The Mouse Action Recognition System (MARS): a software pipeline for automated analysis of social behaviors in mice

Cristina Segalin[1], Jalani Williams[1], Tomomi Karigo[2], May Hui[2], Moriel Zelikowsky[2,3], Jennifer J. Sun[1], Pietro Perona[1], David J. Anderson[2,4], and Ann Kennedy[2,5]

[1]Department of Computing & Mathematical Sciences
California Institute of Technology
Pasadena, CA, 91125 USA

[2]Division of Biology and Biological Engineering 156-29
TianQiao and Chrissy Chen Institute for Neuroscience
California Institute of Technology
Pasadena, CA, 91125 USA

[3] Current Address: Department of Neurobiology and Anatomy
University of Utah
Salt Lake City, UT, 84112 USA

[4]Howard Hughes Medical Institute
California Institute of Technology
Pasadena, CA, 91125 USA

[5]Author for correspondence. Email: kennedya@caltech.edu

Abstract

The study of social behavior requires scoring the animals' interactions. This is generally done by hand— a time consuming, subjective, and expensive process. Recent advances in computer vision enable tracking the pose (posture) of freely-behaving laboratory animals automatically. However, classifying complex social behaviors such as mounting and attack remains technically challenging. Furthermore, the extent to which expert annotators, possibly from different labs, agree on the definitions of these behaviors varies. There is a shortage in the neuroscience community of benchmark datasets that can be used to evaluate the performance and reliability of both pose estimation tools and manual and automated behavior scoring.

We introduce the Mouse Action Recognition System (MARS), an automated pipeline for pose estimation and behavior quantification in pairs of freely behaving mice. We compare MARS's annotations to human annotations and find that MARS's pose estimation and behavior classification achieve human-level performance. As a by-product we characterize the inter-expert variability in behavior scoring. The two novel datasets used to train MARS were collected from ongoing experiments in social behavior, and identify the main sources of disagreement between annotators. They comprise 30,000 frames of manual annotated mouse poses and over 14 hours of manually annotated behavioral recordings in a variety of experimental preparations. We are releasing this dataset alongside MARS to serve as community benchmarks for pose and behavior systems. Finally, we introduce the Behavior Ensemble and Neural Trajectory Observatory (Bento), a graphical interface that allows users to quickly browse, annotate, and analyze datasets including behavior videos, pose estimates, behavior annotations, audio, and neural recording data. We demonstrate the utility of MARS and Bento in two use cases: a high-throughput behavioral phenotyping study, and exploration of a novel imaging dataset. Together, MARS and Bento provide an end-to-end pipeline for behavior data extraction and analysis, in a package that is user-friendly and easily modifiable.

Introduction

The brain evolved to guide survival-related behaviors, which frequently involves interaction with other animals. Gaining insight into brain systems that control these behaviors requires recording and manipulating neural activity while measuring behavior in freely moving animals. Recent technological advances, such as miniaturized imaging and electrophysiological devices has enabled the recording of neural activity in freely behaving mice[1-3]—however to make sense of the recorded neural activity, it is also necessary to obtain a detailed characterization of the animals' actions during recording, usually via manual scoring of animals' actions[4-6]. A typical study of freely behaving animals can produce from tens of hours to several days' worth of video[7-9]. Scoring for social behaviors often takes human annotators 3-4x the video's duration to annotate; for long recordings there is high risk of "drift" in the style and rigor of annotation. It is also unclear to what extent different human annotators within and between neuroscience labs agree on the definitions of behaviors, especially the precise timing of behavior onset/offset. When behavior is being analyzed alongside neural recording data, it is also usually unclear whether the set of behaviors that were chosen to annotate are a good fit for explaining the activity of a neural population, or whether other, unannotated behaviors with clearer neural correlates may have been missed.

An accurate, sharable, automated approach to scoring animals' behavior is thus needed, both to enable behavioral measurements in large scale experiments, and to help make behavior scoring objective and reproducible. Automation of behavior classification using machine learning methods poses a potential solution to both the time demand of annotation and to the risk of inter-individual differences in annotation style. We present the Mouse Action Recognition System (MARS), and its accompanying data analysis interface, the Behavior Ensemble and Neural Trajectory Observatory (Bento), a pair of tools

enabling high-performance automated behavior classification in pairs of interacting mice. Our system includes: 1) a novel computer vision system to track pairs of interacting mice and estimate their poses in terms of a set of anatomical keypoints, 2) a set of trained supervised classifiers capable of detecting bouts of attack, mounting, and social investigation with high accuracy, 3) an annotation interface for scoring new behaviors of interest and training additional behavior classifiers, and 4) a graphical interface for visualizing behavior annotations and MARS output alongside behavior video and recorded neural activity and audio. Taken together, MARS and Bento provide an end-to-end system for extracting behaviorally meaningful signals from video data, and interpreting these signals alongside neural recordings.

Related Work

Automated tracking and behavior classification can be broken into a series of computational steps, which may be implemented separately, as we do, or combined into a single module. First, animals are detected, producing a 2D/3D centroid, blob, or bounding box that captures the animal's location, and possibly its orientation. When animals are filmed in an empty arena, a common approach is to use background subtraction to segment animals from their environments[9]. Deep networks for object detection (such as Inception Resnet[10], Yolo[11], or Mask R-CNN[12]) may also be used. Some behavior systems, such as Ethovision[13], MoTr[14], idTracker[15], and previous work from our group[16], classify behavior from this location and movement information alone.

If multiple animals are tracked, each animal identity must be detected, located, and identified consistently over the duration of the video. Altering the appearance of individuals using paint or dye, or by selecting animals with differing coat colors, helps this task[14,17]. In cases where these manipulations are not possible, animal identity can in some cases be tracked by identity-matching algorithms such as the Hungarian method[9].

Second, the posture (pose) of the animal is computed and tracked across frames. A pose estimate reflects the position and identity of multiple tracked body parts, either in terms of a set of anatomical "keypoints"[18], shapes[19,20], or a dense 2D or 3D mesh[21]. Keypoints are typically defined based on anatomical landmarks (nose, ears, paws, digits) and their selection is determined by the experimenter depending on the recording setup and type of motion being tracked.

Animal tracking and pose estimation systems have evolved in step with the field of computer vision. Early computer vision systems relied on specialized data acquisition setups using multiple cameras and/or depth sensors[16], and were sensitive to sensitive to changes in experimental conditions. More recently, systems for pose estimation based on machine learning and deep neural networks, including DeepLabCut[22], LEAP[23], and DeepPoseKit[24], have emerged as a flexible and accurate tool in behavioral and systems neuroscience. These networks are more accurate and more adaptable to recording changes than their predecessors[25], although they typically require an initial investment in creating labeled training data before they can be used.

Pose estimation remains only a partial solution to the problem of behavior analysis: once raw animal pose data is acquired, a third step is to produce an interpretation of behavior (**Fig. 1A**). Several methods have been introduced for analyzing the actions of animals in an unsupervised or semi-supervised manner, in which behaviors are identified by extracting features from the animal's pose and performing clustering or temporal segmentation based on those features, including Moseq[26], MotionMapper[27], and multiscale unsupervised structure learning[28]. Unsupervised techniques are said to identify behaviors in a "user-unbiased" manner (although the behaviors identified do depend on how pose is pre-processed prior to clustering). Thus far they are most successful when studying individual animals in isolation.

In multi-animal experiments, some systems detect user-defined functions of the animals' poses, such as head-to-head or head-to-tail contact[29,30]. An alternative to this approach is supervised learning, in which manually scored examples ("training data") are used to train a classifier to detect social or nonsocial behaviors[31-33]. The creation of a supervised classifier requires collection of a large number of training examples, however supervised learning has the advantage that the experimenter has more control over what actions are detected. Our goal with MARS was to detect complex and temporally structured social behaviors that were previously determined to be of interest to experimenters—attack, mounting, and close investigation—therefore MARS takes a supervised learning approach. Supervised behavior classification can also be performed directly from video frames, forgoing the animal detection and pose estimation steps [34,35]. This is usually done by adopting variations of CNN to classify frame by frame actions or combining CNN-RNN architectures that classify the full video as an action or behavior. We did not opt for this approach with MARS, because we find that the intermediate step of pose estimation is useful in its own right for analyzing finer features of animal behavior and is more interpretable than features extracted by CNNs directly from video frames.

The Mouse Action Recognition System (MARS) and the Behavior Ensemble and Neural Trajectory Observatory (Bento)

We introduce the Mouse Action Recognition System (MARS), a software platform for high-throughput behavioral profiling and data analysis in pairs of socially interacting mice. The MARS pipeline includes deep neural networks for mouse detection and pose estimation, followed by custom Python code for automated detection of social behaviors by an ensemble of trained classifiers. Our pose model is robust to occlusions caused by animal interactions and neural recording devices and cables, and is trained on a new, high-quality pose annotation dataset. The MARS pipeline is accompanied by a user interface for the annotation and visualization of joint neural and behavioral datasets, the Behavior Ensemble and Neural Trajectory Observatory (Bento).

Contributions

The contributions of this paper are as follows:

Data. MARS pose estimators are trained on a novel corpus of manual pose annotations in top- and front-view video of pairs of mice engaged in a standard resident-intruder assay[36]. These data include a variety of experimental manipulations of the resident animal, including mice that are unoperated, cannulated, or implanted with fiberoptic cables, fiber photometry cables, or a head-mounted microendoscope, with one or more cables leading from the animal's head to a commutator feeding out the top of the cage.

Anatomical landmarks (`keypoints' in the following) in this training set are manually annotated by five human annotators, whose labels are combined to create a "consensus" keypoint location for each image. Nine anatomical keypoints are annotated on each mouse in the top view, and thirteen in the front view (two keypoints, corresponding to the midpoint and end of the tail, are included in this dataset but were omitted in training MARS due to high annotator noise.) We investigate the dependence of inter-annotator variability on experimental conditions, and benchmark MARS pose estimators against human performance. The MARS pose annotation dataset will be made publicly available as a community tool for benchmarking of pose estimation algorithms.

MARS also includes three supervised classifiers trained to detect attack, mounting, and close investigation behaviors in tracked animals. These classifiers were trained on 7.3 hours of behavior video, 4 hours of which were obtained from animals with a cable-attached device such as a microendoscope. Separate evaluation (3.65 hours) and test (3.37 hours) sets of videos were used to constrain training and evaluate

MARS performance, giving a total of over 14 hours of video (ED Fig. 1). All videos were manually annotated on a frame-by-frame basis by a single trained human annotator. To evaluate inter-annotator variability in behavior classification, we collected frame-by-frame manual labels of animal actions by eight trained human annotators on a dataset of ten 10-min videos. We examine inter-annotator variability in behavior identification, and benchmark MARS behavior classifiers against human performance. Videos used for training and testing of MARS behavior classifiers, including the ten-video dataset and all accompanying annotations, will also be made publicly available.

MARS pose and behavior datasets can be found at: https://neuroethology.github.io/MARS/ under "datasets".

Approach. The MARS pipeline follows a modular design that separates behavior classification into four stages: animal (mouse) detection, pose estimation, pose feature extraction, and behavior classification (**Fig 1B**). This design allows the mouse detection, pose estimation, and behavior classification elements of the MARS pipeline to be retrained separately, simplifying troubleshooting and improving generalizability of MARS to novel experimental setups. We train separate binary classifiers for behavior of interest (attack, mount, and close investigation), to allow for easier future addition of classifiers for other behaviors. MARS's custom-designed pose features expand on our previously designed feature set[16] to provide a rich description of animals' movements and postures. These features are the basis of MARS's supervised behavior classifiers, and can also be visualized in Bento to gain further insight into animals' behavior. We demonstrate the power of MARS as a system for quantifying social behavior by presenting trained classifiers for three social behaviors of interest: attack, mounting, and close investigation.

Software. MARS is an open-source, Python-based software suite that can be run on a Linux desktop computer equipped with Tensorflow and a graphical processing unit (GPU). It is distributed as a conda environment with an accompanying code library, and supports both Python command-line and GUI-based usage (**ED Fig 2**). The MARS GUI allows users to select a directory containing videos, and will produce as output a folder containing bounding boxes, pose estimates, features, and predicted behaviors for each video in the directory. The Python command line interface allows added flexibility, including an option for users to crop videos to one or more regions containing behavior arenas, an option to specify the version of a pose or detection model to use, and an option to provide MARS with previously extracted bounding boxes produced by other tracking software.

Bento is a Matlab-based GUI that allows users to easily synchronize the display of neural recording data, multiple videos, human/automated behavior annotations, spectrograms of recorded audio, pose estimates, and 270 "features" extracted from MARS pose data—such as animals' velocities, joint angles, and relative positions. It features an interface for fast frame-by-frame manual annotation of animal behavior, as well as a tool to create annotations programmatically by applying thresholds to combinations of the MARS pose features. Bento also provides tools for exploratory neural data analysis, such as PCA and event-triggered averaging, as well as interfaces for manual and semi-automated data annotation. A separate Python implementation of Bento is currently in preparation to allow more fluid interaction of Bento with MARS.

Github repositories for MARS and Bento can be found on the MARS project website at: https://neuroethology.github.io/MARS/

Paper organization

We first outline the MARS pipeline for behavior classification, and our procedures for collecting training data for pose estimation and behavior classification. We then characterize inter-annotator variability in

mouse pose estimation, and contrast this with the performance of MARS's pose estimator. We similarly characterize inter-annotator variability in frame-by-frame labeling of mouse social behaviors, which we find to be far more subjective. We evaluate the performance of MARS for the detection of three social behaviors of interest: attack, mounting, and sniffing.

