The representational geometry of emotional states in basolateral amygdala

Pia-Kelsey O'Neill[1,2,*,†], Lorenzo Posani*[2,3], Jozsef Meszaros[4,5], Phebe Warren[1,6], Carl E. Schoonover[1,2,7], Andrew J.P. Fink[1,2,7], Stefano Fusi[1-3,8], C. Daniel Salzman[1-5,8,†]

[1] The Mortimer B. Zuckerman Mind Brain Behavior Institute, New York, NY, USA
[2] Department of Neuroscience, Columbia University, New York, NY, USA
[3] Center for Theoretical Neuroscience, Columbia University, New York, NY USA
[4] Department of Psychiatry, Columbia University, New York, NY, USA
[5] New York State Psychiatric Institute, New York, NY, USA
[6] Present Address: Columbia University Mailman School of Public Health, New York, NY, USA
[7] Howard Hughes Medical Institute, Columbia University, New York, NY, USA
[8] Kavli Institute for Brain Science, Columbia University, New York, NY USA
*These authors contributed equally to this work
†Correspondence: pto2102@columbia.edu (PK.O.), cds2005@columbia.edu (C.D.S.)

SUMMARY

Sensory stimuli associated with aversive outcomes can cause multiple behavioral responses related to an animal's evolving emotional state. We employed chemogenetic inactivation and two-photon imaging to reveal how the basolateral amygdala (BLA) mediates these state changes. Mice were presented stimuli in a virtual burrow, causing two responses reflecting fear and flight to safety: tremble and ingress into the burrow. Inactivation eliminated differential tremble and ingress to aversive and neutral stimuli without eliminating responses themselves. Multiple variables, including stimulus valence and identity, and being in the tremble or ingressed state, typically modulated each neuron's activity (mixed-selectivity). BLA neural ensembles represented these variables even after neurons with apparent specialized selectivity were eliminated from analyses. Thus, implementing different readouts of BLA ensembles comprised of mixed-selectivity neurons can identify distinct emotional states defined by responses, like tremble for fear and ingress for safety. This mechanism relies on BLA's representational geometry, not its circuit specialization.

INTRODUCTION

In everyday life, we commonly ascribe emotional states to individuals by observing their responses to sensory stimuli or events. These responses include various behaviors like approach or avoidance, as well as physiological changes such as autonomic reactivity. In fact, sensory stimuli that elicit emotional states commonly cause individuals to generate not one but multiple different responses, like freezing and fleeing in response to a threat[1,2]. Although the sensory cues that elicit these responses are often predictive of rewarding or aversive outcomes, even novel stimuli can evoke similar responses, highlighting the propensity of emotional behavior to generalize from familiar situations to new ones[3]. Neural mechanisms that underlie emotional behavior must therefore provide the capacity to assign affective significance to sensory stimuli, to generalize affective significance from familiar to novel stimuli, and to connect representations of stimuli to circuitry that implements behavioral and physiological responses. To date, it is still

unclear which mechanisms are employed to satisfy these requirements of emotional processing in the brain.

Dedicated and distinct neural circuitry is required to generate any specific response, but it remains unclear when this specialization first emerges during the processing of stimuli. Many of the behavioral and physiological responses that reflect an aversive emotional state are generated by specialized neural circuitry in the central nucleus of the amygdala (CeA)[4–9]. As an example, genetically and physiologically distinct neuronal populations in CeA mediate freezing and flight[6,8,10]. The CeA receives direct input from the basolateral complex of the amygdala (BLA)[11–14], but it is unknown if both specialized CeA populations receive input from the same BLA neurons. Moreover, the BLA may also influence CeA processing through polysynaptic pathways, such as those involving the intercalated masses in the amygdala, the insula, and a variety of other structures[11]. The existence of polysynaptic pathways indicates that anatomical tracing of the direct BLA to CeA circuit alone will not resolve the degree of BLA circuit specialization with respect to generating any specific response. Physiological assessment of neuronal coding properties is likely needed to shed light on whether specialized BLA circuits are engaged during different emotional states associated with distinct responses.

To study how an aversive conditioned stimulus (CS) can elicit multiple responses, we adopted a recently established experimental platform, the virtual burrow assay[15]. This approach enabled us to characterize complex and evolving defensive behaviors in head-fixed mice while we simultaneously recorded or inactivated neural activity in BLA. In response to an aversive olfactory CS, mice enact two defensive behaviors in the burrow without training or operant conditioning: tremble and ingress into the burrow. Tremble and ingress possess distinct dynamics, and mice transition between these responses during the constant delivery of a conditioned odor. Thus, changes in internal state, not changes in external cues, likely drive transitions between these responses. Tremble and ingress into the burrow can be used to ascribe distinct emotional states to mice, states that are analogous to those observed when a threatening stimulus causes freely moving mice to initially freeze, reflecting fear, and then flee to a location where they can feel safe[1,2,16,17].

To study the role of BLA in processing aversive stimuli presented in the burrow, we first used chemogenetic inactivation to determine if BLA activity is required to generate tremble and ingress. Inactivation did not eliminate either response, but it did establish that BLA activity is necessary for generating differential tremble and ingress responses to an aversive as compared to a neutral CS. However, these results do not illuminate whether circuit specialization in BLA underlies the tremble and ingress responses. We therefore turned to two-photon imaging of cellular activity to characterize the physiological properties of BLA neurons as conditioned stimuli were presented to mice in the burrow.

Understanding the physiological profiles of neurons can help to elucidate the degree of circuit specialization in BLA because if neurons exhibit mixed selectivity for multiple variables[19], then the same underlying neural ensemble could contribute to more than one input-output computation[18–20], such as computations related to both tremble and ingress. By contrast, if BLA

neurons predominantly exhibit specialized selectivity for one variable, it would suggest the existence of distinct BLA circuits for different responses associated with distinct emotional states, as specialized neurons can only perform one type of input-output operation[19,20].

Although the characterization of individual neurons' physiological response properties is an important first step for understanding the functional role of a neural ensemble, recent work has emphasized that an analysis of representational geometry reveals the distinctive computational capabilities of a neural ensemble[21–23]. The representational geometry of a neural ensemble is defined by the configuration of points when the activity of every recorded neuron is plotted against each other for every condition[18]. In our study, conditions were defined by the variables describing stimulus valence (aversive vs. neutral odors), stimulus identity, and whether a mouse is in a trembling, ingressed or stationary state. Here we employ analyses of representational geometries to ascertain which input-output computations can be performed by a linear classifier that reads out the activity from a single BLA neural ensemble, and to determine if observed representational geometries in BLA could account for the phenomenon of emotional generalization. Only particular representational geometries can support this type of generalization[18]. Further, to illuminate mechanisms underlying these computational capabilities, we ask if eliminating neurons exhibiting apparent specialized selectivity from analyses impacts encoding of all variables[19,20,24]. If eliminating these neurons from analyses diminishes encoding, it would imply that specialized BLA circuits are required to confer computational capabilities. Our analyses demonstrate, however, that BLA neurons with apparent specialized selectivity are not commonly observed. Moreover, eliminating these neurons from analyses does not diminish any of the identified computational capabilities of BLA ensembles. As a result, the distinct emotional states elicited by an aversive CS - a fear state as assessed by tremble, and a state of safety as assessed by ingress into the burrow – can be identified by implementing different readouts of the same underlying BLA neural ensemble.

RESULTS

Aversively conditioned stimuli cause mice to tremble and ingress into the burrow

Mice underwent a Pavlovian aversive conditioning paradigm using four distinct olfactory conditioned stimuli, two that predicted footshock (CS+s), and two other stimuli that were not reinforced (CS-s) (Figure 1a,b). Following a conditioning session, the four odors were presented randomly to head-fixed mice in the virtual burrow assay; in the burrow, the CS+ was not paired with reinforcement. Linear displacement of the burrow - its forward and backward movement - was measured with a laser sensor. During most of the testing session, mice displace the burrow negligibly (less than 0.5mm). We refer to such epochs as 'stationary'. However, interspersed with stationary epochs, there are epochs when these small amplitude displacements fluctuate with a more regular periodicity, which we refer to as 'tremble', and other epochs in which mice induce a large amplitude forward displacement that draws them into the burrow, which we refer to as 'ingress' (Figure 1c). While tremble mirrors a freezing-like state (see below), ingress is a flight-like behavior into the protected environment of the burrow.[15]

