High-throughput unsupervised quantification of patterns in the natural behavior of marmosets

William Menegas[1,2,*], Erin Corbett[1,*], Kimberly Beliard[1], Haoran Xu[1], Shivangi Parmar[1], Robert Desimone[1], Guoping Feng[1,2]

Affiliations

1 Hock E. Tan and K. Lisa Yang Center for Autism Research, Yang Tan Collective, McGovern Institute for Brain Research, Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA

2 Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA

* Equal contribution

Corresponding author: William Menegas (menegas@mit.edu)

Abstract

Recent advances in genetic engineering have accelerated the production of nonhuman primate models for neuropsychiatric disorders. To use these models for preclinical drug testing, behavioral screening methods will be necessary to determine how the model animals deviate from controls, and whether treatments can restore typical patterns of behavior. In this study, we collected a multimodal dataset from a large cohort of marmoset monkeys and described typical patterns in their natural behavior. We found that these behavioral measurements varied substantially across days, and that behavioral state usage was highly correlated to the behavior of cagemates and to the vocalization rate of other animals in the colony. To elicit acute behavioral responses, we presented animals with a panel of stimuli including novel, appetitive, neutral, aversive, and social stimuli. By comparing these behavioral conditions, we demonstrate that outlier detection can be used to identify atypical responses to a range of stimuli. This data will help guide the study of marmosets as models for neuropsychiatric disorders.

INTRODUCTION

Expanding the number of species across the phylogenetic tree that can be used to test potential treatments for neuropsychiatric disorders should increase the probability that these treatments can generalize to humans[1]. Marmosets have recently gained significant attention as a nonhuman primate model species[2–6] that displays several types of social behaviors which are disrupted in human neuropsychiatric disorders such as vocal communication[7,8], face processing[9,10], and social learning[11–13]. We applied methods for studying natural behavior, vocalizations, responses to stimuli, and cognitive ability to a large cohort of marmosets to find patterns characteristic of typically developing animals. This data will help determine what symptoms are present in neuropsychiatric disorder model animals, and whether treatments can revert the symptoms.

The application of deep learning to the quantification of video data has recently expanded the potential scope of behavioral measurements. Specifically, methods for extracting animal posture from video data[14–16] have been transformative because they allow high-dimensional video data to be distilled into much lower-dimensional postural time series data that retains information about the animals' overt movements. Furthermore, novel approaches to supervised[17,18] and unsupervised[19–22] modeling have demonstrated that patterns in these large new datasets can be identified and interpreted. These computational approaches to ethology have already been applied to the study of neuropsychiatric disorder model mice to reveal similarities and differences across different mouse models of autism spectrum disorder (ASD)[23]. Deep learning has also been applied to the annotation and study of patterns in audio data[24,25]. This has allowed for high-dimensional audio data (spectrograms) to be compressed into low-dimensional time series data, which can be modeled to understand the basic patterns underlying these complex social processes[26].

Because the phenotypes of neuropsychiatric disorders are defined in relation to the behavioral patterns of typically developing controls, we collected a large baseline video and audio dataset from group-housed animals living in captivity comprised of 5,400 sessions from 112 animals living in 36 different family units. We found that the animals' natural behavioral state usage was highly correlated with environmental variables such as their cagemates' behavior and the vocalizations of other animals in the colony. To elicit acute behavioral responses, we presented the animals with a set of stimuli including novel, appetitive, neutral, aversive, and social stimuli. We found that patterns in the acute responses to these stimuli were consistent across animals, meaning that this data could be used to benchmark methods for characterizing the response to each stimulus.

Human patients with ASD can display diverse symptoms[27] including differences in social affect, repetitive behaviors, restricted interests, lower verbal ability, and impaired cognitive ability. Importantly, this diversity is observed even in monogenic forms of ASD, such as in Phelan-McDermid syndrome[28–30], which can be studied in animal models with alterations in the Shank3 gene[31]. These patients have a range of symptoms, and individual patients can exhibit different

combinations of symptoms. For this reason, we aimed to develop a framework for identifying outliers on an individual basis (by comparing data from each animal to the whole-population distribution) rather than on a group basis (comparing across pre-defined groups of animals).

To benchmark our methods, we measured the distinguishability between nine stimulus-response conditions. Surprisingly, we found that most stimuli did not cause large changes in the total usage of each behavioral state. Instead, the responses to these stimuli were better differentiated by comparing the correlation patterns between the usage of each behavioral state and the time spent interacting with the stimuli. Using these patterns, we showed that outlier detection could distinguish between conditions with similar accuracy as 2-class classification. Crucially, we found that outlier detection was also robust to heterogeneity, whereas 2-class classification required homogeneity in the training and test data. This demonstrates the importance of using outlier detection in cases where experimental groups may be heterogeneous, such as in ASD models.

Overall, these results will help guide studies of how marmoset models for neuropsychiatric disorders differ from typical animals, and whether treatments can revert these changes.

RESULTS

Extracting behavioral data from video

Marmosets form stable breeding pairs that cooperate to raise infants during their development, which takes over a year. Because marmosets often give birth to 2 sibling animals at a time, they usually live in a family of at least 4 total animals when raised in captivity[32]. Once the adult female gives birth to another set of siblings, older siblings are paired with new partners in different cages to form their own family units. To study marmosets living in this controlled semi-natural environment, we trained a neural network to label 21 postural key points of multiple animals (Fig. 1a-b) and developed a classifier to distinguish animals based on the color of the hair tuft near the ear, which we artificially dyed to make animals easily distinguishable to a human observer (Fig. 1c-d). This allowed us to extract postural time series data from 36 cages with similar family structure (two adult parents and two juvenile animals).

Principal component analysis of the posture time series data revealed that most of the variance was explained by the animal's X position, Y position, and body angle (Fig. 1e, Fig. S1). This is an expected result from video data where the animals are small compared to the field of view. To extract more detailed information about the animals' body movements, we created an egocentric version of the data that was normalized with respect to X position, Y position, and body angle. In this egocentric data, the principal components of variance reflect differences in posture, such as head angle and limb position (Fig. 1f, Fig. S1). Because the Z position of the animal can vary dramatically as animals climb on the cage walls, we used stereo vision to estimate the animals' depths in the cage (Fig. S2). Recent work has shown that is also possible to estimate 3D postures directly from 2D postures[33], and this will be a promising future direction for enabling similar types of high-throughput video data collection with even fewer cameras. In this study, we used stereo vision to estimate depth since we collected video from multiple cameras above each cage. Notably, we temporarily installed smooth transparent ceilings during the weeks of recording to restrict the animals' ability to climb on the ceiling, and temporarily placed horizontal cage dividers to reduce the total cage size during data collection.

Labeling natural behavior on a frame-by-frame basis is a difficult task because behaviors can have different timescales and levels of description. Additionally, an animal may simultaneously perform multiple behaviors at once. Our approach to this problem was to train a recurrent neural network to use long input sequences of features (25 seconds) to predict the following 5 seconds (Fig. 2a-b). After training the network, we hierarchically clustered the latent space using t-SNE while systematically varying the alpha and perplexity parameters[34]. We then traced cluster identities across hierarchical levels (Fig. 2c, S3). We found that the lowest level of clustering (level 5) resulted in sub-second labels with very specific human descriptions (Fig. 2e-f, S4). The highest level of clustering (level 1) divided the data based on broad categories that humans described as "climbing", "active", "social", and "alone" (Fig. S4). To increase the human-interpretability of these clusters, we plotted the distribution of features across clusters at every

hierarchical level (Fig. S4). This unsupervised hierarchical approach to labeling behavioral states allowed us to study behavioral responses at multiple levels of description.

To understand how the input time series data was weighed in the bottleneck layer of the model, we measured the average attention allocation to previous time points (Fig. S5). We found that most attention was concentrated in the 1-second period prior to the current frame, but that attention was more evenly distributed across the prior 20-second period when animals' behavioral states remained similar across those frames (Fig. S5). This demonstrates a key benefit of using a deep neural network model to describe the ongoing behavioral state. Compared to using features with pre-determined durations (which would specifically set the timescale of the analysis based on the kernel size of those features), this approach allowed the model to access information from different timescales.

Comparing striatal neural data to video data

To determine whether the organization of the behavioral latent space was relevant to the neural representation of ongoing behavior, we recorded striatal neuron activity from 4 animals using silicon probes mounted on movable drives (Fig. 3a-c). Each animal was implanted with two 64 channel probes. We recorded neural activity in the striatum because it is known to contain diverse signals related to movement[35] and social interactions[36,37]. To allow animals to have a full range of movement and to have natural interactions with cagemates, we used on-head amplifiers and an on-head data logger to wirelessly record from these neurons while animals moved freely in their home cages.

Because the number of concurrently recorded neurons was relatively low, we did not use population-based methods to analyze the data – as is possible in studies where hundreds or thousands of neurons are recorded simultaneously[38]. Instead, to estimate whether individual neuronal activity was related to the behavioral states defined by our clustering method, we compared each neuron's firing rate distribution to shuffled data based on the overall frequency of each behavior during the session. We did not track the same neurons across multiple sessions, but instead moved the probes between sessions to record new neurons in subsequent sessions. We pooled the data from 234 sessions (from 4 animals) and hierarchically clustered neurons according to their firing rate enrichment maps (Fig. 3d). We found that the highest-level clusters of striatal neurons had average activity patterns that were clearly related to the behavioral state clustering. Neural cluster 1 had increased activity when animals were in the "social" state (Fig. 3e), cluster 2 had increased activity when animals were in the "alone" state (Fig. 3e) and cluster 3 has increased activity when animals were in the "climbing" state (Fig. 3e). Interestingly, cluster 4 had higher activity when neurons were "alone" or "active" (Fig. 3e). This class of neurons is likely related to movement, because transitions between these two states would be characteristic of an animal alternating between movement and temporary pauses between movements (Fig. S3). Finally, we also found that individual neurons could have similar patterns of enrichment as the group averages (Fig. 3f).

Large-scale data collection reveals variability across days and correlation between cagemates

To treat the study of marmoset behavior as a quantitative data science problem[39,40], we collected a large dataset of natural behavioral data. We recorded video data from 112 individual animals living in 36 families at time points corresponding to the siblings being aged 3 months, 6 months, 9 months, 12 months, and 15 months. Some parents were observed with different sets of siblings, as the data collection took several years to complete. In total, we recorded 5400 sessions (defined as the data from one animal in one day) and sorted the sessions using PCA to visualize the largest sources of variance (Fig. S6). We found that behavioral state usage varied dramatically across days (Fig. 4a-b). Notably, there was a large amount of variability across all sessions (Fig. 4c), and individual animals measured longitudinally had large within-animal variability across sessions as well (Fig. 4d). To determine whether this variability could be caused by non-uniform sleep quality, we used motion watches to measure activity across all hours. We found that animals in our study had very consistent sleep patterns and that sleep quality was similar across days (Fig. S7). Because of the variability in behavioral state usage across days, it would be difficult to find stable phenotypes of individual animals without collecting weeks or months of data per animal. To determine whether animals' stable behavioral traits could be measured more efficiently, we aimed to better understand the sources of the observed variability.

When measuring the time spent in "active" behavioral states across 45 sessions (non-continuous days) in an example family, we found that cagemates displayed highly similar patterns of activity (Fig. 4e-f). We pooled data from all simultaneously measured animals and found a high correlation between the time spent active and the cagemates' time spent active (Fig. 4e). To better understand the timescale of this effect, we compared the average correlation to cagemates across timescales ranging from seconds to hours (Fig. S6). For this analysis, we calculated correlation based on the average usage of all behavioral states. We found that both the correlations to cagemates and the correlation to shuffled data increased as a function of bin size. This is because the distribution of behavioral states becomes more similar to the mean as the sampling window size increases. However, the correlation to cagemates was higher than the correlation to shuffled data across all bin sizes (Fig. S6). Interestingly, measuring an animal's average time spent active relative to cagemates gave an accurate estimate of the animal's average time spent active relative to the colony average (Fig. 4g). Overall, these results indicate that measurements of animals' behavioral traits depend heavily on their social environment.

Correlation between total call rate in a housing room and behavioral state usage

Families of marmosets are housed in separate cages, but animals can make loud calls that likely influence the behavioral state of neighboring cages despite the lack of visual access. Because we found that animals' behavioral state usage was highly correlated to their cagemates' behaviors, we also wanted to measure the effect that these loud vocalizations from neighboring cages could have on behavioral state usage. To do this, we installed microphones outside of each cage in a room, such that the signal would be loudest when a call originated from a nearby cage (Fig. 5a,

S8), but loud calls in the room could still be detected by each microphone. As is standard in this colony, visual barriers between cages prevented animals from directly observing their neighbors. We trained a neural network to identify segments of audio data that contained vocalizations using human labels (Fig. 5b) and trained a separate network to classify the vocalizations according to their call type (Fig. S9).

Similar to behavioral state usage (Fig. 5c), we found that the total call rate varied dramatically across recordings (Fig. 5d). We split the audio data into 5-minute (non-continuous) segments and found that the average call rate in these segments varied from 0 calls per minute to 60 calls per minute (Fig. 5d-e). We measured the correlation between call rate and behavioral state usage and found a strong negative correlation between call rate and the high-level "social" behavioral state (Fig. 5f-g). To better understand which specific behaviors were anticorrelated with high call rate, we plotted the correlation across all 80 low-level behavioral states (Fig. 5h) and found that call rate was specifically anti-correlated with social behavioral states in which the animals are at rest and not moving (Fig. 5h, S4). This implies that a high rate of calls in a room can interrupt animals during social resting states and cause them to preferentially exhibit more active/attentive states. To confirm the generality of this effect, we compared data from 52 animals living in families of 4 animals (two parents and two siblings) to data from 43 animals living in families of 2 animals (two unrelated adults). In both datasets, a similar overall pattern emerged (Fig. S10).

Measuring behavioral response to a panel of stimuli

Because the animals' natural behavior was highly variable and highly correlated to their cagemates' behavior and facility conditions, we aimed to measure species-typical responses to a set of stimuli that could be robustly elicited by presenting animals with these stimuli. We selected a range of stimuli including positive, neutral, negative, novel, and social stimuli (Fig. 6a, Fig. S11). These stimuli were presented to the animals over the course of a 3-week recording session, which was repeated every 3 months based on the age of the juveniles in the cage. We found that most stimuli did not cause large changes in behavioral state usage rates (Fig. 6a, Fig. S12). Interestingly, we also found that some stimuli which did not cause a large change in behavioral state usage did still cause a significant change in the time spent near the stimulus (Fig. 6b). To better understand this data, we looked for correlations between behavioral state usage and the amount of time spent interacting with each stimulus (Fig. S11, Fig. S12).

As an example of this type of analysis, we compared the "toy response" and "tablet response" conditions (Fig. 6c-f). We found that the usage of most behavioral states was similar between these conditions, though a few were different (Fig. 6c-d). This allowed the conditions to be weakly discriminated based on behavioral state usage (Fig. 6d). By contrast, the correlation patterns between behavioral state usage and time spent interacting with each stimulus were very different between conditions (Fig. 6e-f). For example, social behavioral state usage was much more highly correlated with tablet interaction time than with toy interaction time. This resulted in high discriminability between the conditions using the behavioral state correlation data (Fig. 6f).

Overall, these results suggest that some stimuli elicit large responses that can be clearly distinguished based on behavioral state usage, but some stimuli required the behavioral state correlation measurements to be discriminable from other stimulus response conditions (Fig. S12). For the following analyses, we use both behavioral state usage and behavioral state correlation measurements as features for studying the species-typical responses to each stimulus condition.

Outlier detection

Given the heterogeneity across patients in human neuropsychiatric disorders, we aimed to develop a phenotyping method to determine whether each animal had a "typical" or "atypical" response to each stimulus. To validate this methodology, we measured the distinguishability of the behavioral responses to each stimulus based on patterns in the behavioral state usage and behavioral state correlations (Fig. 7).

We first estimated a probability distribution for each feature in each condition (Fig. 7a). Next, we assigned outlier scores for new data, which came either from held-out animals responding to the same stimulus ("Condition 1"), or from animals responding to another stimuli ("Condition 2"). We assigned an outlier score for each feature based on the probability distributions derived from animals in the training set (Fig. 7a). As expected, data from different conditions had higher outlier scores (Fig. 7b). To benchmark the number of animals needed to robustly discriminate between conditions using outlier detection, we compared classification accuracy between 2-class classification using SVM models (Fig. 7c) and 1-class outlier detection (Fig. 7d).

We found that 2-class classification using SVM could accurately discriminate between conditions even with very few animals used as training data (Fig. 7e-f). This approach could be useful in cases where the number of animals available is small, and the phenotype is very consistent across animals because SVM models can assign greater importance to variables that differ between pre-defined classes. By contrast, 1-class outlier detection methods must weigh all variables equally. In spite of this, we found that outlier detection scores were able to accurately discriminate between conditions if the number of animals used in the training set was sufficient (Fig. 7g-h). For studies of nonhuman primates, where the total number of animals used per study must be minimized for ethical and logistical reasons, large datasets of typically developing animals will be valuable to use as control data for studies of disordered states that can use outlier detection to determine the type and severity of the disorders. In this study, we were able to use outlier scores to distinguish between stimulus response conditions with similar accuracy as 2-class classification in cases where a large number of age-matched typically developing animals were used in the training set (Fig. 7). One caveat to this approach is that outlier detection was less accurate when non-age-matched animals were used in the training set (Fig. S13). This is likely because of large scale differences between the behavioral state usage distributions of juvenile and adult animals (Fig. S14).