Finally, we apply MARS and Bento in two example use-cases. First, we used MARS to quantify expression of social behaviors in freely interacting mice from five different strains, including three strains with mutations engineered in genes identified by a genome-wide association study (GWAS) of autism-spectrum patients. Our automated, MARS-based analysis of this 45-hour dataset revealed multiple differences in behavioral phenotypes between these strains. Second, we used MARS to track social behavior and features of animal pose in a standard resident-intruder assay, while simultaneously recording neural population activity in a hypothalamic nucleus using a head-mounted miniature microscope to measure fluorescence changes reported by a Genetically Encoded Calcium Indicator (GECI) expressed in a specific cell type. Inspection of this dataset using Bento revealed a population of neurons responsive during MARS-identified mounting behavior, which we isolated and displayed using Bento's behavior-triggered average plugin.

Results

Characterizing variability of human pose annotations

The quality of automated pose estimation is determined by the quality of the data that was used to train the algorithm. We therefore investigated the degree of variability in human annotations of animal pose for two camera placements—filming animal interactions from above and from the front—using our previously published recording chamber[16] (**Fig 1C**). We collected 93 pairs of top- and front-view behavior videos (over 1.5 million frames per view) under a variety of lighting/camera settings, bedding conditions, and experimental manipulations of the recorded animals (**Fig 1D**). A subset of 15,000 frames were extracted from each of the top- and front-view datasets and manually labeled by trained human annotators for a set of anatomically defined keypoints on the bodies of each mouse (**Fig. 2A-B**, see Methods for description of annotator workforce). 5,000 frames in our labeled dataset are from experiments in which the black mouse was implanted with either a microendoscopic, fiber photometry, or optogenetic system attached to a cable of varying color and thickness. This focus on manipulated mice allowed us to train pose estimators to be robust to the presence of devices or cables, making our system usable in a variety of experimental applications.

To assess annotator reliability, each keypoint in each frame was annotated by five individuals; the geometric median across annotations was taken to be the ground truth location of that keypoint, as we found this approach to be most robust to outliers (**Fig 2C-F**). We quantified annotator variance in terms of the fraction of human-produced pose labels within a given radius of ground truth (**Fig. 2G-H**). We observed much higher inter-annotator variability for front view videos compared to top view videos: 86.2% of human-annotated keypoints fell within a 5mm radius of ground truth in top-view frames, while only 52.3% fell within a 5mm radius of ground truth in front-view frames (scale bar in **Fig. 2E-F**). Higher inter-annotator variability in the front view likely arises from the much higher incidence of occlusion in this view, as can be seen in the sample frames in **Fig 1D**.

Pose estimation of unoperated and device-implanted mice in the resident-intruder assay

We used our human-labeled pose dataset to train a machine learning system for pose estimation in interacting mice. While multiple pose estimation systems exist for laboratory mice[22-24], we chose to include a novel pose estimation system within MARS for three reasons: 1) to produce an adequately detailed representation of the animal's posture, 2) to allow integration of MARS with existing tools that detect mice (in the form of bounding boxes) but do not produce detailed pose estimates, and 3) to ensure high quality pose estimation in cases of occlusion and motion blur during social interactions. Pose estimation in MARS is carried out in two stages: MARS first detects the body of each mouse (**Fig 3**), then crops the video frame to the detected bounding box and estimates the animal's pose (as collection of keypoints locations) within the cropped image (**Fig 4**).

MARS's detector performs MSC-Multibox detection[37] using the Inception ResNet v2[10] network architecture. Briefly, this method produces a set of possible object detection bounding boxes in an image, each with accompanying confidence score; the model is then trained to select the box that best matches annotated ground truth, which we define to be the geometric median of bounding boxes produced by five human annotators (**Fig 3A**). We evaluated detection performance using the Intersection Over Union (IoU) metric[38] for both top- and front-view datasets (**Fig 3B, D**). Plotting Precision-Recall (PR) curves for various IoU cutoffs revealed a single optimal performance point for both the black and white mouse, in both the top and front view (seen as an "elbow" in plotted PR curves, **Fig 3C, E**).

Following detection, MARS estimates the pose of each mouse using a Stacked Hourglass network architecture, which was previously shown to achieve high performance in human pose estimation[39]. The Stacked Hourglass network produce a likelihood map for the location of each keypoint, the global maximum of which we take to be that keypoint's location (**Fig 4A**). MARS pose estimates reach the upper limit of human accuracy for both top and front view frames, suggesting that quality of human pose annotation is a limiting factor of the model's performance (**Fig. 4B-C**). In the top view, 92% of estimated keypoints fell within 5mm of ground truth, while in the front view 67% of estimates fell within a 5mm radius of ground truth (scale bar in **Fig 2E**). Because of the poor performance of front-view pose estimation, we opted to use only the top-view video and pose in our supervised behavior classifiers (described below).

Importantly, the presence of a microendoscope or other cabled device only marginally affected the performance of pose estimation in both top and front views, despite the substantial increase in occlusion caused by the scope and cable. The Stacked Hourglass network architecture pools information across multiple spatial scales of the image to infer the location of keypoints, producing high quality pose estimates even in cases of occlusion and motion blur (**Fig 4D**).

Quantifying inter-annotator variability in scoring of social behaviors

As in pose estimation, different human annotators can show substantial variability in their annotation of animals' social behaviors, even when those individuals are trained in the same lab. To better understand the variability of human behavioral annotations, we collected annotation data from eight experienced annotators on a common set of 10 behavior videos. Human annotators included three senior laboratory technicians, two postdocs with experience studying mouse social behavior, two graduate students with experience studying mouse social behavior, and one graduate student with previous experience studying fly social behavior. All annotators were instructed to score the same three social behaviors: close investigation, mounting, and attack, and given written descriptions of each behavior (see Methods). Two

of the eight annotators showed a markedly different annotation "style" with far longer bouts of some behaviors, and were omitted from further analysis (see **ED Fig 3**).

We noted several forms of inter-annotator disagreement, consisting of 1) the precise timing of initiation of close investigation, 2) at what point aggressive investigation behavior transitioned to attack, and 3) the extent to which annotators merged together multiple consecutive bouts of the same behavior (**Fig 5A**). Other inter-annotator differences which we could not characterize could be ascribed to random variation. Examining the onset of close investigation annotations relative to the group median reveals separate contributions of noise and bias to inter-annotator differences, where noise is reflected in the variance of bout start times, and bias in a shift in the mean bout start time, estimated relative to the group median for each bout (**ED Fig 4**). We also found that as a result of inter-annotator differences in merging vs splitting multiple bouts (point 3 above), inter-annotator differences in behavior quantification were more pronounced when behavior was reported in terms of total bouts rather than cumulative behavior time, particularly for the two omitted annotators (**Fig 5B-C, ED Fig 2**).

Importantly, the Precision and Recall of annotators was highly dependent on the dataset used for evaluation (**ED Fig 5**). To quantify this, for each approximately ten-minute video we computed the mean F1 score (the harmonic mean of Precision and Recall) across six annotators, with each annotator evaluated relative to the median of the other five. For close investigation, single video F1 scores ranged dramatically, from 0.543 to 0.959 (**ED Fig 5**). We found that the mean annotator F1 score was well predicted by the mean bout duration of annotations in a video, with shorter bout durations leading to lower annotator F1 scores (**ED Fig 5**). This suggests that annotator disagreement over the start and stop times of behavior bouts may be a primary form of inter-annotator variability. Furthermore, this finding shows the importance of standardized datasets for evaluating the performance of different automated annotation approaches.

Because of observed inter-annotator differences in style, we selected one annotator with good self-consistency and from whom a large number of manually scored videos were available, Human #1, as our sole source of labeled data to train MARS's automated behavior classifiers. Accuracy of the five other annotators with respect to Human #1 was very high for mounting, with Precision (or Specificity) of 96.7±0.9 (mean ± SEM) and mean Recall (or Sensitivity) of 96.9±0.6, while Precision and Recall for close investigation and attack were lower (attack Precision 83.7±1.3, Recall 86.2±2.5; close investigation Precision 89±0.9, Recall 80.5± 1.8) (**Fig 5D**).

MARS achieves high accuracy in automated classification of three social behaviors

To create a training set for automatic detection of social behaviors in the resident-intruder assay, we collected manually annotated videos of social interactions between a male resident (black mouse) and a male or female intruder (white mouse), from existing experimental datasets (source: Anderson lab, Caltech). Even for videos that were scored by only one individual we found variability in the degree of granularity of behavior annotations: for example, some videos were scored for face-, body-, and anogenital-directed investigation separately, while others were scored for close investigation regardless of body part investigated (**ED Fig 1**). To allow for combination of training data across videos, we assigned annotated behaviors to categories of close investigation, mounting, or attack (or to "other") based on visual inspection (see Methods for behavior definitions and assignments).

Because of the lower quality of MARS pose estimates in front-view video, we performed behavior classification using only information from the top-view video. To detect social behaviors from animals' pose, we designed a set of 270 custom spatio-temporal features from the tracked poses of the two mice

on each frame of the top-view video (full list of features in **Suppl. Table 1**). For each feature, we then computed the feature mean, standard deviation, minimum, and maximum over windows of 3, 11, and 21 frames, to capture how features evolved in time. We trained a set of binary supervised classifiers to detect each behavior of interest using the XGBoost algorithm[40], then smoothed classifier output and enforced one-hot labeling (ie, one behavior/frame only) of behaviors with a Hidden Markov Model (HMM), to produce a most-likely estimate of the behavior on each frame (**Fig 6A**).

In preliminary experiments, we found that classifiers trained with multiple annotators' labels of the same actions were less accurate than classifiers trained on a smaller pool of annotations from a single individual. To determine a "best-case" accuracy of MARS classifiers given our selected pose features, we trained classifiers on 7.3 hours of video annotated by a single individual (Human #1 in **Fig 5**) for attack, mounting, and close investigation behavior. These videos included experiments both in unoperated mice and in mice with head-mounted microendoscopes or other cable-attached devices. To avoid overfitting to the training data, we implemented early stopping of training based on performance on a separate validation set of 3.65 videos. Distributions of annotated behaviors in the training, evaluation, and test sets are reported in **ED Fig 5**.

When tested on the 10 videos previously scored by multiple human annotators ("test set 1", 1.7 hours of video, behavior breakdown in **ED Fig 1**), Precision and Recall of MARS classifiers was comparable to that of human annotators for both attack and close investigation, and slightly below human performance for mounting (**Fig 6B-C**, humans and MARS both evaluated with respect to Human #1). Varying the threshold of a given binary classifier in MARS produces a Precision-Recall curve (PR curve) showing the trade-off between the classifier's true positive rate and its false positive rate (**Fig 6B-D**, black lines). Interestingly, the Precision and Recall scores of different human annotators often fell quite close to this PR curve.

False positive/negatives in MARS output could be due to mistakes by MARS, however they may also reflect noise or errors in the human annotations serving as our "ground truth." We therefore also computed the Precision and Recall of MARS output relative to the pooled (median) labels of all six annotators. To pool annotators, we scored a given frame as positive for a behavior if at least three out of six annotators labeled it as such. Precision and Recall of MARS relative to this "denoised ground truth" was further improved, particularly for the attack classifier (**Fig 6D-E**).

Strikingly, the Precision and Recall of MARS on individual videos was highly correlated with the average Precision and Recall of individual annotators with respect to the annotator median (**ED Fig 6**). Hence, as for human annotators, Precision and Recall of MARS are correlated with the average duration of behavior bouts in a video (see **ED Fig 5B**), with longer behavior bouts leading to higher Precision and Recall values.

We next tested MARS classifiers on a second set of videos of mice with head-mounted microendoscopes or other cable-attached devices ("test set 2", 1.66 hours of video, behavior breakdown in **ED Fig 1**). While Precision and Recall curves differ on this test set (**ED Fig 7A**), we do not observe a difference on individual videos with vs without cable when controlling for mean bout length in the video (**ED Fig 7B**). We therefore conclude that MARS classifier performance is robust to occlusions and motion artifacts produced by head mounted recording devices and cables.

Integration of video, annotation, and neural recording data in a user interface

Because one objective of MARS is to accelerate the analysis of behavioral and neural recording data, we developed an open-source interface to allow users to more easily navigate neural recording, behavior video, and tracking data (**Fig 7A**). This tool, called the Behavior Ensemble and Neural Trajectory

Observatory (Bento) allows users to synchronously display, navigate, analyze, and save movies from multiple behavior videos, behavior annotations, MARS pose estimates and features, audio recordings, and recorded neural activity. Bento is currently Matlab-based, although a Python version is in development.

Bento includes an interface for manual annotation of behavior, which can be combined with MARS to train, test, and apply classifiers for novel behaviors. Bento also allows users to access MARS pose features directly, to create hand-crafted filters on behavioral data (**Fig 7B**). For example, users may create and apply a filter on inter-animal distance or resident velocity, to automatically identify all frames in which feature values fall within a specified range.

Bento also provides interfaces for several common analyses of neural activity, including event-triggered averaging, 2D linear projection of neural activity, and clustering of cells by their activity. Advanced users have the option to create additional custom analyses as plugins in the interface. Bento is freely available on GitHub, and is supported by documentation and a user wiki.