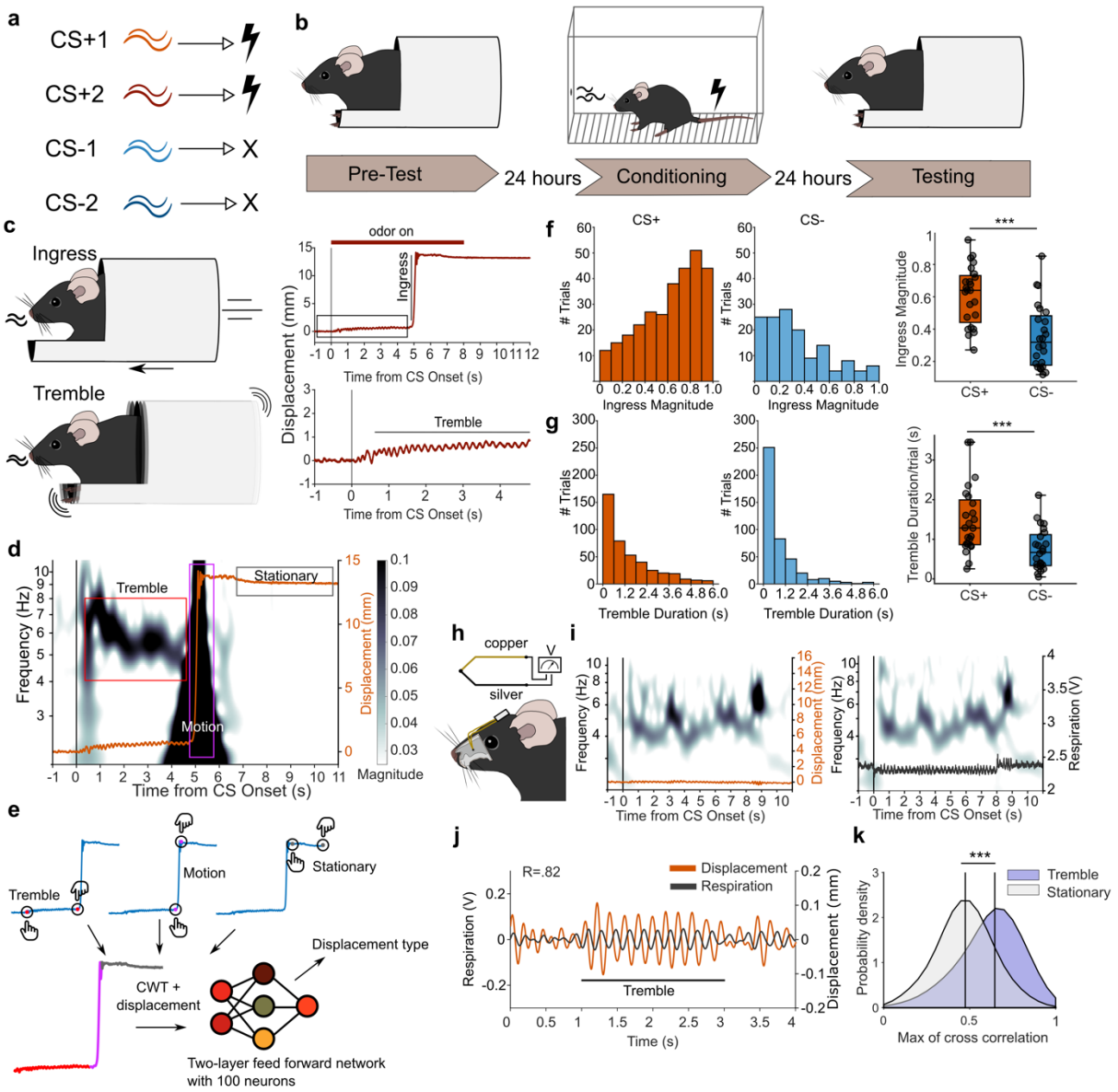


Figure 1. Aversive conditioned stimuli elicit tremble and ingress into the virtual burrow. a, Experiments employed four initially neutral odors that undergo Pavlovian aversive conditioning before testing. During conditioning, two odors co-terminate with footshock (CS+1 and CS+2), and two odors are unpaired (CS-1 and CS-2). **b**, Time course of experiment. Left, Pre-Test Session: all four odors are delivered unpaired to mice in the virtual burrow. Middle, Conditioning Session: 24 hours after Pre-Test, CS+ and CS- odors are presented in random order, with CS+ odors co-terminating with footshock. Right, Test Session: 24 hours after conditioning, the four odors are again presented to mice in the VBA and responses are measured. **c**, Mice display two behaviors in response to odor presentation in the Testing Session. Top left: A large forward displacement of the burrow enclosure constitutes Ingress. Top right, burrow displacement as a function of time relative to CS+ onset from one trial in one mouse. Rectangle captures time period depicted in zoomed-in graph below. Bottom left, A low amplitude oscillation of burrow position constitutes Tremble. Bottom right, zoomed in view of displacement as a function of time from the trial above during the tremble period. **d**, Example cwt scalogram (frequency vs. time) showing 4-8Hz tremble, motion (ingress) and stationary periods in one CS+ trial. **e**, Diagram of neural network trained to predict tremble,

motion, or stationary periods in the displacement from a subset of labeled trials. **f**, Normalized ingress magnitude for each trial in response to CS+ (left) and CS- (middle) odors. Boxplots (right) show median normalized ingress magnitude in CS+/- for each mouse (n=24 mice, ***p<0.001, paired t-test). **g**, Same as (f) for tremble duration. Boxplot (right) is median tremble duration in CS+/- for each mouse (n=24 mice, ***p<0.001, paired t-test). **h**, Schematic showing thermocouple implant in nasal cavity to measure respiration in mice. **i**, Example scalograms of displacement (orange, left) and respiration (black, right) from one trial in one mouse. **j**, Example displacement and respiration during transition from a stationary to a tremble period. **k**, Probability density distributions of the maximum correlation between displacement and respiration during tremble (purple) and stationary (light gray) periods (vertical lines depict median of distributions, tremble vs. respiration: 0.63, stationary vs. respiration: 0.4, ***p<0.001, Kolmogorov-Smirnov test).

We developed an unbiased approach for classifying the displacement data by using a wavelet transform to decompose the data into frequency components over time. Tremble periods are enriched in spectral power within the 4-8 Hz range (Figure 1d). By comparison, an ingress, which appears as a rapid, step-like shift in displacement, predictably contains spectral power within all frequencies (Figure 1d). To automatically label each frame as either tremble, ingress, or stationary displacement across experiments, we trained a neural network to classify every timepoint as one of the three categories (Figure 1e). This classification allowed us to calculate, with millisecond resolution, the magnitude and the onset and offset times of both tremble and ingress responses, extending the previous characterization of mouse behavior in the virtual burrow[15]. The presentation of CS+ odors increased the probability that a mouse would ingress (average number of ingresses: CS+, 12/20 trials; CS-, 6/20 trials; n=24 mice, p<0.001; Suppl. Figure 1a). We computed the magnitude of ingress as the maximum displacement into the burrow attained on a trial relative to the maximal possible ingress (where a magnitude of 1.0 is maximal ingress). On average, ingress magnitudes on CS+ trials were higher than on CS- trials (CS+: 0.61; CS-: 0.35; n=24 mice, p<0.001; Figure 1f). Furthermore, ingress on CS- trials was characterized by more fluctuations in displacement after ingress, and a logistic function therefore fit ingress less well on CS- than CS+ trials (CS+ r2: 0.85, CS- r2: 0.74, n=24 mice, p<0.01; Suppl. Figure 1b). Meanwhile, tremble duration, maximum frequency, and average number of trembles per trial were all increased in response to a CS+ relative to a CS- (average time trembling per trial CS+: 1.48s, CS-: 0.75s; max frequency: CS+: 5.4Hz, CS-: 5.0Hz; average number of trembles: CS+: 0.93, CS-: 0.61, n=24 mice, p<0.001; Figure 1g, Suppl. Figure 1d,e). The latency to ingress or tremble did not significantly differ between CS+ and CS- trials (ingress onset CS+: 3.5s CS-: 3.1s, tremble onset CS+: 2.3s CS-: 2.4s; n=24 mice, p > 0.05; Suppl. Figure 1c,f). Additionally, there was no significant difference between the spectral power of tremble on CS+ and CS- trials (CS+: 3.3, CS-: 2.7; n=24 mice, p> 0.05; Suppl. Figure 1g).

In behavioral paradigms in which mice are not head-fixed, freezing is accompanied by the modulation of autonomic responses such as respiration and heart rate[25–29]. To assess whether trembling, like freezing, is related to changes in breathing, we implanted a thermocouple in the mouse nasal cavity to monitor changes in respiration (Figure 1h). Breathing rate during tremble was within the 4-8Hz frequency range, overlapping the frequency range of displacement changes in tremble (Figure 1i, Suppl. Figure 2a,b). Spectrograms of breathing and tremble oscillations were remarkably phase-coupled, and tremble and respiration bouts were highly correlated

(tremble vs. respiration: r=0.63, stationary vs. respiration: r=0.4, p<0.001; Figure 1j,k, Suppl. Figure 2c). The tight correlation between breathing and tremble suggests that the displacement fluctuations during tremble may reflect inertial motions due to inhalation-exhalation cycles of sufficient strength to appreciably shift the animal's center of mass. The minimal friction of the burrow assay enables us to capture these displacements. Overall, these data are consistent with studies in freely moving mice that have demonstrated a relationship between freezing elicited by an aversively conditioned stimulus and respiration[26,27]. Trembling therefore bears similarity to freezing, which is normally thought of as the absence of movement and not a coordinated motor response. Previously, optogenetically activating BLA neurons responsive to footshock were shown to modulate breathing and heart rate, a finding consistent with autonomic reactivity caused by changes in a mouse's emotional state[30]. Indeed, optogenetically activating the same footshock-responsive neurons can induce associative fear learning[30].

According to predator imminence theory, an animal's defensive strategy shifts from freezing to flight behaviors as the animal's perception of threat escalates[31–33]. In an analogous manner, trembles largely occur prior to ingress on individual CS+ trials. When a mouse trembled on a CS+ trial, ingress followed tremble on 51% of trials (165/319 tremble CS+ trials). In contrast, only 15% of trembles on CS- trials were followed by ingress (33/215 tremble CS- trials; Figure 2a,b).

Given the observation that ingress tends to follow tremble, we wondered whether measurable aspects of a tremble could predict a subsequent ingress. We trained a decoder to distinguish between trembles that transitioned to ingress within 1 second or less of tremble offset compared to trembles that did not lead to ingress. Features related to tremble quality (the maximum frequency of the tremble and spectral power) or to the timing of trembles (tremble latency, duration, and the number of trembles in a trial) accurately predicted transition to ingress on a trial (58% accuracy using tremble features, p<0.001; 64% accuracy using timing, p<0.001, n=24 mice; Figure 2c). Trial number was not predictive of ingress (Figure 2c, 51% accuracy, p>0.5, n=24; Figure 2c). Since ingress can be predicted by specific features of a tremble bout, tremble and ingress are linked behaviors in a response sequence. Tremble, in combination with ingress thus forms a behavioral cascade analogous to freezing followed by flight to a safe environment, behaviors that can be difficult to elicit in a controlled manner in experiments without using compound stimuli or explicit training[8,34,35].

