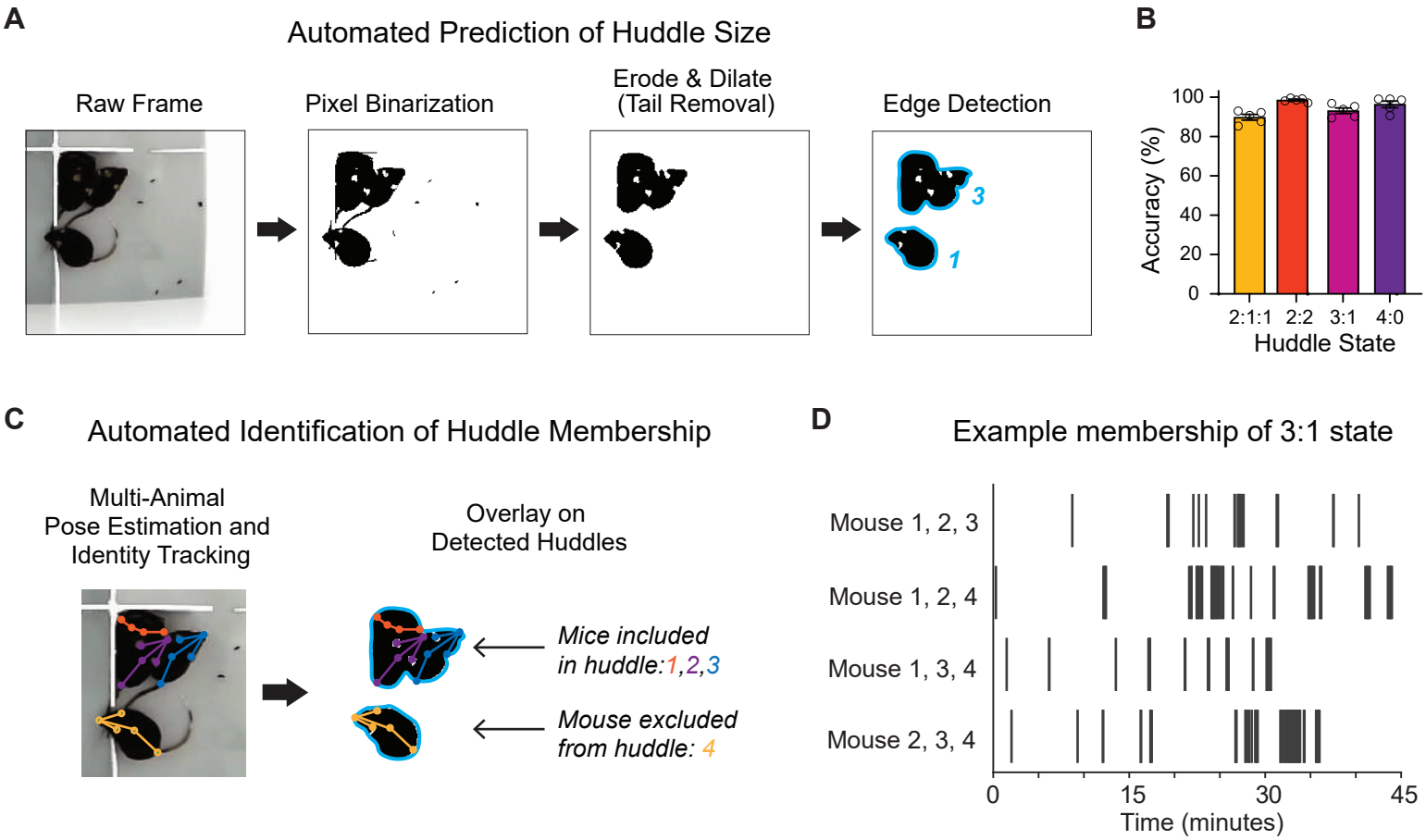
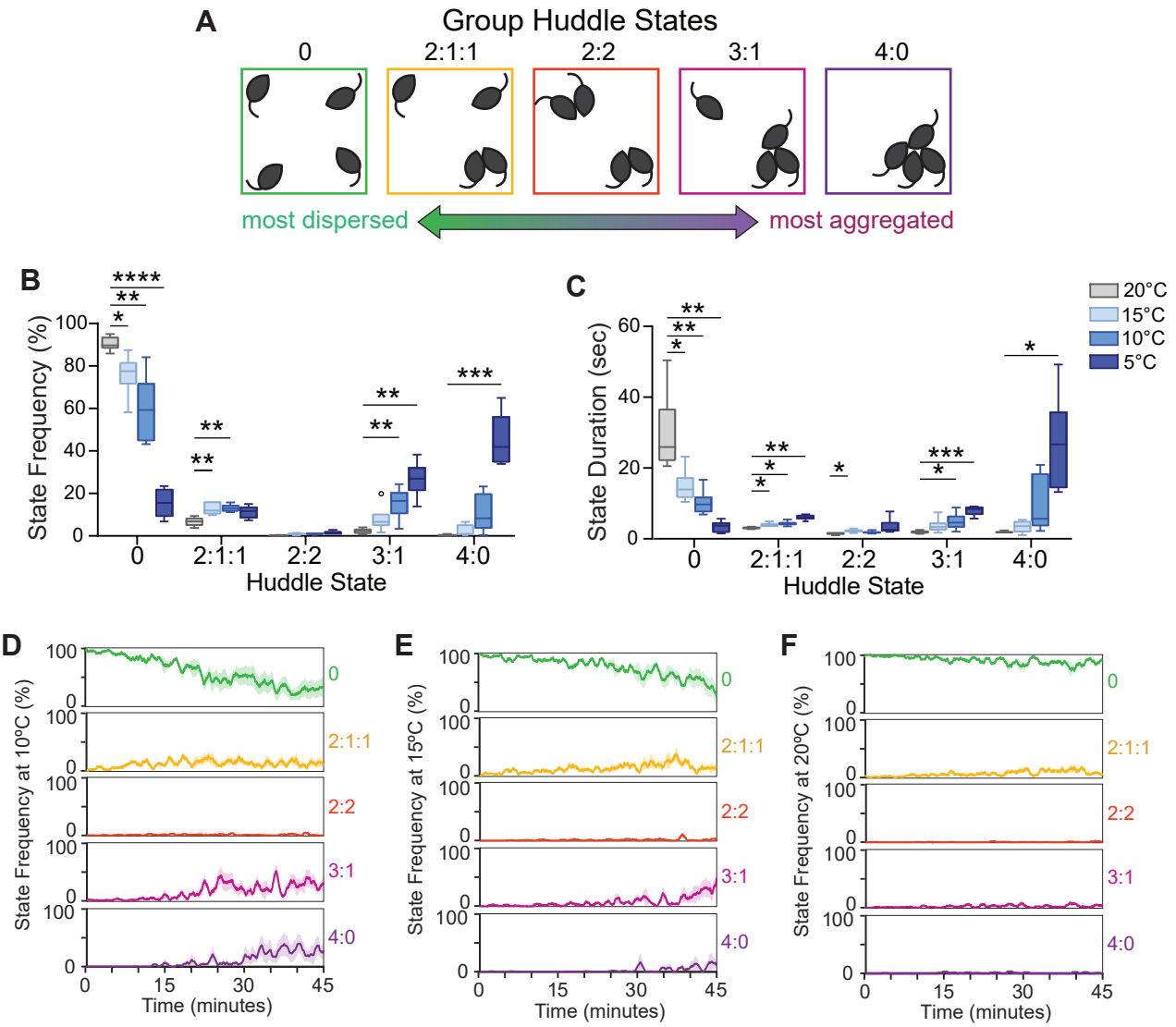
Supplementary Figure 1



Supplementary Figure 1: Automated pipeline for group huddle behavior analysis. **A.** Pipeline for automated detection of huddle size. Raw frames are binarized into black and white pixels. Opening (erosion followed by dilation) is performed to removed tails and fecal artifacts. Edge detection is performed to identify connected groups of animals. **B.** Percent accuracy of detected huddle state compared to manual human annotation **C.** Automated Identification of huddle membership is achieved by tracking raw behavior videos with a trained neural network (Social Leap Estimates Animal Poses) to identify individual nodes and identities. Tracked poses and identities are overlayed on top of detected huddles to identify the membership. **D.** Example raster plot for one group demonstrating membership configurations for huddles of three throughout one behavior session.