Use case 1: high-throughput social behavioral profiling of multiple genetic mouse model lines

Advances in human genetics, such as genome-wide association studies (GWAS), have led to the identification of multiple gene loci that may increase susceptibility to autism[41]. The laboratory mouse has been used as a system for studying "genocopies" of allelic variants found in humans, and dozens of mouse lines containing engineered autism-associated mutations have been produced in an effort to understand the effect of these mutations on neural circuit development and function[5,42]. While several lines show atypical social behaviors, it is unclear whether all lines share a similar behavioral profile, or whether different behavioral phenotypes are associated with different genetic mutations.

We collected and analyzed a 45-hour dataset of male-male social interactions using mice from five different lines: three lines that genocopy autism-associated mutations (*Chd8*, *Cul3*, and *Nlgn3*[43]), one inbred line that has previously been shown to exhibit atypical social behavior and is used as an autism "model" (BTBR), and a C57Bl/6J control line. For each line, we collected behavior videos during ten-minute social interactions with a male intruder, and quantified expression of attack and close investigation behaviors using MARS. In autism lines, we tested heterozygotes vs age-matched wild-type (WT) littermates; BTBR mice were tested alongside age-matched C57Bl/6J mice. Due to the need for contrasting coat-colors to distinguish interacting mouse pairs, all mice were tested with BalbC intruder males. Each mouse was tested using a repeated measures design (**Fig. 8A**): first in their home cage after group-housing with heterozygous and wild-type littermates, and again after two weeks of single-housing.

Consistent with previous studies, we observed increased aggression in group-housed *Chd8*$^{+/-}$ mice relative to WT littermate controls (**Fig. 8B**). *Nlgn3*$^{+/-}$ mice were more aggressive than C57Bl/6J animals, consistent with previous work[44], and showed a significant increase in aggression following single-housing. But interestingly, there was not a statistically significant difference in aggression between single-housed *Nlgn3*$^{+/-}$ mice and their WT littermates, which were also aggressive. The increased aggression of WT littermates of *Nlgn3*$^{+/-}$ mice may be due to their genetic background (C57Bl6-SV129 hybrid rather than pure C57Bl/6), or could arise from the environmental influence of these mice being co-housed with aggressive heterozygote littermates[45].

We also confirmed previous findings[16] that BTBR mice spend less time investigating intruder mice than C57Bl/6J control animals (**Fig. 8C**), and that the average duration of close investigation bouts was reduced (**Fig. 8E, left**). Using MARS's estimate of the intruder mouse's pose, we defined two anatomical boundaries on the intruder mouse's body: one (the "face-body boundary") midway between the nose and neck

keypoints, and the other (the "body-genital boundary") midway between the tail and the center of the hips. We then used these boundaries to relabel frames of close investigation as either face-, body-, or genital-directed investigation (**Fig 8D**). Interestingly, this relabeling revealed that in addition to showing shorter investigation bouts in general, the BTBR mice showed shorter bouts of face- and genital-directed investigation compared to C57Bl/6J controls, while the duration of body-directed investigation bouts was not significantly different from controls (**Fig 8E**). This finding may suggest a loss of preference for investigation of, or sensitivity to, socially relevant pheromonal cues in the BTBR inbred line.

Without automation of behavior annotation by MARS, analysis of body part-specific investigation would have required complete manual reannotation of the dataset, a prohibitively slow process. Our findings in BTBR mice therefore demonstrates the power of MARS behavior labels and pose features as a resource for exploratory analysis of behavioral data.

Use case 2: finding neural correlates of mounting behavior

The sensitivity of electrophysiological and 2-photon neural imaging methods to motion artifacts has historically required the recording of neural activity to be performed in animals that have been head-fixed or otherwise restrained. However, head-fixed animals cannot perform many naturalistic behaviors, including social behaviors. The emergence of novel technologies such as microendoscopic imaging and silicon probe recording has enabled the recording of neural activity in freely moving animals[46], however these techniques still require animals to be fitted with a head-mounted recording device, typically tethered to an attached cable (**Fig. 9A-B**).

To demonstrate the utility of MARS and Bento for these data, we analyzed data from a recent study in the Anderson lab, in which male Esr1+ Cre mouse was implanted with a microendoscopic imaging device targeting the medial preoptic area (MPOA), a hypothalamic nucleus implicated in social and reproductive behaviors. We first used MARS to automatically detect bouts of close investigation and mounting while this mouse freely interacted with a female conspecific. Next, video, pose, and annotation data was loaded into Bento, where additional social behaviors of interest were manually annotated. Finally, we re-loaded video, pose, and annotation data in Bento alongside traces of neural activity extracted from the MPOA imaging data.

Using Bento's Behavior-Triggered Average plugin, we visualized the activity of individual MPOA neurons when the animal initiated mounting behavior (**Fig. 9C**), and identified a subset of 28 imaged neurons whose activity was modulated by mounting. Finally, using a subset of these identified cells, we exported mount-averaged activity from the Behavior-Triggered Average plugin and visualized their activity as a heatmap (**Fig. 9D**). This analysis allowed us to quickly browse this imaging dataset and determine that multiple subtypes of mount-modulated neurons exist within MPOA, with all analysis except for the final plotting in Figure 9D performed from within the Bento user interface.

Discussion

Automated systems for accurate pose estimation are increasingly available to neuroscientists and have proven to be useful for measuring animal pose and motion in a number of studies. However, pose alone does not provide sufficient insight into an animal's behavior. Together, MARS and Bento provide an end-to-end tool for automated pose estimation and supervised social behavior classification in the widely-used resident-intruder assay, and links these analyses with a graphical user interface for the quick exploration and analysis of joint neural and behavioral datasets. MARS allows users to perform high-throughput screening of social behavior expression, and is robust to occlusion and motion from head-

mounted recording devices and cables. While the pre-trained version of MARS does not require users to collect and annotate their own keypoint annotations as training data, an additional software package, which will be available in a future release of MARS, will allow users to fine-tune MARS's pose estimator with their own pose annotations.

MARS operates without requiring multiple cameras or specialized equipment such as a depth camera, unlike our previously published system[16]. MARS is also computationally distinct from our previous work: while Hong et al used a Matlab implementation of cascaded pose regression[20] to fit an ellipse to the body of each mouse (a form of blob-based tracking), MARS is built using the deep learning package Tensorflow and performs true pose estimation, in that it predicts the location of individual anatomical landmarks on the bodies of the tracked mice. In terms of performance, MARS is also much more accurate and invariant to changes in lighting and to the presence of head-mounted cables. Eliminating the IR-based depth sensor simplifies data acquisition and speeds up processing, and also allows MARS to be used without creating IR artifacts during microendoscopic imaging.

Comparing pose and behavior annotations from multiple human annotators, we were able to quantify the degree of inter-human variability in both tasks, and found that in both cases MARS performs comparably to the best-performing human. This suggests that improving the quality of training data, for example by providing better visualizations and clearer instructions to human annotators, could help to further improve the accuracy of pose estimation and behavior classification tools such as MARS. Conversely, the inter-human variability in behavior annotation may reflect the fact that animal behavior is too complex and heterogeneous for behavior labels of "attack" and "close investigation" to be applied reliably. If so, future behavior classification efforts with more granular action categories may both reduce inter-human variability and lead to higher performance by automated classifiers.

Unsupervised approaches are a promising alternative to behavior quantification, and may eventually bypass the need for human input during behavior discovery[26-28]. While current efforts in unsupervised behavior discovery have largely been limited to single animals, the pose estimates and features produced by MARS could potentially prove useful for future investigations that identify behaviors of interest in a user-unbiased manner. Alternatively, unsupervised analysis of MARS pose features may help to reduce redundancy among features, potentially leading to a reduction in the amount of sample data required to train classifiers to detect new behaviors of interest. Tools for faster training of behavior classifiers, as well as an interface for novel classifier training and testing within Bento are currently being developed for inclusion in future version of the software.

1 Remedios, R. *et al.* Social behaviour shapes hypothalamic neural ensemble representations of conspecific sex. *Nature* **550**, 388-392, doi:10.1038/nature23885 (2017).
2 Li, Y. *et al.* Neuronal Representation of Social Information in the Medial Amygdala of Awake Behaving Mice. *Cell* **171**, 1176-1190 e1117, doi:10.1016/j.cell.2017.10.015 (2017).
3 Falkner, A. L. *et al.* Hierarchical representations of aggression in a hypothalamic-midbrain circuit. *Neuron* (2020).
4 Yang, M., Silverman, J. L. & Crawley, J. N. Automated three-chambered social approach task for mice. *Current protocols in neuroscience* **56**, 8.26. 21-28.26. 16 (2011).
5 Silverman, J. L., Yang, M., Lord, C. & Crawley, J. N. Behavioural phenotyping assays for mouse models of autism. *Nature Reviews Neuroscience* **11**, 490-502 (2010).
6 Winslow, J. T. Mouse social recognition and preference. *Current protocols in neuroscience* **22**, 8.16. 11-18.16. 16 (2003).
7 Zelikowsky, M. *et al.* The Neuropeptide Tac2 Controls a Distributed Brain State Induced by Chronic Social Isolation Stress. *Cell* **173**, 1265-1279 e1219, doi:10.1016/j.cell.2018.03.037 (2018).
8 Shemesh, Y. *et al.* High-order social interactions in groups of mice. *Elife* **2**, e00759 (2013).
9 Branson, K., Robie, A. A., Bender, J., Perona, P. & Dickinson, M. H. High-throughput ethomics in large groups of Drosophila. *Nat Methods* **6**, 451-457, doi:10.1038/nmeth.1328 (2009).
10 Szegedy, C., Ioffe, S., Vanhoucke, V. & Alemi, A. A. in *Thirty-first AAAI conference on artificial intelligence.*
11 Redmon, J., Divvala, S., Girshick, R. & Farhadi, A. in *Proceedings of the IEEE conference on computer vision and pattern recognition.* 779-788.
12 He, K., Gkioxari, G., Dollár, P. & Girshick, R. in *Proceedings of the IEEE international conference on computer vision.* 2961-2969.
13 Noldus, L. P., Spink, A. J. & Tegelenbosch, R. A. EthoVision: a versatile video tracking system for automation of behavioral experiments. *Behavior Research Methods, Instruments, & Computers* **33**, 398-414 (2001).
14 Ohayon, S., Avni, O., Taylor, A. L., Perona, P. & Roian Egnor, S. E. Automated multi-day tracking of marked mice for the analysis of social behaviour. *J Neurosci Methods* **219**, 10-19, doi:10.1016/j.jneumeth.2013.05.013 (2013).
15 Pérez-Escudero, A., Vicente-Page, J., Hinz, R. C., Arganda, S. & De Polavieja, G. G. idTracker: tracking individuals in a group by automatic identification of unmarked animals. *Nature methods* **11**, 743-748 (2014).
16 Hong, W. *et al.* Automated measurement of mouse social behaviors using depth sensing, video tracking, and machine learning. *Proceedings of the National Academy of Sciences* **112**, E5351-E5360 (2015).
17 Gal, A., Saragosti, J. & Kronauer, D. J. C. anTraX: high throughput video tracking of color-tagged insects. *bioRxiv*, 2020.2004.2029.068478, doi:10.1101/2020.04.29.068478 (2020).
18 Toshev, A. & Szegedy, C. in *Proceedings of the IEEE conference on computer vision and pattern recognition.* 1653-1660.
19 Dankert, H., Wang, L., Hoopfer, E. D., Anderson, D. J. & Perona, P. Automated monitoring and analysis of social behavior in Drosophila. *Nat Methods* **6**, 297-303, doi:10.1038/nmeth.1310 (2009).
20 Dollár, P., Welinder, P. & Perona, P. in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition.* 1078-1085 (IEEE).
21 Alp Güler, R., Neverova, N. & Kokkinos, I. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 7297-7306.