BLA inactivation eliminates differential tremble and ingress responses to aversive and neutral stimuli

We hypothesized that readouts of activity in BLA neural ensembles would be required for the presence of differential tremble and ingress responses to CS+ and CS- presentations. We therefore tested whether decreasing BLA activity alters differential tremble and ingress by using a pharmacogenetic approach to inhibit BLA activity before measuring behavioral responses in the burrow. An AAV (adeno-associated virus) harboring the inhibitory DREADD (Designer Receptor Exclusively Activated by Designer Drug) receptor hM4Di and tdTomato was bilaterally injected into BLA (Figure 2d). Prior to testing in the burrow, we separated mice into two groups and

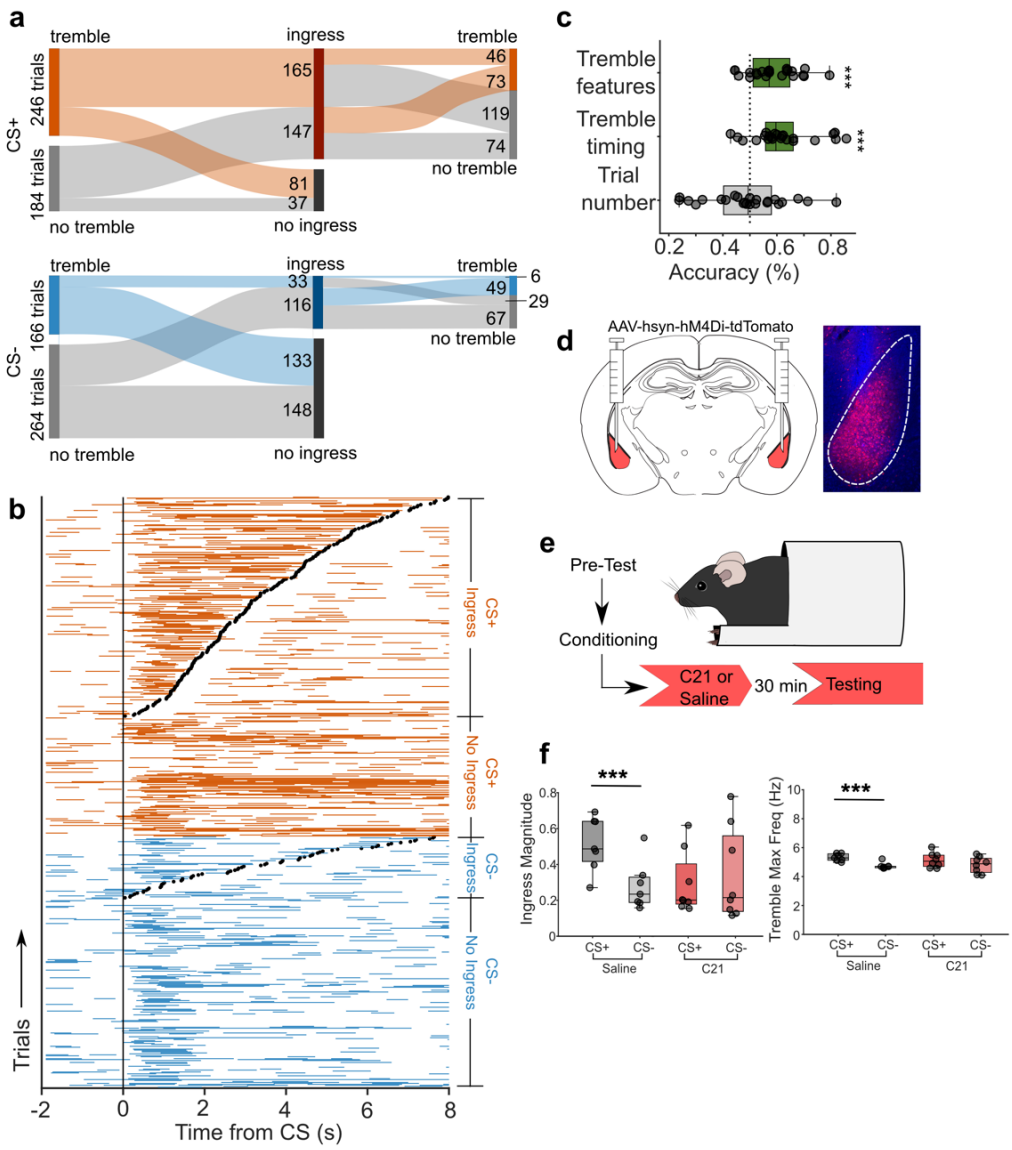


Figure 2. BLA activity is required for differential expression of tremble and ingress. a, Alluvial diagram showing the relationship between tremble and ingress for CS+/- (860 total trials, 430 trials per CS type, n=24 mice). **b**, Tremble bout(s) on ingress and no ingress trials for CS+ (orange) and CS- (blue) odors. Trials with ingress are ordered based on ingress onset time (dots) following CS onset (vertical line). **c**, Logistic regression models trained to classify trembles that transition to ingress within 1 sec of tremble offset vs. other trembles not immediately followed by ingress. Models include either tremble features (maximum frequency, spectral power), tremble timing variables (duration, number of trembles on trial, tremble latency), or trial number. Box plots depict median accuracies across mice (n=24 mice, dashed vertical line indicates 50% chance accuracy, *** p<.001, t-test from chance accuracy). **d**, Bilateral injections in BLA of an adeno-associated virus harboring the inhibitory DREADD hM4Di and the fluorophore tdTomato (AAV-hsyn-hM4Di-tdTomato). Example injection site in one mouse, white dashed line outlines BLA. **e**, Prior to testing, mice were given an i.p. injection of the DREADD agonist, Compound 21 (C21), or saline (n=8 mice C21, n=7 mice saline). **f**, Box plots showing normalized ingress magnitude and tremble maximum frequency to CS+ and CS- in Saline and C21 groups (***p<.001, paired t-test).

administered the DREADD agonist Compound 21 (C21) or saline through intraperitoneal injection

(Figure 2e). In animals injected with C21, ingress magnitude was diminished in response to the CS+, becoming similar to levels observed in response to the CS- (ingress magnitude CS+/saline: 0.51, CS-/saline: 0.28, n=7, $p<0.001$; ingress magnitude CS+/C21: 0.29, CS-/C21: 0.34, n=8, $p>0.05$; Figure 2f). Similarly, the maximum frequency of tremble during the CS+ was significantly reduced in C21 mice (max frequency CS+/saline: 5.32Hz, CS-/saline: 4.73Hz, n=7, $p<0.001$; max frequency CS+/C21: 5.14Hz, CS-/C21: 4.82Hz, n=8, $p>0.05$; Figure 2f). Thus, like freezing and avoidance responses in freely moving animals[36,37], both differential tremble and ingress response to a CS+ and CS- depend on activity in the BLA. Despite these results, it is unlikely that BLA provides a motor signal that generates tremble and ingress responses. Trembling and ingress still occur following BLA inactivation; it is differential responses to the CS+ and CS- that are eliminated.

BLA representational geometry following CS onset

The fact that mice tremble and ingress more frequently to CS+ as compared to CS- stimuli suggests that a computation of the affective significance, or valence, of stimuli impacts the propensity to generate tremble and ingress responses. Prior studies in rodents and monkeys have observed that BLA provides a representation of stimulus valence[12,14,35,38–42]. To test whether the same is true in the burrow assay, we conducted two-photon calcium imaging in a subset of mice while presenting stimuli in the burrow (Figure 3a, Suppl. Figure 3). To analyze neural activity, we first used a linear classifier to decode stimulus valence from BLA neural activity within the 2-second window following CS onset (Figure 3b). Our experimental design used four different odor CSs (CS+1, CS+2, CS-1, and CS-2; Figure 1a). A variable describing valence corresponds to a dichotomy that divides these four conditions into two equally sized groups (CS+1/CS+2 vs. CS-1/CS-2). Training and testing a classifier on this dichotomy resulted in high decoding accuracy for valence both across the BLA neuronal population, and on average across individual mice (population: 100% accuracy, $p<0.001$; average individual decoding accuracy: 74%, $p<0.05$; Figure 3c, Suppl. Figure 4a). The encoding of valence was robust, as decoding performance became asymptotic once only ~25% of randomly selected neurons are included in the analysis (Suppl. Figure 4c). However, valence is not the only information represented in BLA during this time. The four CS conditions can be divided into two other equally sized dichotomies (CS+1/CS-2 vs. CS+2/CS-1, and CS+2/CS-2 vs. CS+1/CS-1; Figure 3b). Variables describing each of these other dichotomies could also be decoded at high levels (99% and 96% accuracy, $p<0.001$; average individual decoding accuracy: 65%, $p<0.05$; Figure 3c, Suppl. Figure 4b). This analysis indicates that BLA ensemble activity differs for all four CS conditions. As a result, CS identity is represented, consistent with observations made previously in studies in monkeys[38,43,44].

Novel stimuli, and not just previously experienced ones, can elicit approach or defensive responses, even when the novel stimuli are neutral. This phenomenon is called generalization, a cardinal feature of emotions observed across species[3]. To our knowledge, no prior work has established whether the representation of stimulus valence within BLA can support this type of generalization. We used a recently described analytic approach, cross-condition generalization performance (CCGP)[18], to determine if the observed representational geometry in BLA could support this form of generalization. To compute CCGP, a decoder is trained to classify CS+ vs.