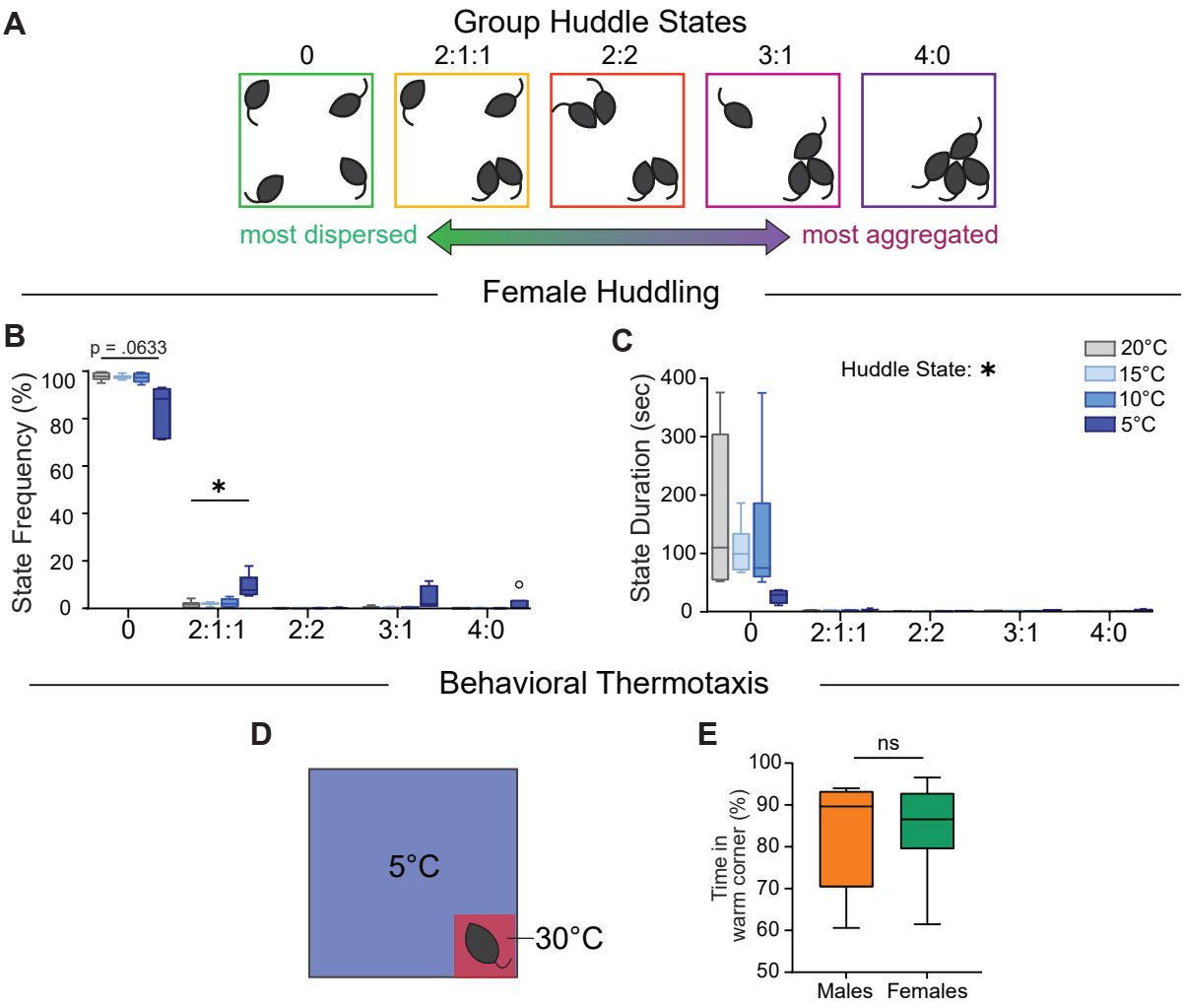
Supplementary Figure 2



Supplementary Figure 2: Titration of ambient temperature during thermal challenge assay.

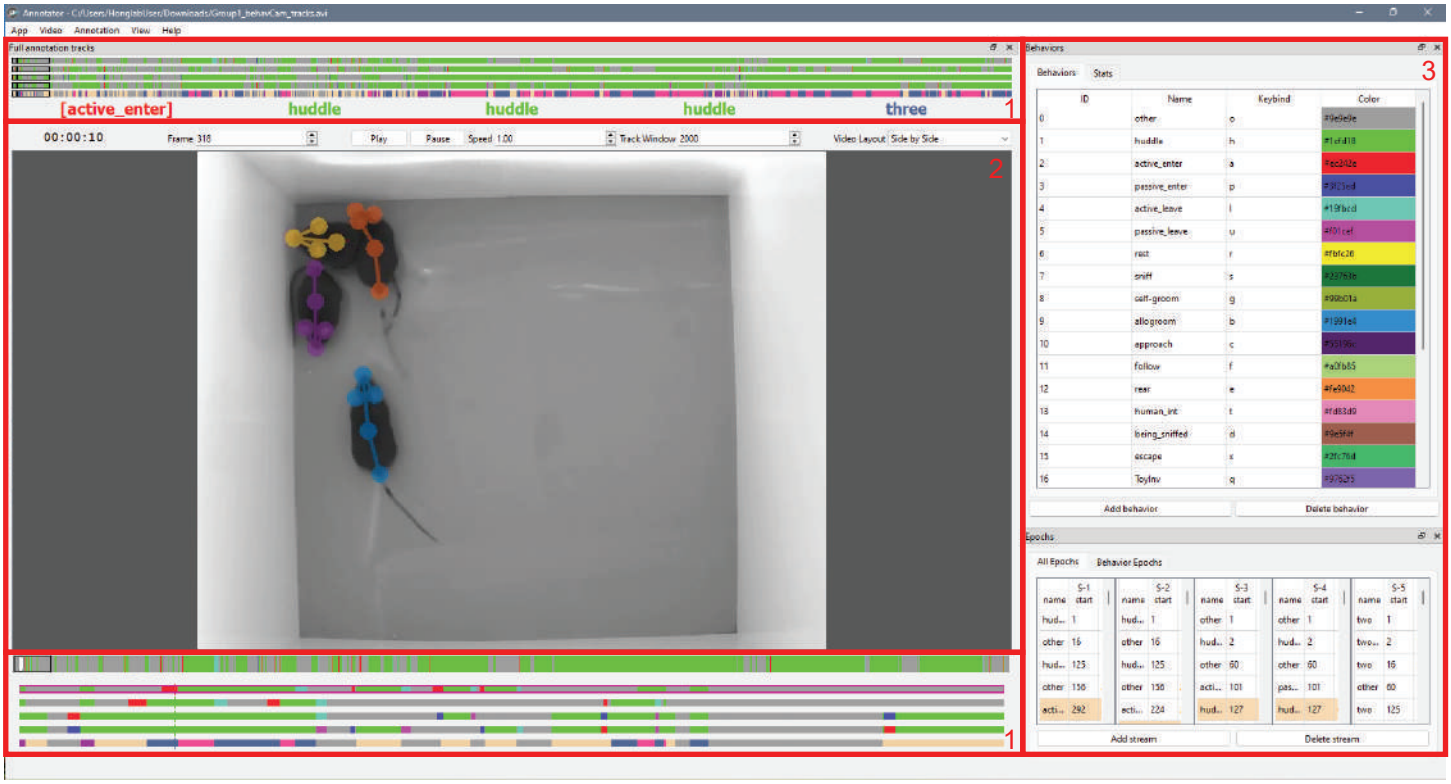
A. Schematics illustrating 5 unique group states derived via automated SLEAP pose estimation and identity tracking, ranging from most dispersed to most aggregated. **B.** Frequency of group states observed at 20°C, 15°C, 10°C, or 5°C during thermal challenge assay (n = 6 groups of 4 individuals). **C.** Mean group state duration in seconds observed at 20°C, 15°C, 10°C, or 5°C during thermal challenge assay (n = 6 groups of 4 individuals). **D.** Moving average (mean ± SEM) of percent time of all five group states plotted over time at 10°C (n = 6 groups of 4 individuals). **E.** Moving average (mean ± SEM) of percent time of all five group states plotted over time at 15°C (n = 6 groups of 4 individuals). **F.** Moving average (mean ± SEM) of percent time of all five group states plotted over time at 20°C (n = 6 groups of 4 individuals). Box and whisker plots indicate the following: center line – median; box limits – upper and lower quartiles; whiskers – minimum and maximum values. Statistical tests include two-way repeated measures analysis of variance (ANOVA) with Bonferroni post-hoc tests (**B**,**C**). **P*<.05, ***P*<.01, ****P*<.001, *****P*<.0001. See Supplementary Table 1 for details of statistical analyses.