In the preceding analysis, we used data from typically developing animals that had similar responses to each stimulus. To simulate increased heterogeneity in the dataset, we artificially mixed data from multiple conditions so that we could measure the effect of heterogeneity on the discriminability of conditions (Fig. 8). Using 2-class classification, we found that the score given to these artificial outliers varied significantly based on the number of outliers artificially added to the training dataset (Fig. 8a). By contrast, outlier detection was not dependent on the number of outliers added to the training dataset (Fig. 8b). We measured this property across all pairs of conditions and found that classification-based scores varied based on the number of outliers artificially added to the training set across many pairs of conditions (Fig. 8c-d). By contrast, outlier detection was robust to changes in the heterogeneity of the training data across pairs of conditions (Fig. 8e-f). This demonstrates the potential usefulness of our outlier detection strategy for studying neuropsychiatric disorder model animals that could have heterogeneous phenotypes.

DISCUSSION

Methods for quantifying natural behavior

Finding patterns in natural behavior is a complex problem, and several different approaches may yield meaningful descriptions of video data. Broadly, the analysis of video data can be either pixel-based[19,20] or keypoint based[41,42]. We chose to use a keypoint-based method because of the large variety of cage decorations, objects, stimuli, and other animals present in the study. Pixel-based methods may be more appropriate when the conditions are more constant across video frames, such as in the quantification of facial expression in head-fixed mice[43]. Another key decision is whether to use supervised or unsupervised learning approaches. Supervised learning can be beneficial in cases where a small number of human-recognizable behavioral modules are the subject of the investigation, such as in studies of aggressive responses[44] or escape responses[45] in mice. Unsupervised learning can be more difficult to implement because the patterns identified by the model may not have much biological relevance. For instance, models using raw x/y/z coordinates rather than egocentric coordinates may label different positions in the cage as key behavioral states, rather than different motions. In cases where input data contains missing values or large jumps, the model may identify these features as the main sources of variance. We chose to use an unsupervised learning approach because our goal was to describe the entire distribution of the animals' daily behaviors in a relatively unbiased manner rather than selecting human-identifiable behaviors manually and quantifying the frequency of those behaviors.

Aside from questions regarding methodology, the study of natural behavior is also complicated by the lack of a specific definition of what constitutes a behavioral module. Importantly, behaviors can occur on different timescales, and it is not always clear (a priori) which timescale will be relevant for each study. For instance, some neuropsychiatric disorder model animals may have motor skill deficits requiring a detailed kinematic analysis of sub-second behaviors[46] and others may have deficits in complex social interactions that occur across several seconds or minutes. Many studies have focused on sub-second units of behavior because they are the shortest segments that can be clearly differentiated into units, often referred to as syllables[19,20]. Another complication is that behavior can be described at many levels of specificity. At higher levels of specificity, more data is needed to find patterns. But at lower levels of specificity, patterns may not be apparent due to mixed effects within a single category. We chose a method where each frame would have a single label, while preserving hierarchical information. This allowed us to identify patterns at multiple levels of specificity.

Large-scale data collection and reproducibility

Logistical constraints and ethical considerations limit the ability to collect large behavioral datasets from nonhuman primates, as is common in several other model organisms. NHP research is comparatively much more expensive, due to the larger living spaces required, longer generation times, and more complex veterinary care needed to maintain high-quality living

conditions. We opportunistically collected data from a large colony of marmosets that was primarily used for transgenic model production. This open-source dataset will be useful for future studies that can use this data as a large normative control group to achieve strong statistical power even in cases where the number of animals in experimental groups is lower. Additionally, sharing this large dataset will allow computational neuroscientists to develop new methods for analysis without conducting their own large-scale data collection efforts. This is crucial because it will allow their models to be trained on data from a variety of animals and conditions, rather than on example datasets comprised of a small number of videos.

When comparing natural behavioral data across facilities, challenges can arise due to environmental variables such as the behavioral distribution of family members, number of vocalizations in the room, room configuration, and stimuli present. Although these conditions may change the exact distribution of behavioral state usage, we expect that the general principles for marmoset natural behavioral analysis identified in this study should apply across facilities.

High dimensional phenotyping for neuropsychiatric disorders

Neuropsychiatric disorders can have a wide range of symptoms, and disease model animals will need to be tested in a variety of contexts to study their natural behavior, responses to stimuli, cognitive ability, and performance in other assays. People with autism can have a variety of symptoms, such as differences in social affect, vocal communication, cognitive ability, responses to novel or high-intensity sensory stimuli, restricted interests, or repetitive behavior. Some of these symptoms may only manifest in certain conditions, such as in novel environments or when meeting new people. Some symptoms may also be secondary effects, such as differences in organ health[47]. To test which phenotypes are present in disease model animals, high throughput and high dimensional methods will be essential. This is particularly important for testing whether treatments can rescue multiple deficits or only a subset of the deficits.

Outlier detection for individual phenotyping

Several physical traits are routinely measured in developing human infants to ensure that they are within normal ranges. For instance, height and weight charts are used to determine how each infant compares to the overall population. Identifying outliers using these measurements is important, because they may need additional care. However, assessing normality in social and cognitive domains is more difficult because of the complexity of these behaviors. Model animals can serve a crucial purpose for preclinical screening in these areas, because measurements of complex behaviors can be collected more reproducibly in well-controlled lab environments. Additionally, even patients with the same monogenic form of autism can exhibit substantial heterogeneity. For instance, one ASD patient could have typical behavioral patterns with respect to sensory stimuli, but atypical behavioral patterns related to social stimuli. Because of this, outlier detection in model animals should be done on an individual basis, to identify whether each individual is an outlier in each condition or assay.

Many studies of disease model mice have focused on group-based differences in specific assays[31,48]. Although bulk differences between heterogeneous classes can be detectible with large sample sizes (as is typical in mouse studies), this is logistically impractical for NHP research. Rather than producing large cohorts of control animals for each study, a much more efficient use of animals is to collect large datasets of control data that can be used to power future studies with smaller cohorts. This could be done for data related to natural behaviors, responses to stimuli, cognitive tests, and other assays. These datasets could be standardized and shared across institutes, as is becoming more common for mouse research[49]. The normative models from these datasets could be used to test whether smaller cohorts of animals (controls, disease models, or treated animals) are outliers relative to the dataset. The open-source data from this study will contribute to these ongoing efforts.

METHODS

Animals

For this study, we opportunistically collected behavioral data from 112 animals that were part of a large pre-existing colony used for transgenic animal production. No animals were obtained or bred specifically for use in this project. Additionally, we used data from 4 animals that had been implanted with silicon probes for electrophysiology. The animals were all group housed in family units consisting of 2 adult animals and juveniles up to the age of 15 months. After 15 months, the juveniles would be paired with other animals (in new cages). The lights in the facility were turned on at 7:00 am and turned off at 7:00 pm. Veterinary checks and other work in the rooms (such as cage changes) were done between 7:00 am and 10:00 am. Each session of data was collected between 10:00 am and 4:00 pm. After 5:00 pm, the animals were not disturbed to ensure that their sleep cycle would not be affected.

All procedures were performed in accordance with the National Institutes of Health Guide for the Care and Use of Laboratory Animals and approved by the MIT Committee on Animal Care (IACUC). The MIT Division of Comparative Medicine (DCM) routinely monitored animal welfare and provided them with veterinary care as needed.

Home cage video recording

Cages were scheduled for recording sessions based on the ages of the juvenile animals living in the cage, such that they would be recorded for 3 weeks at a time when the animals were 3 months, 6 months, 9 months, 12 months, and 15 months old. Recording sessions were scheduled for weekdays to avoid variance in environmental conditions between weekdays and weekend days. Once juveniles were 15 months old, they were transferred to new cages to begin their own families with new partners. These measurements began once the animals were 3 months old, because that was when the hair tufts near their ears was long enough to dye (to differentiate between animals) and also because the animals began to behave more independently from their parents (rather than clinging to one of their parents) at this time. One week before each recording session, cages were changed to accommodate a transparent ceiling panel to facilitate top-down video recording. The smooth transparent ceiling panels prevented animals from climbing on the ceiling during the recording sessions, and horizontal dividers were added to further restrict the total cage size during the recordings.

One day before the first recording session, cameras were placed above the cages, the ceiling was replaced with a transparent panel, and animals' hair was dyed. The hair colors were planned such that each cage would have a red-haired animal, white-haired animal, blue-haired animal, and yellow-haired animal. Red, blue, and yellow were chosen as the dye colors to maximize distinguishability between animals, according to human observers. Human distinguishability is essential to this approach since we used a supervised model to classify the animals based on hair color. To apply hair dye, one experimenter caught and restrained the animal while a second

experimenter applied the dye into the hair. Dye was applied liberally to the tufts of hair near the ears so that there was no white hair remaining. Animals were then left in a temporary holding cage for 20 minutes so that the dye could dry before they were transferred back to their home cage. This was done to prevent family members from rubbing the dye onto themselves and to prevent the dye from marking items/surfaces in the home cage. After animals were returned to their cages, we waited at least one day before any recordings were performed to allow the dye to dry completely and allow the other animals to habituate to the dye.

Video recording was done using ZED stereo cameras mounted above the cages and collected using the ZED SDK developed by the manufacturer. Video was collected from 2 cameras at 60 frames per second, and with 720p resolution. Each image had a size of 1280 pixels by 720 pixels. Camera settings (i.e. brightness/contrast/exposure/gain) were manually adjusted after each camera installation to avoid image quality issues caused by oversaturation or insufficient gain. Because light sources directly above the cages would cause glare in the video, we either moved the cages or moved the light sources to ensure that cages were not directly below light bulbs. After exporting the proprietary "svo" videos generated by the cameras into the "avi" format using the ZED SDK, further processing was done using python and Matlab.

Pose extraction

Postural keypoints were extracted from the raw video data using DeepLabCut (DLC)[14] and sorted using a two-step method we developed for this specific use case. First, tracklets were joined based on kinematic features (as in the typical DLC pipeline). Then, tracklets were sorted based on the average distance between each tracklet and the labels of each hair color. To train a network to recognize hair color, we placed one label on the center of the head of each animal, rather than on each hair tuft. This approach leveraged the kinematic data to join nearby points and used the hair color data to ensure that animal identities were not mixed after animals had close interactions or were occluded by other animals. Egocentric postures were calculated by setting the middle point on the body to (0,0) and then rotating all points based on the angle formed by the front point on the body and the back point on the body. Features were then calculated based on both the raw time series data and the egocentric time series data. All code and posture time series data will be made available online alongside the publication of this paper. Raw data will be available upon request, due to logistical issues with publicly hosting the entire unprocessed dataset.

Behavioral state labeling

A recurrent neural network was trained to use 25 seconds of input data (features) to predict the next 5 seconds of data (features). This network had a bottleneck size of 256, and the data from this bottleneck was used for hierarchical t-SNE clustering. As suggested in tree-SNE[34], we systematically varied the alpha (exaggeration) and perplexity (# of neighbors) in the t-SNE clustering algorithm. At high levels of alpha/perplexity, we observed fewer distinct clusters. At

lower levels of alpha/perplexity, we observed more distinct clusters. Initial clustering was done using Density-Based Spatial Clustering of Applications with Noise (DBSCAN)[50]. Cluster labels were assigned to new data (obtained after initial model training) using openTSNE[51] to embed new points into the existing embedding, and then using k-nearest neighbors (KNN) to assign a state label. This was done at the lowest hierarchical level only, because higher levels could be inferred from this information. All code and model output data will be made available online alongside the publication of this paper.

Home cage audio recording

Microphones were placed outside of each cage under study and also outside of each neighboring cage. Small wireless microphones were used so that they could be placed around the room without adding any wires or large equipment that could potentially draw the animals' attention. Audio data was recorded at 48,000 samples per second (48 kHz) and logged locally on each wireless recording device. The data from each device was synchronized to the video data using a series of 5 claps from a clapper board at the beginning and end of each video. These claps were clearly visible in the video and easily identifiable in the audio data by a human observer. The clap onset time was manually annotated for each video and audio recording.

Vocalization labeling

Audio data was converted into spectrograms using SciPy[50] with nperseg = 700, noverlap = 525, and nfft = 2000. These spectrograms were segmented using a supervised neural network trained to perform binary classification where 0 = no call and 1 = call. To determine the average call rate (Fig. 5), a threshold of 0.5 was applied to the output of the neural network, such that all outputs below 0.5 were set to 0 and all outputs above 0.5 were set to 1. Spectrograms were segmented into discrete calls using this binary classification. Next, the discrete calls were classified by a separate neural network trained to perform classification based on human categorization of calls in the training set into 8 classes. The types of calls labeled by this model were limited to the types of calls that were commonly observed in our training data. The audio data will be made available online alongside the publication of this paper.

Neural data acquisition

Animals were implanted with silicon probes mounted on drives that could be lowered into the striatum. Data was collected using on-head amplifiers and on-head data loggers. Two silicon probes with 64 channels each were implanted into each of the four animals according to the protocol published by Cambridge Neurotech[52] and lowered between sessions using drives made by Cambridge Neurotech. The full description of this surgical procedure will be published in a separate paper. Data was logged on-head during each session and the logger was removed so that it could be charged, and the data could be downloaded between sessions. This reduced the total weight on the head of the animal during the non-recording time. During non-recording time, a lightweight black cover was attached in place of the data logger. Data was acquired at a rate of

25,000 samples per second (25 kHz) using data loggers made by White Matter. Spike sorting was performed using Kilosort[53] version 3 followed by further manual curation. Units with missing data (either due to drift or due to lapses in the connection between the amplifiers and the logger) were not included for further analysis.

Neural data analysis

Neural data was aligned to video data using a flashing light on the top of the data logger that could be seen by the camera at the beginning and end of each session. Each data format (video, audio, and neural) had a different sampling rate. Video was sampled at 60 Hz, audio was sampled at 48,000 Hz, and neural data was sampled at 25,000 Hz. We converted all data into bins of 60 samples per second to match the rate of the video recording, which was the slowest of the three data modalities. We compared the observed distribution of spikes across the behavioral states to a simulated distribution based on the frequency of behavioral events in each session. We clustered neurons based on these normalized spike distributions across the behavioral latent space, and manually inspected the 4 highest-level clusters using post-hoc statistical tests to determine whether the neurons in each cluster had increased activity during each of the highest-level behavioral states.

Stimulus presentation

Stimuli were presented over a course of 3 weeks, and the same series of stimuli were presented to each cage every 3 months. The first 5 sessions were "control" sessions. In this condition, a horizontal divider was placed in the middle of the cage to restrict the total cage size and simplify analysis. Ledges, branches, and hammocks were removed so that animals would not be occluded from view. Nest boxes and other cage attachments were also removed, so that animals could not leave the field of view. A bed was placed in the rear left corner of the cage so that animals had a comfortable place to sit when resting. Food was placed into two bowls that fit into slots designed to hold them so that they remained in place. Water was freely available through spouts on the side of the cages. At the beginning of each session, cameras were turned on and audio/neural data was synchronized to the cameras as described above (using a clapper board to synchronize the audio data to the video and using the lights on top of the data logger to synchronize the neural data to the video). After the completion of each session, cages were restored to their normal state.

After the 5 days of these control sessions, stimuli were presented once per day. During these sessions, the cages were set up as described above and then one of the 8 types of stimuli was added. Including the control sessions, this means that there were 9 total stimulus-response conditions (one control condition, and 8 stimulus conditions). Because the control condition was performed 5 times per set of recordings, and the toy conditions was performed 3 times per set of recordings, there was not an even number of each condition in the total dataset. The conditions are described in the following section.

1) Control. The only items present in the cage were a bed that animals could rest in, two bowls of food placed in slots designed to hold the bowls, and water-dispensers. These items were also present in the other stimulus conditions.

2) Novel object. A large and brightly colored plastic dog toy was placed in the cage. Importantly, this type of object is not typically placed in the cage. Marmosets cautiously approach these objects but have low overall interest in them.

3) Treat. A mixture of foraging material and foods high in sugar were placed in a bowl. Foods such as banana chips, marshmallows, pudding, craisins, and dried mango were mixed into the foraging materials. Importantly, these foods were not regularly given as part of the standard diet. Marmosets quickly approach and consume these highly appetitive items.

4) Noise. A movie played by a computer outside of the cage. Marmosets do not have visual access to the movie, and only hear the dialogue and soundtrack. This "background noise" stimulus is not meant to have a large impact on their behavior.

5) Mask. A gorilla mask was placed outside the cage in the same position as the computer playing the "background noise" stimulus. The mask stimulus is similar to "human intruder" experiments. Marmosets have a very large negative response to this stimulus, often climbing to the furthest corner of the cage in a group, and sometimes making angry calls.

6) Toys. A set of colorful wicker balls that animals frequently chew on and interact with. The toys are presented for 3 consecutive days. The wicker balls are attached to the side of the cage using zip ties so that they cannot be moved to different positions in the cage, which could complicate the interpretation of the data.

7) Intruder animal. A novel animal from a different family is placed in a transfer box inside the cage. Animals were allowed to have auditory/visual access to the animal, but not able to physically interact (they were separated by a transparent plastic panel). This was done to prevent injury of either the family or the intruder animals. This is a highly salient stimulus and causes a large increase in activity.

8) Tablet 1. Animals were presented with a neutral stimulus (blue square) in the center of the screen. They could touch the screen to receive a syrup reward that was delivered from a metal spout attached to the bottom of the tablet. Further cognitive tasks used the same apparatus, but had increased task demands (forced choice, following a moving target, remembering past trials, or foraging) and took place individually rather than in the group setting. This cognitive data will be fully explored in an accompanying study.

9) Tablet 2. Animals were presented with a social stimulus (marmoset face) in the center of the screen. They could touch the screen to receive a syrup reward that was delivered from a metal spout attached to the bottom of the tablet.

2-class classification

For 2-class classification, we used a support vector machine (SVM) approach to distinguish the behavioral responses to different stimuli. First, we normalized features using z-scoring (so that each feature would have a mean of 0 and a standard deviation of 1). Then, we trained 2-class SVM models to discriminate between each pair of stimulus-response conditions using these normalized features (derived from both behavioral state usage and behavioral state correlations). We evaluated these models by calculating scores for held-out data from each class.