22 Mathis, A. *et al.* DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nat Neurosci* **21**, 1281-1289, doi:10.1038/s41593-018-0209-y (2018).
23 Pereira, T. D. *et al.* Fast animal pose estimation using deep neural networks. *bioRxiv*, doi:10.1101/331181 (2018).
24 Graving, J. M. *et al.* DeepPoseKit, a software toolkit for fast and robust animal pose estimation using deep learning. *Elife* **8**, doi:10.7554/eLife.47994 (2019).
25 Sturman, O. *et al.* Deep learning based behavioral analysis enables high precision rodent tracking and is capable of outperforming commercial solutions. *bioRxiv* (2020).
26 Wiltschko, A. B. *et al.* Mapping sub-second structure in mouse behavior. *Neuron* **88**, 1121-1135 (2015).
27 Berman, G. J., Choi, D. M., Bialek, W. & Shaevitz, J. W. Mapping the stereotyped behaviour of freely moving fruit flies. *Journal of The Royal Society Interface* **11**, 20140672 (2014).
28 Vogelstein, J. T. *et al.* Discovery of brainwide neural-behavioral maps via multiscale unsupervised structure learning. *Science* **344**, 386-392 (2014).
29 de Chaumont, F. *et al.* Computerized video analysis of social interactions in mice. *Nat Methods* **9**, 410-417, doi:10.1038/nmeth.1924 (2012).
30 Giancardo, L. *et al.* Automatic visual tracking and social behaviour analysis with multiple mice. *PLoS One* **8**, e74557, doi:10.1371/journal.pone.0074557 (2013).
31 Hong, E. J. & Wilson, R. I. Simultaneous encoding of odors by channels with diverse sensitivity to inhibition. *Neuron* **85**, 573-589, doi:10.1016/j.neuron.2014.12.040 (2015).
32 Kabra, M., Robie, A. A., Rivera-Alba, M., Branson, S. & Branson, K. JAABA: interactive machine learning for automatic annotation of animal behavior. *Nature methods* **10**, 64 (2013).
33 Nilsson, S. R. *et al.* Simple Behavioral Analysis (SimBA): an open source toolkit for computer classification of complex social behaviors in experimental animals. *BioRxiv* (2020).
34 Monfort, M. *et al.* Moments in time dataset: one million videos for event understanding. *IEEE transactions on pattern analysis and machine intelligence* **42**, 502-508 (2019).
35 Tran, D., Bourdev, L., Fergus, R., Torresani, L. & Paluri, M. in *Proceedings of the IEEE international conference on computer vision.* 4489-4497.
36 Thurmond, J. B. Technique for producing and measuring territorial aggression using laboratory mice. *Physiology & behavior* **14**, 879-881 (1975).
37 Szegedy, C., Reed, S., Erhan, D., Anguelov, D. & Ioffe, S. Scalable, high-quality object detection. *arXiv preprint arXiv:1412.1441* (2014).
38 Lin, T.-Y. *et al.* in *European conference on computer vision.* 740-755 (Springer).
39 Newell, A., Yang, K. & Deng, J. in *European Conference on Computer Vision.* 483-499 (Springer).
40 Chen, T. & Guestrin, C. in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining.* 785-794.
41 Grove, J. *et al.* Identification of common genetic risk variants for autism spectrum disorder. *Nature genetics* **51**, 431-444 (2019).
42 Moy, S. & Nadler, J. Advances in behavioral genetics: mouse models of autism. *Molecular psychiatry* **13**, 4-26 (2008).
43 Tabuchi, K. *et al.* A neuroligin-3 mutation implicated in autism increases inhibitory synaptic transmission in mice. *science* **318**, 71-76 (2007).
44 Burrows, E. L. *et al.* A neuroligin-3 mutation implicated in autism causes abnormal aggression and increases repetitive behavior in mice. *Mol Autism* **6**, 62, doi:10.1186/s13229-015-0055-7 (2015).
45 Kalbassi, S., Bachmann, S. O., Cross, E., Roberton, V. H. & Baudouin, S. J. Male and Female Mice Lacking Neuroligin-3 Modify the Behavior of Their Wild-Type Littermates. *eNeuro* **4**, doi:10.1523/ENEURO.0145-17.2017 (2017).

46 Resendez, S. L. *et al.* Visualization of cortical, subcortical and deep brain neural circuit dynamics during naturalistic mammalian behavior with head-mounted microscopes and chronically implanted lenses. *Nat Protoc* **11**, 566-597, doi:10.1038/nprot.2016.021 (2016).

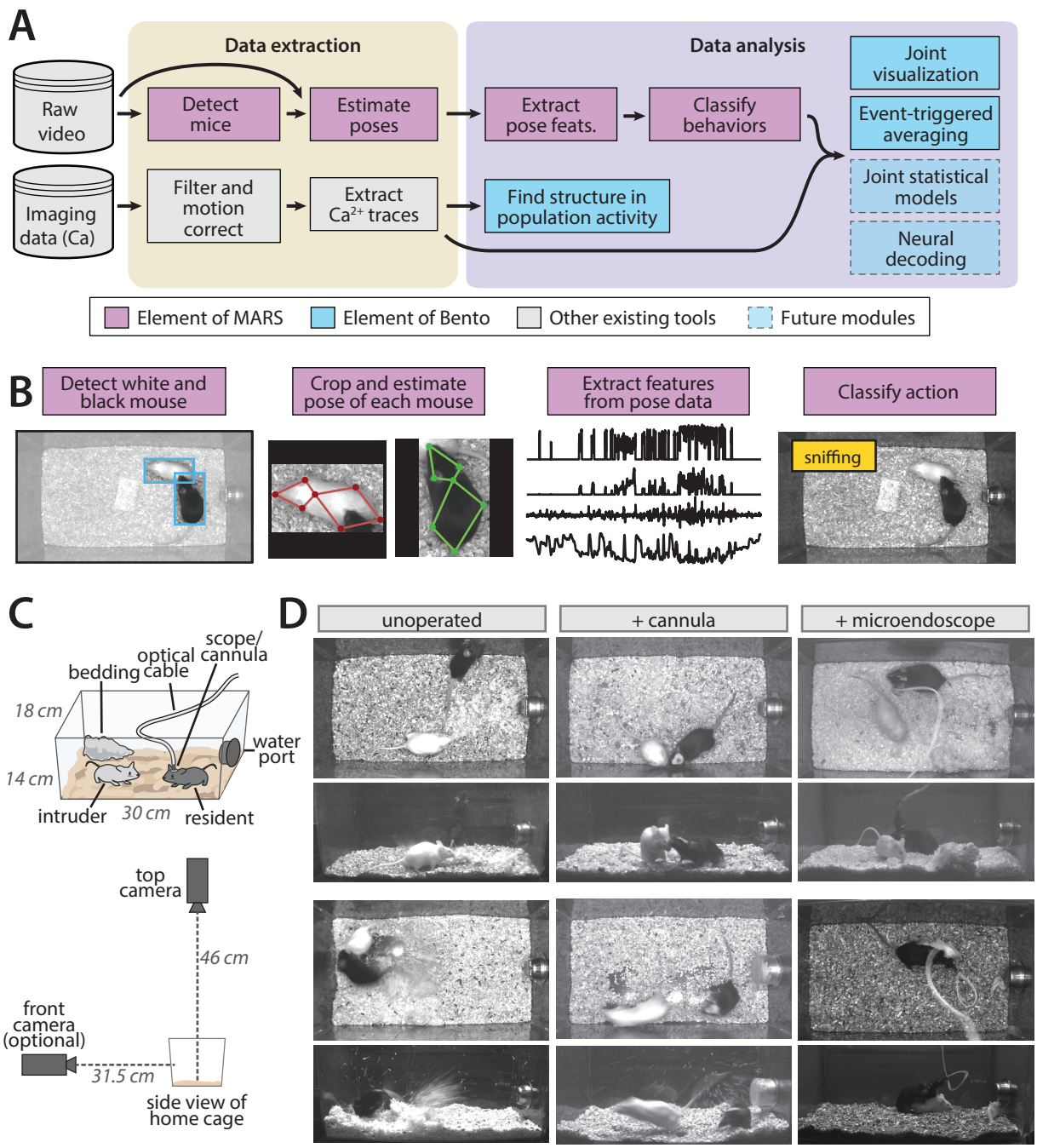

Figure 1. The MARS data pipeline. A) Overview of data extraction and analysis steps in a typical neuroscience experiment, indicating contributions to this process by MARS and Bento. **B)** Illustration of the four stages of data processing included in MARS. **C)** Above: schematic of standard mouse home cage. Below: placement of front- and top-view cameras relative to home cage. **D)** Example top and front view video frames from training set, showing range of contrast, illumination, and motion blur/occlusion present in the dataset.

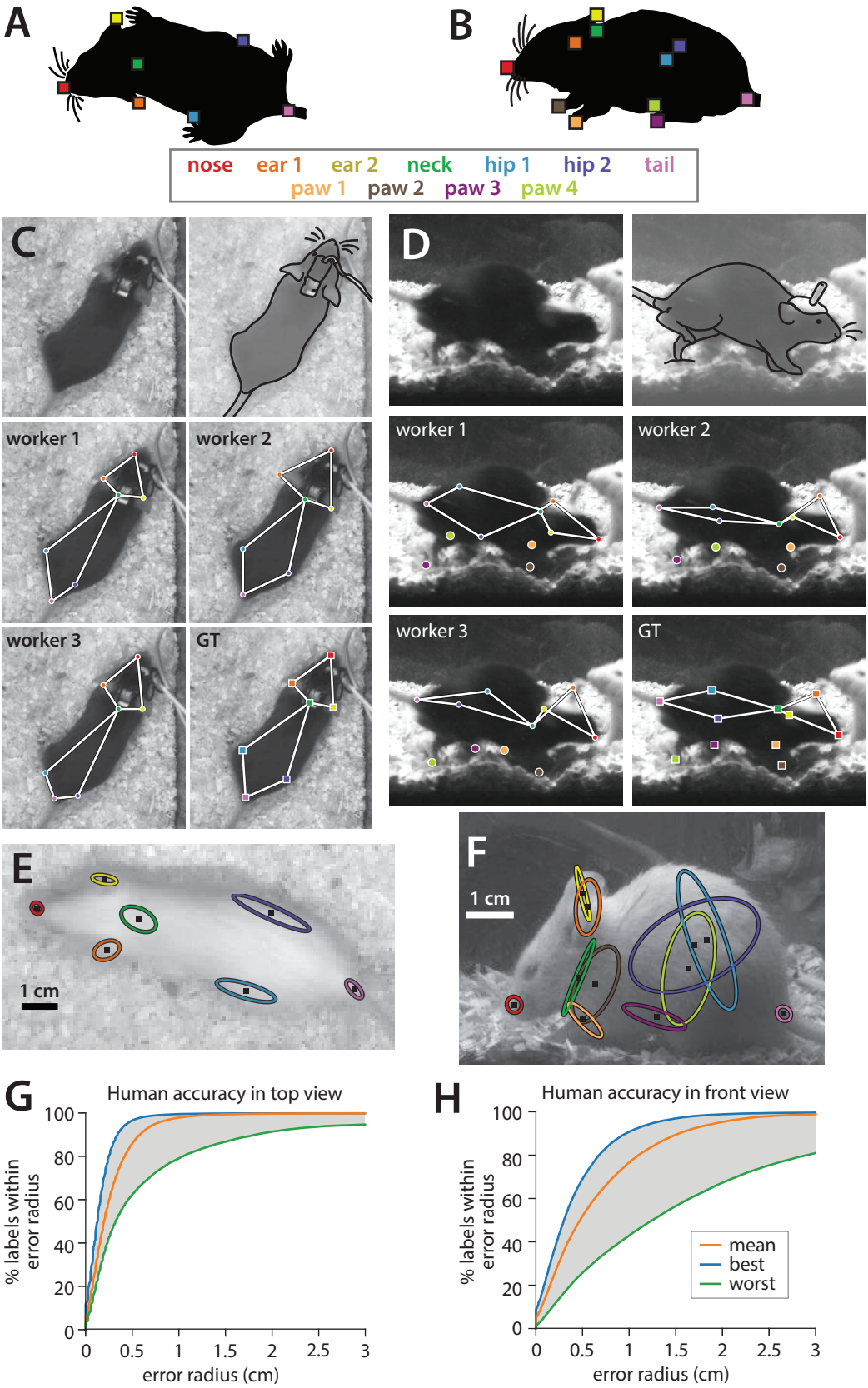

Figure 2. Quantifying human variability in top- and front-view pose estimates. A-B) Anatomical keypoints labeled by human annotators in A) top-view and B) front-view movie frames. **C-D)** Comparison of annotator labels in **C)** top-view and **D)** front-view frames. Top row: left original image shown to annotators (cropped), right approximate figure of mouse. Middle-bottom rows: Keypoint locations provided by three example annotators, and the extracted "ground truth" from the median of all annotations. **E-F)** Ellipses showing variability of human annotations of each keypoint in one example frame from **E)** top view and **F)** front view (N=5 annotators, one standard deviation ellipse radius.) **G-H)** Variability in human annotations of mouse pose for the top-view video, plotted as the percentage of human annotations falling within radius X of ground truth, for **G)** top-view and **H)** front-view frames.

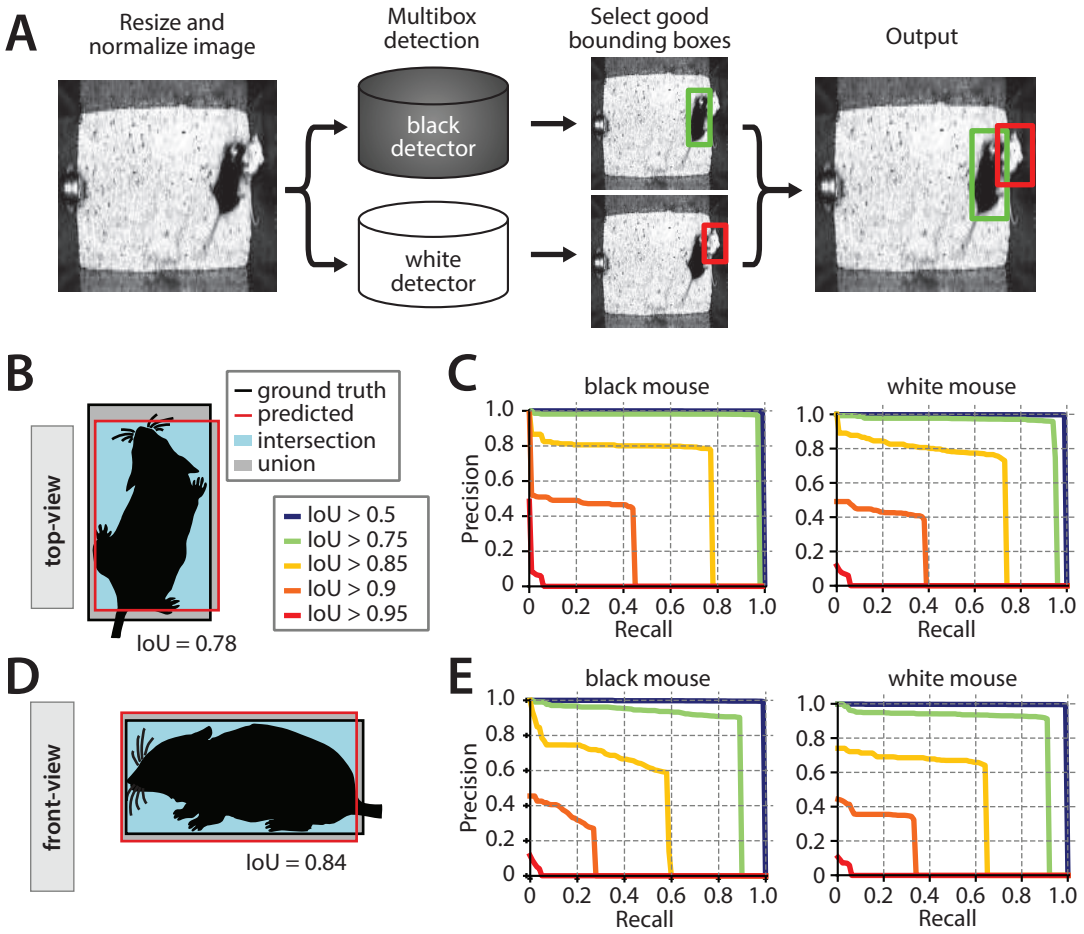

Figure 3. Performance of the mouse detection step. A) Processing stages of mouse detection pipeline. **B)** Illustration of Intersection over Union (IoU) metric for the top-view video. **C)** PR curves for multiple IoU thresholds for detection of the two mice in the top-view video. **D)** Illustration of IoU for the front-view video. E) PR curves for multiple IoU thresholds in the front-view video.