Supplementary Figure 3



Supplementary Figure 3: Huddling during thermal challenge in females.
A. Schematics illustrating 5 unique group states derived via automated SLEAP pose estimation and identity tracking, ranging from most dispersed to most aggregated. **B.** Frequency of group states observed at 20°C, 15°C, 10°C, or 5°C during thermal challenge assay in females (n = 6 groups of 4 individuals). **C.** Mean group state duration in seconds observed at 20°C, 15°C, 10°C, or 5°C during thermal challenge assay in females (n = 6 groups of 4 individuals). **D.** Schematic illustrating behavioral thermotaxis assay. Animals are placed in a behavioral chamber at 5°C with free access to a 30°C warm corner. **E.** Comparison of percent time spent in warm corner during thermotaxis assay in males and females (n = 8 males, 8 females). Box and whisker plots indicate the following: center line – median; box limits – upper and lower quartiles; whiskers – minimum and maximum values. Statistical tests include two-way repeated measures analysis of variance (ANOVA) with Bonferroni post-hoc tests (**B**,**C**), and Wilcoxon matched pairs tests (**E**). **P*<.05, ***P*<.01, ****P*<.001, *****P*<.0001. See Supplementary Table 1 for details of statistical analyses.

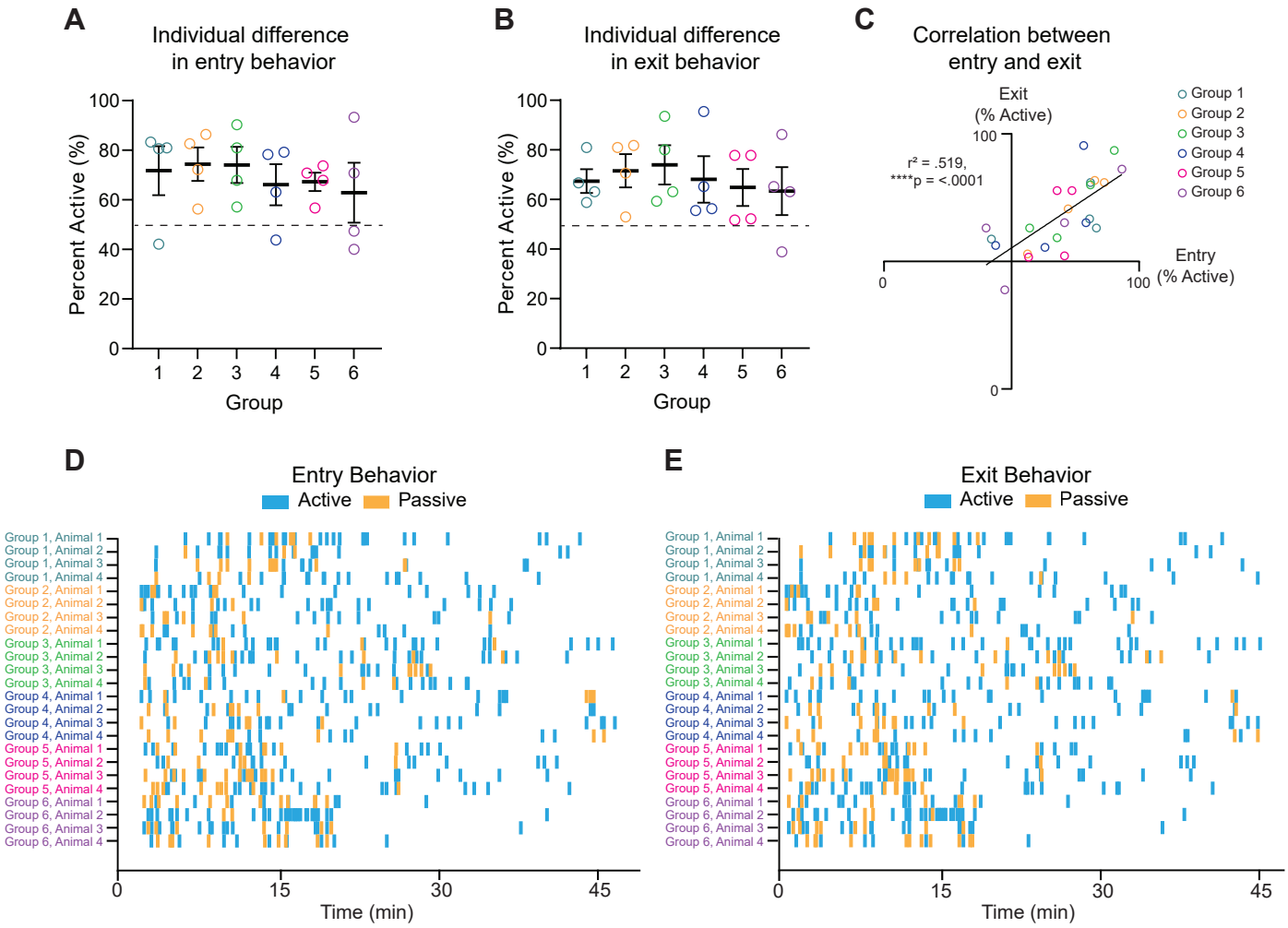
Supplementary Figure 4



Supplementary Figure 4: Graphical user interface for BehaviorAnnotator, a custom software for manual annotation and analysis of multi-animal behavior.

