 43. https://doi.org/10.1038/s41467-017-00061-1 (2017).
40. Whatley, M. *et al.* Usher syndrome: Genetics and molecular links of hearing loss and directions for therapy. *Front. Genet.* **11**, 565216. https://doi.org/10.3389/fgene.2020.565216 (2020).
41. Alagramam, K. N., Stahl, J. S., Jones, S. M., Pawlowski, K. S. & Wright, C. G. Characterization of vestibular dysfunction in the mouse model for Usher syndrome 1F. *J. Assoc. Res. Otolaryngol.* **6**, 106–118. https://doi.org/10.1007/s10162-004-5032-3 (2005).
42. Kikkawa, Y. *et al.* Mutations in a new scaffold protein Sans cause deafness in Jackson shaker mice. *Hum. Mol. Genet.* **12**, 453–461. https://doi.org/10.1093/hmg/ddg042 (2003).
43. Sheppard, K. *et al.* Stride-level analysis of mouse open field behavior using deep-learning-based pose estimation. *Cell Rep.* **38**, 110231. https://doi.org/10.1016/j.celrep.2021.110231 (2022).
44. Vidal, P.-P., Degallaix, L., Josset, P., Gasc, J.-P. & Cullen, K. E. Postural and locomotor control in normal and vestibularly deficient mice. *J. Physiol.* **559**, 625–638. https://doi.org/10.1113/jphysiol.2004.063883 (2004).
45. Gabriel, C. J. *et al.* BehaviorDEPOT is a simple, flexible tool for automated behavioral detection based on markerless pose tracking. *Elife* **11**, e74314. https://doi.org/10.7554/eLife.74314 (2022).
46. Lauer, J. *et al.* Multi-animal pose estimation, identification and tracking with DeepLabCut. *Nat. Methods* **19**, 496–504. https://doi.org/10.1038/s41592-022-01443-0 (2022).
47. Blumstein, D. T. *et al.* Acoustic monitoring in terrestrial environments using microphone arrays: Applications, technological considerations and prospectus. *J. Appl. Ecol.* **48**, 758–767. https://doi.org/10.1111/j.1365-2664.2011.01993.x (2011).
48. Kahl, S., Wood, C. M., Eibl, M. & Klinck, H. BirdNET: A deep learning solution for avian diversity monitoring. *Ecol. Inform.* **61**, 101236. https://doi.org/10.1016/j.ecoinf.2021.101236 (2021).
49. Gardner J., Kusner M., Zhixiang X., Weinberger K., Cunningham J. Bayesian optimization with inequality constraints. In *Proceedings of the 31st International Conference on Machine Learning, Beijing, China* (2014).

50. Carpenter, J. & Bithell, J. Bootstrap confidence intervals: When, which, what? A practical guide for medical statisticians. *Stat. Med.* **19**, 1141–1164. https://doi.org/10.1002/(SICI)1097-0258(20000515)19:93.0.CO;2-F (2000).

Acknowledgements

The authors wish to thank Celia Fernandez-Brillet, Olivia Leavitt Brown, Robyn Mildren, Kiara Quinn, and Kantapon Pum Wiboonsaksakul for their insightful feedback on the manuscript, Dale Roberts for assistance with Python, and Ryan Riegel for assistance with recording the initial behavior videos. This work was supported by NIH grants R01DC012564 (ZMA) and 2T32DC000023 (KEC).

Author contributions

O.R.S., A.S., E.W., and K.E.C. designed the study. Z.M.A provided animals and experimental facilities. A.S. and E.W. performed the experiments. O.R.S, A.S, and E.W. served as behavior raters. O.R.S, A.S., and C.B. wrote data analysis code. O.R.S. analyzed the data and prepared the figures and tables. O.R.S and K.E.C. wrote the paper with input from A.S., E.W., C.B., and Z.M.A. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to K.E.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.