Several other model types could be used for 2-class classification such as Gaussian Naïve Bayes (GNB) or Linear discriminant analysis (LDA). Because we obtained high discriminability between conditions using SVM models even when using a small number of animals as training data, we did not seek to further optimize our 2-class classification approach.

Outlier detection

For 1-class outlier detection, we first generated a probability density estimate (PDE) for each input feature based on the distribution of that feature in the training data. This PDE for behavioral state usage features was defined in the range of (0,1) and the PDE for behavioral state correlation features was defined in the range of (-1,1). Both types of distributions were made with a kernel bandwidth of 0.2. Outlier scores were calculated by evaluating the PDE at a single point to obtain a density value. Importantly, this is a relative measurement (i.e. if x_1 has a higher density value than x_2, then x_1 is relatively more likely to have been drawn from the distribution than x_2). As such, these measurements are not probability measurements because we evaluated the PDE at a fixed point rather than finding the probability that the data would fall within a certain interval. We calculated this relative density measurement for each input feature (including behavioral state usage features and behavioral state correlation features), and then averaged across these values to generate a single outlier score.

To estimate the number of animals needed, we randomly sampled a set number of animals (i.e. 4, 8, 16, 32) from the training set and compared the scores of held-out animals from the same condition to scores obtained from other conditions. As the number of animals in the training set increases, the resulting PDE should more accurately estimate the true distribution, and outliers should be more discriminable from inliers. In line with this reasoning, we observed that when the number of observations (animals) used as training data was low, all data used in the test set had a high outlier score – regardless of whether it came from the same condition or different conditions. In these cases (with few observations in the training set) each new observation would appear to be an outlier. By contrast, when we increased the number of observations (animals) used as training data, the outlier scores could accurately discriminate between conditions (as measured by auROC). For auROC analysis, we calculated the area under the ROC curve, such that a score of 0.5 would indicate a model with no discriminative ability and a score of 1.0 would indicate perfect discrimination between the conditions.

Creating artificial heterogeneity to test 2-class classification and outlier detection

To determine whether each model was robust to outliers, we artificially added "outlier" data to the training set by mixing data from different stimulus-response conditions. To do this, we pooled data from all other stimulus-response conditions and randomly selected a number of datasets (i.e. 0, 2, 4, 8, 16) to add into one of the conditions. We then calculated either 2-class classification scores or outlier scores for these artificial outliers as described above. The "change in score" metric indicates the difference between the average score given to outliers when the number of outliers added was 0 and when the number of outliers added was 2, 4, 8, or 16.

Data and code availability

Data and code associated with this paper will be available online upon publication. Processed data types such as postural time series data, behavioral state data, and vocalization time series data will be publicly hosted. Large data types such as raw multi-view video data will be available upon request, due to logistical issues with publicly hosting the entire unprocessed dataset.

REFERENCES

1. Barron, H. C., Mars, R. B., Dupret, D., Lerch, J. P. & Sampaio-Baptista, C. Cross-species neuroscience: closing the explanatory gap. *Philosophical Transactions of the Royal Society B* **376**, (2021).

2. Miller, C. T. *et al.* Marmosets: A Neuroscientific Model of Human Social Behavior. *Neuron* **90**, 219–233 (2016).

3. Sasaki, E. *et al.* Generation of transgenic non-human primates with germline transmission. *Nature 2009 459:7246* **459**, 523–527 (2009).

4. Bendor, D. & Wang, X. The neuronal representation of pitch in primate auditory cortex. *Nature 2005 436:7054* **436**, 1161–1165 (2005).

5. Okano, H. Current Status of and Perspectives on the Application of Marmosets in Neurobiology. *Annu Rev Neurosci* **44**, 27–48 (2021).

6. Miller, C. T. *et al.* Marmosets: A Neuroscientific Model of Human Social Behavior. *Neuron* **90**, 219–233 (2016).

7. Miller, C. T., Beck, K., Meade, B. & Wang, X. Antiphonal call timing in marmosets is behaviorally significant: interactive playback experiments. *J Comp Physiol A Neuroethol Sens Neural Behav Physiol* **195**, 783–789 (2009).

8. Li, J., Aoi, M. C. & Miller, C. T. Representing the dynamics of natural marmoset vocal behaviors in frontal cortex. *bioRxiv* (2024) doi:10.1101/2024.03.17.585423.

9. Hung, C.-C. *et al.* Functional mapping of face-selective regions in the extrastriate visual cortex of the marmoset. *J Neurosci* **35**, 1160–1172 (2015).

10. Mitchell, J. F. & Leopold, D. A. The marmoset monkey as a model for visual neuroscience. *Neurosci Res* **93**, 20–46 (2015).

11. Takahashi, D. Y., Liao, D. A. & Ghazanfar, A. A. Vocal Learning via Social Reinforcement by Infant Marmoset Monkeys. *Curr Biol* **27**, 1844-1852.e6 (2017).

12. Meisner, O. C. *et al.* Development of a Marmoset Apparatus for Automated Pulling (MarmoAAP) to Study Cooperative Behaviors. *bioRxiv* (2024) doi:10.1101/2024.02.16.579531.

13. Tyree, T. J., Metke, M. & Miller, C. T. Cross-Modal Representation of Identity in Primate Hippocampus. *Science* **382**, 417 (2023).

14. Lauer, J. *et al.* Multi-animal pose estimation, identification and tracking with DeepLabCut. *Nat Methods* **19**, 496–504 (2022).

15. Pereira, T. D. *et al.* SLEAP: A deep learning system for multi-animal pose tracking. *Nat Methods* **19**, 486–495 (2022).

16. Dunn, T. W. *et al.* Geometric deep learning enables 3D kinematic profiling across species and environments. *Nat Methods* **18**, 564 (2021).

17. Kabra, M., Robie, A. A., Rivera-Alba, M., Branson, S. & Branson, K. JAABA: interactive machine learning for automatic annotation of animal behavior. *Nat Methods* **10**, 64–67 (2013).

18. Goodwin, N. L. *et al.* Simple Behavioral Analysis (SimBA) as a platform for explainable machine learning in behavioral neuroscience. *Nat Neurosci* (2024) doi:10.1038/S41593-024-01649-9.

19. J., B. G., M., C. D., William, B. & W., S. J. Mapping the stereotyped behaviour of freely moving fruit flies. *J R Soc Interface* **11**, 20140672 (2014).

20. Wiltschko, A. B. *et al.* Mapping Sub-Second Structure in Mouse Behavior. *Neuron* (2015) doi:10.1016/j.neuron.2015.11.031.

21. Hsu, A. I. & Yttri, E. A. B-SOiD, an open-source unsupervised algorithm for identification and fast prediction of behaviors. *Nat Commun* **12**, (2021).

22. Chaumont, F. de *et al.* Live Mouse Tracker: real-time behavioral analysis of groups of mice. *bioRxiv* 345132 (2018) doi:10.1101/345132.

23. Klibaite, U. *et al.* Deep phenotyping reveals movement phenotypes in mouse neurodevelopmental models. *Mol Autism* **13**, 1–18 (2022).

24. Coffey, K. R., Marx, R. G. & Neumaier, J. F. DeepSqueak: a deep learning-based system for detection and analysis of ultrasonic vocalizations. *Neuropsychopharmacology* **44**, 859–868 (2019).

25. Fonseca, A. H. O., Santana, G. M., Bosque Ortiz, G. M., Bampi, S. & Dietrich, M. O. Analysis of ultrasonic vocalizations from mice using computer vision and machine learning. *Elife* **10**, (2021).

26. Grijseels, D. M., Fairbank, D. A. & Miller, C. T. A model of marmoset monkey vocal turn-taking. *Proceedings of the Royal Society B: Biological Sciences* **291**, 20240150 (2024).

27. Buch, A. M. *et al.* Molecular and network-level mechanisms explaining individual differences in autism spectrum disorder. *Nat Neurosci* **26**, 650–663 (2023).

28. Frank, Y. The Neurological Manifestations of Phelan-McDermid Syndrome. *Pediatr Neurol* **122**, 59–64 (2021).

29. Schön, M. *et al.* Definition and clinical variability of SHANK3-related Phelan-McDermid syndrome. *Eur J Med Genet* **66**, 104754 (2023).

30. Phelan, K. *et al.* Phelan-McDermid syndrome: a classification system after 30 years of experience. *Orphanet J Rare Dis* **17**, (2022).

31. Peça, J. *et al.* Shank3 mutant mice display autistic-like behaviours and striatal dysfunction. *Nature* **472**, 437–442 (2011).

32. Del Rosario, R. C. H. *et al.* Sibling chimerism among microglia in marmosets. *bioRxiv* (2023) doi:10.1101/2023.10.16.562516.

33. Gosztolai, A. *et al.* LiftPose3D, a deep learning-based approach for transforming two-dimensional to three-dimensional poses in laboratory animals. *Nat Methods* **18**, 975–981 (2021).

34. Robinson, I. & Pierce-Hoffman, E. Tree-SNE: Hierarchical Clustering and Visualization Using t-SNE. (2020).

35. Markowitz, J. E. *et al.* The Striatum Organizes 3D Behavior via Moment-to-Moment Action Selection. *Cell* **174**, 44-58.e17 (2018).

36. Baéz-Mendoza, R. & Schultz, W. Performance error-related activity in monkey striatum during social interactions. *Sci Rep* **6**, (2016).

37. Klein, J. T. & Platt, M. L. Social information signaling by neurons in primate striatum. *Curr Biol* **23**, 691–696 (2013).

38. Stringer, C., Pachitariu, M., Steinmetz, N., Carandini, M. & Harris, K. D. High-dimensional geometry of population responses in visual cortex. *bioRxiv* 374090 (2018) doi:10.1101/374090.

39. Anderson, D. J. & Perona, P. Toward a Science of Computational Ethology. *Neuron* **84**, 18–31 (2014).

40. Mobbs, D. *et al.* Promises and challenges of human computational ethology. *Neuron* **109**, 2224–2238 (2021).

41. Weinreb, C. *et al.* Keypoint-MoSeq: parsing behavior by linking point tracking to pose dynamics. *bioRxiv* (2023) doi:10.1101/2023.03.16.532307.

42. Schneider, S., Lee, J. H. & Mathis, M. W. Learnable latent embeddings for joint behavioural and neural analysis. *Nature* **617**, 360–368 (2023).

43. Dolensek, N., Gehrlach, D. A., Klein, A. S. & Gogolla, N. Facial expressions of emotion states and their neuronal correlates in mice. *Science (1979)* **368**, (2020).

44. Nair, A. *et al.* An approximate line attractor in the hypothalamus encodes an aggressive state. *Cell* **186**, 178 (2023).

45. Yilmaz, M. & Meister, M. Rapid Innate Defensive Responses of Mice to Looming Visual Stimuli. *Current Biology* **23**, 2011–2015 (2013).

46. Preisig, D. F. *et al.* High-speed video gait analysis reveals early and characteristic locomotor phenotypes in mouse models of neurodegenerative movement disorders. *Behavioural Brain Research* **311**, 340–353 (2016).

47. Tian, Y. E. *et al.* Evaluation of Brain-Body Health in Individuals With Common Neuropsychiatric Disorders. *JAMA Psychiatry* **80**, 567 (2023).

48. Welch, J. M. *et al.* Cortico-striatal synaptic defects and OCD-like behaviours in Sapap3-mutant mice. *Nature* **448**, 894–900 (2007).

49. Aguillon-Rodriguez, V. *et al.* Standardized and reproducible measurement of decision-making in mice. *Elife* **10**, (2021).

50. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods 2020 17:3* **17**, 261–272 (2020).

51. Poličar, P. G., Stražar, M. & Zupan, B. openTSNE: A Modular Python Library for t-SNE Dimensionality Reduction and Embedding. *J Stat Softw* **109**, 1–30 (2024).

52. Jacobs, T., De Zeeuw, C. & Romano, V. Standard Operating Protocol: Implantation of Cambrige NeuroTech chronic silicon probe and mini-amp-64 digital headstage in mice. (2023) doi:10.21203/RS.3.PEX-2188/V2.

53. Pachitariu, M., Sridhar, S. & Stringer, C. Solving the spike sorting problem with Kilosort. *bioRxiv* 2023.01.07.523036 (2023) doi:10.1101/2023.01.07.523036.

ACKNOWLEDGEMENTS

We thank Alex Mathis and Mackenzie Mathis for the helpful discussions and access to beta versions of their software. We thank MIT DCM for support with animal care. This work was funded by the Yang-Tan Center for Molecular Therapeutics in Neuroscience and the Tan-Yang Center for Autism Research of the Yang Tan Collective at MIT, the Stanley Center for Psychiatric Research at Broad Institute of MIT and Harvard, the Poitras Center for Psychiatric Disorders Research at MIT, and the Simons Center for the Social Brain at MIT.

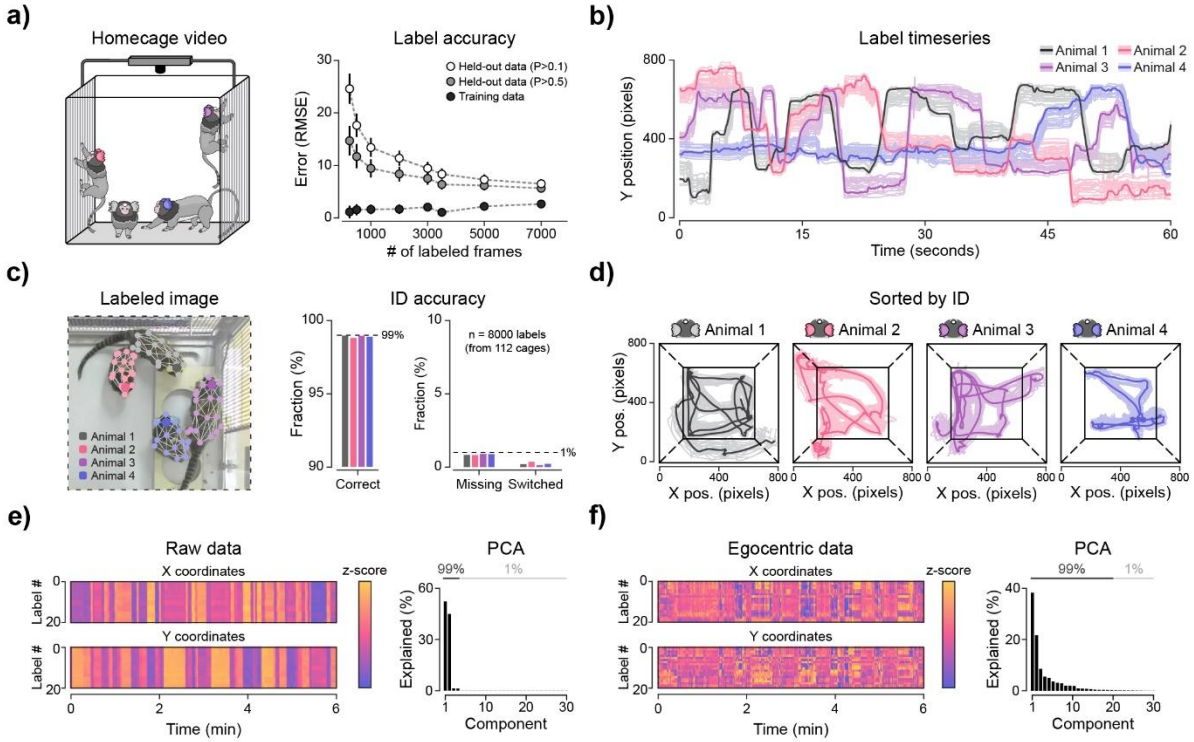
AUTHOR CONTRIBUTIONS

W.M. and E.C. designed the study and equipment. W.M., E.C., and K.B. collected the behavioral datasets. W.M., S.P., and H.X. collected the neural dataset. W.M. performed the analysis and wrote the manuscript. All authors reviewed the manuscript.

DECLARATION OF INTERESTS

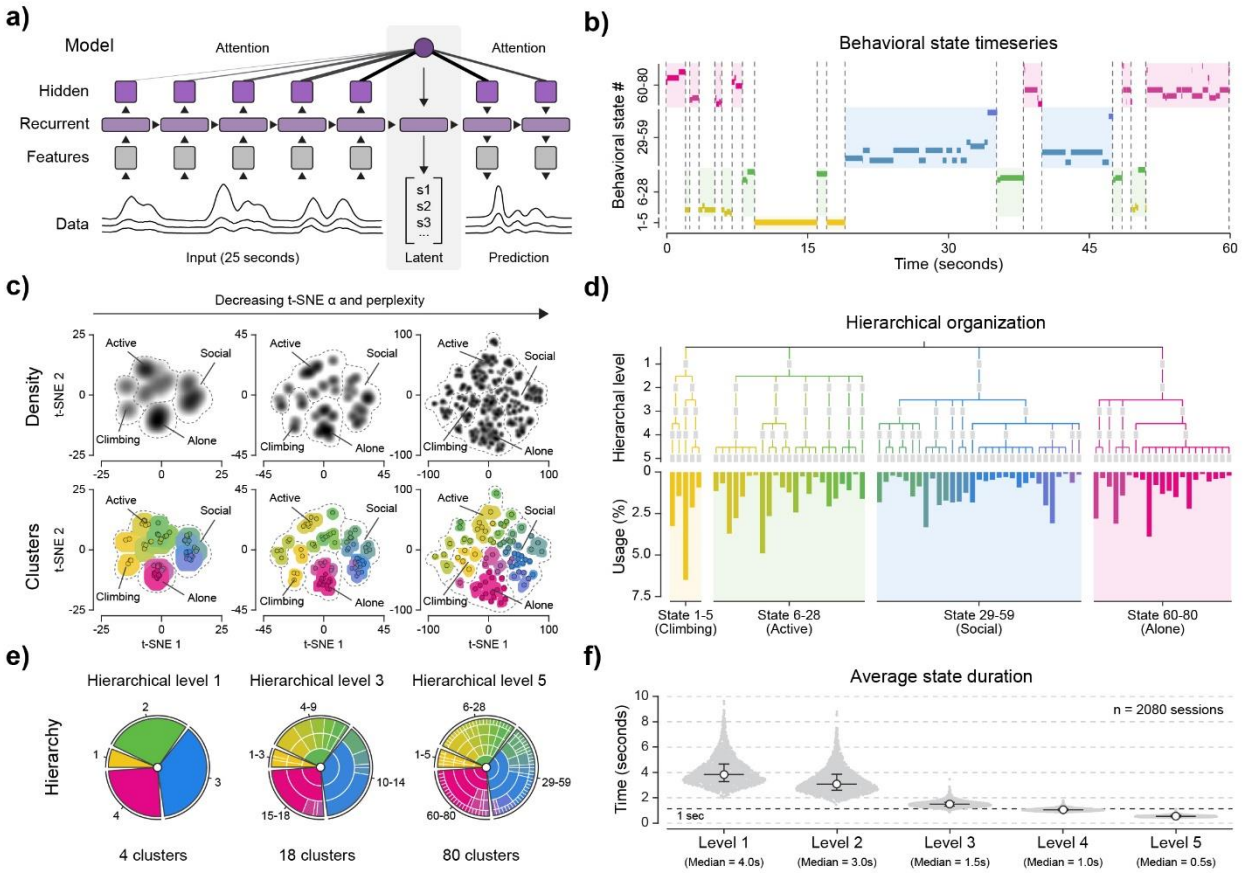
The authors declare no competing interests.