CS- using only 1 exemplar of each type of CS condition (e.g. CS+1, CS-1), and then tested on held out types of conditions (e.g. CS+2, CS-2). CCGP is the average decoding performance

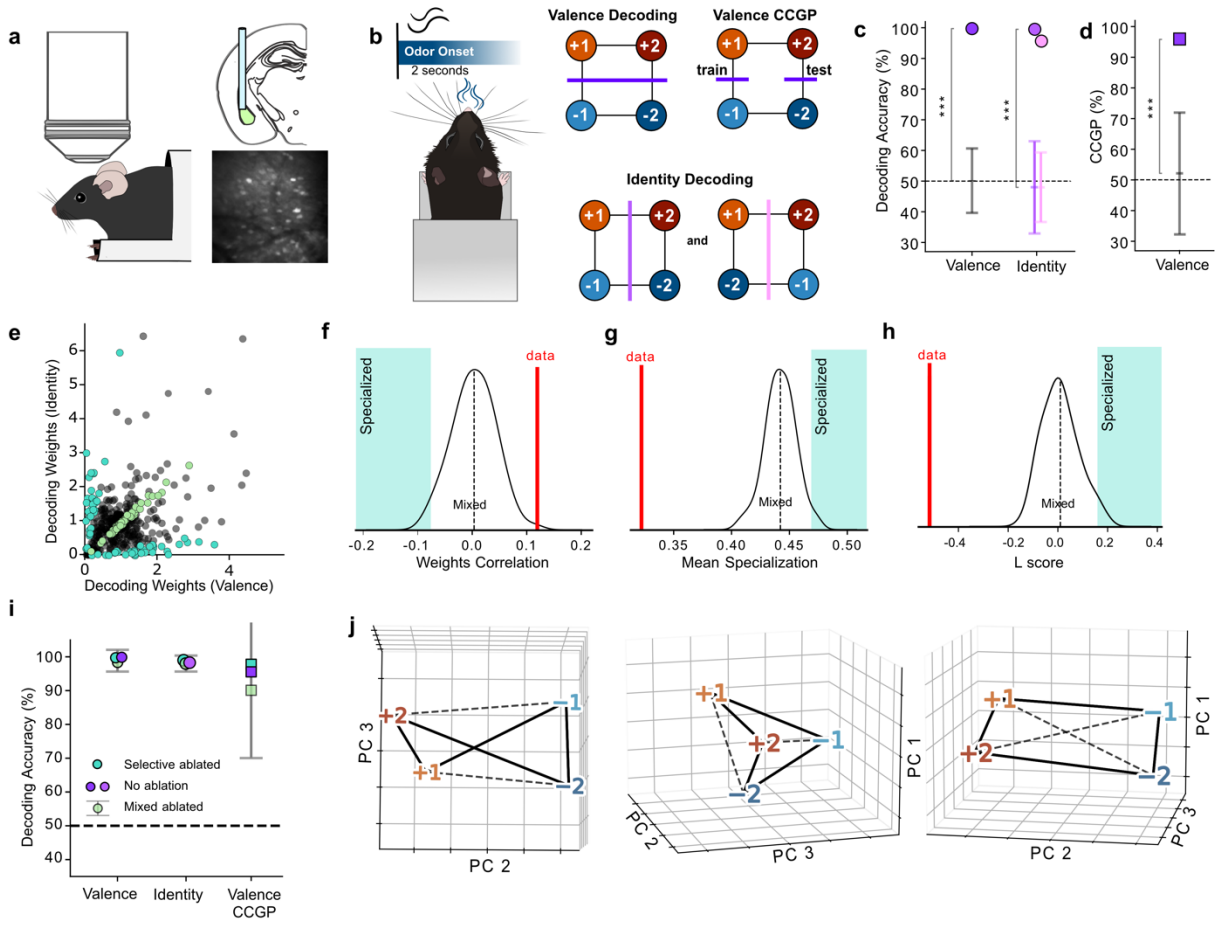


Figure 3. BLA representational geometry of CS valence and identity following CS onset. a, Schematic of two-photon imaging setup during testing in the virtual burrow (left), examples of location of GRIN lens in BLA (top right) and field of view from one mouse (bottom right). **b**, Neural activity during the 2-second window of time following CS onset was selected for decoding analysis. The four CS conditions (CS+1, CS+2, CS-1, CS-2) could be combined into two equally sized groups in three ways (three dichotomies). A variable called CS valence describes one dichotomy; the other two dichotomies together relate to a variable describing CS identity. Decoding involves implementing a linear classifier to distinguish the conditions on each side of a dichotomy; this decoding involves using randomly selected trials from all conditions for training, and testing occurs on held-out trials. Cross Condition Generalization Performance (CCGP) involves training a decoder on a subset of conditions, and testing on held-out conditions, as exemplified schematically for the variable valence (e.g. CS+1 v CS-1 used for training; CS+2 v CS-2 used for testing) **c**, Decoding performance of valence and identity (valence: 100% accuracy, identity dichotomies: 99% and 96%; ***p<0.001; error bars, 95% confidence interval of null model). **d** CCGP for valence (96% accuracy, ***p<0.001; error bars, 95% confidence interval of null model)**. e**, For each neuron, the absolute value of its decoding weight for valence is plotted against its decoding weight for decoding one of the two identity dichotomies shown in (c). Mint color, 'specialized' selectivity neurons artificially ablated in (i). Light green, mixed selectivity neurons artificially ablated in one iteration of the null model in (i). **f**, Spearman's correlation calculated between the decoding weights for valence and identity plotted derived from the data in (e) (red

line) and null model (distribution) (null: -0.00±0.05, data: 0.12, **p<0.01). **g**, Mean specialization of neurons calculated from normalized decoding weights from the data in (e) (red line) and null model (distribution) (null: 0.44±0.01, data: 0.32, ***p<0.001). **h**, L-score of decoding weights for valence and identity from the data in (e) (red line) and null model (distribution) (null: 0.01±0.07, data: -0.52, ***p<0.001). **i**, Decoding performance for valence, identity and valence CCGP for valence and one of the other dichotomies after ablating neurons in (e) with specialized selectivity (mint),with mixed selectivity (light green, one iteration of ablation), or with no ablation of neurons (purple). Grey error bars, 95% confidence interval for null model generated from iterations of ablating mixed selectivity neurons. **j**, Three views of PCA plot depicting BLA representational geometry following CS onset (+1: CS+1, +2: CS+2, -1: CS-1, -2: CS-2).

across all ways of choosing the testing and training CS conditions[18]. High CCGP is observed only with certain representational geometries, in particular, geometries in which pairs of conditions have parallel coding directions[18]. For the variable valence in our experiments, this type of geometry would be observed if the coding direction for pairs of conditions with opposing valences (CS+ vs. CS-) tend to be parallel. Indeed, in our data, during the 2-second interval after CS onset, CCGP for valence was significantly different from chance computed from the null model (96% accuracy; p<0.001 Figure 3d; see Methods). Since CCGP reports decoding accuracy on CS+ and CS- trial types that are not used for training, high decoding performance indicates that the decoder can correctly classify a stimulus as a CS+ or CS- even if the stimulus is “new” to the decoder. High CCGP indicates that the geometry of a neural representation has been sculpted by a process of abstraction[18] to create a lower dimensional scaffold exhibiting parallel coding directions for the variable valence.

Conceivably, BLA neurons that encode CS valence could be specialized, representing valence equivalently for all stimuli. If neurons have specialized selectivity for a variable, then CCGP for that variable will be high. We sought to assess the contribution of neurons with mixed as opposed to specialized selectivity for readouts of valence and CS identity (Figure 3e, Suppl. Figure 5a,b). To accomplish this, we took two approaches to quantify the selectivity of a neuron. One approach uses the population decoding weight of a neuron as a proxy for its selectivity for the variable being decoded, and the other approach uses a non-parametric method, ROC analysis, to characterize selectivity based on firing rates inferred from calcium signals of single neurons[38]. ROC analysis computes the area under the ROC curve (AUC), with larger AUC values indicating more selectivity.

Three statistical methods were employed to quantify whether the recorded population was comprised predominantly of specialized neurons (see Methods). First, we calculated the Spearman correlation between the decoding weights or between the computed AUCs. A highly negative correlation would indicate that neurons that strongly encode one variable are typically not relevant for the other variable, suggesting some degree of specialization. Second, we calculated a mean specialization index, defined as the absolute value of the normalized difference of weights or AUCs across the neuronal population, using data from both methods of assessing selectivity. Third, we developed an index called the L-index for quantifying the degree to which datapoints within a scatterplot (“selectivity plot”) lie along the axes. Of course, datapoints lying along the axes modulate activity with respect to only one variable (specialized selectivity). For all three statistical measures, we compared the observed values to null distributions created by

performing iterations of a solid random rotation of the patterns in the neural activity space, breaking the alignment of neural responses with specific neural axes and thus implementing mixed selectivity (see Methods). For each iteration, we compute a correlation coefficient, the average selectivity index for each variable, and the L-index value. If an observed correlation coefficient is significantly more negative than 95% of the values obtained from many iterations of the null model, and if a computed selectivity index or L-index is greater than 95% of values from the same random rotations, it would indicate that the population predominantly exhibited specialized selectivity. However, the Spearman correlation coefficient, selectivity index and L-index values all indicated that neurons predominantly exhibited mixed selectivity, whether using decoding weights or ROC analyses to quantify selectivity (Figure 3f-h, Suppl. Figure 5c-e).

Consistent with the predominance of mixed selectivity neurons, decoding performance and high CCGP for valence did not depend on neurons with specialized selectivity. Artificial ablation of those neurons that appeared to exhibit specialized selectivity for variables describing valence and one of the other dichotomies (i.e., eliminating from analyses neurons that were close to either axis in the selectivity plot, Figure 3e) had minimal effects on decoding performance, indistinguishable from artificial ablations of a similar number of mixed selectivity neurons (Figure 3i).

These results indicate that the representational geometry in BLA following CS onset has two main features. It possesses a low-dimensional scaffold that represents CS valence, thereby accounting for high CCGP and the propensity of emotional valence to generalize. At the same time, the representation retains sufficiently high dimensionality to allow all stimulus identities to be decodable This representational geometry can be visualized using a dimensionality reduction technique. Here the coding directions for CS valence (i.e. the lines connecting different CS+ and CS- conditions) tend to be parallel, but the geometry is not planar and all binary variables describing the four CS conditions can be read out (Figure 3j). This higher dimensionality allows for input-output computations specific for the identity of stimuli.

BLA representational geometry during tremble

We next examined neural activity when mice were in the trembling state. Neural ensemble activity was analyzed during two time epochs: trembling bouts that lasted at least 500ms and time epochs of at least 500ms in which mice were stationary (Figure 4a). Since tremble occurs more commonly on CS+ compared to CS- trials, we balanced the conditions included for decoding by subsampling the data such that decoding performance could not derive from correlations between

CS valence and tremble. The conditions used for decoding tremble therefore contained a 50-50 blend of CS+ and CS- trials for the trembling and stationary conditions (Figure 4a).