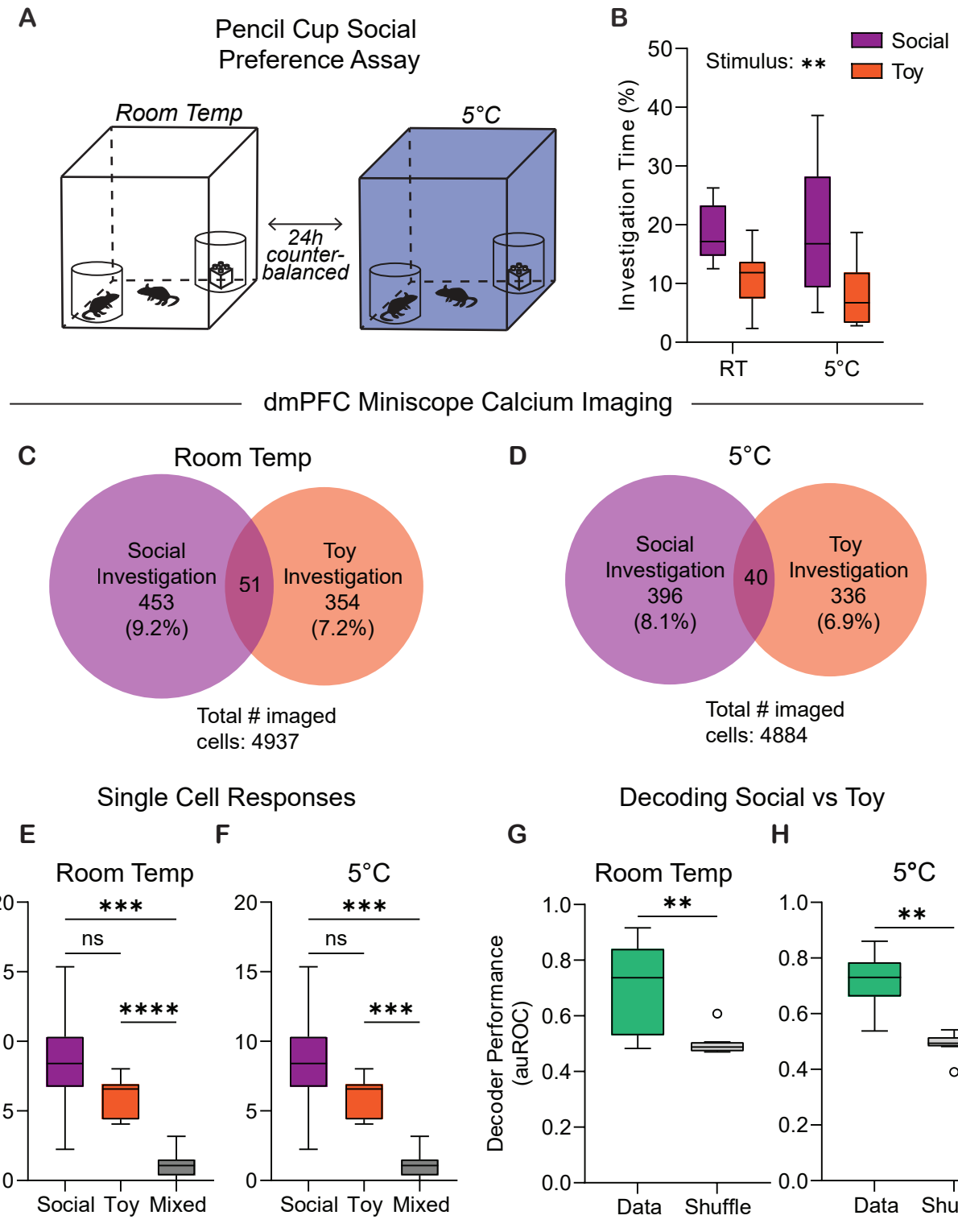
The graphic user interface of the annotator has 3 panels. Panel 1 displays the annotation streams containing user defined behaviors for all four animals, and a fifth stream which denotes the aggregate huddle size when a huddle is present. Panel 2 displays the behavior video(s). Panel 3 displays the list of user-defined behaviors and labeled behavior epochs.

Supplementary Figure 5



Supplementary Figure 5: Individual difference in active vs passive decisions.
A. Percent of entry decisions (mean ± SEM) that are active plotted for all four individuals in six groups. **B.** Percent of exiting decisions (mean ± SEM) that are active plotted for all four individuals in six groups. **C.** Correlation between percent of entry decisions and percent of exiting decisions that are active (n = 24 individuals from 6 groups). **D.** Raster plot illustrating active and passive entry events throughout full behavioral session (n = 24 individuals from 6 groups). **E.** Raster plot illustrating active and passive exiting events throughout full behavioral session (n = 24 individuals from 6 groups). Statistical tests include linear regression (**C**). *P<.05, **P<.01, ***P<.001, ****P<.0001. See Supplementary Table 1 for details of statistical analyses.

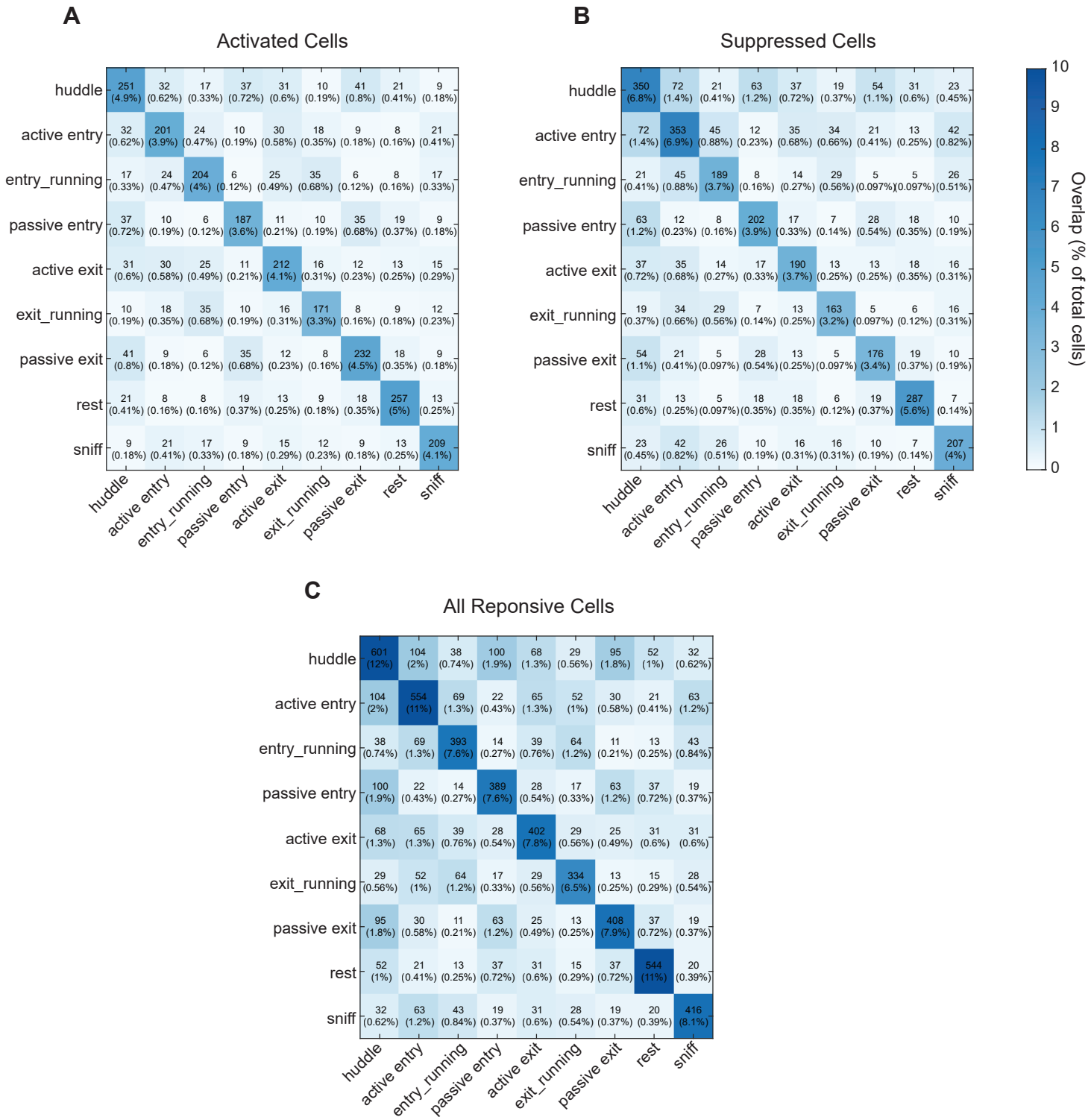
Supplementary Figure 6



Supplementary Figure 6: Cold ambient temperature does not alter general social preference or dmPFC encoding of social stimuli.

A. Schematic illustrating pencil cup social preference assay. Animals were tested for 30 minutes at room temperature or 5°C to determine preference for wired pencil cup containing a conspecific vs a toy. **B.** Quantification of investigation time directed towards social cup vs toy cup at room temperature (RT) or 5°C (n = 10 animals). **C.** Venn diagram showing dmPFC cells responsive to social and toy investigation at room temperature. Total # of imaged cells = 4937 from animals. **D.** Venn diagram showing dmPFC cells responsive to social and toy investigation at room temperature. Total # of imaged cells = 4884 from 10 animals. **E.** Percent of dmPFC cells that are social responsive, toy responsive, or mixed responsive at room temperature (n = 10 animals). **F.** Percent of dmPFC cells that are social responsive, toy responsive, or mixed responsive at 5°C (n = 10 animals). **G.** Support vector machine (SVM) decoder performance to decode social vs toy investigation at room temperature (n = 10 animals). **H.** Support vector machine (SVM) decoder performance to decode social vs toy investigation at 5°C (n = 10 animals). Box and whisker plots indicate the following: center line – median; box limits – upper and lower quartiles; whiskers – minimum and maximum values. Statistical tests include one-way (**E-F**) and two-way (**B**) repeated measures analysis of variance (ANOVA) with Bonferroni post-hoc tests, and Wilcoxon matched pairs tests (**G-H**). **P*<.05, ***P*<.01, ****P*<.001, *****P*<.0001. See Supplementary Table 1 for details of statistical analyses.