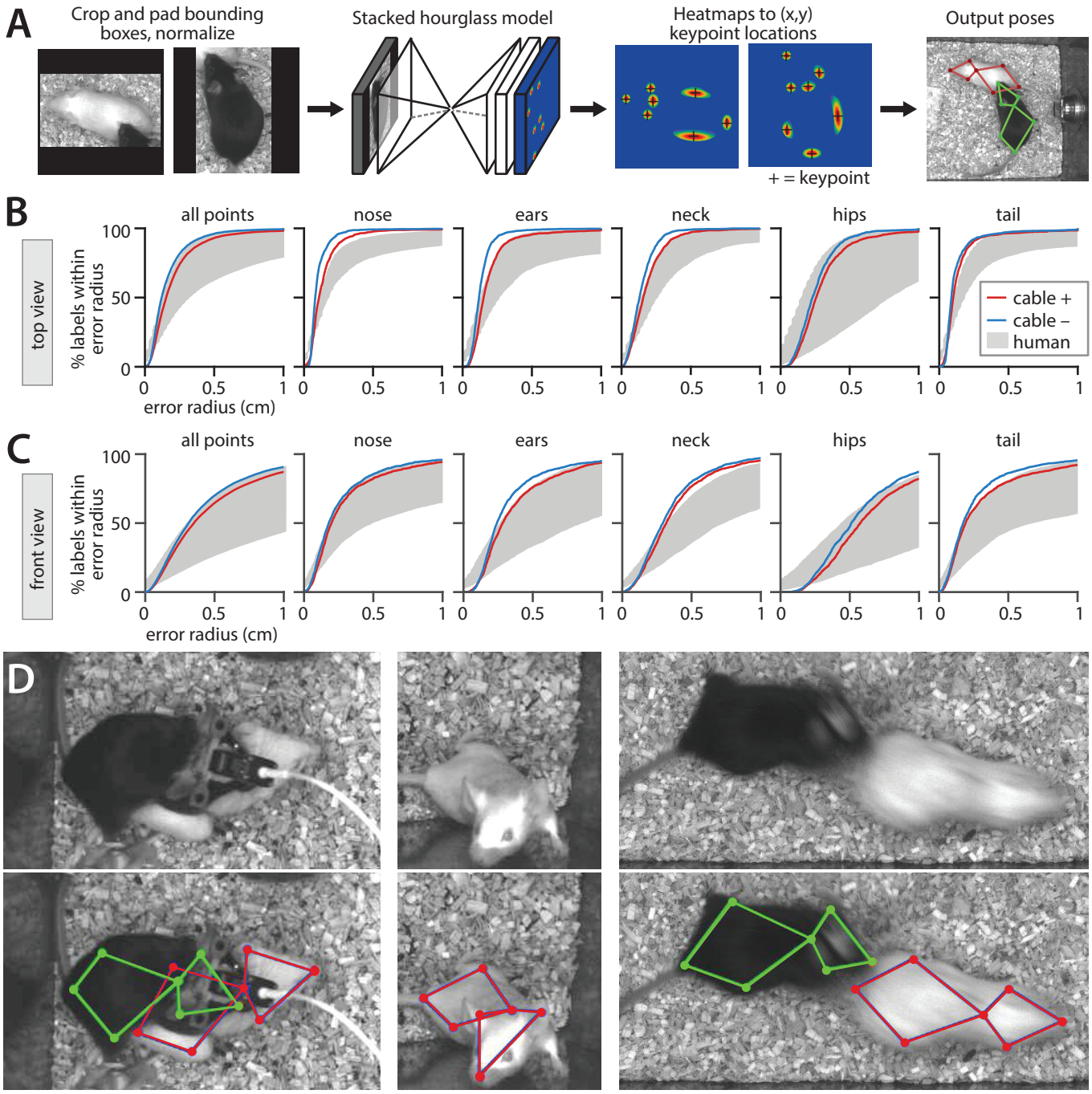

Figure 4. Performance of the stacked hourglass network for pose estimation. A) Processing stages of pose estimation pipeline. **B)** MARS accuracy for individual body parts, showing performance for videos with vs without a head-mounted microendoscope or fiberphotometry cable on the black mouse. Gray envelop shows the accuracy of the best vs worst human annotations. **C)** MARS accuracy for individual body parts in front-view videos with vs without microendoscope or fiberphotometry cables. **D)** Sample video frames (above) and MARS pose estimates (below) in cases of occlusion and motion blur.

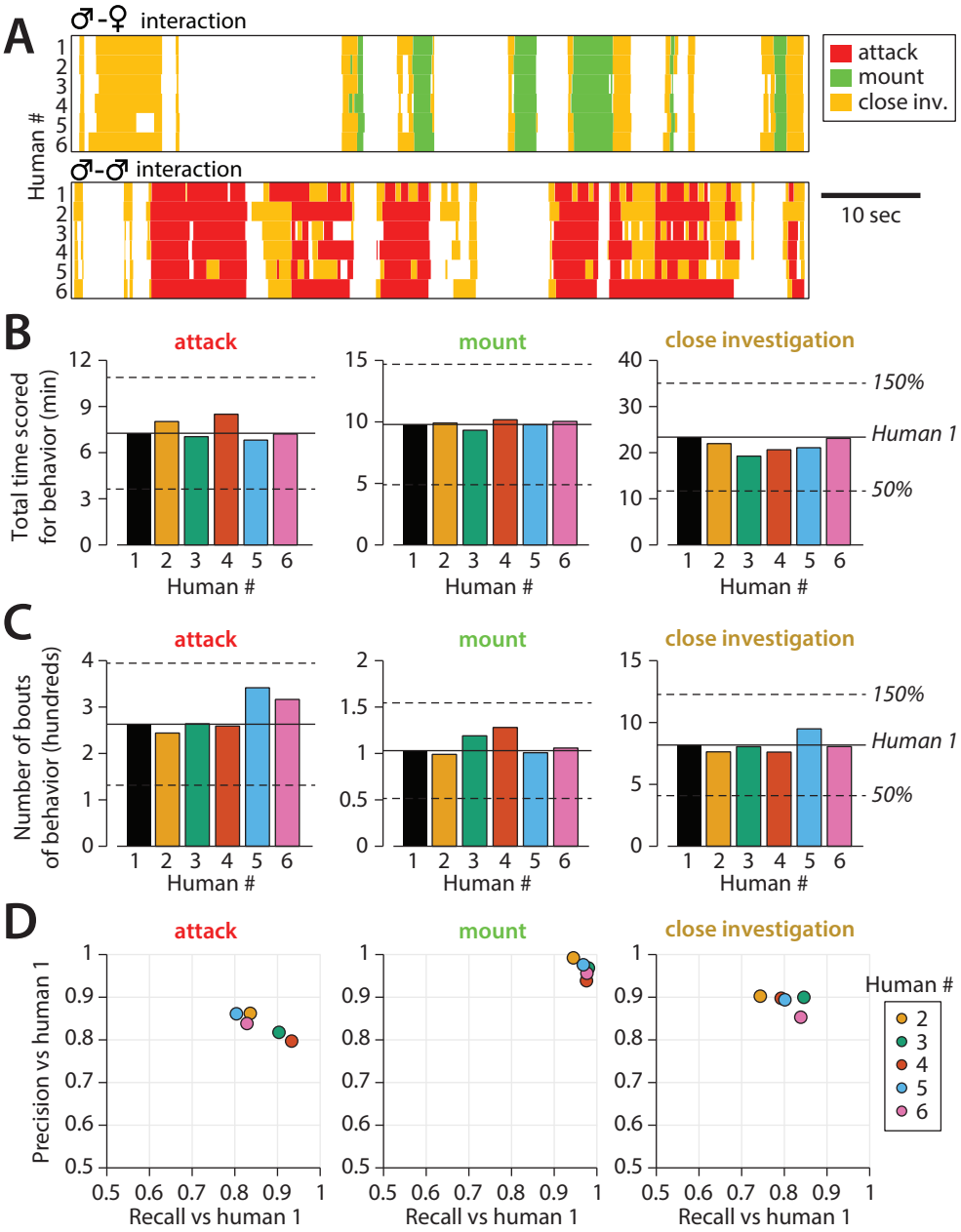

Figure 5. Quantifying inter-annotator variability in behavior annotations. A) Example annotation for attack, mounting, and close investigation behaviors by six trained annotators on segments of male-female (top) and male-male (bottom) interactions. **B)** Inter-annotator variability in the total reported time mice spent engaging in each behavior. **C)** Inter-annotator variability in the number of reported bouts (contiguous sequences of frames) scored for each behavior. **D)** Precision and recall of annotators (humans) 2-6 with respect to annotations by human 1.

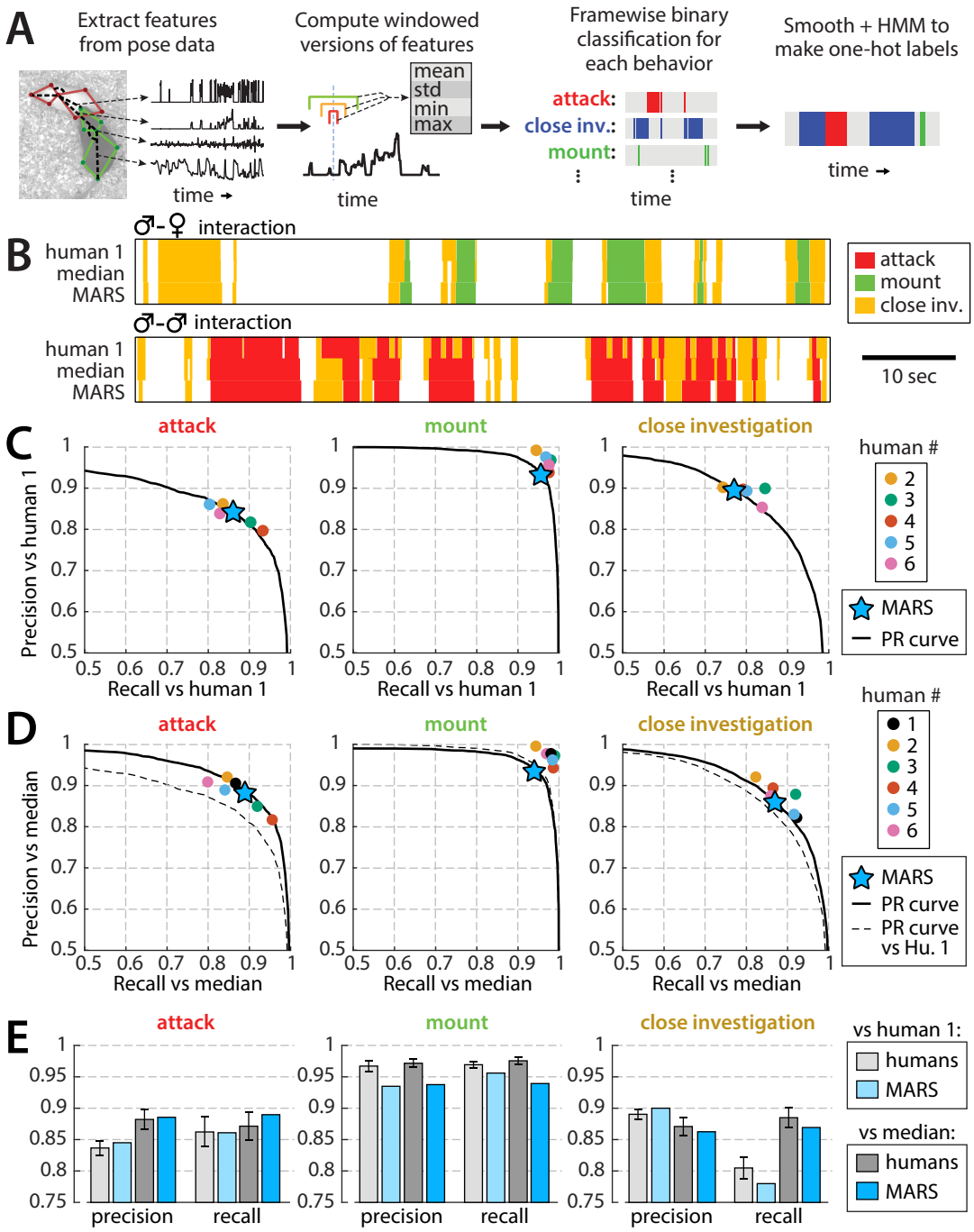

Figure 6. Performance of behavior classifiers. A) Processing stages of estimating behavior from pose of both mice. **B)** Example output of the MARS behavior classifiers on segments of male-female and male-male interactions, compared to annotations by human 1 (source of classifier training data) and to the median of the six human annotators analyzed in **Figure 5. C)** Precision, recall, and PR curves of MARS with respect to human 1 for each of the three behaviors. **D)** Precision, recall, and PR curves of MARS with respect to the median of the six human annotators (precision/recall for each human annotator was computed with respect to the median of the other five.) **E)** Mean precision and recall of human annotators vs MARS, relative to human 1 and relative to the group median (mean ±SEM).