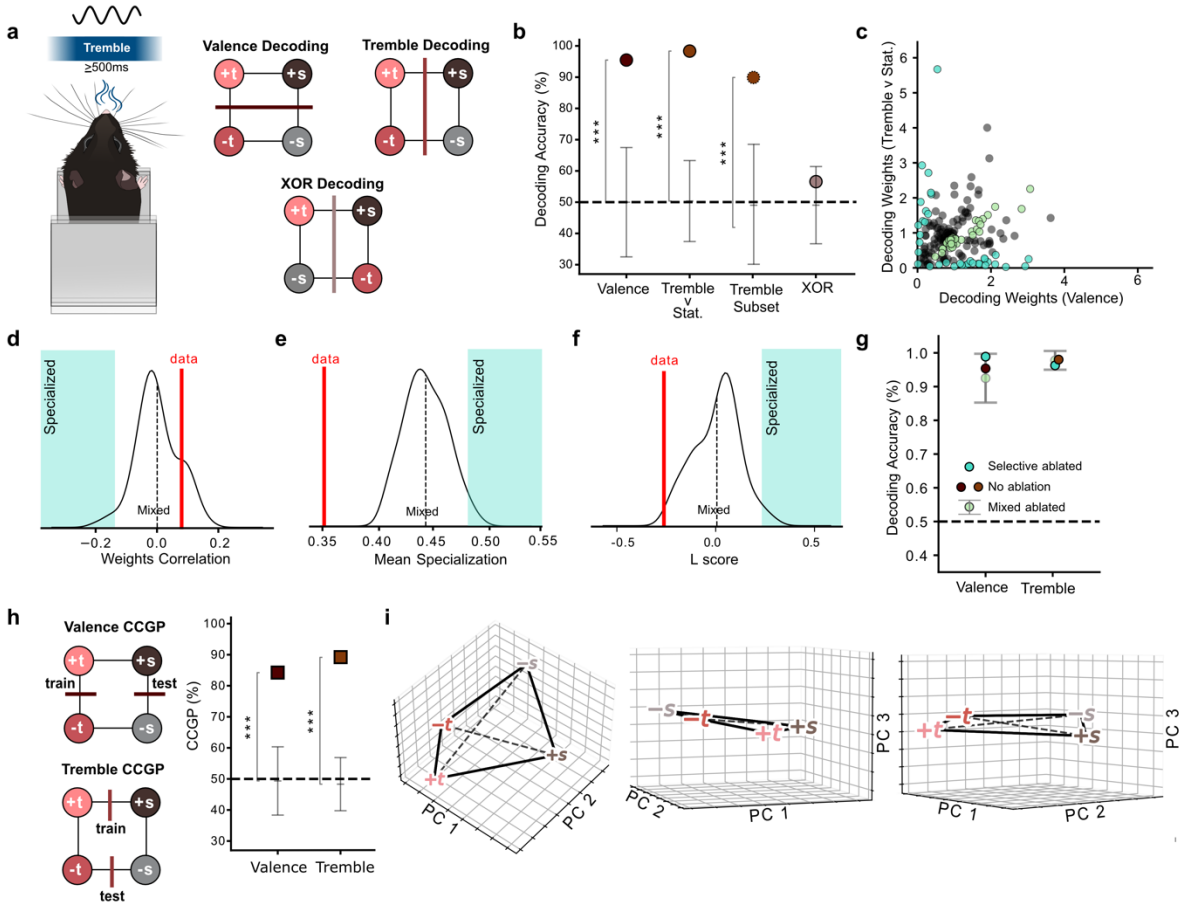


Figure 4. BLA representational geometry of valence and tremble. a, Neural activity within a window of tremble or stationary behavior (lasting at least 500ms) were selected for decoding analysis. Since odor delivery remains constant during these time periods, conditions used for decoding can be defined by whether the animal is trembling or not and whether the odor is a CS+ or CS-. These four conditions are t+, t-, s+, and s- and can be combined into three dichotomies. One of the dichotomies corresponds to valence (t+/s+ vs t-/s-). Another corresponds to tremble vs. stationary (t+/t- vs s+/s-). The third corresponds to the exclusive OR (XOR) dichotomy (t+/s- vs t-/s+). **b**, Decoding performance is plotted for valence, tremble v. stationary, tremble v. stationary when a subset of tremble epochs are selected that possess indistinguishable burrow displacement ranges to stationary periods, and the 3rd (XOR) dichotomy (valence: 95% accuracy, tremble: 98% accuracy, tremble subset: 90% accuracy, ***p<0.001 for all; XOR: 57% accuracy, p > 0.05). **c**, For each neuron, the absolute value of its decoding weight for valence is plotted against its decoding weight for tremble. Mint color, 'specialized' selectivity neurons artificially ablated in (g). Light green, mixed selectivity neurons artificially ablated in one iteration of the null model in (g). **d**, Spearman's correlation calculated between the decoding weights for valence and tremble from the data in (c) (red line) and null model (distribution) (null: 0.00±0.07, data: 0.08, p=.12). **e**, Mean specialization of neurons calculated from normalized decoding weights from the data in (c) (red line) and null model (distribution) (null: 0.44±0.02, data: 0.35, ***p<0.001). **f**, L-score of decoding weights for valence and tremble calculated from the data in (e) (red line) and null model (distribution) (null: 0.01±0.12, data: -0.26,

***p<0.01). **g**, Decoding performance for valence and tremble vs stationary after ablating neurons in (c) with specialized selectivity (mint), with mixed selectivity (light green, average performance across ablation iterations), or with no ablation of neurons (brown). Grey error bars, 95% confidence interval for null model generated from iterations of ablating mixed selectivity neurons. **h**, CCGP for valence and tremble (valence: 84% accuracy, tremble: 89% accuracy, ***p<0.001) **i**, Three views of PCA plot depicting BLA representational geometry during tremble (trial conditions labeled according to tremble/stationary and valence, t: tremble, s: stationary, +: CS+, -: CS-).

Our analysis revealed that both tremble and valence were robustly decodable across the neuronal population and across individual mice (valence 95% accuracy, tremble 98% accuracy, p<0.001; Figure 4b, Suppl. Figure 6a). Decoding performance was asymptotic when ~30% of neurons were included (Suppl. Figure 6b). Decoding of the third dichotomy (CS+/stationary and CS-/tremble vs. CS-/stationary and CS+/tremble), which corresponds to an exclusive OR (XOR) computation, was not different from chance (57% accuracy, p>0.05; Figure 4b). The ability to decode tremble was not due to a correlation between tremble and ingress. If we restricted the decoding analysis of tremble vs. stationary to only trials where mice later ingressed, decoding performance remained high (100% accuracy, p<0.001). Thus, even though ingress and tremble are correlated, and even though a tremble often leads to an ingress, a neural representation of the trembling state exists independent of ingress behavior.

The ability to decode tremble states from stationary states also could not be accounted for by differential displacement movements of the burrow. To demonstrate this, we selected tremble and stationary bouts where the displacement range of the burrow was less than 0.5mm. The distributions of displacement ranges for these selected stationary and tremble bouts were similar. Decoding accuracy of this subset of trembles vs. stationary states remained high despite the similarity in the movements of the burrow during the two states (90% accuracy, p<0.001, Figure 4b).

We considered the possibility that different readouts of the same BLA ensemble could be used to decode both valence and tremble. To assess this possibility, we again analyzed the selectivity of neurons using both of our methods: decoding weights and AUCs computed using ROC analyses. Many neurons exhibited high weights for decoding of both tremble and valence, as well as high AUC values for both variables (Figure 4c, Suppl. Figure 7a). The majority of the neurons did not fall along the axes but instead fell at intermediate locations between the axes on the selectivity plots (Figure 4c, Suppl. Figure 7a,b). Using both decoding weights and the ROC AUC methods for quantifying selectivity, all of our measures of specialization (Spearman correlation, mean specialization index, and the L-index) failed to reject the hypothesis that the neural population exhibited mixed selectivity for valence and tremble (Figure 4d-f; Suppl. Figure 7c-e). Further, artificial ablation of neurons whose decoding weights suggested coding specialization for either valence or tremble did not diminish decoding performance (Figure 4g). The effects of excluding these neurons from decoding also did not differ from artificial ablations of equivalent numbers of mixed selectivity neurons (Figure 4g).

We then analyzed the representational geometry to quantify to what extent valence and tremble are represented in a manner that generalizes across conditions. CCGP for valence across

the trembling and stationary states was high (84% accuracy, $p<0.001$; Figure 4h). In addition, we analyzed neural activity only during trembles, separating data into conditions, one for each CS+ and CS-, and found that CCGP for valence was high (71% accuracy, $p<0.05$). Thus, during trembles, which often occur not long after CS onset, valence is still represented in a manner that generalizes across different stimuli, just like immediately after CS onset (Figure 3). Of note, the representation of tremble also has high CCGP, indicating a similar readout of the trembling state for both valences of stimuli (89% accuracy, $p<0.001$; Figure 4h). The high CCGP for valence and tremble, combined with a low XOR decoding performance[23], suggest that valence and tremble are represented in a low-dimensional planar geometry in the neural activity space, which can be visualized using dimensionality reduction methods (Figure 4i). Here the variables for valence and tremble allow one to visualize parallel coding directions across both behavioral states (for valence) and across both valences of stimuli (for tremble). Thus, both valence and tremble variables appear to be represented with a lower dimensional scaffold.

BLA representational geometry during ingress

We next examined neural representations in BLA during the ingressed state, after mice have entered the burrow enclosure. We compared stationary periods (as used for the analysis of tremble) to the 2-second time epoch after mice had ingressed into the burrow (a time epoch that follows the ingress response itself; Figure 5a). Again, we balanced the data for valence by including an equal number of CS+ and CS- trials, eliminating the possibility that CS valence coding could account for ingress decoding performance. Decoding accuracy of valence and the ingressed state were both significantly greater than 50% across the recorded population and across individual mice (valence 72% accuracy, ingress 89% accuracy, $p<0.001$; Figure 5b, Suppl. Figure 8a).