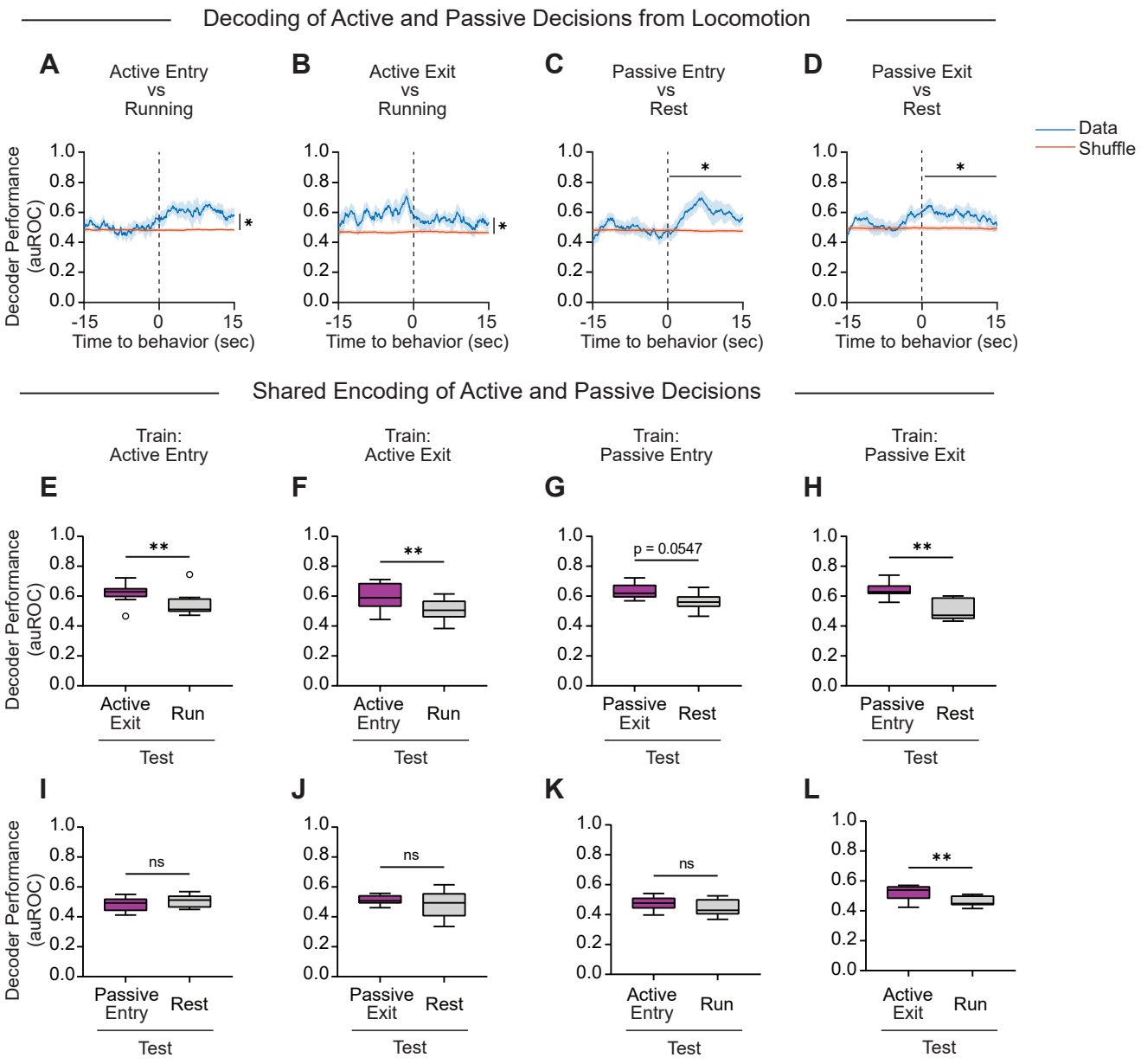
Supplementary Figure 7



Supplementary Figure 7: Matrices showing overlap of dmPFC cells responsive to various behaviors.

A. Matrix showing number of cells activated by behaviors on x and y axis. Percentages correspond to percent of total imaged cells. **B.** Matrix showing number of cells suppressed by behaviors on x and y axis. Percentages correspond to percent of total imaged cells. **C.** Matrix showing number of all cells responsive to behaviors on x and y axis. Percentages correspond to percent of total imaged cells. Total # of imaged cells = 5141 from 11 animals.

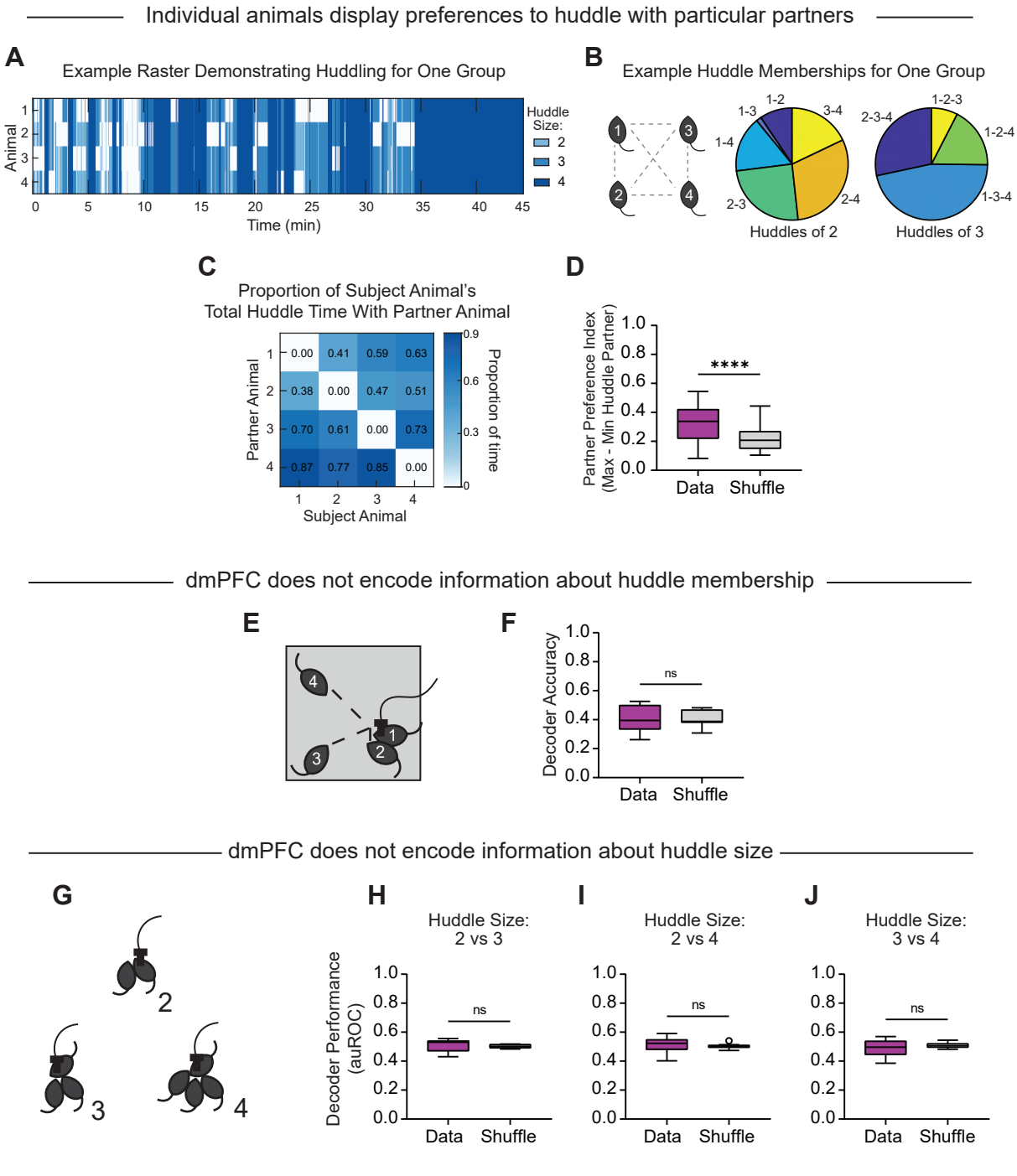
Supplementary Figure 8



Supplementary Figure 8: Additional dmPFC decoding of active and passive decisions.