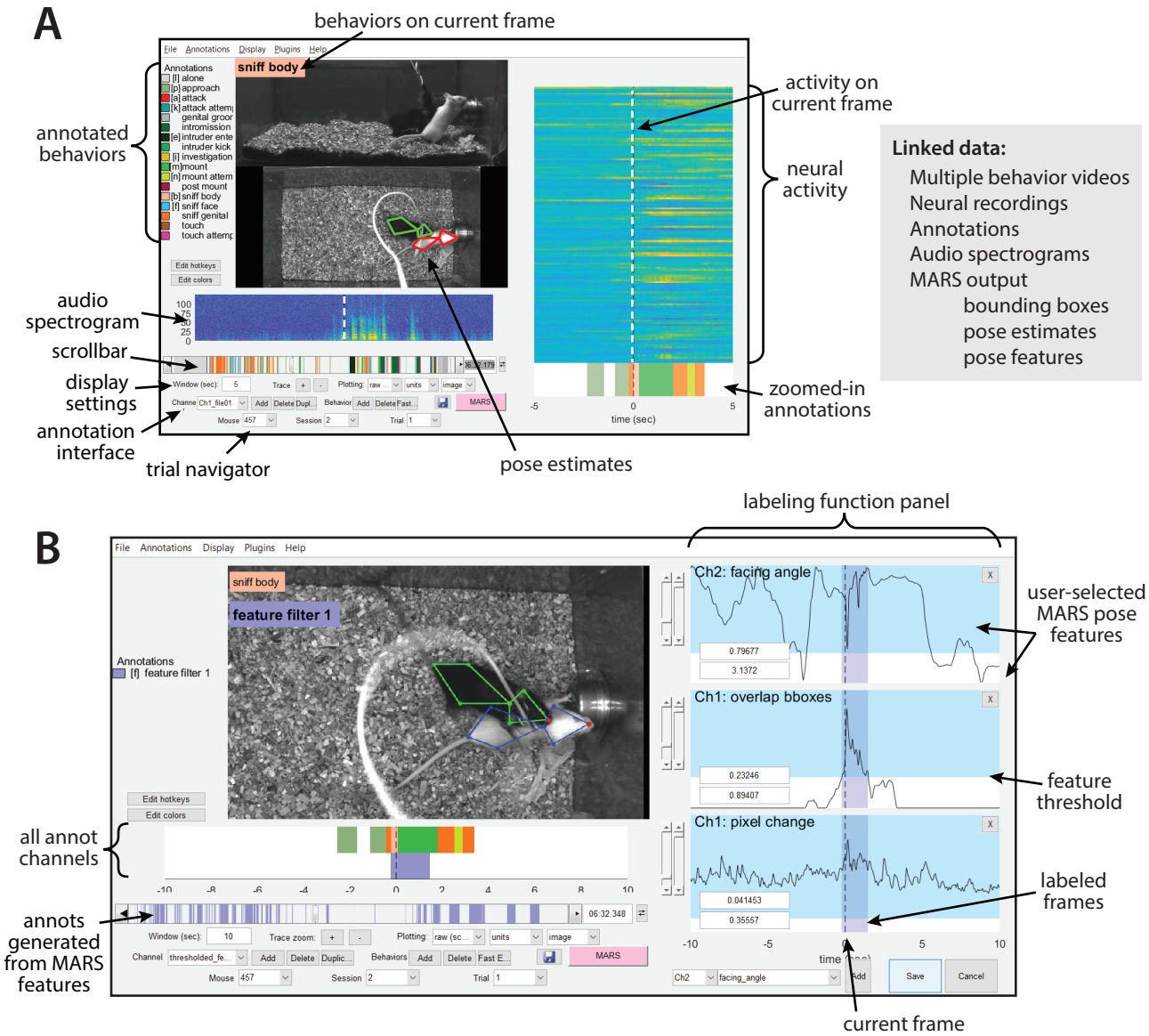

Figure 7. Cartoon of the Bento user interface. A) (*Left*) the main user interface showing synchronous display of video, pose estimation, neural activity, and pose feature data. (*Right*) list of data types that can be loaded and synchronously displayed within Bento. **B)** Bento interface for creating annotations based on thresholded combinations of MARS pose features.

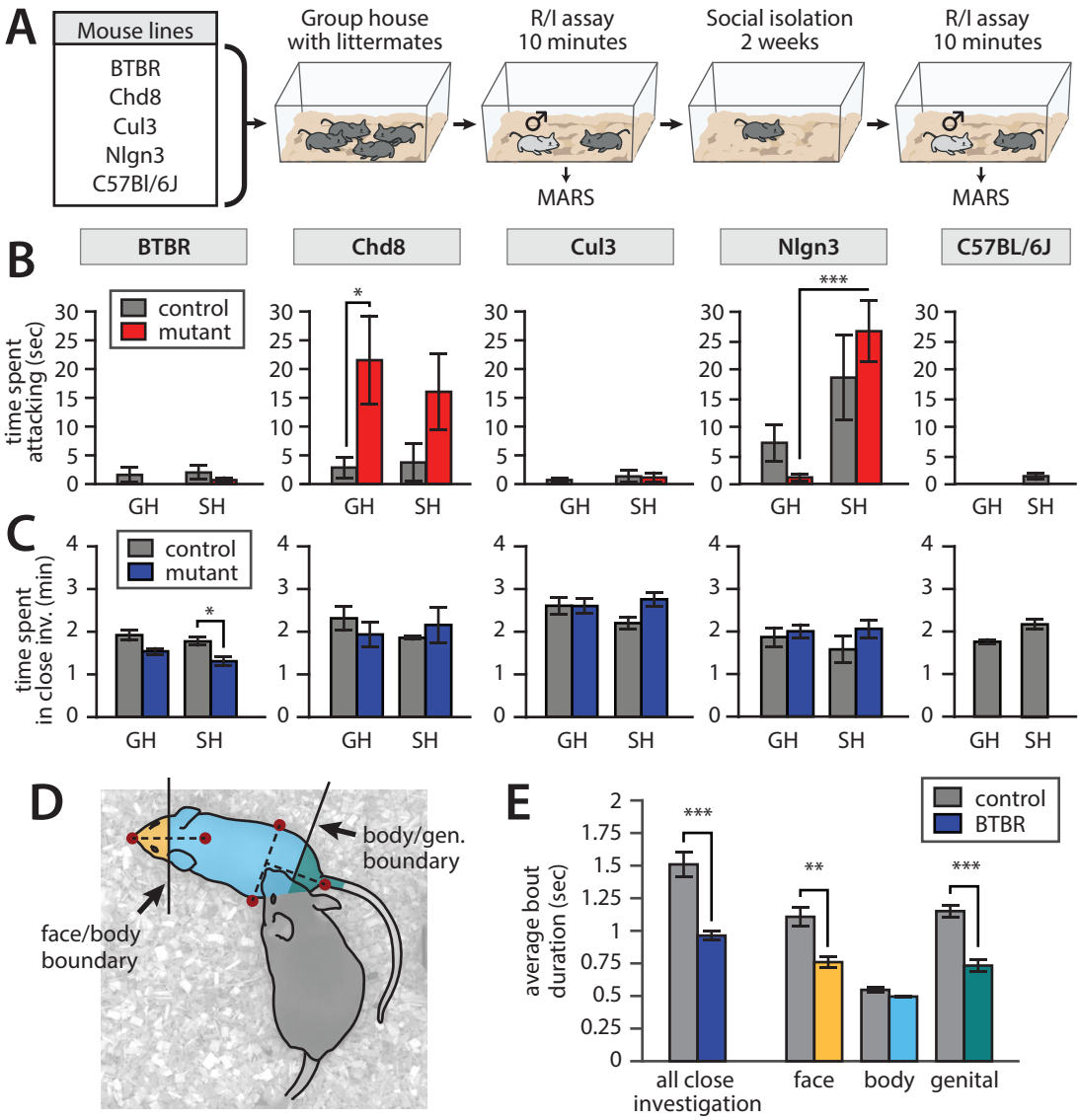

Figure 8. Application of MARS in a large-scale behavioral assay. All plots show mean +/- SEM, N=8-10 mice per line. **A)** Assay design. **B)** Time spent attacking by group-housed (GH) and single-housed (SH) mice from each line, compared to controls. **C)** Time spent engaged in close investigation by each condition/line. **D)** Cartoon showing segmentation of close investigation bouts into face-, body-, and genital-directed investigation. Frames are classified based on the position of the resident's nose relative to a boundary midway between the intruder mouse's nose and neck, and a boundary midway between the intruder mouse's hips and tail base. **E)** Average duration of close investigation bouts in BTBR mice, for investigation as a whole and broken down by the body part investigated.

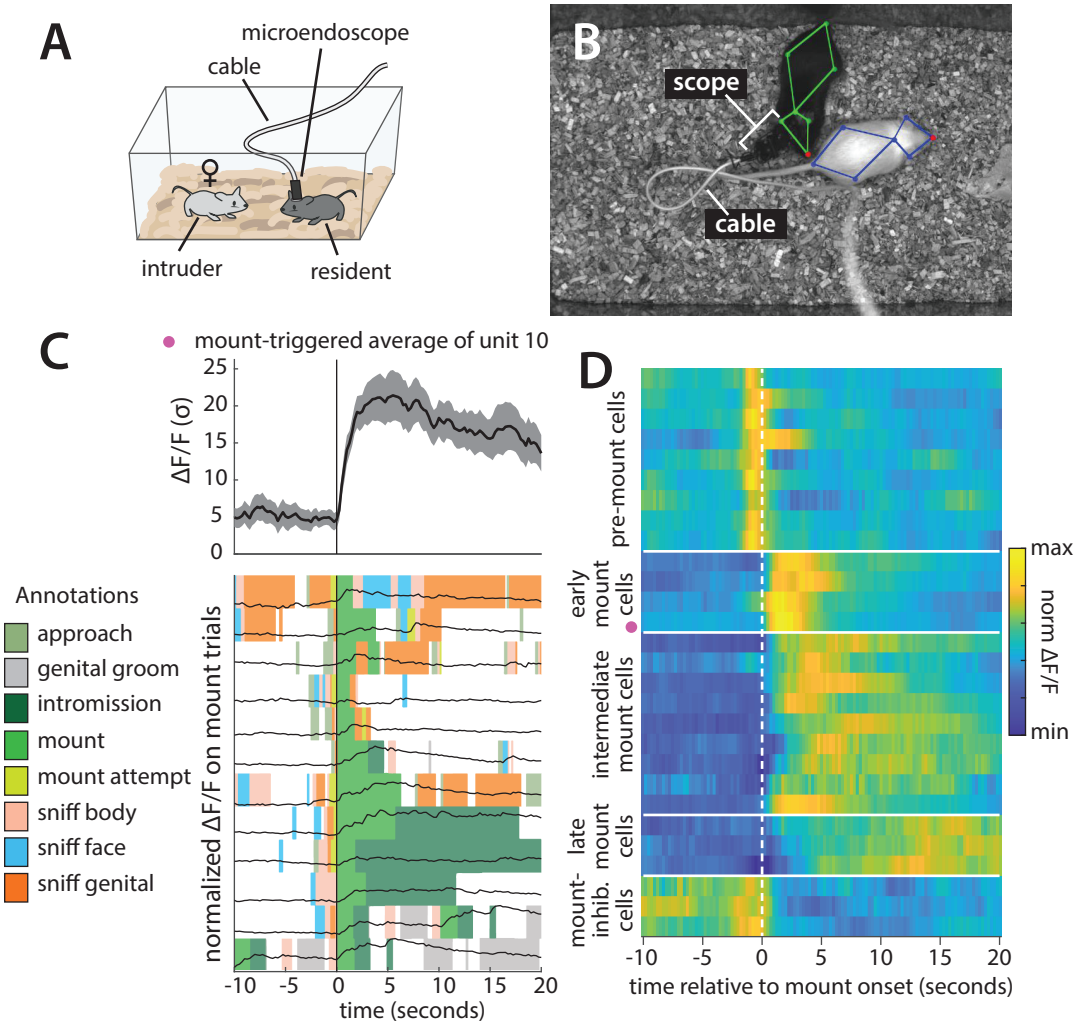

Figure 9. Analysis of a microendoscopic imaging dataset using MARS and Bento. A) Cartoon of the imaging setup, showing head-mounted microendoscope. **B)** Sample video frame with MARS pose estimate, showing appearance of the microendoscope and cable during recording. **C)** Sample behavior-triggered average figure produced by Bento. (*Top*) mount-triggered average response of one example neuron within a 30-second window (mean ± SEM). (*Bottom*) individual trials contributing to mount-triggered average, showing animal behavior (colored patches) and neuron response (black lines) on each trial. The behavior-triggered average interface allows the user to specify the window considered during averaging (here 10 seconds before to 20 seconds after mount initiation), whether to merge behavior bouts occurring less than X seconds apart, whether to trigger on behavior start or end, and whether to normalize individual trials before averaging; results can be saved as a pdf or exported to the Matlab workspace. **D)** Normalized mount-triggered average responses of 28 example neurons in the medial preoptic area (MPOA), identified using Bento. Grouping of neurons reveals diverse subpopulations of cells responding at different times relative to the onset of mounting. (Pink dot = neuron shown in panel **C**.)