Figure 1: Pose estimation for marmosets in group housing



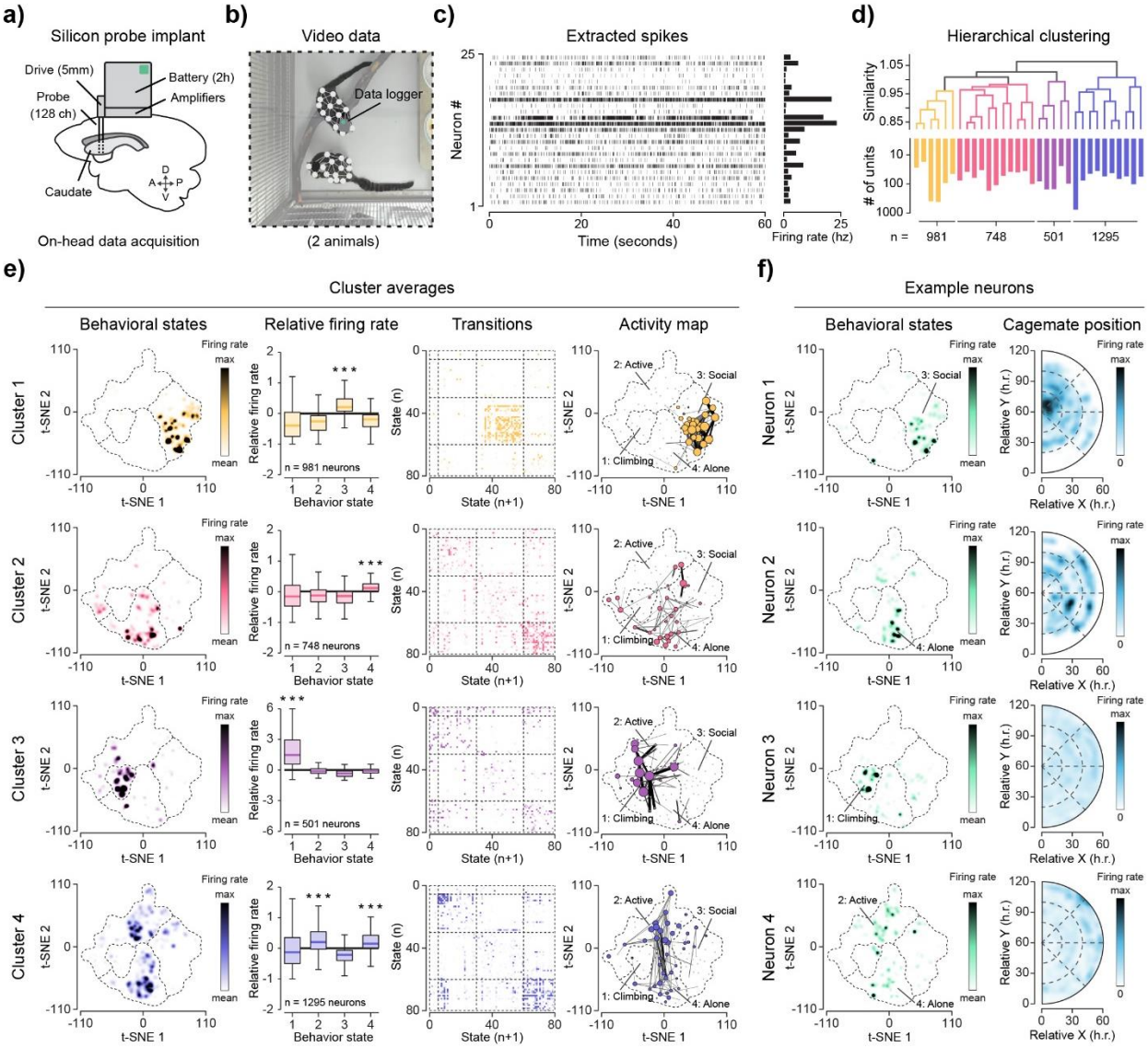
a) Schematic of home cage video recording, and keypoint label accuracy (pixel-based root mean square error (RMSE) across an increasing number of labeled frames). Bars indicate standard deviation across training sessions. **b)** Example time series of all keypoints (Y coordinate), colored based on animal hair color. **c)** Example labeled image, and ID accuracy based on 8000 labeled animals drawn from 112 different camera installations. “Missing” indicates that an animal was not labeled and “Switched” indicates that an animal was labeled with the wrong identity. **d)** Example time series of all keypoints (X and Y coordinates), colored based on animal hair color. **e)** Raw data time series (z-scored) across 6 minutes (left), and PCA of raw data (right). **f)** Egocentric data time series (z-scored) after normalization using X/Y position and orientation (left), and PCA of egocentric data (right).

Figure 2: Hierarchical description of natural behavior using deep learning



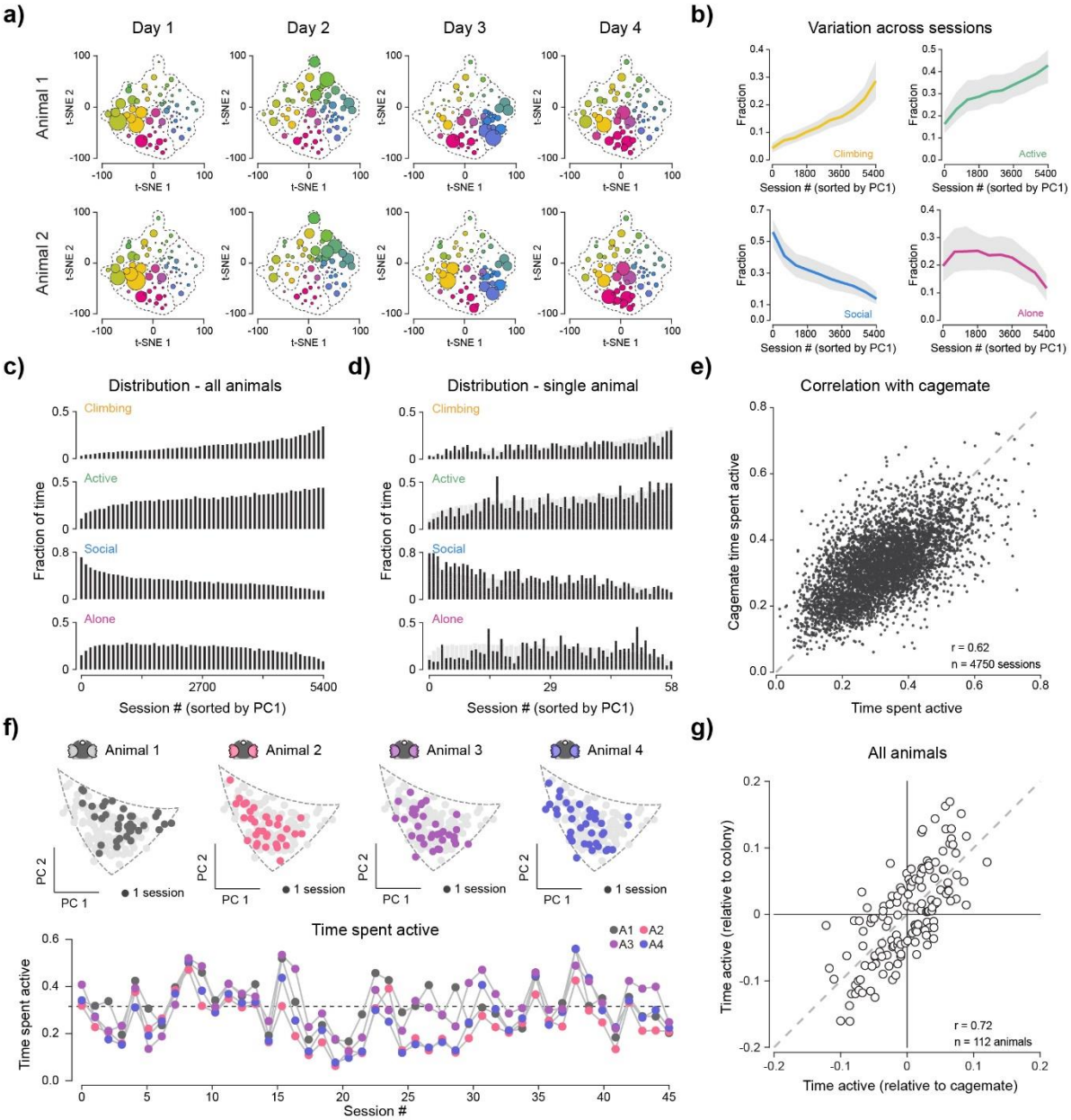
a) Schematic of network structure. Latent state at each time point used for downstream analysis. **b)** Example ethogram across 60 seconds with colors indicating behavioral state as defined by the cluster ID. **c)** Latent space clustering using t-SNE with decreasing values of α and perplexity to produce a hierarchical clustering (with colors representing different clusters). **d)** Hierarchical organization of behavioral states across levels of analysis. with the average usage for each state (% of total time) shown as a bar graph. **e)** Hierarchical organization of behavioral states, with colors representing cluster ID. **f)** Average state duration at each level of analysis, with each point representing the average state duration from one video session. Bars indicate quartiles (25%, 50%, 75%) and open circles indicate the median across sessions.

Figure 3: Striatal neurons reflect behavioral states defined by deep learning



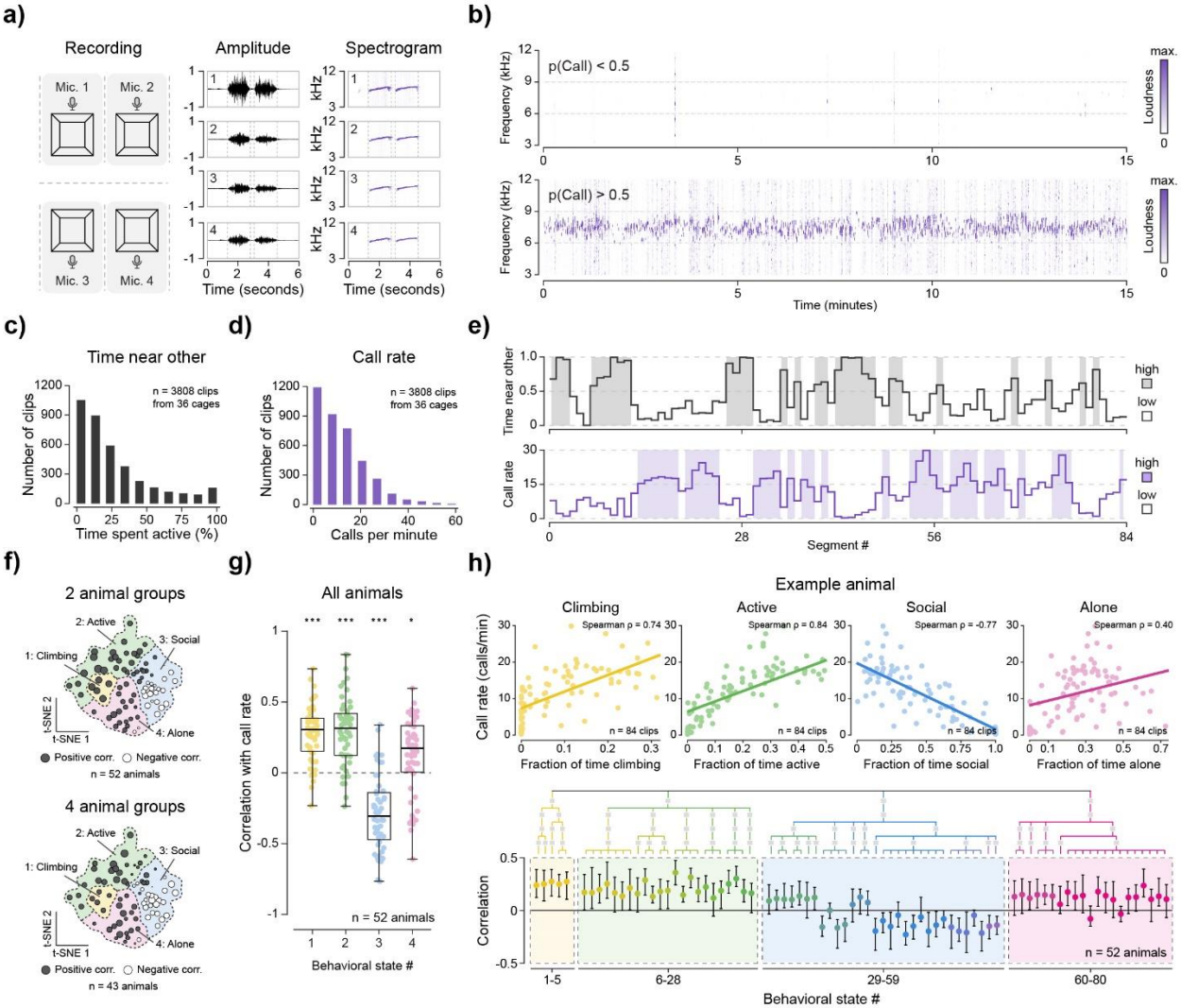
a) Schematic of silicon probe implant design and on-head wireless recording. **b)** Example image from video of two animals living in the same cage, with one of those animals wearing a wireless logger. **c)** Example of spike data from 1 minute of recording from 128 channels (spikes sorted into 25 units). **d)** All recorded neurons clustered by similarity, with colors indicating the highest-level clusters of neurons. **e)** For each of the 4 high level clusters: the regions in the behavioral latent space with enriched activity (left), the normalized change in firing rate in each of the highest-level behavioral clusters (mid-left), the transitions with enriched activity (mid-right), and summary of the enriched states and transitions (right). *** indicates $p < 10^{-20}$ using a one-sample t-test to determine if neurons had a higher firing rate during each behavioral state. **f)** Data from four example neurons showing their firing rate enrichment (left) and the normalized change in firing rate based on the position of the other animal (right). Position is displayed as a distance relative to the head radius (h.r.) of the animal. One example neuron was selected from each of the 4 highest-level clusters.

Figure 4: Behavioral distribution varies across days and is highly correlated to cagemates



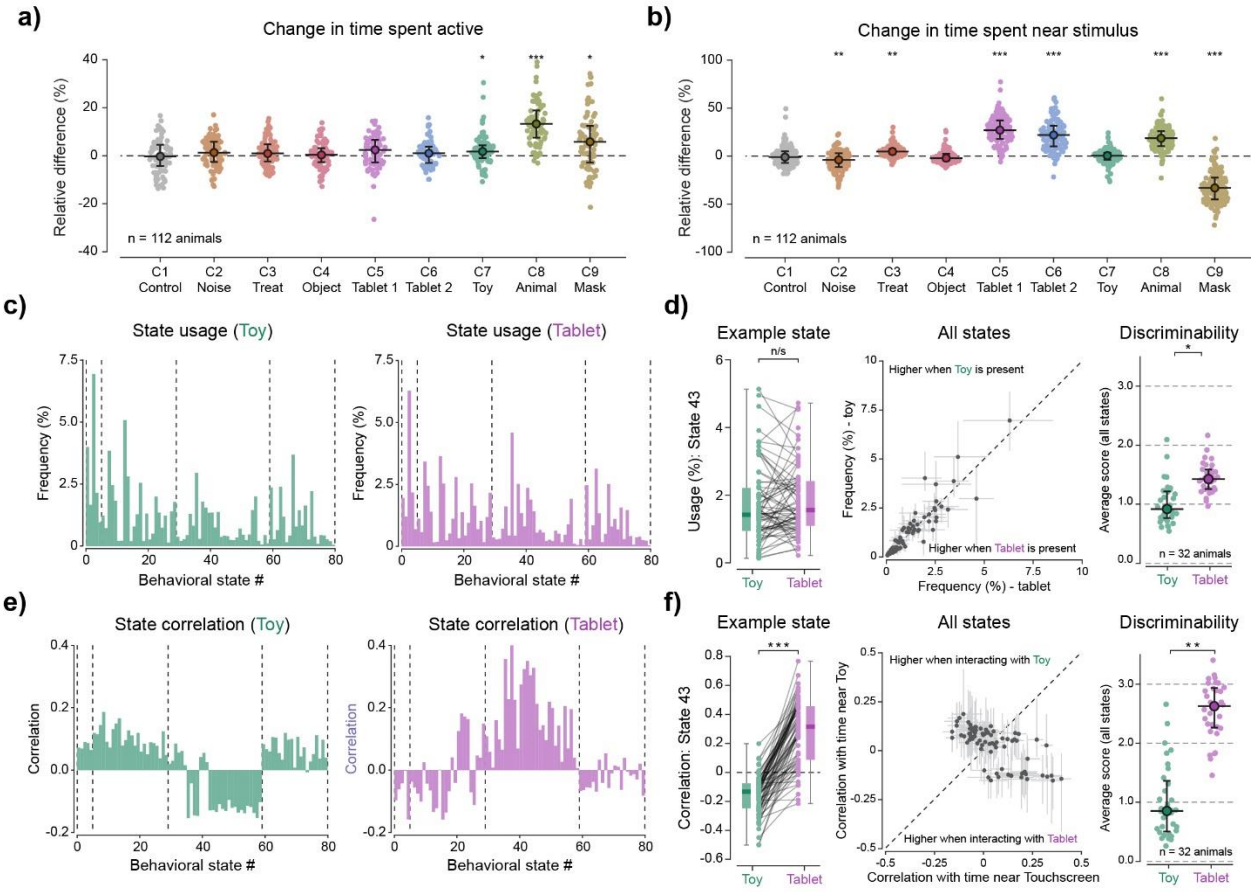
a) Example behavioral state usage from 4 different days sampled from two cagemates. Circle diameter indicates the frequency of each behavioral state. **b)** Distribution across all animals segmented into 10 quantiles with error bars representing the 25%, 50%, and 75% boundaries in the data. **c)** Distribution of behavioral states across 5400 sessions sampled from 112 animals, sorted based on PC1 from principal component analysis across sessions. **d)** Example distribution of behavioral states across 58 sessions sampled from a single animal, also sorted based on PC1. **e)** Correlation of time spent active across all animals (each dot represents one paired observation of an animal and one cagemate. **f)** Top: distribution of PC1 and PC2 across 45 sessions from an example family of 4 marmosets. Bottom: total time spent in active behavioral states across the 45 sessions, with colors representing animal ID. **g)** Correlation between measurements of activity normalized by cagemate activity (X-axis) or normalized by average colony activity (Y-axis). Each dot represents single animal average.