The ability to decode the ingressed state from the stationary state did not derive from differential displacements of the burrow in the two states. When in the ingressed position, mice remain inside the burrow but still sometimes displace the burrow more than that observed when they are in the stationary state. To ensure that these somewhat larger displacements while inside the burrow did not account for decoding performance, we selected a 600ms time epoch in which mice remained in the ingressed position but did not displace the burrow more than 0.5mm (a displacement range similar to stationary epochs). Decoding accuracy of BLA activity using this subset of ingress epochs compared to stationary epochs remained high even when burrow displacements were minimal (94% accuracy, $p<0.001$; Figure 5b). The ability to decode whether

mice are in the ingressed state compared to states when mice were stationary therefore does not rely on movements that cause different magnitudes of displacements of the burrow.

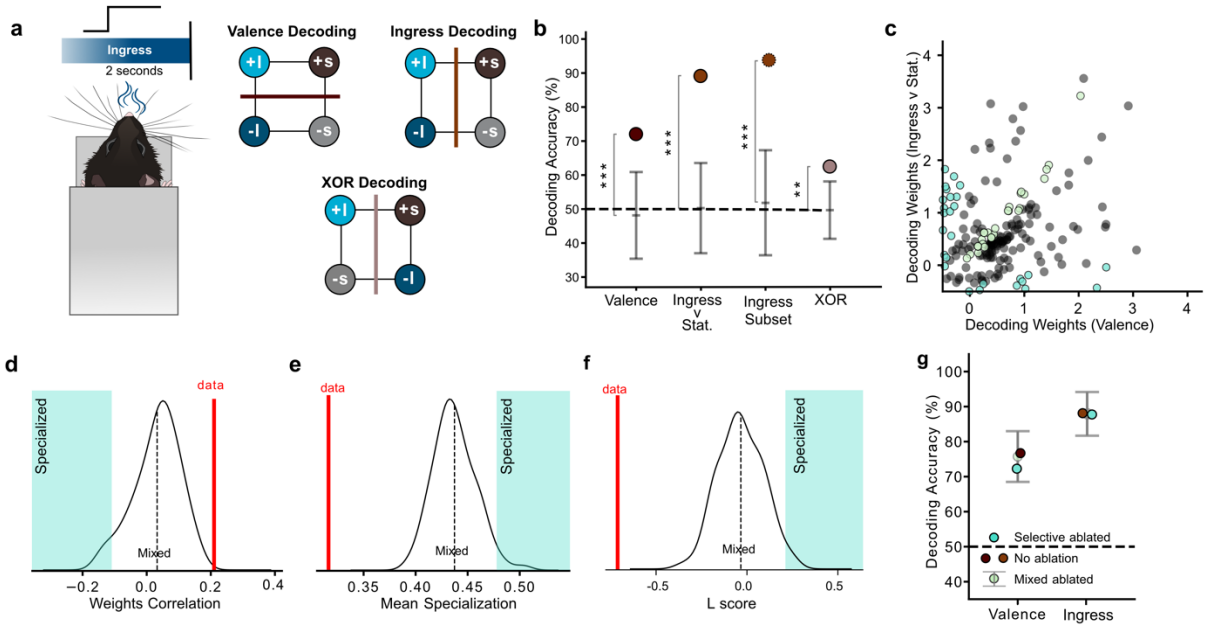


Figure 5. BLA representational geometry of valence and ingress. a, Neural activity within the 2-second time window after mice reach the ingressed position or when they are stationary were selected for decoding analysis. Since odor delivery remains constant during these time periods, conditions used for decoding can be defined by whether the animal is ingressed or not and whether the odor is a CS+ or CS-. These four conditions are I+, I-, s+, and s-, and they can be combined into three dichotomies. One of the dichotomies corresponds to valence (I+/s+ vs I-/s-). Another corresponds to ingressed vs. stationary (I+/I- vs s+/s-). The third corresponds to the exclusive OR (XOR) dichotomy (I+/s- vs I-/s+). **b**, Decoding performance is plotted for valence, ingress v. stationary, ingress v. stationary when a subset of ingress time epochs are selected possessing displacement ranges similar to stationary periods, and the XOR dichotomy (valence: 72% accuracy, ingress: 89% accuracy, ingress subset: 94% accuracy, XOR: 62% accuracy, ***p<0.001). **c**, For each neuron, the absolute value of its decoding weight for valence is plotted against its decoding weight for ingress. Mint color, 'specialized' selectivity neurons artificially ablated in (g). Light green, mixed selectivity neurons artificially ablated in one iteration of the null model in (g). **d**, Spearman's correlation calculated between the decoding weights for valence and ingress from the data in (c) (red line) and null model (distribution) (null: 0.02±0.07, data: 0.25, ***p<0.001). **e**, Mean specialization of neurons calculated from normalized decoding weights from the data in (c) (red line) and null model (distribution) (null: 0.44±0.02, data: 0.31, ***p<0.001). **f**, L-score of decoding weights for valence and ingress calculated from the data in (c) (red line) and null model (distribution) (null: -0.03±0.12, data: -0.71, ***p<0.001).**g**, Decoding performance for valence and ingress vs. stationary after ablating neurons in (c) with specialized selectivity (mint), with mixed selectivity (light green, average performance across ablation iterations), or with no ablation of neurons (brown). Grey error bars, 95% confidence interval for null model generated from iterations of ablating mixed selectivity neurons.

We next analyzed the selectivity properties of neurons and found that, similar to what we observed for tremble and valence, the recorded population exhibited mixed selectivity for the

ingressed vs. stationary states and for valence. Many neurons exhibited high weights for decoding of both ingress and valence, as well as high AUC values for both variables, with the majority of neurons falling at intermediate locations between the axes on scatterplots of these indices of selectivity (Figure 5c; Suppl. Figure 9a,b). The Spearman correlation, the mean specialization index, and the L-index, each computed using both decoding weights and AUCs to quantify selectivity, enabled us to reject the hypothesis that neurons exhibited significantly more specialized selectivity for valence and ingress than expected by chance as inferred from the null model that implements mixed selectivity ($p<0.001$; Figure 5d-f; Suppl. Figure 9c-e). Further, artificial ablation of neurons with decoding weights suggesting coding specialization of valence and ingress did not affect decoding performance significantly, nor did artificial ablations of equivalent numbers of mixed selectivity neurons (Figure 5g). Thus, the distributed nature of the neural code for valence and the ingressed state makes it so that activity from the *same* ensemble of BLA neurons can be used to discriminate the two variables by different downstream readouts.

A single BLA neural ensemble can encode both the tremble and ingressed states

In principle, the ability to decode the trembling and ingressed states from the stationary state could each be mediated by distinct BLA sub-populations, one specialized for tremble and the other for ingress. In this case, neurons would exhibit specialized selectivity for one or the other response. To assess the contribution of neurons with mixed as opposed to specialized selectivity for readouts of tremble and ingress, we examined the decoding weights and AUC of each neuron for each of the two variables (Figure 6a). This analysis could only be applied to those neurons included in both the analysis of tremble vs. valence, and the analysis of ingress vs. valence (see Methods). Of course, trembles and ingress did not occur at the same time, so this entailed examining decoding weights from the same neurons at different timepoints. To eliminate the possibility that activity related to a tremble extended beyond the tremble itself, thus affecting the decoding of both variables, we restricted this analysis such that any trial could only contribute data for the decoding of ingress or tremble but not both. This restriction was achieved by randomly assigning all trials containing both trembles and ingress to the decoding of only one of the two responses.

We observed that many neurons exhibited a high degree of response to both variables, as shown on scatterplots of decoding weights and AUCs for both variables (Figure 6b; Suppl. Figure 10a,b). Our measures of specialization again indicated that neurons predominantly exhibited mixed selectivity using both decoding weights and ROC analysis ($p<0.001$; Figure 6d-f; Suppl. Figure 10c-e). Consistent with this observation, obtaining high decoding performance for tremble or ingress did not rely on the presence of neurons with specialized selectivity, as artificial ablation of neurons with specialized selectivity compared to randomly selected mixed selectivity neurons had equivalent and minimal effects on decoding performance for both tremble and ingress (Figure 6c).