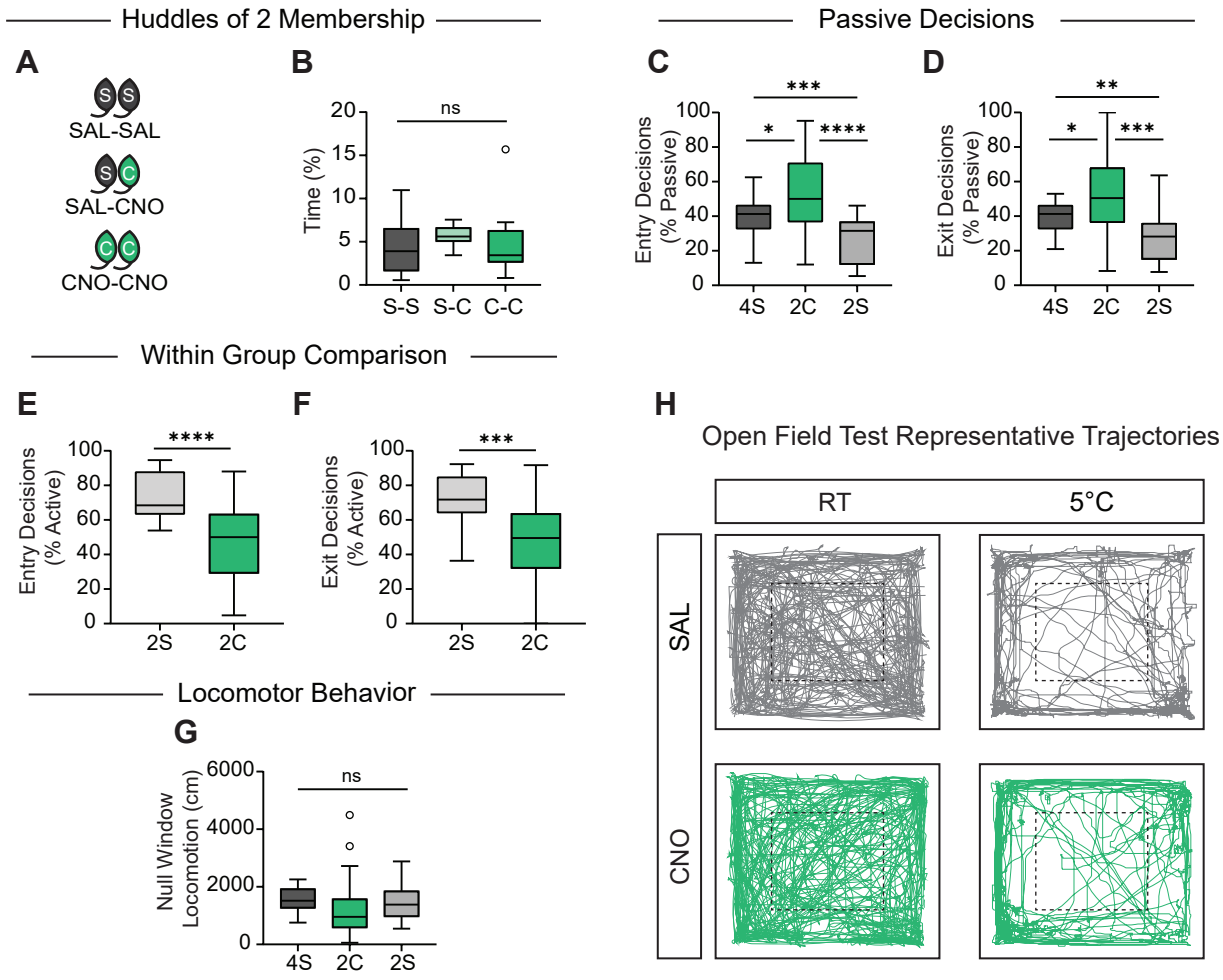
A. Performance of SVM decoders trained to classify active entry from speed-matched running. **B.** Performance of SVM decoders trained to classify active exit from speed-matched running. **C.** Performance of SVM decoders trained to classify passive entry from rest. **D.** Performance of SVM decoders trained to classify passive exit from rest. **E.** Performance of SVM decoders trained to classify active entry from baseline in predicting active exit from speed-matched running. **F.** Performance of SVM decoders trained to classify active exit from baseline in predicting active entry from speed-matched running. **G.** Performance of SVM decoders trained to classify passive entry from baseline in predicting passive exit from rest. **H.** Performance of SVM decoders trained to classify passive exit from baseline in predicting passive entry from rest. **I.** Performance of SVM decoders trained to classify active entry from baseline in predicting passive entry from rest. **J.** Performance of SVM decoders trained to classify active exit from baseline in predicting passive exit from rest. **K.** Performance of SVM decoders trained to classify passive entry from baseline in predicting active entry from speed-matched running. **L.** Performance of SVM decoders trained to classify passive exit from baseline in predicting active exit from speed-matched running. Box and whisker plots indicate the following: center line – median; box limits – upper and lower quartiles; whiskers – minimum and maximum values. Statistical tests include two-way repeated measures analysis of variance (ANOVA) with Bonferroni post-hoc tests (**A-D**) and Wilcoxon matched pairs tests (**E-L**). **P*<.05, ***P*<.01, ****P*<.001, *****P*<.0001. See Supplementary Table 1 for details of statistical analyses.

Supplementary Figure 9



Supplementary Figure 9: dmPFC does not encode huddle size or membership
A. Example raster plot demonstrating huddling behavior for all four animals in one session, color coded by huddle size. **B.** Example pie charts showing proportion of time for various huddle configurations for huddles of two and three for one group. **C.** Matrix demonstrating proportion of subject animal's (x-axis) total huddle time with partner animals (y-axis) for one session. Sum of proportions for one animal can exceed 1 because subjects can huddle with more than one animal at a time in a larger huddle of two or three. **D.** Partner preference index (maximum preferred partner – minimum preferred partner) for real data versus a shuffled variation of the data in binary vectors containing individual huddle behaviors are circularly shifted relative to each other. **E.** Schematic illustrating potential huddle memberships for huddles of two during a miniscope imaging session. **F.** Performance of multi-class linear discriminant analysis (LDA) decoders trained to classify huddle membership for huddles of two from dmPFC population activity. Note that baseline is .33 because there are three possible memberships. **G.** Schematic illustrating potential huddle sizes during a miniscope imaging session. **H.** Performance of SVM decoders trained to classify huddle size of 2 from 3. **I.** Performance of SVM decoders trained to classify huddle size of 2 from 4. **J.** Performance of SVM decoders trained to classify huddle size of 3 from 4. Box and whisker plots indicate the following: center line – median; box limits – upper and lower quartiles; whiskers – minimum and maximum values. Statistical tests include Wilcoxon matched pairs tests (**D**,**F**,**H-J**). **P*<.05, ***P*<.01, ****P*<.001, *****P*<.0001. See Supplementary Table 1 for details of statistical analyses.