DATA COLLECTION

Animals

Chd8$^{+/-}$ and *Cul3*$^{+/-}$ mice were obtained from Dr. Mark Zylka, BTBR and *Nlgn3*$^{+/-}$ mice were obtained from Jackson Labs (BTBR Stock No 2282, Nlgn3 Stock No 8475), and wild-type C57Bl/6J and BALB/c mice were obtained from Charles River. All mice were received at 6-10 weeks of age, and were maintained in the Caltech animal facility, where they were housed with three same-sex littermates (unless otherwise noted) on a reverse 11-hour dark 13-hour light cycle with food and water *ad libitum*. Behavior was tested during the dark cycle. All experimental procedures involving the use of live animals or their tissues were performed in accordance with the NIH guidelines and approved by the Institutional Animal Care and Use Committee (IACUC) and the Institutional Biosafety Committee at the California Institute of Technology (Caltech).

The resident-intruder assay

Testing for social behaviors using the resident-intruder assay[1] was performed as in [2-5]. Experimental mice ("residents") were transported in their homecage (with cagemates removed) to a behavioral testing room, and acclimatized for 5-15 minutes. Homecages were then inserted into a custom-built hardware setup[3] with infrared video captured at 30 fps from top- and front-view cameras (Point Grey Grasshopper3) recorded at 1024x570 (top) and 1280x500 (front) pixel resolution using StreamPix video software (NorPix). Following two further minutes of acclimatization, an unfamiliar group-housed male or female BALB/c mouse ("intruder") was introduced to the cage, and animals were allowed to freely interact for a period of approximately 10 minutes. BALB/c mice are used as intruders for their white coat color (simplifying identity tracking), as well as their relatively submissive behavior, which reduces the likelihood of intruder-initiated aggression. Social behaviors were manually scored on a frame-by-frame basis, as described in the "Mouse behavior annotation" section below.

Videos for training of MARS detector, pose estimator, and behavior classifier were selected from previously performed experiments in the Anderson lab. In approximately half of videos in the training data, mice were implanted with a cranial cannula, or with a head-mounted miniaturized microscope (nVista, Inscopix) or optical fiber for optogenetics or fiber photometry, attached to a cable of varying color and thickness. Surgical procedures for these implantations can be found in [2,6,7].

Screening of autism-associated mutation lines

Group-housed heterozygous male mice from mouse lines with autism-associated mutations (*Chd8*$^{+/-}$, *Cul3*$^{+/-}$, BTBR, and *Nlgn3*$^{+/-}$ mice, plus C57Bl/6J control mice), were first tested in a standard resident-intruder assay as outlined above. To control for differences in rearing between lines, an equal number of wild-type littermate male mice from each line were tested on the same day; for the inbred BTBR strain, C57Bl/6J mice were used as controls. All mice were tested at 35-50 weeks of age. Following the resident-intruder assay, mice were housed in social isolation (one mouse per cage, all cage conditions otherwise identical to those of group-housed animals) for at least 2 weeks, and then tested again in the resident-intruder assay with an unfamiliar male intruder. Videos of social interactions were scored on a frame-by-frame basis for mounting, attack, and close investigation behavior using MARS; select videos were also manually annotated for these three behaviors to confirm the accuracy of MARS classifier output.

Mouse pose annotation

Part keypoint annotations are common in computer vision, and are included in datasets such as Microsoft COCO[8], MPII human pose[9], and CUB-200-2011[10]; they also form the basis of markerless pose estimation systems such as DeepLabCut[11], LEAP[12], and DeepPoseKit[13]. For MARS, we defined nine anatomical keypoints in the top-view video (the nose, ears, base of neck, hips, and tail base, midpoint, and endpoint), and 13 keypoints in the front-view video (top-view keypoints plus the four paws). The tail mid- and endpoint annotations were subsequently discarded for training of MARS, leaving seven keypoints in the top-view and 11 in the front-view (as in **Figure 2a-b**). We used the crowdsourcing platform Amazon Mechanical Turk (AMT) to obtain manual annotations of pose keypoints on a set of video frames. AMT workers were provided with written instructions and illustrated examples of each keypoint, and instructed to infer the location of occluded keypoints. To compensate for annotation noise, each keypoint was annotated by five AMT workers, and a "ground truth" location for that keypoint was defined as the geometric median across annotators (**Figure 2c-d**). Annotations of individual workers were also post-processed to correct for common mistakes, such as confusing the left and right sides of the animals. Another common worker error was to mistake the top of the head-mounted microendoscope for the resident animal's nose; we visually screened for these errors and corrected them manually. Python scripts for creation of AMT labeling jobs and post-processing of labeled data are included in the released code for MARS.

To create a data set of video frames for labeling, we sampled 64 videos from several years of experimental projects in the Anderson lab. While all videos were acquired in our standardized hardware setup, we observed some variability in lighting and camera contrast across the dataset; examples are shown in **Figure 1b**. We extracted a set of 15,000 individual frames each from the top- and front-view cameras, giving a total of 2,700,000 individual keypoint annotations (15,000 frames x (7 top-view + 11 front-view keypoints per mouse) x 2 mice x 5 annotators). 5,000 of the extracted frames included resident mice with a fiberoptic cable, cannula, or head-mounted microendoscope with cable.

Bounding box annotation

For both top- and front-view video, we estimated a bounding box by finding the minimal rectangle that contained all seven (top) or eleven (front) pose keypoints. (For better accuracy in the detection and pose estimation we discarded the middle and end keypoints of the tail.) We then padded this minimal rectangle by a constant factor to prevent cutoff of body parts at the rectangle border.

Mouse behavior annotation

Behaviors were annotated on a frame-by-frame basis by a trained human expert in the Anderson lab. Annotators were provided with simultaneous top- and front-view video of interacting mice, and scored every video frame for close investigation, attack, and mounting, using the criteria described below. In some videos, additional behaviors were also annotated- when this occurred, these behaviors were assigned to one of close investigation, attack, mounting, or "other" for the purpose of training classifiers. Definitions of these additional behaviors are listed underneath the behavior to which they were assigned. Annotation was performed either in Bento or using a previously developed custom Matlab interface[14].

Close investigation: resident (black) mouse is in close contact with the intruder (white) and is actively sniffing the intruder anywhere on its body or tail. Active sniffing can usually be distinguished from passive orienting behavior by head bobbing/movements of the resident's nose.

Other behaviors converted to the “close investigation” label:

- **Sniff face:** resident investigation of the intruder’s face (typically eyes and snout).
- **Sniff genitals:** resident investigation of the intruder’s anogenital region, often occurs by shoving of the resident’s snout underneath the intruder’s tail.
- **Sniff body:** resident investigation of the intruder’s body, anywhere away from the face or genital regions.

Attack: high-intensity behavior in which the resident is biting or tussling with the intruder, including periods between bouts of biting/tussling during which the intruder is jumping or running away and the resident is in close pursuit. Pauses during which resident/intruder are facing each other (typically while rearing) but not actively interacting should not be included.

Other behaviors converted to the “attack” label:

- **Attempted attack:** intruder is in a defensive posture (standing on hind legs, often facing the resident) to protect itself from attack, and resident is close to the intruder, either circling or rearing with front paws out towards/on the intruder. Typically follows or is interspersed with bouts of actual attack.

Mount: copulatory behavior in which the resident is hunched over the intruder, typically from the rear, and grasping the sides of the intruder using forelimbs (easier to see on the Front camera). Early-stage copulation is accompanied by rapid pelvic thrusting, while later-stage copulation (sometimes annotated separately as intromission) has a slower rate of pelvic thrusting with some pausing: for the purpose of this analysis, both behaviors should be counted as mounting, however periods where the resident is climbing on the intruder but not attempting to grasp the intruder or initiate thrusting should not.

Other behaviors converted to the “mount” label:

- **Intromission:** late stage copulatory behavior that occurs after mounting, with a slower rate of pelvic thrusting. Occasional pausing between bouts of thrusting are still counted as intromission.
- **Attempted mount:** attempts to mount another animal in the absence of intromission. Palpitations with forepaws and pelvic thrusts may be present, but the resident is not aligned with the body of the intruder mouse or the intruder mouse may be unreceptive and still moving.

Other: behaviors that were annotated in some videos but not included in any of the above categories.

- **Approach:** resident orients and walks toward a typically stationary intruder, typically followed by periods of close investigation. Approach does not include hore high-intensity behavior during which the intruder is attempting to evade the resident, which is instead classified as chase.
- **Chase:** resident is closely following the intruder around the home cage, while the intruder attempts to evade the resident. Typically interspersed with attempted mount or close investigation. In aggressive encounters, short periods of high intensity chasing between periods of attack are still labeled as attack (not chase), while longer periods of chasing that do not include further attempts to attack are labeled as chasing.
- **Grooming:** usually in a sitting position, the mouse will lick its fur, groom with the forepaws, or scratch with any limb.

Behavior annotation by multiple individuals

For our analysis of inter-annotator variability in behavior scoring, we provided a group of graduate students, postdocs, and technicians in the Anderson lab with the descriptions of close investigation, mounting, and attack given above, and instructed them to score a set of ten resident-intruder videos, all taken from unoperated mice. Annotators were given front- and top-view video of social interactions, and scored behavior using either Bento or the Caltech Behavior Annotator[14], both of which support simultaneous display of front- and top-view video and frame-by-frame browsing and scoring. All but one annotator (Human 7) had previous experience scoring mouse behavior videos; Human 7 had previous experience scoring similar social behaviors in flies.

THE MARS PIPELINE

Mouse detection using the Multi-Scale Convolutional Multibox detector

We used the multi-scale convolutional MultiBox (MSC-MultiBox)[15,16] approach to train a pair of deep neural networks to detect the black and white mice using our 15,000 frame bounding box annotation dataset. Specifically, we used the Inception-Resnet-v2 architecture[17] with ImageNet pre-trained weights, trained using a previously published implementation of MSC-MultiBox for Tensorflow (https://github.com/gvanhorn38/multibox). Briefly, MSC-MultiBox computes a short list of up to K possible object detections proposal (bounding boxes) and associated confidence scores denoting the likelihood of that box containing a target object, in this case the black or white mouse. During training, MSC-MultiBox seeks to optimize location and maximize confidence scores of predicted bounding boxes that best match the ground truth, while minimizing confidence scores of predicted bounding boxes that do not match the ground truth. Bounding box location is encoded as the coordinates of the box's upper-left and lower-right corners, normalized with respect to the image dimensions; confidence scores are scaled between 0 (lowest) and 1 (highest). Once we have the predicted bounding box proposals and confidence score we used NMS (non-maximum suppression) to select the bounding box proposal that best matches with the ground truth.

Detector evaluation

We evaluated performance of the MSC-Mutibox detector by computing the intersection-over-union (IoU) between estimated and human-annotated bounding boxes. The IoU is computed as the area of the intersection of the human-annotated and estimated bounding boxes, divided by the area of their union [cite]. Precision/recall curves (PR curves) were plotted based on the fraction of MSC-Multibox detected bounding boxes with an IoU > X, for X in (0.5, 0.75, 0.85, 0.9, 0.95). Detectors were trained on 13,500 frames from our pose annotation data set, and evaluated on the remaining 1,500 held-out frames, which were randomly sampled from the full data set.

Pose estimation

Following the detection step (above), we use the Stacked Hourglass Network architecture[18] to estimate the pose of each mouse in terms of a set of anatomical keypoints (seven keypoints in top-view videos and eleven keypoints in front-view videos) on our 15,000 frame keypoint annotation dataset. We selected the Stacked Hourglass architecture for its high performance on human pose estimation tasks. The network's

repeated "hourglass" modules shrink an input image to a low resolution, then up-samples it while combining it with features passed via skip connections; representations from multiple scaled versions of the image are thus combined to infer keypoint location. We find that the Stacked Hourglass Network is robust to partial occlusion of the animals, using the visible portion of a mouse's body to infer the location of parts that are hidden.

To construct the input to the Stacked Hourglass Network, MARS crops each video frame to the bounding box of a given mouse plus an extra 65% width and height, pad the resulting image with zero-value pixels to make it square, and resize to 299x299 pixels. To improve generalizability of MARS pose estimation, we used several data augmentation manipulations to expand the effective size of our training set, including random variations in bounding box padding, random blurring, noise, color distortion, and jpeg artifacts, random rotation and flipping, and random perturbations of the bounding box.

Because the Stacked Hourglass Network converts an input image to a heatmap predicting the probability of a target keypoint being present at each pixel, during network training we constructed training heatmaps as 2D Gaussians with standard deviation of 1px centered on each annotator-provided keypoint. During inference on user data, MARS takes the maximum value of the generated heatmap to be the keypoint's location. We trained a separate model for pose estimation in front- vs top-view videos, but for each view the same model was used for both the black and the white mouse.