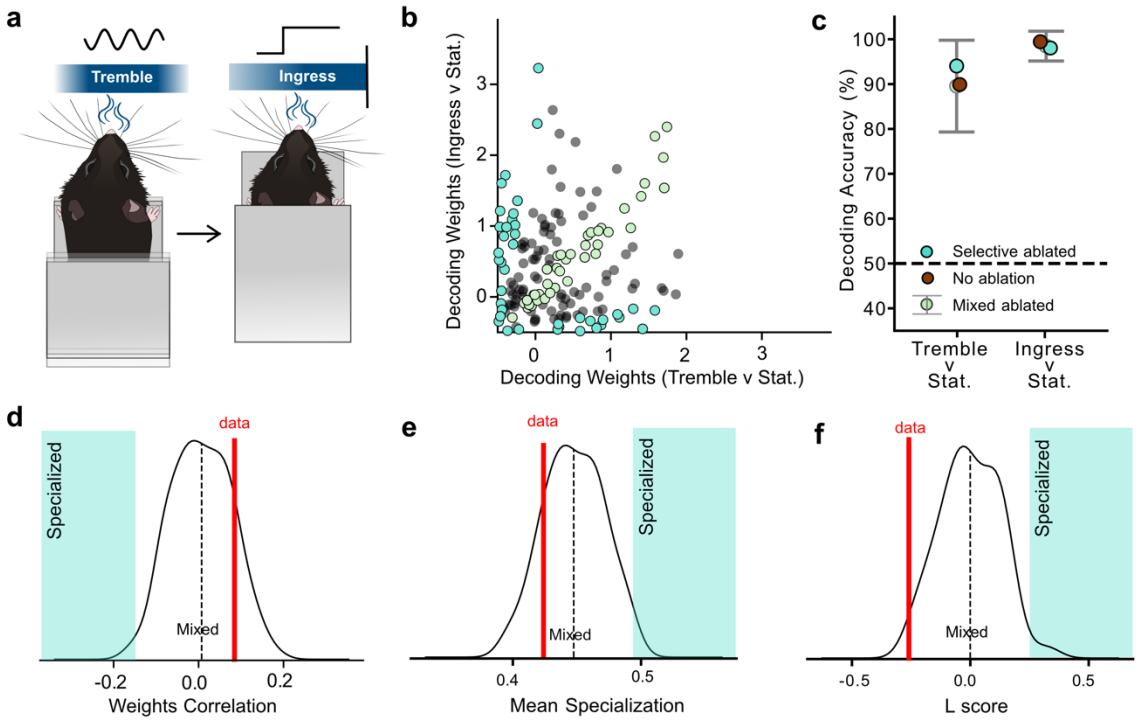


Figure 6. BLA neurons exhibit mixed selectivity for tremble and ingress. a, Tremble and ingress epochs were selected from different trials. **b**, For each neuron, the absolute value of its decoding weight for tremble is plotted against its decoding weight for ingress. Mint color, 'specialized' selectivity neurons artificially ablated in (c). Light green, mixed selectivity neurons artificially ablated in one iteration of the null model in (c). **c**, Decoding performance for tremble and ingress (vs stationary) after ablating neurons in (c) with specialized selectivity (mint), with mixed selectivity (light green, average performance across iterations of the null model), or with no ablation of neurons (brown). Grey error bars, 95% confidence interval for null model generated from iterations of ablating mixed selectivity neurons. **d**, Spearman's correlation calculated between the decoding weights for each neuron for ingress and tremble from the data in (b) (red line) and null model (distribution) (null: 0.01±0.07, data: 0.08, p=0.13). **e**, Mean specialization of neurons calculated from normalized decoding weights from the data in (b) (red line) and null model (distribution) (null: 0.45±0.02, data: 0.42, p=.15). **f**, L-score of decoding weights for tremble and ingress calculated from the data in (b) (red line) and null model (distribution) (null: -0.00±0.13, data: -0.26, *p<0.05).

Given that the same neural ensemble, which we found is composed of neurons exhibiting mixed selectivity, could be read out to decode both tremble and ingress, we wondered whether the brain states corresponding to the tremble and ingressed states were distinct. We found that we could decode the tremble compared to the ingressed states from each other using BLA ensemble activity (81% accuracy, p<0.01). Consistent with this finding, a linear classifier trained to decode the tremble vs. stationary states did not generalize to decode the ingressed vs. stationary states (52% accuracy, p>0.05). Thus, BLA exhibits distinct representational geometries during the stationary, tremble, and ingressed states.

The representational geometry of safety

Ingress is a behavioral state that can be conceptualized as the result of a flight to safety. Consequently, once in the ingressed position, the valence of a CS may no longer be behaviorally relevant. Consistent with this possibility, when we analyzed neural activity only when mice were in the ingressed state (CS+/ingressed vs. CS-/ingressed), decoding performance for valence was not different from chance (43% accuracy, p>0.05; Figure 7a). Since ingress typically occurs well within the 8-second presentation of odors, the inability to decode valence when mice are ingressed likely reflects the fact that the mouse has now entered a 'safe' emotional state, and the valence of stimuli are no longer important to track even though the stimuli are still present. We therefore considered the possibility that the BLA provides a representation of a variable describing safety during the ingressed state, and not CS valence. Since the concept of safety may have been learned from prior experience, we also hypothesized that prior experience invoked an abstraction process that, in turn, sculpted the representational geometry to allow for an abstract, generalized coding of the variable safety.

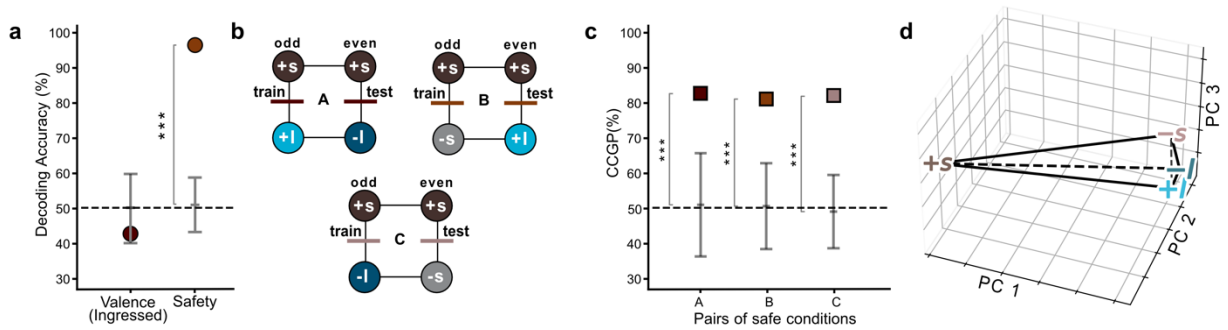


Figure 7. BLA representational geometry of safety. a, Decoding accuracy of valence while mice are in the ingressed position, and decoding accuracy of the S+ condition vs. the other three conditions (I+, I-, S-), labeled as the safety variable (valence: 43%; safety: 96%, ***p<0.001). **b**, Adapted CCGP analyses to test whether a decoder can generalize across pairs of "safe" and "unsafe" trial conditions. CS+/stationary (unsafe) trials (+s) were divided into odd and even trials. A, B, C: three different dichotomies comprised of safe (+I, -I, -s) vs. unsafe (+s) trial conditions, having divided +s conditions into two. CCGP is performed by training a decoder to distinguish between one set of +s conditions and a safe condition (e.g. +s odd trials v. +I) and then testing on the two held out conditions (in this case, +s even trials v. -I). **c**, CCGP for the A, B, and C dichotomies (decoding accuracy: 0.83, 0.81, 0.82, respectively; ***p< 0.001). **d**, PCA plot depicting BLA representational geometry during ingress (trial conditions labeled according to ingress/stationary and valence, I: ingress, s: stationary, +: CS+, -: CS-).

Within our experiment, we posited that those conditions where the animal is either presented with a neutral stimulus (CS-) or in an ingressed state could be characterized as "safe". This combination is found in three out of four conditions (CS-/stationary, CS-/ingressed, CS+/ingressed). By contrast, the remaining condition, where the animal is presented with a shock-associated stimulus and is not ingressed in the burrow (CS+/stationary) could be perceived as "unsafe" by the animal. To test the hypothesis that BLA represents the variable safety that describes these conditions, we used population activity to decode CS+/stationary vs. the other three conditions combined. Decoding accuracy was high (96%, p<0.001; Figure 7a), indicative of

strong coding of safety by BLA activity. Next, to determine the generalizability of the variable safety, we analyzed the parallelism between the three coding directions for CS+/stationary against each of the three safe conditions by an adapted version of our CCGP analysis. In this analysis, we divided the CS+/stationary condition into two groups such that even and odd trials are considered separate conditions (Figure 7b). CCGP for safety was high in all cases (mean accuracy across the three pairs of safe conditions: 82%, $p < 0.001$; Figure 7c), indicating that the coding direction between the Stationary/CS+ condition and all of the three other conditions tended to be parallel. Thus, a classifier that separates the S+ condition from one of the safe conditions can successfully separate S+ from each of the remaining 'safe' conditions. We then visualized the representational geometry of the four conditions using dimensionality reduction of neural ensemble activity. In accordance with the CCGP finding, we observed a geometry that was strikingly elongated along the safe-unsafe axis (Figure 7d). These data suggest that a process of abstraction has sculpted the BLA representational geometry so that the variable safety generalizes across different types of safe conditions.

DISCUSSION

In this paper, we used chemogenetic inactivation and two-photon imaging of BLA to elucidate mechanisms underlying how an aversive conditioned stimulus elicits multiple responses that reflect evolving emotional states. A virtual burrow was employed to assay emotional states. Two response measures —tremble and ingress into the burrow— were used to infer distinct states. Trembles are small periodic movements correlated with respiration, similar to the rhythmic respiration observed when mice freeze during aversive conditioning[26,27]. Ingress can be viewed as a flight response, whereupon mice then reach a position of safety inside the burrow. Thus, presentation of an aversive conditioned stimulus in the virtual burrow elicits a response traditionally used to infer a fearful state (trembling, or freezing), followed by a second response (ingress, or fleeing) that allows mice to attain a more protected and thereby safer state, the ingressed position. Tremble and ingress occur more frequently in response to presentation of an aversive CS+ than to a neutral CS-. Inactivation of BLA eliminated differential tremble and ingress responses to these stimuli. Thus, readouts of BLA neural ensemble activity are required for the differential expression of tremble and ingress responses to different stimuli with distinct affective significance. However, BLA activity is not required for the movement generation characteristic of either tremble or ingress, as both ingress and tremble still occur upon BLA inactivation. Just as inactivation of a brain area that provides a sensory representation and lacks motor signals can eliminate differential responses to two sensory stimuli, inactivation of BLA can eliminate differential responses to two stimuli, a CS+ and a CS-, that normally elicit distinct emotional states and their associated responses.

Although the inactivation experiments establish that BLA mediates differential tremble and ingress responses depending upon the affective significance of a stimulus, it does not determine if these effects are mediated by distinct BLA circuits for each of the two responses. Two-photon imaging data was thereby used to assess the degree of specialization within BLA for encoding variables that describe the distinct behavioral states defined by tremble and ingress responses. BLA neural ensemble activity was used to train and test multiple linear classifiers that robustly

decoded whether mice were in the trembling or ingressed state compared to time epochs when they were stationary. In addition, after stimulus onset, BLA ensemble activity encoded the valence of stimulus, and valence encoding persisted during trembling and stationary states, but not during the ingressed state. Analyses of neuronal selectivity demonstrated that BLA neural ensembles are predominantly comprised of neurons exhibiting mixed selectivity for two or more of these variables during all the time epochs we examined. Eliminating neurons (i.e. artificially ablating neurons) that appear to have specialized selectivity from every decoding analysis did not diminish decoding performance for stimulus valence, identity, or for whether mice were in the trembling or ingressed state.