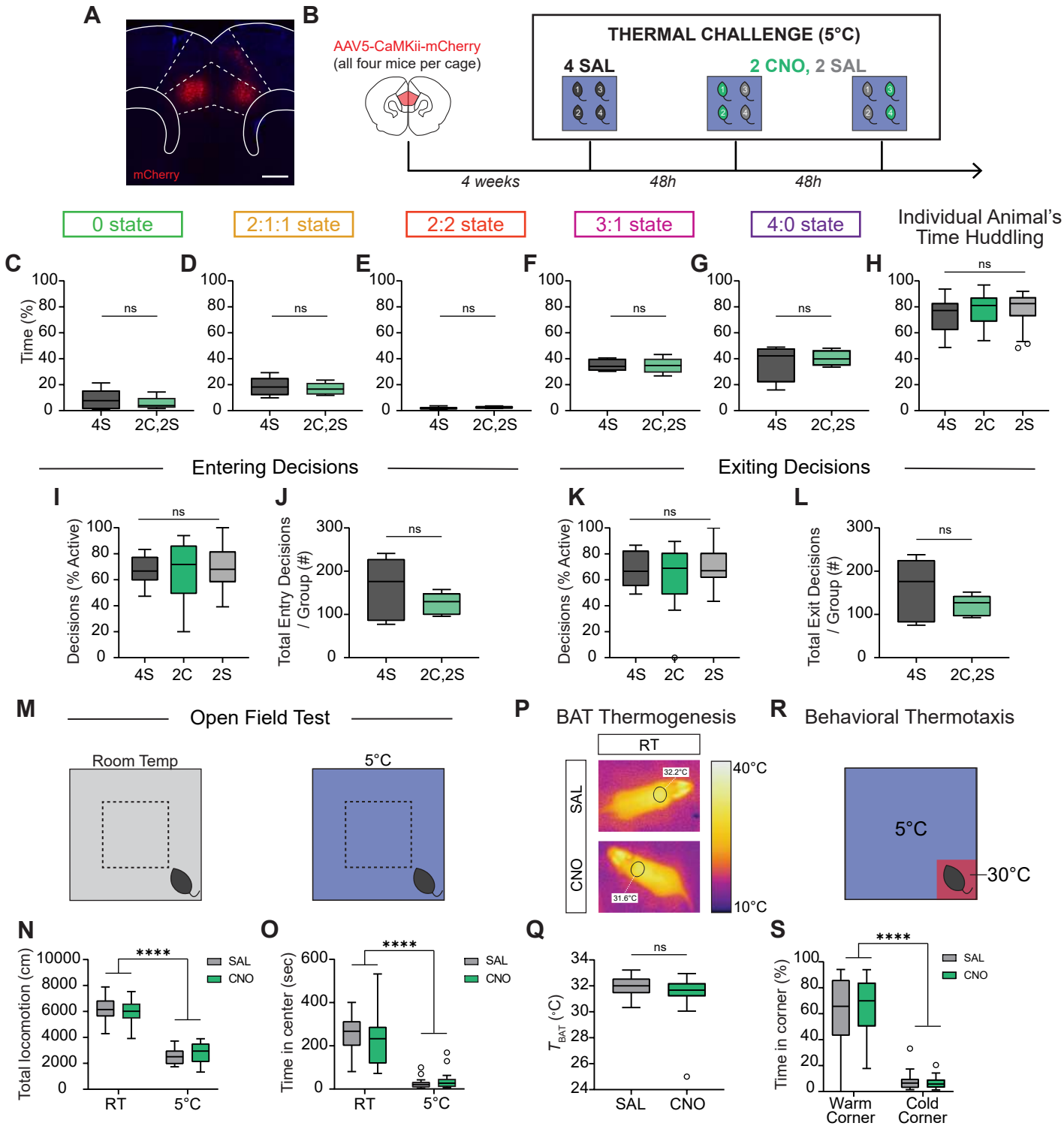
Supplementary Figure 10



Supplementary Figure 10: Additional data related to chemogenetic silencing in Figure 5.

A. Schematic illustrating potential composition of membership for huddles of 2: SAL-SAL, SAL-CNO, and CNO-CNO during thermal challenge in 2C,2S condition. **B.** Percent of total time observed for possible membership compositions for huddles of two (n = 24 individuals from 6 groups). **C.** Within animal comparison of percent of entry decisions that are passive during 4S, 2C, and 2S conditions (n = 24 individuals from 6 groups). **D.** Within animal comparison of percent of exiting decisions that are passive during 4S, 2C, and 2S conditions (n = 24 individuals from 6 groups). **E.** Within group comparison of percent of entry decisions that are active for 2C, 2S condition (n = 24 individuals from 6 groups). Data in main figure shown as within animal comparisons. **F.** Within group comparison of percent of exiting decisions that are active for 2C, 2S condition (n = 24 individuals from 6 groups). Data in main figure shown as within animal comparisons. **G.** Individual animals' total locomotion during null windows when no active, passive, or huddle behaviors are annotated during 4S, 2C, and 2S conditions (n = 24 individuals from 6 groups). **H.** Representative open field test trajectories at room temperature and 5°C after SAL or CNO injection. Box and whisker plots indicate the following: center line – median; box limits – upper and lower quartiles; whiskers – minimum and maximum values. Statistical tests include one-way repeated measures analysis of variance (ANOVA) with Bonferroni post-hoc tests (**B-D**,**G**) and Mann-Whitney tests (**E-F**). **P*<.05, ***P*<.01, ****P*<.001, *****P*<.0001. See Supplementary Table 1 for details of statistical analyses.

Supplementary Figure 11



Supplementary Figure 11: mCherry controls for chemogenetic silencing experiments.
A. Example image showing AAV-mCherry expression in the dmPFC. Scale bar, 500 μm. **B.** Schematic illustrating experimental paradigm for mCherry chemogenetic control during thermal challenge. 4 SAL refers to condition in which all four animals are injected with saline. 2 CNO, 2 SAL refers to condition in which two animals are injected with CNO, and two with saline. **C-G.** Percent time in huddle states observed for all five group states during 4S and 2C,2S conditions (n = 5 groups). **H.** Individual animal's total percent time spent huddling in 4S, 2C, and 2S conditions (n = 20 individuals from 5 groups). **I,K.** Within animal comparison of percent of entry or exit decisions that are active during 4S, 2C, and 2S conditions (n = 20 individuals from 5 groups). **J,L.** Total number of entry or exit decisions per group (active and passive from all four animals) during 4S and 2C,2S conditions (n = 20 individuals from 5 groups). **M.** Schematic illustrating open field test at room temperature (RT) and 5°C. **N.** Within animal comparison of total locomotion or time in center during open field test at both room temperature and 5°C after SAL or CNO injection (n = 20 animals). **P.** Representative infrared thermal images demonstrating temperature above BAT (brown adipose tissue, black circles) after SAL or CNO injection at room temperature. **Q.** Quantification of thermography images in regions above BAT after SAL or CNO injection (n = 20 animals). **R.** Schematic illustrating behavioral thermotaxis assay. **S.** Within animal comparison of percent time spent in warm corner versus the average of three cold corners after SAL or CNO injection (n = 20 animals). Box and whisker plots indicate the following: center line – median; box limits – upper and lower quartiles; whiskers – minimum and maximum values. Statistical tests include one-way (**H**,**I**,**K**) and two-way (**N**,**O**,**S**) repeated measures analysis of variance (ANOVA) with Bonferroni post-hoc tests, and Wilcoxon matched pairs tests (**C-G**,**J**,**L**,**Q**) tests. **P*<.05, ***P*<.01, ****P*<.001, *****P*<.0001. See Supplementary Table 1 for details of statistical analyses.

Supplementary Table 1: Details of Statistical Analyses
(see attached Excel File: Supplementary Table 1)

SUPPLEMENTARY METHODS

Female Huddling Behavior

To assess group huddling behaviors, groups of four co-housed female mice were acclimated to behavioral box for 20 minutes per day for three days prior to onset of behavior testing. In addition, mice were habituated to human handling for 5 minutes. On the day of testing, animals were tail-marked with Sharpie pen in order to maintain identities when performing analyses post-hoc. Co-housed groups of four animals were placed in a 40 x 40 cm acrylic behavior box in a temperature controlled chamber (Thermo Scientific, PR505755R Refigerated Incubator) to measure huddling behavior for 45 minutes. After behavioral testing, animals were returned to their home cage. Animals were tested every 24 hours at 5ºC, 10ºC, 15ºC, 20ºC in scrambled order. Analysis of huddling behavior states was carried out in an identical fashion to male animals (see Pipeline for Analysis of Huddling Behavior in Main Methods). Fecal matter was quantified by counting the total number of feces on the last frame of each behavior video, for both males and females at 5°C and 20°C.