Figure 5: Room call rate varies across days and is highly correlated to behavior



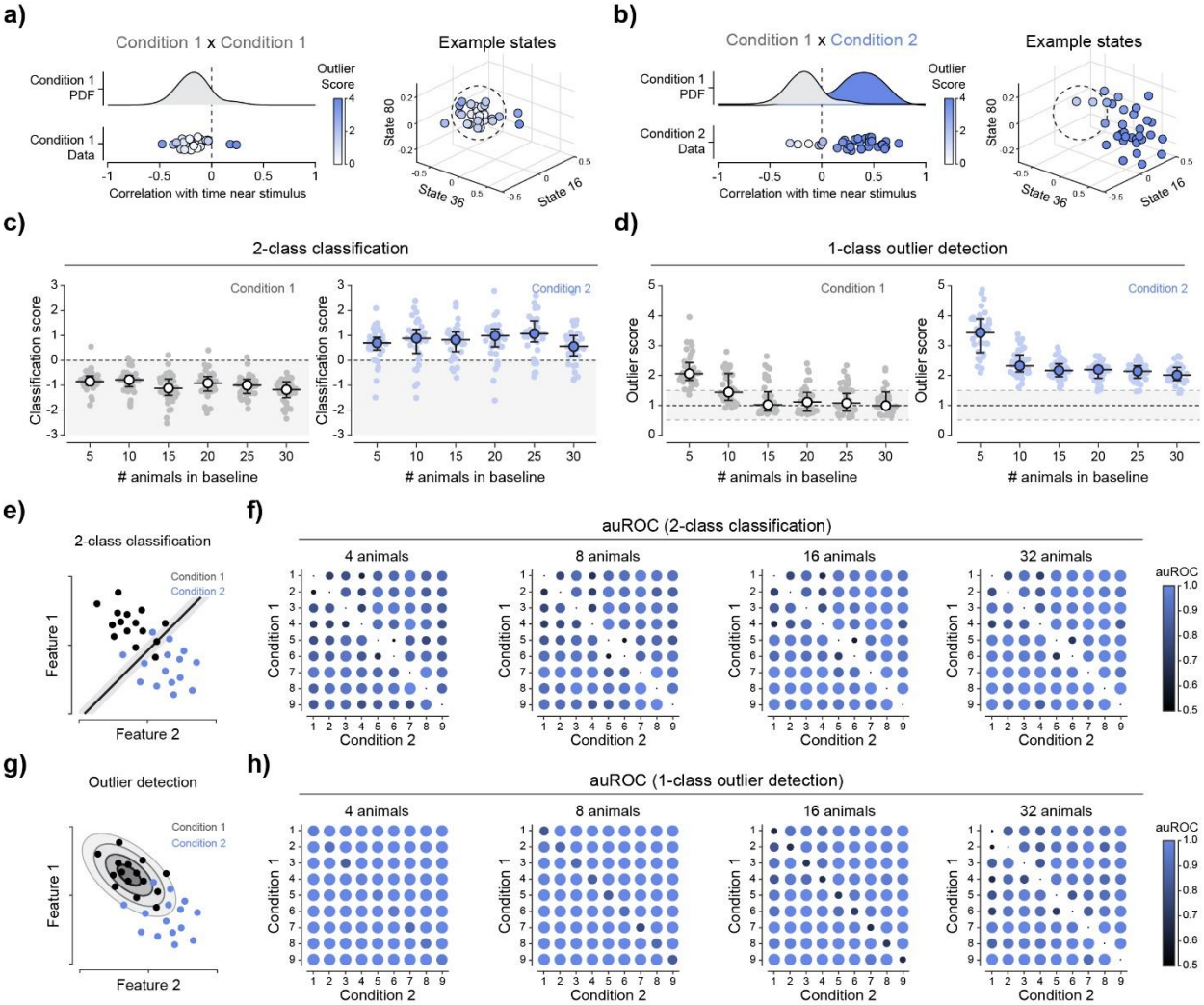
a) Schematic of microphone distribution and example of raw amplitude data converted to spectrogram data. **b)** Example audio data classified by a supervised neural network as time containing no calls (15 minutes total of concatenated spectrograms, top), and containing calls of any type (15 minutes total of concatenated spectrograms, bottom). **c)** Distribution of time spent near other animals from 5 minute clips (n = 3808 clips) from 36 cages. **d)** Distribution of average call rates across all 5 minute clips (n = 3808 clips) from 36 cages. **e)** Example series of non-continuous 5 minute clips comparing the time spent near other animals (top) and total call rate in the room (bottom). **f)** Comparison of cages with 2 animals (top) or 4 animals (bottom). **g)** Correlation between behavioral state usage and call rate across n=52 pair housed animals, with * indicating $p<10^{-2}$ and *** indicating $p<10^{-10}$ using a one-sample t-test to determine if each behavioral state was positively/negatively correlated with call rate across animals. **h)** Top: Correlations between behavioral states and call rate from one example animal, colored based on behavioral state. Each point represents one 5 minute clip. Bottom: Hierarchical organization of the correlation between behavioral state usage and room call rate. Bars indicate the quartiles (25%, 50%, and 75%) and circles indicate the median across animals.

Figure 6: Comparing across behavioral conditions induced by stimuli



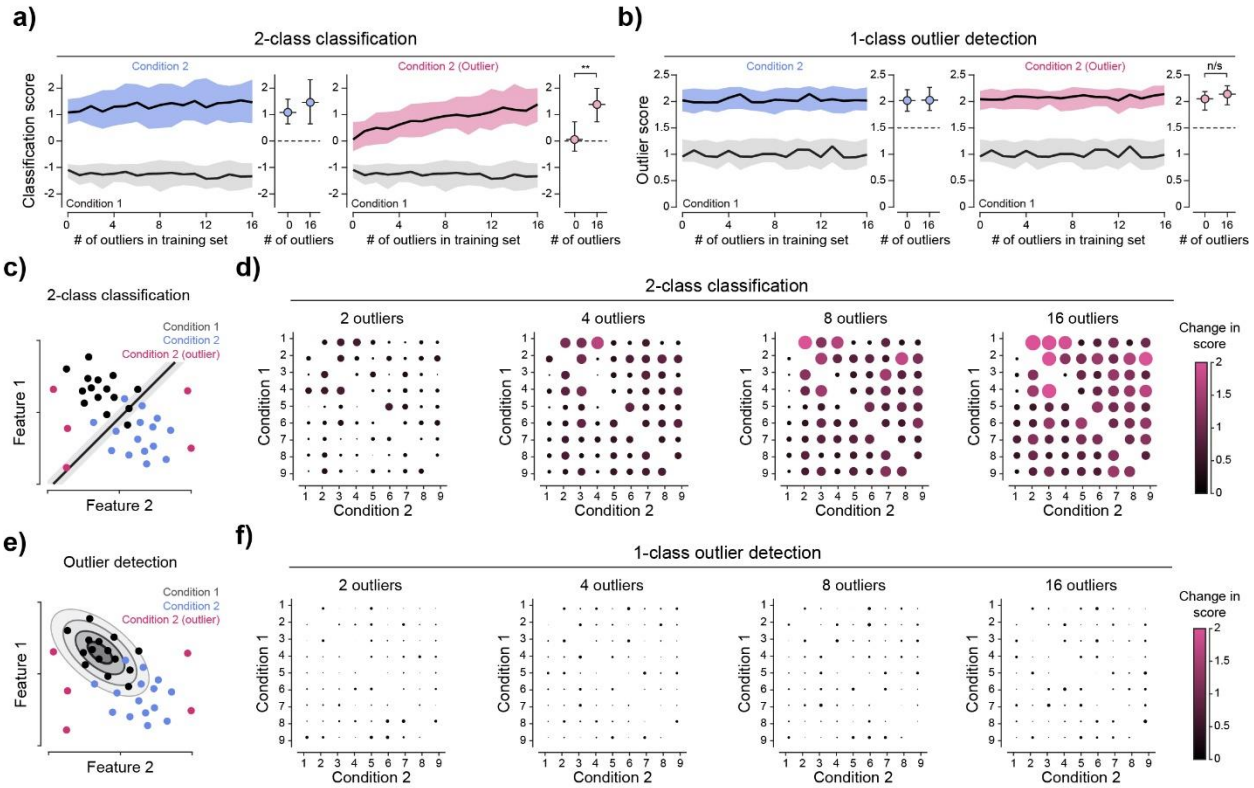
a) Change in total time spent active during the presentation of each stimulus. * indicates $p<10^{-3}$, ** indicates $p<10^{-6}$, and *** indicates $p<10^{-20}$ using a one-sided t-test to determine whether each stimulus increased or decreased the time spent active. **b)** Change in time spent in the region of the cage where the stimulus is placed (for each stimulus). * indicates $p<10^{-3}$, ** indicates $p<10^{-6}$, and *** indicates $p<10^{-20}$ using a one-sided t-test to determine whether each stimulus increased or decreased the time spent near the stimulus. **c)** Left: average usage of behavioral states when Toy or Tablet are presented. **d)** Difference in usage between conditions and discriminability between conditions based on usage. * indicates $p<10^{-3}$, ** indicates $p<10^{-6}$, and *** indicates $p<10^{-20}$ using a paired t-test to determine whether there was a difference in behavioral state usage between conditions. Paired t-tests compare data from the same animal in different conditions. **e)** Left: average correlation between behavioral state usage and time spent interacting with stimuli in each condition. **f)** Difference in correlations between conditions and discriminability between conditions based on these correlations. * indicates $p<10^{-3}$, ** indicates $p<10^{-6}$, and *** indicates $p<10^{-20}$ using a paired t-test to determine whether there was a difference in behavioral state correlations between conditions. Paired t-tests compare data from the same animal in different conditions.

Figure 7: Discriminating between conditions using outlier detection



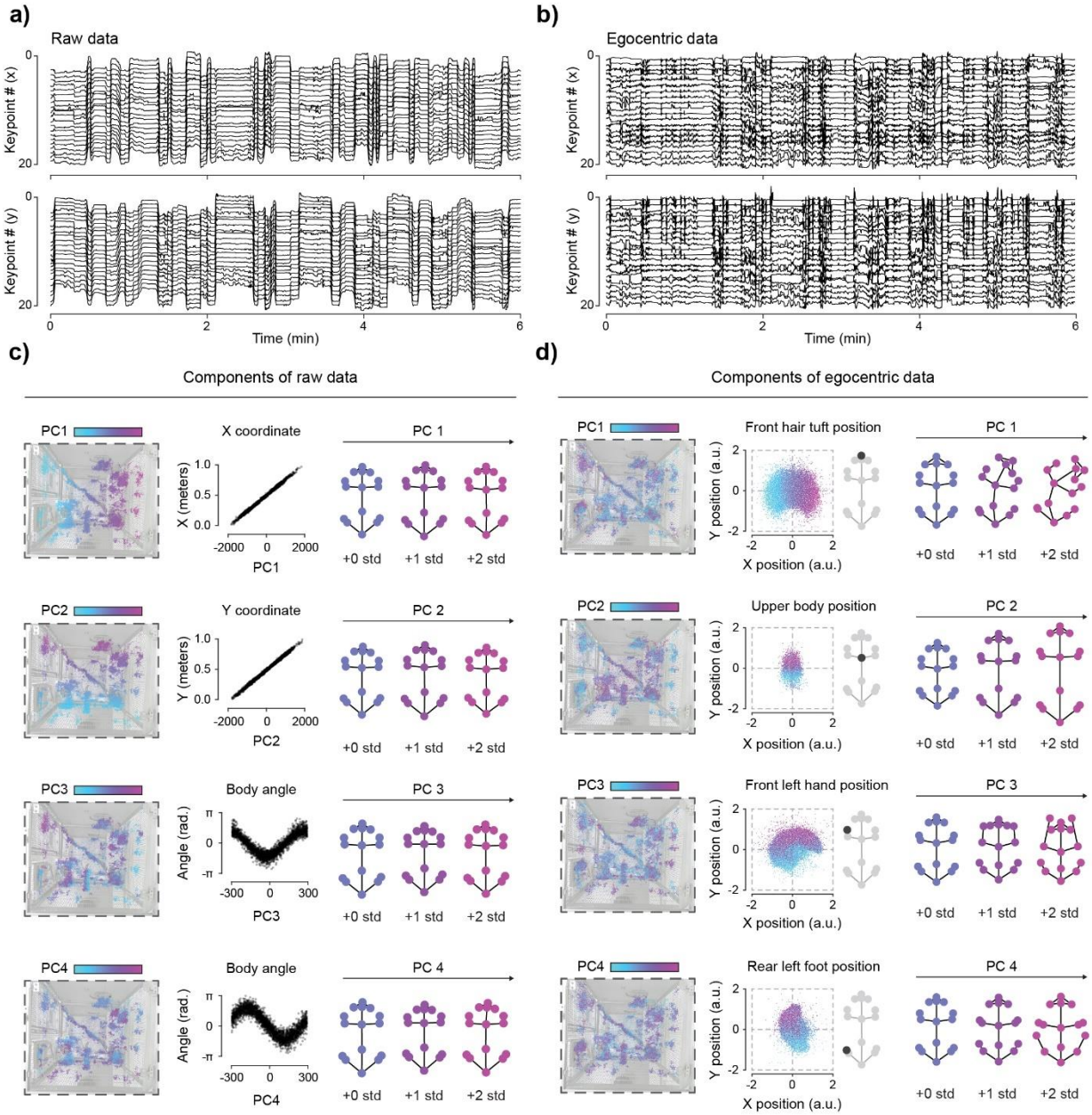
a) Strategy for assigning outlier score based on a probability distribution defined by "Condition 1" and then applied to either held-out data from "Condition 1" or to held-out data from "Condition 2". **b)** Outlier scores for each behavioral state across 32 age-matched animals in Condition 2. **c)** Discriminability between example conditions (Toy and Touchscreen) based on 2-class classification. **d)** Discriminability between example conditions (Toy and Touchscreen) based on 1-class outlier detection. Bars indicate quartiles (25%, 50%, 75%) and open circles indicate the median score for each group. **e)** Schematic of 2-class classification. See methods for full detail. **f)** Comparison of auROC scores (from held-out data) across models trained using a different number of animals (n=4 to n=32). **g)** Schematic of outlier detection. See methods for full detail. **h)** Comparison of auROC scores (from held-out data) across models trained using a different number of animals (n=4 to n=32).