Pose evaluation

Each pose network was trained on 13,500 video frames from our pose annotation data set, and evaluated on the remaining 1,500 held-out frames, which were randomly sampled from the full data set. The same held-out frames were used for both detection and pose estimation steps. We evaluated the accuracy of the MARS pose estimator by computing the fraction of predicted keypoints on test frames that fell within a radius X of "ground truth", which we defined as the geometric median of human annotations of keypoint location.

Pose features

Building on our previous work[3], we designed a set of 270 spatiotemporal features extracted from the poses of interacting mice, to serve as input to supervised behavior classifiers. MARS's features can be broadly grouped into locomotion, position, appearance, social and image-based categories. Position features describe the position of an animal in relation to landmarks in the environment, such as the distance to the wall of the arena. Appearance-based features describe the pose of the animal in a single frame, such as the orientation of the head and body or the area of an ellipse fit to the animal's pose. Locomotion features describe the movement of a single animal, such as speed or change of orientation of the animal. Image-based features describe the change of pixel intensity between movie frames. Finally, social features describe the position or motion of one animal relative to the other, such as inter-animal distance or difference of orientation between the animals. A full list of extracted features and their definitions can be found in **Table 1**. Most features are computed for both the resident and intruder mouse, however a subset of features are identical for the two animals and are computed only for the resident, as indicated in the Table.

Features are extracted for each frame of a movie, then each feature is smoothed by taking a moving average over three frames. Next, for each feature we compute the mean, standard deviation, minimum,

198 and maximum value of that feature in windows of +/-1, +/-5, and +/-10 frames relative to the current
199 frame, as in [19]; this addition allows MARS to capture how each feature is evolving over time. We thus
200 obtain 12 additional "windowed" features for each original feature; we use 11 of these (omitting the mean
201 of the given feature over +/-1 frame) plus the original feature as input to our behavior classifiers, giving a
202 total of 3144 features.

203 In addition to their use for behavior classification, the pose features extracted by MARS can be loaded and
204 visualized within Bento, allowing users to create custom annotations by applying thresholds to any
205 combination of features. MARS features include many measurements that are commonly used in
206 behavioral studies, such as animal velocity and distance to arena walls.

207 Behavior classifiers

208 We compiled a training set of 7.3 hours of video annotated on a frame-by-frame basis by a single individual
209 for close investigation, mounting, and attack behavior. From these annotated videos, for each behavior
210 we constructed a training set (**X**, **y**) where X_i corresponds to the 3144 windowed MARS pose features on
211 frame *i*, and y_i is a binary label indicating the presence or absence of the behavior of interest on frame *i*.
212 We evaluated performance of and performed parameter exploration for multiple classification
213 algorithms, primarily multilayer perceptron with and without bagging and gradient boosting, using a held-
214 out validation set of videos. A common form of error in many of our tested classifiers was to have
215 sequences (1-3 frames) of false negative or false positives that were shorter than the typical behavior
216 annotation bout. To correct these short error bouts, we introduced a post-processing stage following
217 frame-wise classification, in which the classifier prediction is smoothed using a Hidden Markov Model
218 followed by a three-frame moving average.

219 We found that overall, the highest precision and recall values for individual binary behavior classifiers
220 were achieved by gradient boosting using the XGBoost algorithm[20]; we therefore used this algorithm for
221 the three classifiers presented in this paper. Custom Python code to train novel behavior classifiers is
222 included with the MARS software, and supports both XGBoost and multilayer Perceptron classification.
223 Classifier hyperparameters may be set by the user, otherwise MARS will provide default values.

224 Each trained classifier produces a predicted probability that the behavior occurred, as well as a binarized
225 output created by thresholding that probability value. Following predictions by individual classifiers, MARS
226 combines all classifier outputs to produce a single, multi-class label for each frame of a behavior video. To
227 do so, we select on each frame the behavior label that has the highest predicted probability of occurring;
228 if no behavior has a predicted probability of > 0.5, then the frame is labeled as "other" (no behavior
229 occurring.) The advantage of this approach over training multi-class XGBoost is that it allows our ensemble
230 of classifiers to be more easily expanded in the future to include additional behaviors of interest, because
231 it does not require the original training set to be fully re-annotated for the new behavior.

232 Classifier evaluation

233 Accuracy of MARS behavior classifiers was estimated in terms of classifier Precision and Recall, where
234 Precision = (number of true positive frames) / (number of true positive and false positive frames), and
235 Recall = (number of true positive frames) / (number of true positive and false negative frames). Precision
236 and Recall scores were estimated for the set of trained binary classifiers on a held-out test set of videos
237 not seen during classifier training. Precision-Recall (PR) Curves were created for each behavior classifier

by calculating classifier Precision and Recall values as the decision threshold (the threshold for classifying a frame as positive for a behavior) is varied from 0 to 1.

1 Blanchard, D. C., Griebel, G. & Blanchard, R. J. The Mouse Defense Test Battery: pharmacological and behavioral assays for anxiety and panic. *European Journal of Pharmacology* **463**, 97-116, doi:10.1016/s0014-2999(03)01276-7 (2003).

2 Zelikowsky, M. *et al.* The Neuropeptide Tac2 Controls a Distributed Brain State Induced by Chronic Social Isolation Stress. *Cell* **173**, 1265-1279 e1219, doi:10.1016/j.cell.2018.03.037 (2018).

3 Hong, W. *et al.* Automated measurement of mouse social behaviors using depth sensing, video tracking, and machine learning. *Proceedings of the National Academy of Sciences* **112**, E5351-E5360 (2015).

4 Lee, H. *et al.* Scalable control of mounting and attack by Esr1+ neurons in the ventromedial hypothalamus. *Nature* **509**, 627-632, doi:10.1038/nature13169 (2014).

5 Hong, W., Kim, D. W. & Anderson, D. J. Antagonistic control of social versus repetitive self-grooming behaviors by separable amygdala neuronal subsets. *Cell* **158**, 1348-1361, doi:10.1016/j.cell.2014.07.049 (2014).

6 Kennedy, A., Kunwar, P. S., Li, L., Wagenaar, D. A. & Anderson, D. J. Stimulus-specific neural encoding of a persistent, internal defensive state in the hypothalamus. *bioRxiv* (2019).

7 Remedios, R. *et al.* Social behaviour shapes hypothalamic neural ensemble representations of conspecific sex. *Nature* **550**, 388-392, doi:10.1038/nature23885 (2017).

8 Lin, T.-Y. *et al.* in *European conference on computer vision.* 740-755 (Springer).

9 Andriluka, M., Pishchulin, L., Gehler, P. & Schiele, B. in *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition.* 3686-3693.

10 Wah, C., Branson, S., Welinder, P., Perona, P. & Belongie, S. The caltech-ucsd birds-200-2011 dataset. (2011).

11 Mathis, A. *et al.* DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nat Neurosci* **21**, 1281-1289, doi:10.1038/s41593-018-0209-y (2018).

12 Pereira, T. D. *et al.* Fast animal pose estimation using deep neural networks. *bioRxiv*, doi:10.1101/331181 (2018).

13 Graving, J. M. *et al.* DeepPoseKit, a software toolkit for fast and robust animal pose estimation using deep learning. *Elife* **8**, doi:10.7554/eLife.47994 (2019).

14 Dollár, P. (2014).

15 Erhan, D., Szegedy, C., Toshev, A. & Anguelov, D. in *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2147-2154.

16 Szegedy, C., Reed, S., Erhan, D., Anguelov, D. & Ioffe, S. Scalable, high-quality object detection. *arXiv preprint arXiv:1412.1441* (2014).

17 Szegedy, C., Ioffe, S., Vanhoucke, V. & Alemi, A. A. in *Thirty-first AAAI conference on artificial intelligence.*

18 Newell, A., Yang, K. & Deng, J. in *European Conference on Computer Vision.* 483-499 (Springer).

19 Kabra, M., Robie, A. A., Rivera-Alba, M., Branson, S. & Branson, K. JAABA: interactive machine learning for automatic annotation of animal behavior. *Nature methods* **10**, 64 (2013).

20 Chen, T. & Guestrin, C. in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining.* 785-794.

Position Features

name	units	definition	res	intr.
(p)_x, (p)_y	cm	x,y coordinates of each body part, for p in (nose, left ear, right ear, neck, left hip, right hip, tail)	x	x
centroid_x, centroid_y	cm	x,y coordinates of the centroid of an ellipse fit to the seven keypoints representing the mouse's pose.	x	x
centroid_head_x, centroid_head_y	cm	x,y coordinates of the centroid of an ellipse fit to the nose, left and right ear, and neck keypoints.	x	x
centroid_hips_x, centroid_hips_y	cm	x,y coordinates of the centroid of an ellipse fit to the left and right hip and tail base keypoints.	x	x
centroid_body_x, centroid_body_y	cm	x,y coordinates of the centroid of an ellipse fit to the neck, left and right hip, and tail base keypoints.	x	x
dist_edge_x, dist_edge_y	cm	distance from the centroid of the mouse to the closest vertical (dist_edge_x) or horizontal (dist_edge_y) wall of the home cage.	x	x
dist_edge	cm	distance from the centroid of the mouse to the closest of the four walls of the home cage.	x	x

Appearance Features

name	units	definition	res	intr.
phi	radians	absolute orientation of the mouse, measured by the orientation of a vector from the centroid of the head to the centroid of the hips.	x	x
ori_head	radians	absolute orientation of a vector from the neck to the tip of the nose.	x	x
ori_body	radians	absolute orientation of a vector from the tail to the tip of the neck.	x	x
angle_head_body_l, angle_head_body_r	radians	angle formed by the left(right) ear, neck and left(right) hip keypoints.	x	x
major_axis_len, minor_axis_len	cm	major and minor axis of an ellipse fit to the seven keypoints representing the mouse's pose.	x	x
axis_ratio	none	major_axis_len/minor_axis_len (as defined above).	x	x
area_ellipse	cm^2	area of the ellipse fit to the mouse's pose.	x	x
dist_(p1)(p2)	cm	distance between all pairs of keypoints (p1, p2) of the mouse's pose.	x	x

Locomotion Features

name	units	definition	res	intr.
speed	cm/s	mean change in position of centroids of the head and hips (see Position Features), computed across two consecutive frames.	x	x
speed_centroid	cm/s	change in position of the mouse's centroid (see Position Features), computed across two consecutive frames.	x	x
acceleration	cm/s^2	mean change in speed of centroids of the head and hips, computed across two consecutive frames.	x	x
acceleration_centroid	cm/s^2	change in speed of the mouse's centroid, computed across two consecutive frames.	x	x
speed_fwd	cm/s	speed of the mouse in the direction of ori_body (see Appearance Features).	x	x
radial_vel	cm/s	component of the mouse's centroid velocity along the vector between the centroids of the two mice, computed across two consecutive frames.	x	x
tangential_vel	cm/s	component of the mouse's centroid velocity tangential to the vector between the centroids of the two mice, computed across two consecutive frames.	x	x
speed_centroid_w2, speed_centroid_w5, speed_centroid_w10	cm/s	speed of the mouse's centroid, computed as the change in position between frames either 2, 5, or 10 frames apart.	x	x
speed_(p)_w2, speed_(p)_w5, speed_(p)_w10	cm/s	speed of each keypoint (p) of the mouse's pose, computed as the change in position between timepoints 2, 5, or 10 frames apart.	x	x

Image-based features

name	units	definition	res	intr.
pixel_change	none	mean squared value of (pixel intensity on current frame minus mean pixel intensity on previous frame) over all pixels, divided by mean pixel intensity on current frame (as defined in [Hong et al].)	x	
pixel_change_ubbox_mice	none	pixel change (as above) computed only on pixels within the union of the bounding boxes of the detected mice (when bounding box overlap is greater than 0; 0 otherwise).	x	
(p)_pc	none	pixel change (as above) within a 20 pixel-diameter square around the keypoint for each body part (p).	x	x

Social features

name	units	definition	res	intr.
resident_head_toward_intruder_head /body (resh_twd_itrhb)	none	binary variable that is 1 if the centroid of the other mouse is within a -45° to 45° cone in front of the animal.	x	x
rel_angle_social	radians	relative angle between the body of the mouse (ori_body) and the line connecting the centroids of both mice.	x	x
rel_dist_centroid	cm	distance between the centroids of the two mice.	x	
rel_dist_centroid_change	cm	change in distance between the centroids of the two mice, computed across two consecutive frames.	x	
rel_dist_gap	cm	distance between ellipses fit two the two mice along the vector between the two ellipse centers, equivalent to Feature 13 of [Hong et al].	x	
rel_dist_scaled	cm	distance between the two animals along the line connecting the two centroids, divided by length of the major axis of one mouse, equivalent to Feature 14 of [Hong et al].	x	x
rel_dist_head	cm	distance between centroids of ellipses fit to the heads of the two mice.	x	
rel_dist_body	cm	distance between centroids of ellipses fit to the bodies of the two mice.	x	
rel_dist_head_body	cm	distance from the centroid of an ellipse fit to the head of mouse A to the centroid of an ellipse fit to the body of mouse B.	x	x
overlap_bboxes	none	intersection over union of the bounding boxes of the two mice.	x	
area_ellipse_ratio	none	ratio of the areas of ellipses fit to the poses of the two mice.	x	x
angle_between	radians	angle between mice defined as the angle between the projection of the centroids.	x	
facing_angle	radians	angle between head orientation of one mouse and the line connecting the centroids of both animals.	x	x
dist_m1(p1)_m2(p2)	cm	distance between keypoints of one mouse w.r.t to the other, for all pairs of keypoints (p1, p2).	x	