Female and Male Thermotaxis Behavior

Age-matched adult male and female mice were acclimated to the behavioral box following the protocol outlined in the Behavior Assays section in the Main Methods. In addition, mice were habituated to human handling 5 minutes on each day. On the day of testing, animals were placed in a 40 x 40 cm acrylic behavior box in at 5ºC for 15 minutes. One corner of the box was warmed to 30ºC by placing Hand Warmers (HotHands) on the outside of the box for 1 hour prior to behavior. CNO and saline were administered 48 hours apart in a counterbalanced manner. The corner used as a warm corner was also counterbalanced. Animal pose points were tracked using SLEAP (see Analysis of Animal Behavior, above), and the animal's time spent in each corner was measured.

Support Vector Machine Decoding of Active and Passive Decisions from Locomotion

A support vector machine (SVM) decoder was trained to decode active decisions from speed-matched running bouts, and passive decisions from rest bouts, using the z-scored population calcium activities. Behavior classes are decoded using calcium activity along a 30-second time window centered at behavioral onset. Behaviors are first aligned from 15 seconds before onset to 15 seconds after onset and balanced with a random bootstrap, as described in the main methods. For each frame in the time series, an SVM decoder was trained on the z-scored population calcium activities of that frame in all behavioral bouts and tested using a leave-one-out cross-validation (LOOCV) procedure as described above. The decoder accuracy was compared with that of 500 random

circularly shifted activities. Manually annotated rest bouts were supplemented with additional speed-matched immobility bouts identified using SLEAP data. For each analysis, animals that did not have a minimum of 5 behavior bouts were excluded.

Mutual Decoding of Active and Passive Behaviors

Shared encoding of active and passive behaviors was assessed by training an SVM decoder on one behavior and testing on another. For instance, to test a decoder trained with active enter on active leave, a decoder is trained to decode active enter bouts against the same number of randomly drawn null bouts (no annotated behavior). The decoder is then tested to decode active leave versus randomly drawn null bouts. Performance is calculated as auROC of the test scores against true labels. Performance is compared to performance on a speed-matched running control when the test set is an active decision, and a rest control when the test set is a passive decision. Animals that had less than 10 manually annotated rest bouts were supplemented with additional speed-matched immobility bouts identified using SLEAP data. For each analysis, animals that did not have a minimum of 10 behavior bouts were excluded.

Calculation of Partner Preference Index for Huddle Memberships

To assess animals' preference to huddle with other group members, a preference index was calculated for each animal using the following equation: $(T_{max} - T_{min})/T_{total}$. Where T_{max} is the total huddle time with the most preferred member, Tmin is the total huddle time with the least preferred member, and T_{total} is the total huddle time for the subject animal. The preference index is compared with a shuffle that controls for each animal's total huddle time respectively. The binary vector representing each animal's frame-by-frame huddle status are circularly shifted against each other to create temporal misalignment between animals. The time frames where only one animal is engaged in the huddle are randomly matched so that a huddle is composed of at least two animals. One thousand shuffles were created for each method and the averaged shuffle preference index was compared to true preference index.

Multi-class Linear Discriminant Analysis (LDA) of Huddle Membership

A multi-class Linear Discriminant Analysis (LDA) was used to decode membership from calcium activities. For each experiment, huddle bouts when the imaged animal huddles with each of the 3 partners are averaged for decoding, each partner forming one class. For huddles longer than 10 seconds, only the first 10 seconds are averaged. Huddles for the 3 classes are balanced with random bootstrapping, trained with a 3-class linear discriminant analysis classifier and tested using a leave-one-out cross validation (LOOCV) procedure. The decoder accuracy was compared with that of 500 random circularly shifted activities. For each analysis, animals that did not have a minimum of 5 behavior bouts were excluded.

Decoding of Huddle Size

A support vector machine (SVM) decoder was trained to decode huddles of 2, 3, and 4 in a pairwise manner using the z-scored population calcium activities. For each imaging session, average neural activities of each behavioral bout are calculated for the two

behavioral classes. For bouts that are longer than 10 seconds, the first 10 seconds are averaged. Bouts of the two behavioral classes are balanced by randomly drawing from the class with more bouts, such that the number of bouts are equal. Performance of the decoder performance is tested using a leave-one-out cross-validation (LOOCV) procedure, where one bout serves as the test set and the rest as the training set which is repeatedly tested for all bouts. To eliminate contamination, the training samples that are within 15 seconds from the test sample are eliminated from the training set. The test samples' prediction scores are compared against the true labels to produce auROC. To generate the shuffled performance, calcium activities are circularly shifted with random time lag against the behaviors for 500 times, and an auROC is calculated for each shuffle. For each imaging session, the averaged auROC of 500 shuffles is compared to the averaged auROC of 50 runs from the experiment data. For each analysis, animals that did not have a minimum of 10 behavior bouts were excluded.

Social Preference Assay Behavior and Calcium Imaging

For calcium imaging during the social preference assay, animals were outfitted with the head-mounted Miniscope, briefly habituated in their home cage for 2-3 minutes, and then placed in a 40 x 40 cm arena. The arena contained two pencil wire cups in opposing corners. One cup contained an unfamiliar adult male, while the other contained an inanimate toy mouse. The subject animals were allowed to freely move about the environment and investigate social and toy stimuli at will for the duration of the 30 minute session. Subjects were imaged at room temperature and at 5°C 48 hours apart in a counterbalanced manner. We imaged 4937 neurons from 10 animals at room temperature and 4884 neurons from the same 10 animals at 5°C. Stimulus animals were habituated to the pencil cups for 20 minutes per day for 3 days prior to experiments. Subject animal pose points (nose, left ear, right ear, body, tail base) were tracked using SLEAP (see Main Methods). We considered investigation events to be periods where the animal's head was within 3 inches of the center of the cup, and the angle between its head and the cup was < 60 degrees. Behavior annotations were converted into binary vectors that denote precisely which frames the animal is engaged in social vs toy investigation for downstream analysis.

Social Preference Assay Single Cell Analysis

In the social preference assay, we analyzed the responses of individual dmPFC neurons when subjects closely investigated either the social or the toy chamber. Prior to downstream analysis, all calcium traces were z-scored and presented throughout in units of standard deviation. We applied receiver operating characteristic (ROC) analysis to identify neurons that significantly responded during each type of investigation. We applied a binary threshold to the $\Delta F/F$ signal, classifying each time point as either indicating or not indicating a specific event. The true positive rate and false positive rate were computed over a range of binary thresholds that spanned the full range of the neural signal. These rates were used to construct an ROC curve, which depicts the detection capability of the neural signal at various thresholds. The area under the ROC curve (auROC) was then determined to quantify how strongly neural activity was influenced by each event. To evaluate significance, the observed auROC was compared against a null distribution, generated by circularly permuting the calcium signals with random circular

time shifts 2000 times. A neuron was deemed significantly responsive ($\alpha < 0.05$) if its auROC exceeded the 97.5th percentile (indicating activation) or fell below the 2.5th percentile (indicating suppression) of the null distribution. This analysis included only time points marked as the behavior event of interest (social investigation or toy investigation) and baseline points which excludes all annotated events, ensuring that the identification of neurons responsive to a specific behavior was not confounded by their activity during other behaviors.

Decoding of Social vs Toy Investigation During Social Preference Assay

To assess population level decoding of social vs object interaction from dmPFC neural data, we applied Linear Support vector machine (SVM) to identify hyperplanes that best separate the pair of population vectors associated with different events, using a leave-one-out prediction cross-validation (LOOCV) approach. We averaged the mean population activity associated with each independent event bout lasting at least 1.5 seconds. The mean activity for each cell was calculated over the entire bout duration, up to a maximum of 10 seconds post-event onset. The leave-one-out cross-validation method was then applied. For each test sample in each validation fold, we excluded samples within one minute before or after the test event onset to prevent temporal contamination between training and test datasets. we randomly down-sampled the majority class to match the minority class within the remaining training samples. The auROC value was computed for the predicted class probabilities. We generated shuffle controls by circularly shifting the events along the time axis 100 times to establish a chance level performance benchmark. These methods were applied to decode social investigation from toy investigation in a pairwise manner.