Figure 8: Outlier detection has stable performance regardless of heterogeneity in the dataset



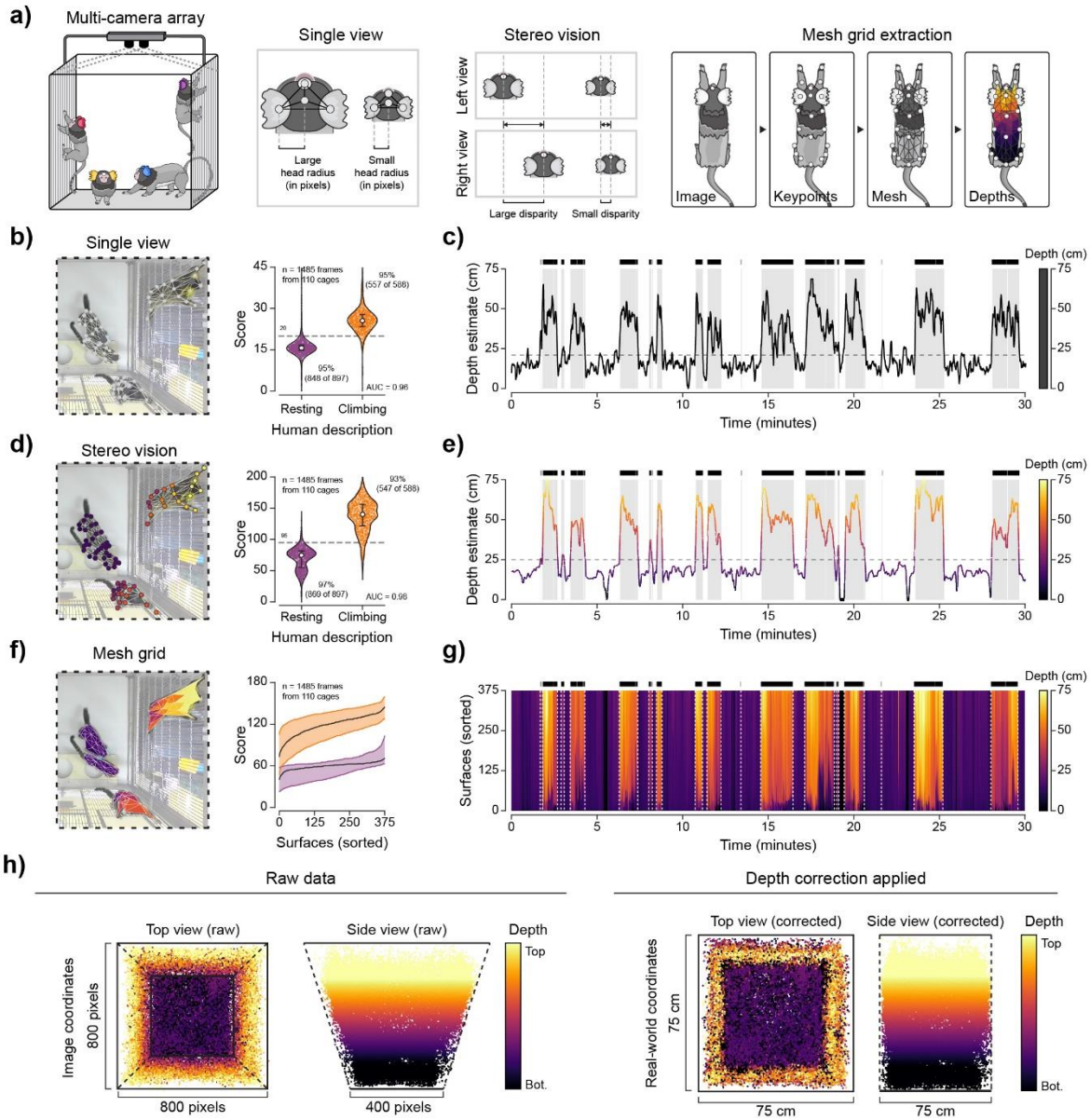
a) Scores assigned to inliers (blue) or outliers added to Condition 2 (red) using 2-class classification, with an increasing number of outliers artificially added to the training set. ** indicates $p<10^{-12}$ using a t-test to determine whether the scores differed when the number of outliers added to the training set was 0 or 16. **b)** Scores assigned to inliers (blue) or outliers added to Condition 2 (red) using outlier detection, with an increasing number of outliers artificially added to the training set. ** indicates $p<10^{-12}$ using a t-test to determine whether the scores differed when the number of outliers added to the training set was 0 or 16. **c)** Schematic of 2-class classification with heterogeneity artificially added to Condition 2. See methods for full detail. **d)** Change in score assigned to outliers as a function of the number of outliers artificially added to the training data. **e)** Schematic of outlier detection with heterogeneity artificially added to Condition 2. See methods for full detail. **f)** Change in score assigned to outliers as a function of the number of outliers artificially added to the training data.

Figure S1: PCA in raw data compared to egocentric data



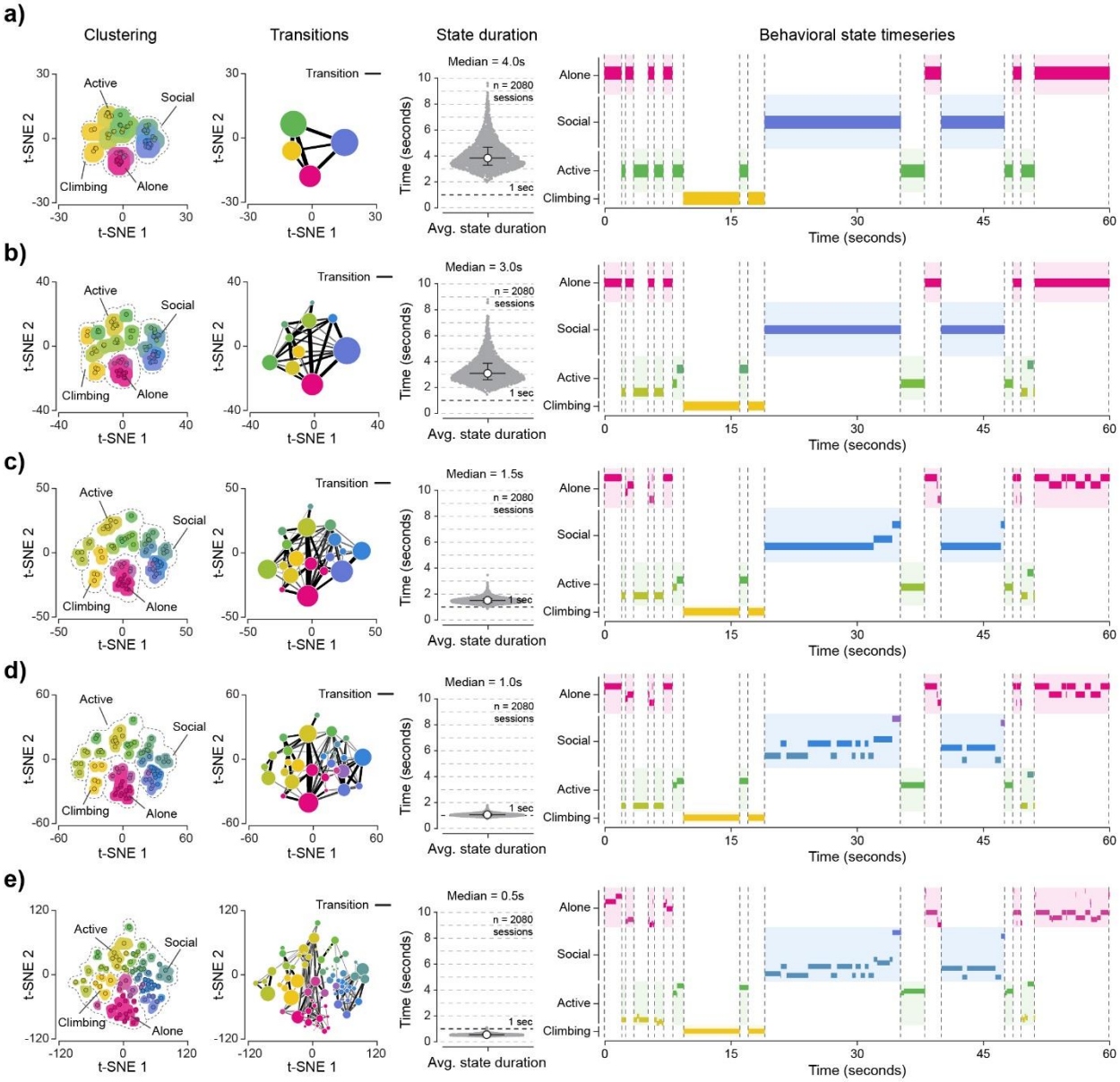
a) Example timeseries of raw postural time series data (21 keypoints) plotted across 6 minutes. **b)** Example time series of egocentric data (normalized by position and body angle) plotted across 6 minutes. **c)** For the first four raw data principal components (PCs), the distribution within the image (left), example regression with features (middle), and average posture at different scores (right). **d)** For the first four egocentric data principal components (PCs), the distribution within the image (left), example egocentric body part coordinate colored based on score (middle), and average posture at different scores (right).

Figure S2: Depth estimates for top-down home-cage video



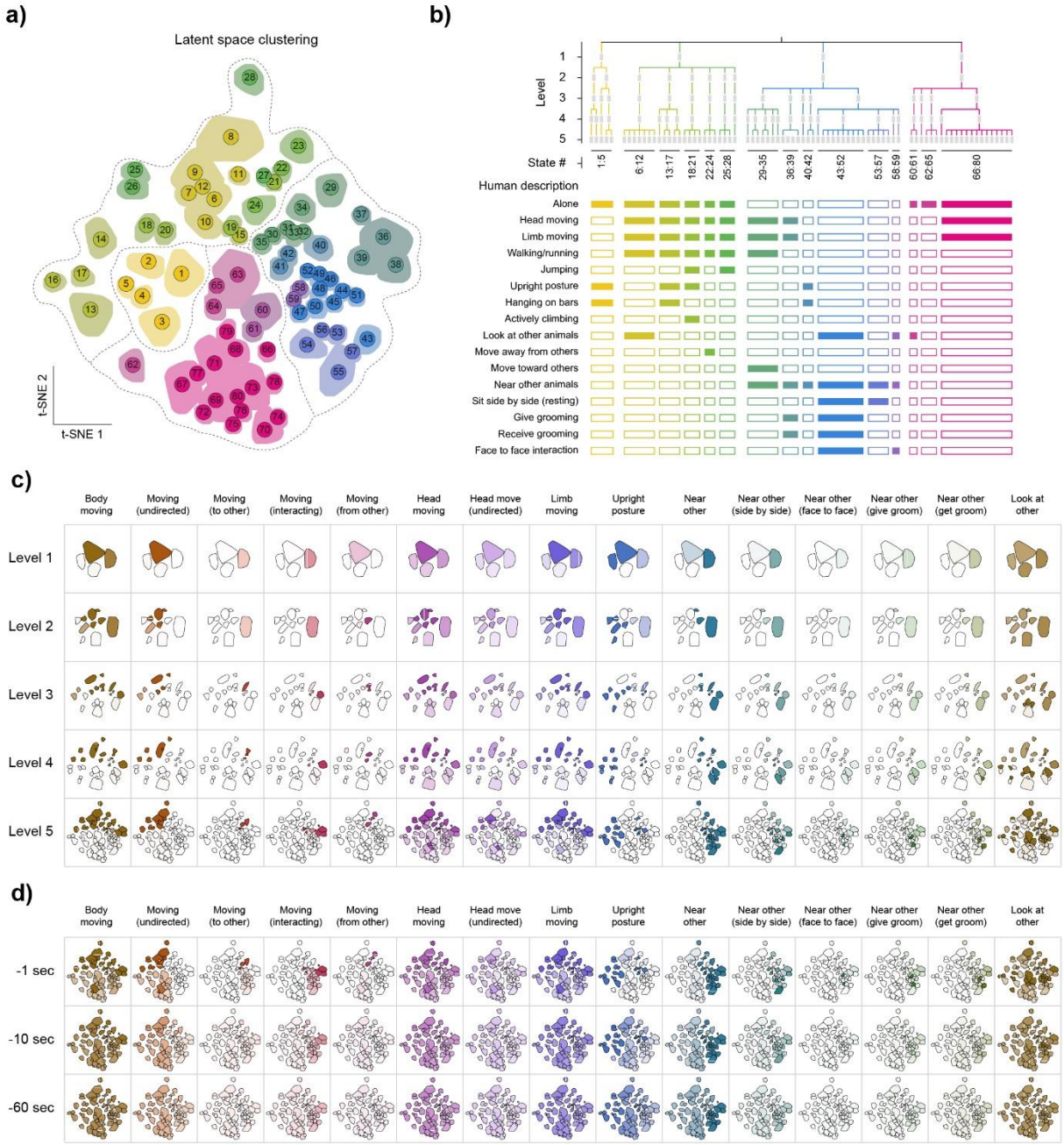
a) Comparison of methods for depth estimation using single view, stereo vision, or stereo vision with mesh grid extraction. **b)** Example single-view image, and comparison of score (pixel-based) between human-annotated frames labeled with binary 0 (not climbing) or 1 (climbing) labels. **c)** Example depth time series based on single-view (30 minute trace). **d)** Example stereo vision image, and comparison of score (depth-based) between human-annotated frames labeled with binary 0 (not climbing) or 1 (climbing) labels. **e)** Example time series (depth estimate) based on stereo vision. **f)** Example mesh grid based on stereo depth estimates, and comparison of score (depth-based) between human-annotated frames labeled with binary 0 (not climbing) or 1 (climbing) labels. **g)** Example mesh grid time series, with depths sorted at each time point. **h)** Example plots of uncorrected position (left) and depth-corrected position (right), with color corresponding to distance from the camera.

Figure S3: Hierarchical behavioral state representation



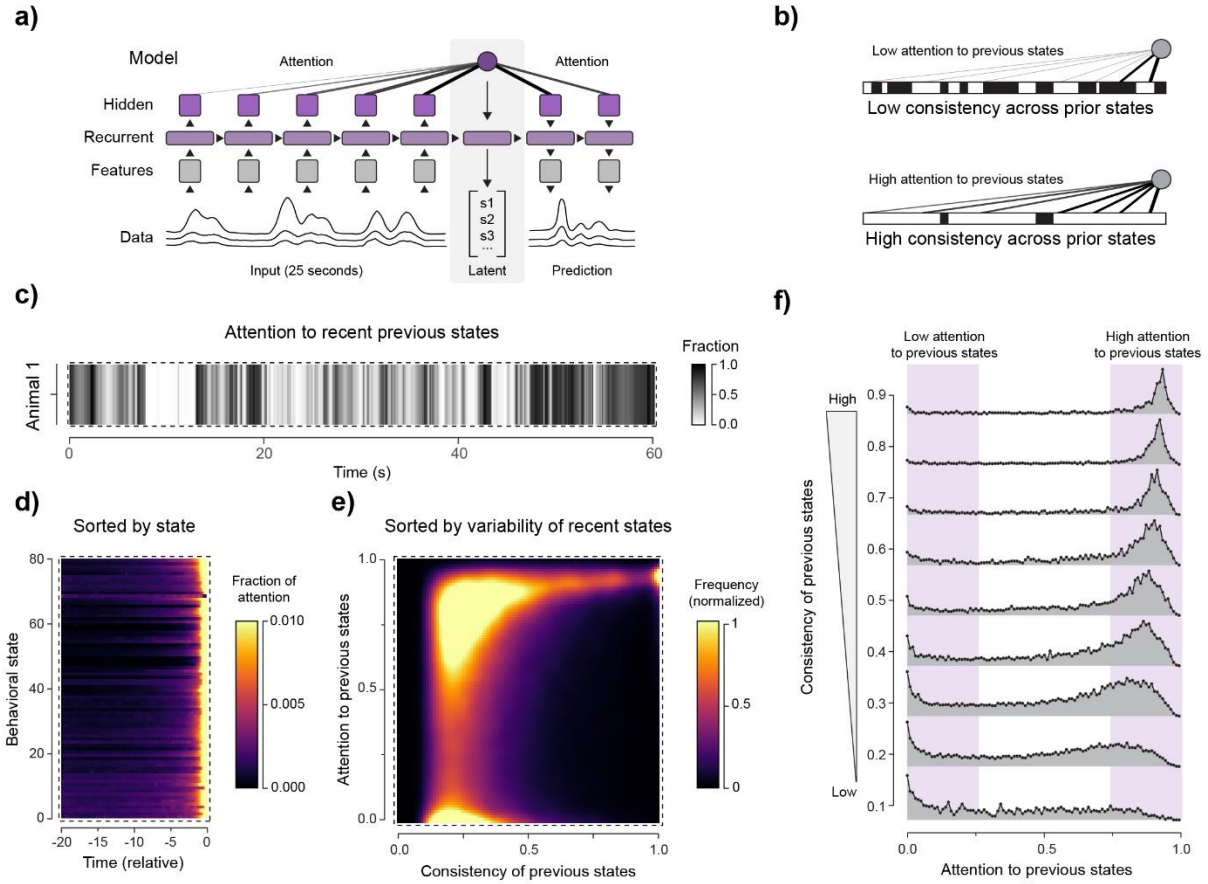
a) Latent space clustered into 4 components with a median duration of 4.0 seconds. **b)** Latent space clustered into 10 components with a median duration of 3.0 seconds. **c)** Latent space clustered into 18 components with a median duration of 1.5 seconds. **d)** Latent space clustered into 30 components with a median duration of 1.0 seconds. **e)** Latent space clustered into 80 components with a median duration of 0.5 seconds.

Figure S4: Mapping human description and features to clusters in latent space



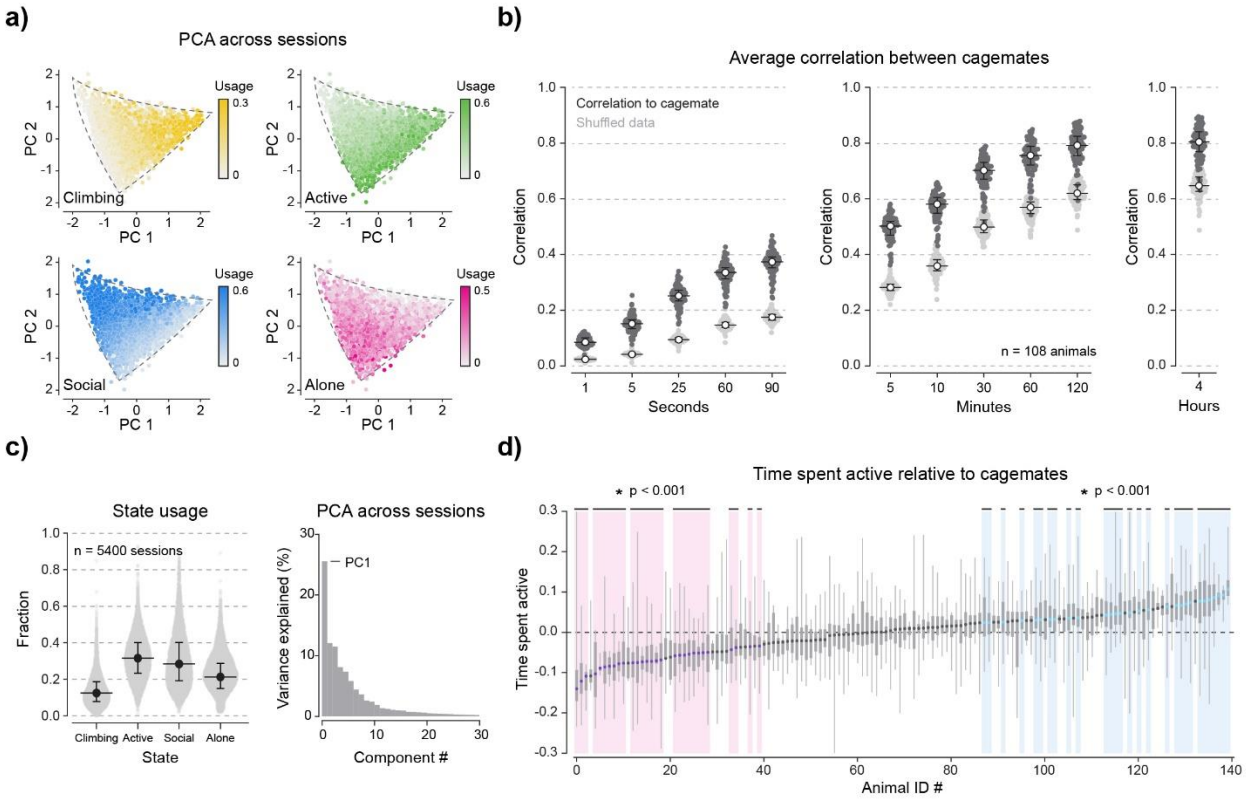
a) Latent space clustering, with each cluster labeled by number. **b)** Post-hoc human description of states. These descriptions were not used in network training. **c)** Distribution of selected example behavioral variables across latent space at each hierarchical level. **d)** Distribution of behavioral variables at the lowest hierarchical level across time. Top row: 1 second prior to the labeled frame. Middle row: 10 seconds prior to the labeled frame. Bottom row: 60 seconds prior to the labeled frame.

Figure S5: Attention distribution varies depending on previous states



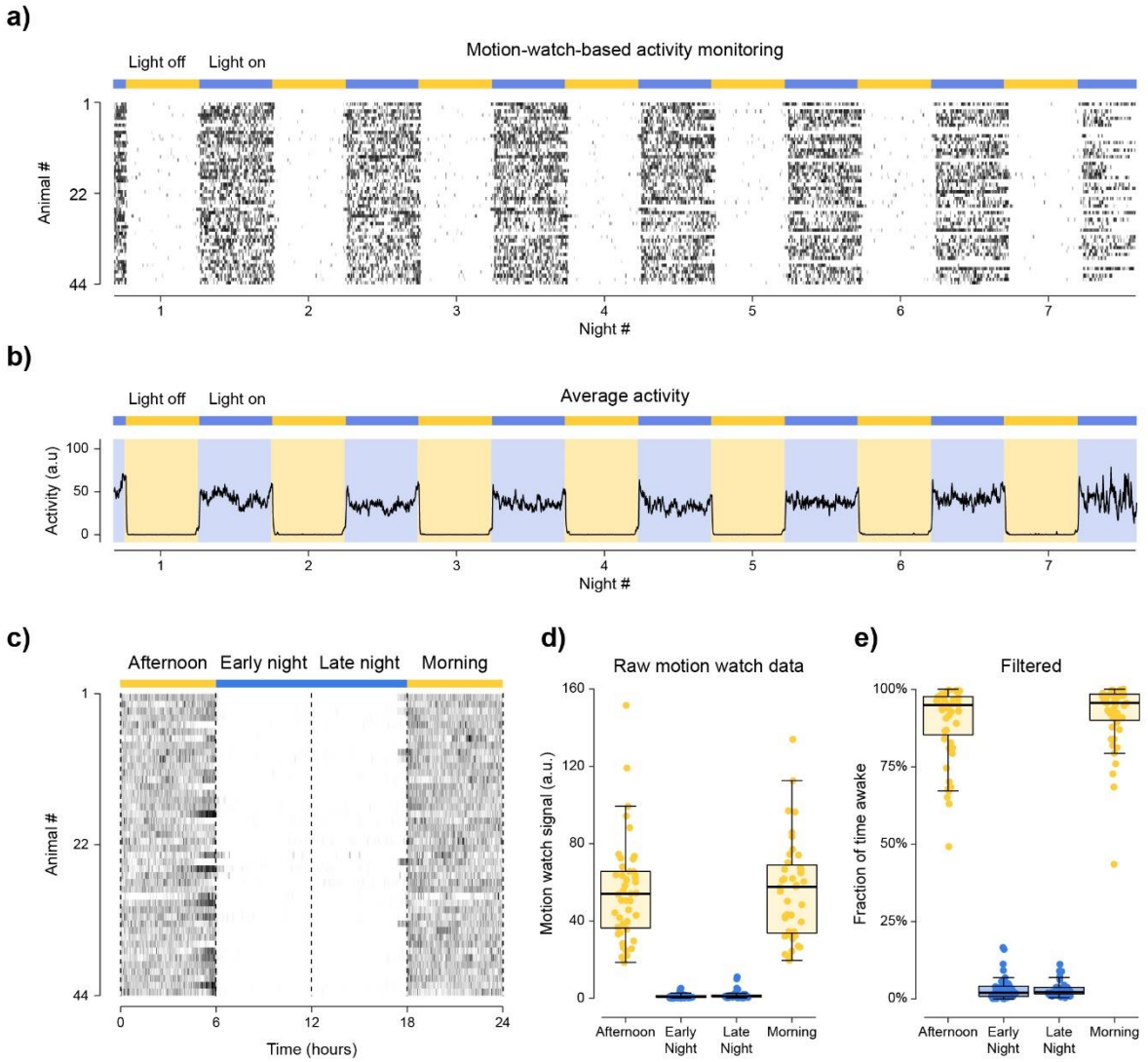
a) Schematic of recurrent neural network model. **b)** Attention distribution concentrated on recent time points when previous states have high variability (top) or attention distribution distributed widely when previous states have low variability (bottom). **c)** Top: Example time series of the fraction of attention allocated to previous states (>1 second prior) over the course of one minute of video. **d)** Average attention allocation across behavioral states where each row represents one behavioral state and color corresponds to average attention. **e)** Attention to previous states plotted as a function of the consistency of previous states. Consistency is defined where 1.0 is the case where all states in the last 20 seconds are identical. **f)** Attention to previous states plotted for discrete ranges of past state consistency. Data is divided into ten groups, starting with the range (0 to 0.1) and ending with the range (0.9 to 1.0).

Figure S6: Measuring correlation between cagemates across timescales



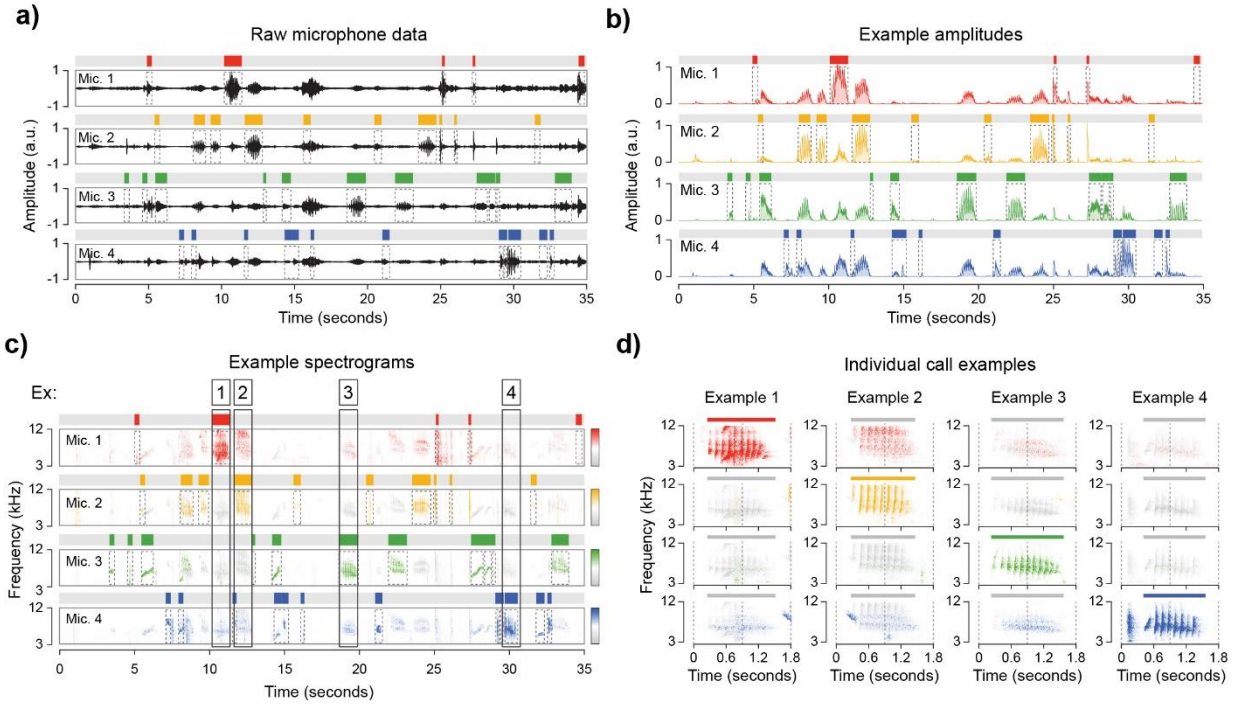
a) All sessions plotted by PC1 and PC2. Colors correspond to the frequency of “Climbing”, “Active”, “Social”, and “Alone” behavioral states. **b)** Correlation measurements with different bin sizes ranging from 1 second to four hours. Gray circles represent correlations obtained using shuffled data (with the same time bin sizes). **c)** Distribution of behavioral state usage, and PCA across sessions. **d)** Time spent active relative to cagemates, with animals significantly less active than cagemates ($p<0.001$) shown in red and animals significantly more active than cagemates ($p<0.001$) shown in blue. Paired t-tests were used to compare simultaneously collected activity between animals’ and their cagemates. Cagemate data was averaged to generate a single measurement for each session.

Figure S7: Sleep cycle monitoring



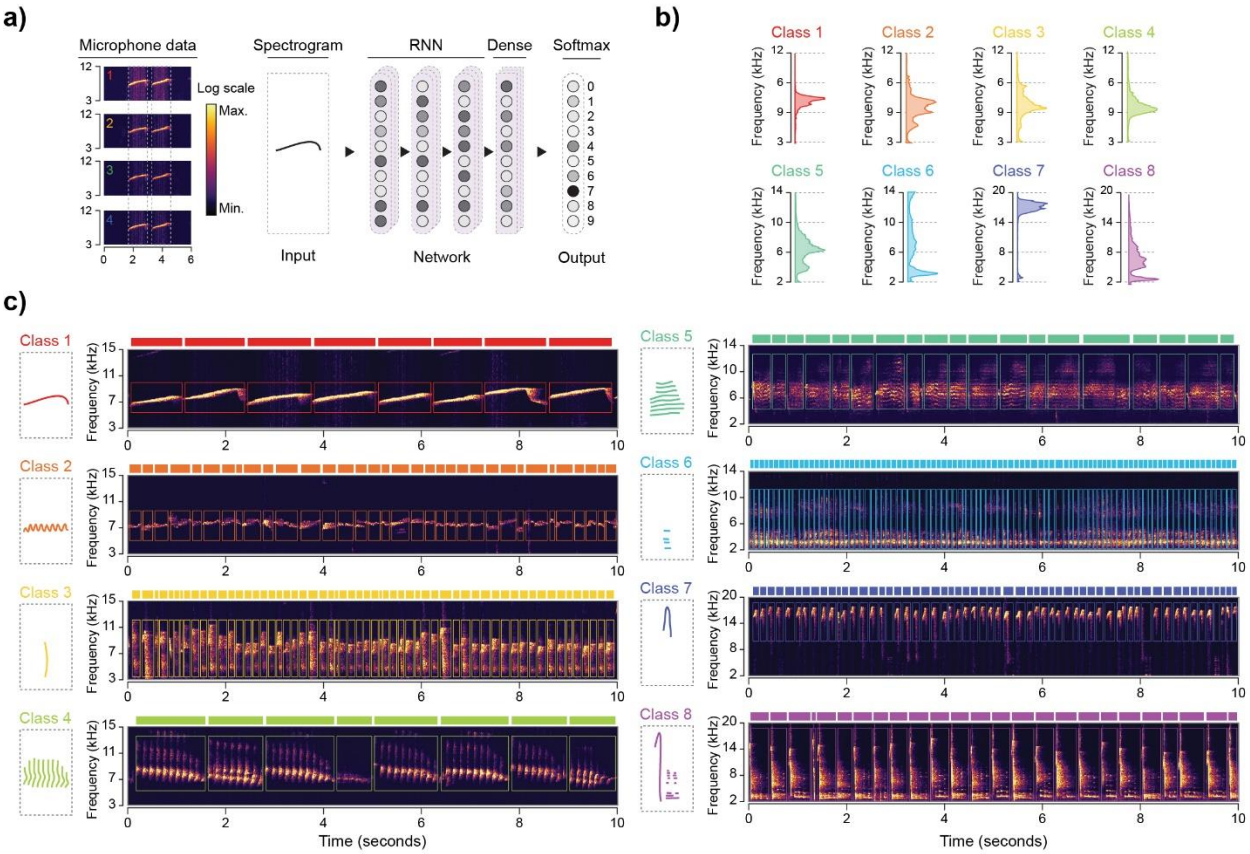
a) Activity monitoring using a motion watch. Continuous data from one week of monitoring is plotted for each animal with n = 44 animals. Blue bars indicate nighttime (lights off) and yellow bars indicate daytime (lights on). Darker areas in the plot indicate higher activity. In cases where animals removed or disabled motion watches prior to the end of the experiment, remaining values were replaced by NaN. **b)** Average activity trace across 44 animals. **c)** Average 24 hour cycle (based on data from 1 week) plotted for each animal. **d)** Raw motion watch data (a.u. = arbitrary units) separated into 4 time windows of 6 hours. **e)** Filtered motion watch data representing an estimate of time spent at rest (%) separated into 4 time windows of 6 hours.

Figure S8: Call localization using array of microphones



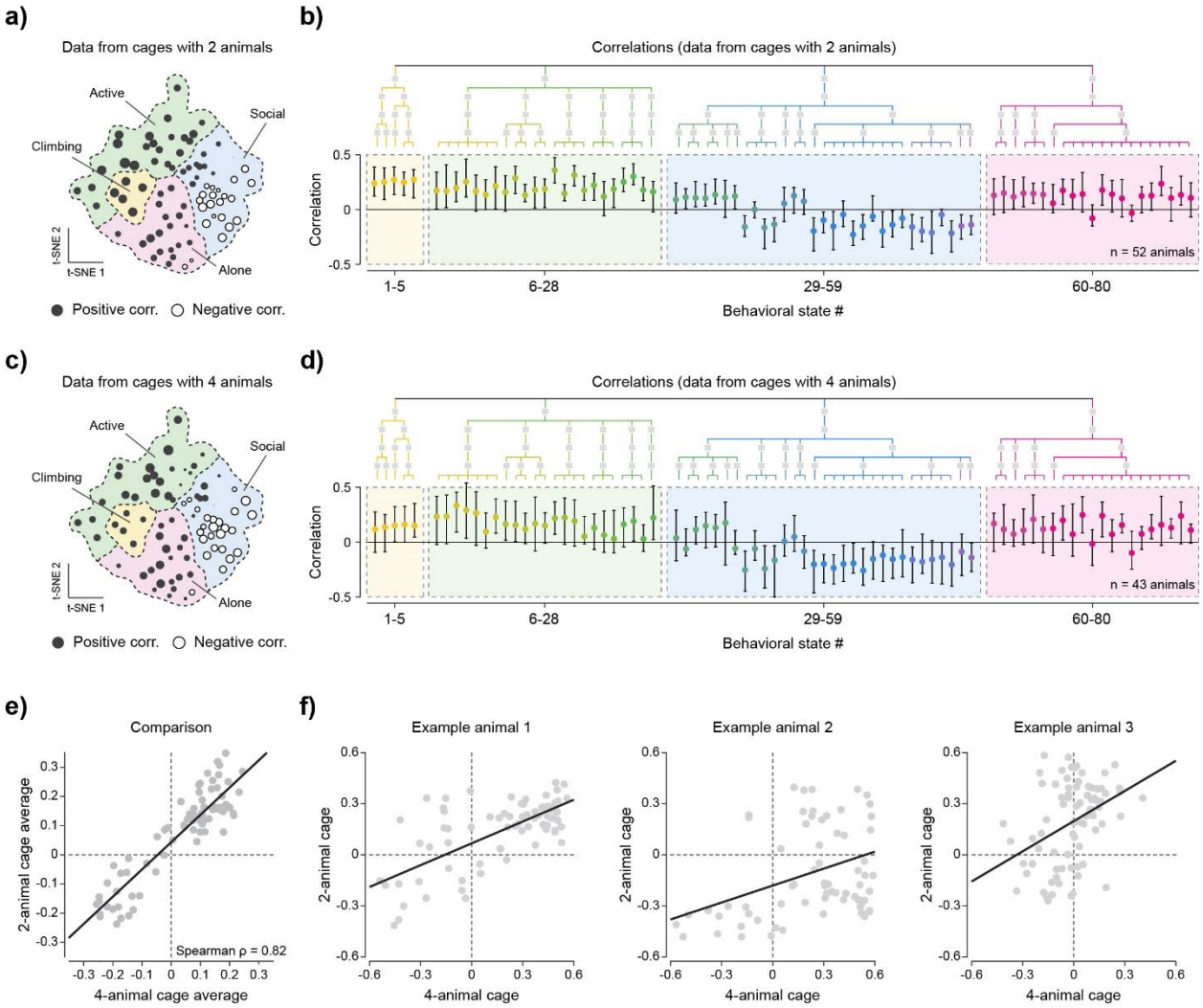
a) Example time series (35 seconds) of raw data from 4 microphones placed near different cages. Dotted lines indicate calls labeled by neural network. **b)** Amplitudes (running average of the absolute value of the spectrogram) used to assign each call to the microphone with the largest amplitude recording. **c)** Spectrograms depicting the distribution of signal across frequencies. **d)** Four individual call spectrograms, where the brightness indicates the amplitude measurement of the call in each microphone, relative to the amplitude measurement from the other microphones.

Figure S9: Supervised categorical audio classification



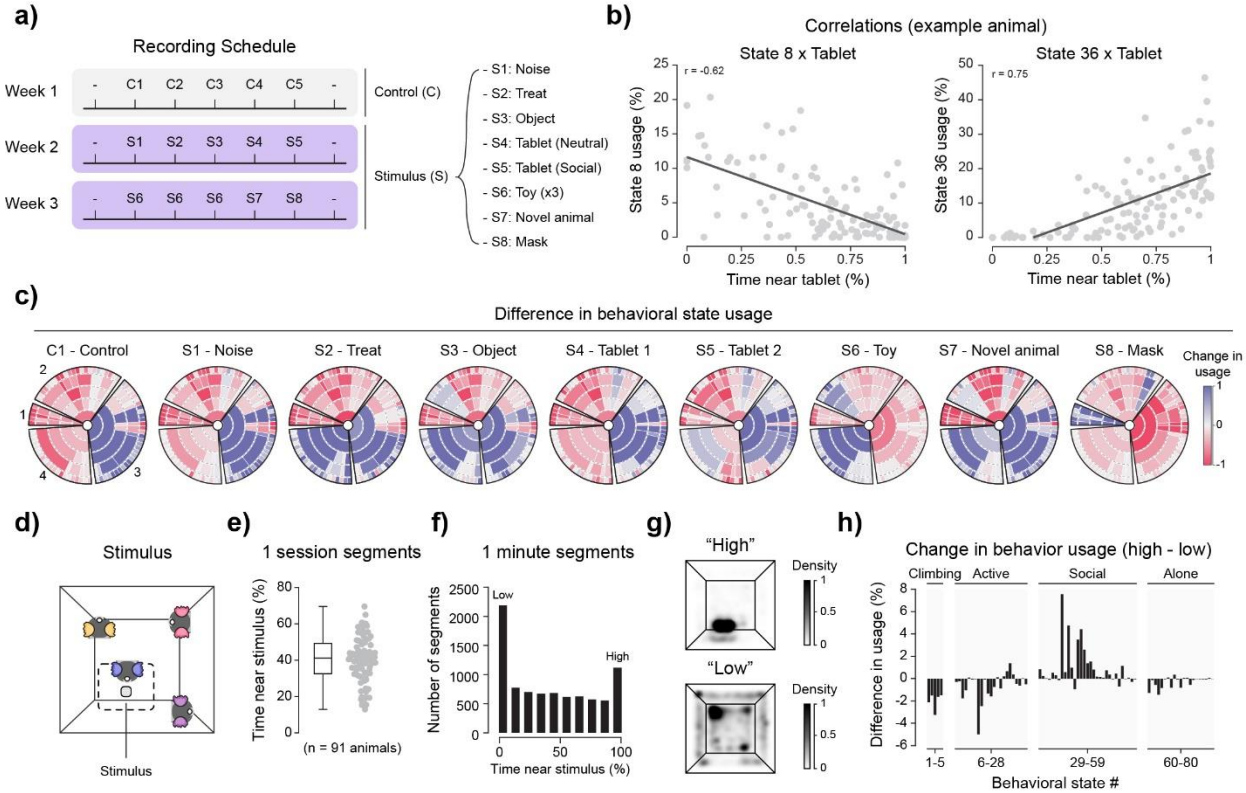
a) Strategy for classifying calls based on call type using a supervised neural network. **b)** Histogram across frequencies for each call type, colored by call type. **c)** Left: A human-drawn representation of calls in each call type. Right: Example calls (10 seconds of concatenated calls) from each category chosen randomly from all calls in the category. Number of calls displayed varies by category due to the difference in call duration across types.

Figure S10: Comparison of 2-animal cage and 4-animal cage correlation patterns



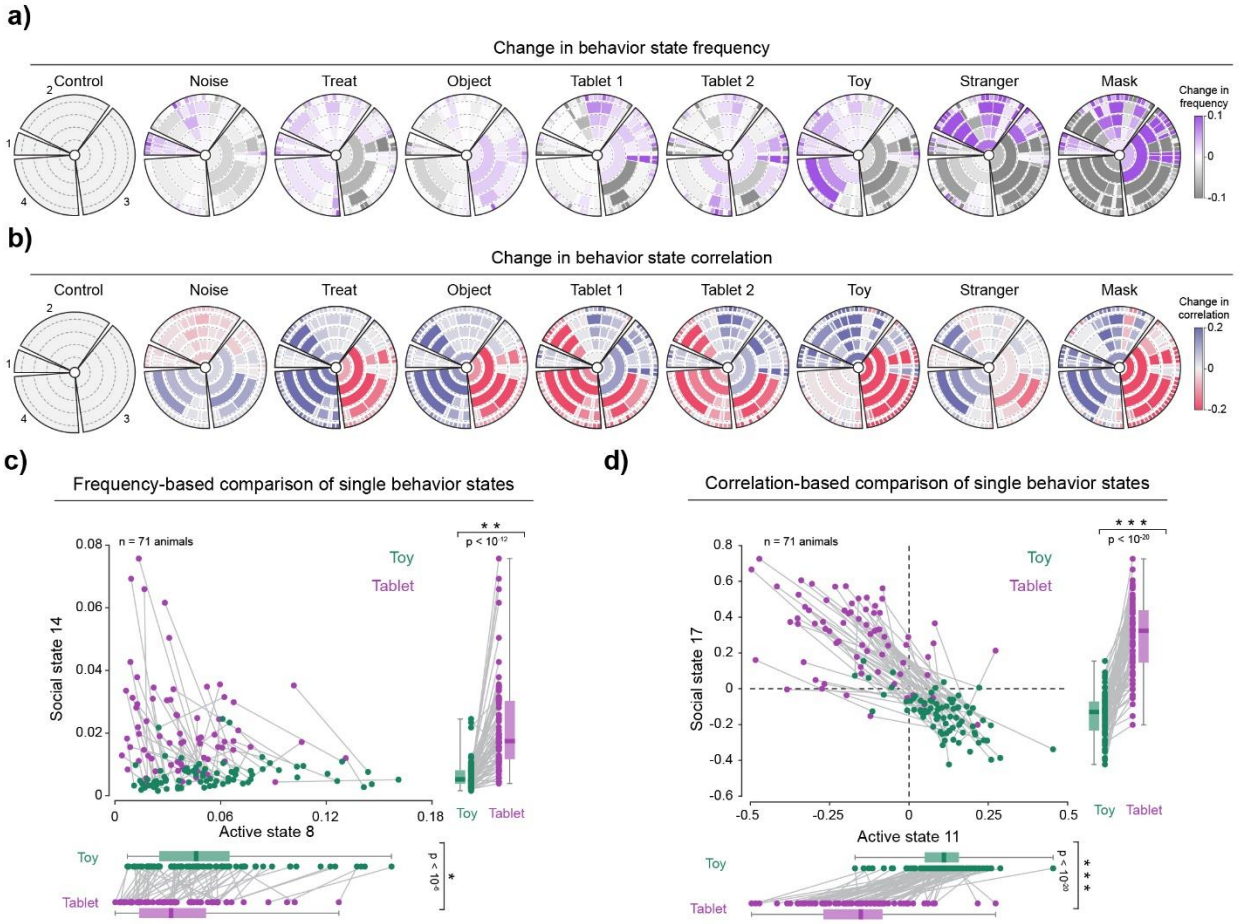
a) Correlation between behavioral state usage and call rate from cages with 2 adult animals (filled circles indicate positive correlation, open circles indicate negative correlation). **b)** Data from n = 52 animals, with bars indicating quartile boundaries (25%, 50%, and 75%) and circles indicating median across animals. **c)** Correlation between behavioral state usage and call rate from cages with 4 animals (filled circles indicate positive correlation, open circles indicate negative correlation). **d)** Data from n = 43 animals, with bars indicating quartile boundaries (25%, 50%, and 75%) and circles indicating median across animals. **e)** Correlation between data from cages with 2 animals and data from cages with 4 animals. **f)** Correlation from three example animals that were recorded in both conditions (both in the 2-animal condition and the 4-animal condition).

Figure S11: Behavioral responses to a panel of stimuli



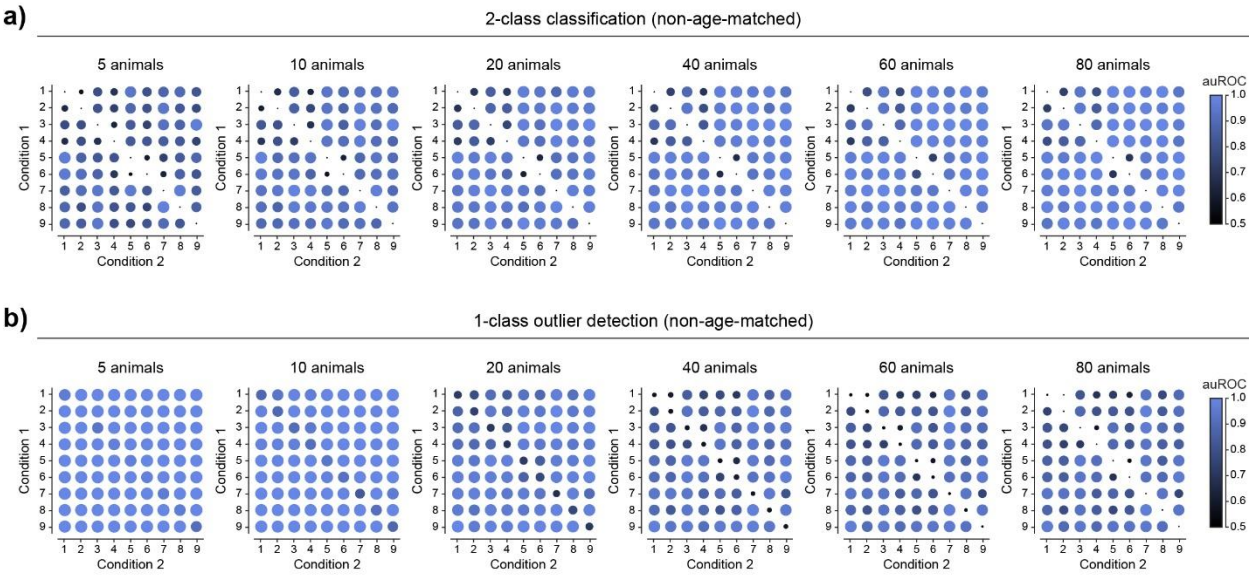
a) Recording timeline, which was repeated every 3 months. **b)** Example correlations between behavioral state usage and time spent near tablet (using 1-minute bins). **c)** Difference in behavioral state usage, comparing periods of "high" interaction (>90% time spent near stimulus per minute) and periods of "low" interaction (<10% time spent near stimulus per minute). **d)** Schematic of stimulus delivery. **e)** Distribution of time spent near tablet across 91 non-age-matched animals (quantified based on % time spent near stimulus per session). **f)** Distribution of time spent near tablet quantified based on % time spent near stimulus per minute. **g)** Comparison of cage location during times of "high" interaction with tablet (>90% per minute) and times of "low" interaction with tablet (<10% per minute). **h)** Difference in behavioral state usage between times of "high" interaction with tablet (>90% per minute) and "low" interaction with tablet (<10% per minute).

Figure S12: Comparing frequency-based metrics and correlation-based metrics



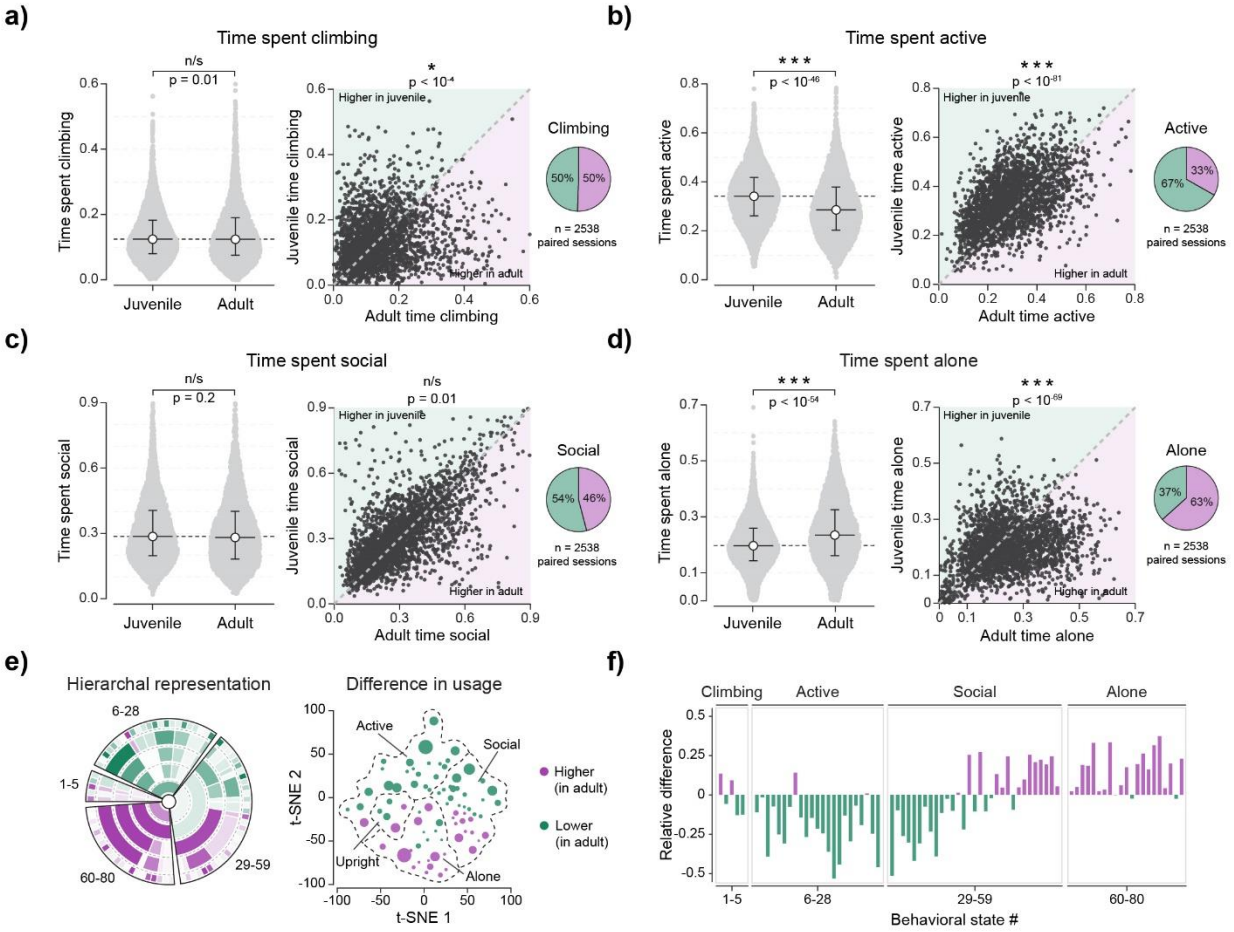
a) Change in behavioral state usage across hierarchical levels (purple indicates an increase in state usage and grey indicates a decrease in state usage). Comparisons are relative to control (bed-only condition with no other stimuli). **b)** Change in behavioral state correlations across hierarchical levels (blue indicates an increase in correlation and red indicates a decrease in correlation). Comparisons are relative to control (bed-only condition with no other stimuli). **c)** Example comparing the usage of two behavioral states when animals were presented with toy (green) or touchscreen (purple). * indicates $p < 10^{-6}$, ** indicates $p < 10^{-12}$, and *** indicates $p < 10^{-20}$ using paired t-tests to determine whether the same animals used each behavioral state more or less in each condition. **d)** Example comparing the correlation between two behavioral states with the toy (green) or touch-screen (purple) stimuli. * indicates $p < 10^{-6}$, ** indicates $p < 10^{-12}$, and *** indicates $p < 10^{-20}$ using paired t-tests to determine whether the same animals had higher or lower correlation measurements for each behavioral state in each condition.

Figure S13: Outlier detection using data from non-age-matched animals



a) auROC scores from 2-class classification across an increasing number of animals (non-age-matched) used in training data. Score indicates the discriminability between each pair of stimulus-response conditions. **b)** auROC scores from 1-class outlier detection across an increasing number of animals (non-age-matched) used in training data. Score indicates the discriminability between each pair of stimulus-response conditions.

Figure S14: Differences in behavioral distribution between juvenile and adult animals



a) Comparison of the fraction of time spent upright/climbing per session between juvenile and adult animals (left), and a paired comparison across cagemates of different ages (right). **b)** Comparison of the fraction of time spent active/running per session between juvenile and adult animals (left), and a paired comparison across cagemates of different ages (right). **c)** Comparison of the fraction of time spent social/near other per session between juvenile and adult animals (left), and a paired comparison across cagemates of different ages (right). **d)** Comparison of the fraction of time spent alone per session between juvenile and adult animals (left), and a paired comparison across cagemates of different ages (right). **e)** Normalized difference in usage of behavioral states at level 5. **f)** Example graphical representation. **g)** Normalized difference in usage of behavioral states across hierarchal levels. **h)** Example graphical representation. Statistical tests were unpaired (left) or paired using data simultaneously collected from cagemates (right).