The prevalence of mixed selectivity neurons in BLA indicate that during the different emotional states elicited by aversive stimuli in the virtual burrow, the same BLA neural ensemble is engaged, albeit with different patterns of activity across neurons and conditions. These patterns of activity define the representational geometry of BLA ensembles. Our data show that the representational geometry in BLA ensemble evolves as mice experience CS onset, and then tremble, ingressed states, an evolution that one can visualize using dimensionality reduction techniques. The observed representational geometries were sufficiently high dimensional to allow the implementation of different readouts of ensemble activity to decode the variables that describe responses associated with a mouse's evolving emotional state. CS valence could also be decoded after CS onset and during trembles, but not in the ingressed state. This observation may reflect the fact that mice may feel safe in the ingressed position, just like they may feel safe upon perceiving a CS-. In support of this notion, we observed that a variable describing safety was represented with high decoding accuracy, and it exhibited a geometry that had high CCGP across different pairs of safe vs. unsafe experimental conditions. Our data thereby reveal that an analysis of BLA representational geometry identifies different patterns of activity across ensembles that define distinct emotional states in the VBA.

The prevalence of mixed selectivity neurons and its implications for BLA circuit specialization

Several studies have emphasized the complexity of the physiological response properties of BLA neurons[24,44–46]. BLA neurons often can respond to different extents to stimuli of both valences[38,47,48]. However, recent work has suggested that BLA contains distinct appetitive and aversive neural circuits, as anatomically segregated and genetically defined populations of BLA neurons appeared to be specialized for each valence (Rspo2-expressing pyramidal neurons and another population expressing Ppp1r1b, Protein phosphatase 1 regulatory inhibitor subunit 1B)[14]. However, subsequent work has shown that Rspo2-expressing neurons are not homogenous, and subpopulations of these neurons have different valence preferences depending upon their projection target[49].

Prior studies have used projection-specific perturbations of BLA output pathways to demonstrate their involvement in valence-specific emotional behavior[12,50,51]. However, these studies often have not characterized neural responses properties across a broad range of parameters, including different valence-specific responses, to assess mixed selectivity. Moreover,

individual BLA neurons have been shown to project to multiple anatomical targets[13,47]. This raises the possibility that the same neurons could contribute to distinct output computations at different target structures. Consequently, while delineating the valence specificity of neural coding in specific populations of neurons is important, the data presented here emphasizes the importance of also examining neural coding in the context of multiple behavioral responses used to infer emotional states. Thus, even if genetically and/or anatomically defined subsets of neurons possess valence-specific coding, they still might exhibit mixed selectivity for other variables, as we have demonstrated to be the case for variables describing the tremble and ingressed states (and for CS identity).

The prevalence of mixed selectivity neurons in BLA, and the computational capabilities therein conferred, have implications for understanding the degree of neural circuit specialization within BLA. Distinct circuits must be engaged at some stage of processing within the brain to generate specific outputs in relation to an emotional state. Aversive stimuli elicit two responses in the virtual burrow, tremble and ingress, each of which would therefore require distinct output circuitry. Our data shows that BLA neurons predominantly exhibit mixed selectivity for tremble and ingress, implying that circuit specialization for these responses may lie downstream from BLA. Genetically defined, distinct circuit elements have been identified in CeA that mediate freezing and fleeing in freely moving animals[8], responses similar to tremble and ingress. Physiological and optogenetic evidence showed that activity in somatostatin-expressing (SOM+) neurons is related to the generation of freezing, while activity in corticotropin-release-factor–expressing (CRF+) neurons mediates flight behaviors. While BLA input to SOM+ and CRF+ neurons in CeA has been demonstrated[8,52], it remains unknown if individual BLA neurons project to both of these distinct neural populations. However, regardless of whether such projections exist, polysynaptic pathways as well as complex processing within BLA microcircuitry likely also occurs that might provide a means for individual BLA neurons to impact multiple output neural populations. For example, activating BLA neurons that exhibit valence preference physiologically and that project to a particular anatomical target can inhibit some nearby neurons and excite others[13]. These locally activated and inactivated neurons may have a variety of response properties. Advancement in multiphoton techniques that allow the simultaneous perturbation and recording of activity-defined ensembles in the BLA can help shed light on mechanisms at the microcircuit level[53].

If the tremble and ingress responses both rely on readouts of BLA neural ensembles comprised of mixed selectivity neurons, these could be coordinated through dynamic synaptic changes within the two previously identified CeA populations for freezing and fleeing[10], or other analogous distinct cell populations mediating output responses. Mechanisms involving dynamic synaptic changes could help coordinate different readouts if BLA directly projects to a target cell population, or if BLA connects to output neuronal populations through polysynaptic pathways. As an example of polysynaptic pathways, there are multiple indirect routes through which a BLA neuron could influence activity in CeA neurons, such as through projections to the intercalated masses, to the insula, and numerous other structures that possess direct connections to CeA[11,54]. Understanding the mechanisms that implement switches between different readouts of neural

ensemble activity is critical for elucidating how neurons with mixed selectivity play a role in generating multiple outputs in response to stimuli that elicit evolving emotional states.

Representational geometry and the generalization of emotional states

Generalization of emotional states to new situations has long been appreciated as an important phenomenon implicated in psychopathology[55–59], but prior studies have typically focused on the role of the amygdala at the cellular/molecular level in stimulus generalization, not generalization to novel stimuli[60,61]. The capacity to generalize in novel situations relies on identifying features shared by different situations. If the brain represents a feature shared across instances, then the representation of that feature can be read out to guide rapid responses in new situations, discarding much of the detailed information describing each instance. For this reason, the process of identifying shared features across instances - a process called 'abstraction' - can provide a more compact and typically dimensionally reduced representational geometry[62,63]. However, if representations discard all details of the environment, fewer variables are represented, which in turn limits the number of input-output operations that can be performed (limiting the number of responses that can be generated by reading out activity)[18–20]. In other words, as an example, responses specific to stimulus identity cannot be generated if all information about stimulus identity is discarded and only valence information saved.

Our data showed that the representational geometry encoding the variable valence both after CS onset, and during the trembling state, allows for high CCGP for valence. Because high CCGP indicates that the coding direction for valence generalizes across different stimuli, this representational geometry suggests a potential mechanism for how novel odors can elicit emotional generalization. When an ensemble of BLA neurons represents a novel neutral stimulus, the pattern of activity will be located somewhere along the coding direction for valence that lies within a lower-dimensional scaffold. Depending upon the location with respect to the valence coding direction, the novel stimulus will therefore be decoded as either aversive or appetitive even though it has no prior emotional meaning. In principle, a variety of factors could modulate where a novel neutral stimulus is represented with respect to the valence coding direction in BLA. These factors include an individual's pre-existing affective state, the similarity of a stimulus to other previously experienced stimuli possessing emotional significance, and the environmental situation. Regardless of the factors that can modulate how a novel stimulus is represented in BLA, we propose that the assignment of valence in BLA to a novel stimulus could underlie what is referred to as emotional generalization[3,55–59]. A mechanism involving the readout of valence for a novel stimulus in BLA would thereby produce emotional states and their associated valence-specific behavioral and physiological responses.

Our data also reveal that input-output computations specific for CS identity can still be performed given the observed representational geometry in BLA. Although the fact that CCGP for valence is high implies that valence is represented within a lower dimensional 'scaffold', this scaffold must be embedded within a higher dimensional geometry that still enables additional input-output computations beyond reading out valence. This type of compromise, between being

able to perform many input-output computations and representing one or more variables in a manner that can support generalization, has been observed previously in monkeys[18].

Emotional states through the lens of representational geometries

Historically, investigations of the neurobiology of emotion have strived to characterize the brain areas and circuit mechanisms that mediate the specific behaviors used to ascribe emotional states to individuals, such as using freezing to ascribe fear. Here we sought to provide insight into the neural mechanisms that mediate how emotional states, and the responses that reflect those states, evolve. Identifying the neural underpinnings of evolving emotional states may seem like an inherently elusive endeavor. Our results suggest that one can understand BLA ensemble activity in terms of evolving representational geometries that constitute different brain states. These brain states map onto behavioral states that can be elicited by aversively conditioned stimuli. Our data show that the BLA neural ensemble activity that instantiates these brain states is largely composed of mixed selectivity neurons. These neurons confer computational capabilities that account for the observed repertoires of responses (tremble and ingress) associated with emotional states in the burrow, as well as the propensity of emotions to generalize.

Our findings form the basis of a conceptual framework whereby mixed selectivity neurons in BLA provide a high-dimensional representational geometry, one where the value of many variables that describe emotional states can be read out. The representational geometry with respect to some of these variables, such as valence, can be sculpted through abstraction so that a lower dimensional scaffold enables generalization. Critically, this representational geometry still allows for the value of other variables to be read out. Thus, different readouts of an underlying representation, rather than engagement of distinct and non-overlapping circuits within BLA, can account for how BLA representations are used to compute the numerous output responses characteristic of emotional behavior. This framework is appealing because it implies that the BLA does not need to encode redundant information in many distinct circuits to generate many different outputs indicative of emotional states. However, it also faces an enduring challenge, as we still must understand how the brain implements a switch between different readouts to express the full range of responses associated with evolving emotional states.

Acknowledgements
We thank Richard Axel for his generosity and numerous contributions. PK.O. was supported by NIMH T32MH015144, BBRF Young Investigator Award, and NIMH K01MH123783. C.D.S. and S.F. were supported by R01-MH082017 and the Simons Foundation.

METHODS

Mice
Male C57BL/6J mice (Jackson Laboratories) were singly housed on a 12-h light-dark cycle with ad libitum access to food and water. All procedures were performed in accordance with standard ethical guidelines approved by Columbia University Institutional Animal Care and Use Committee. 24 mice were used for initial experiments characterizing behaviors. A subset of 9 mice underwent

two photon imaging. A cohort of 15 mice were used for pharmacogenetic experiments. All experiments were performed during the light cycle on three consecutive days. Surgical procedures were performed on mice at 8-10 weeks of age. Mice included in all behavioral, inactivation and imaging experiments were between 12-16 weeks of age.

Virus injection surgical procedure

Animals were anesthetized with isoflurane and secured in a stereotactic apparatus (Stoelting). Internal body temperature was regulated through a heating pad placed under the body. Ophthalmic ointment was applied to the eyes to prevent corneal drying (Patterson Veterinary) and Buprenoprhine (0.05mg/kg) was administered as an analgesic. The surgical site was clipped and cleaned with iodine followed by ethanol washes. Lidocaine was applied topically to the scalp to numb the area of incision. The skull was exposed and cleaned with sterile cotton swabs. One machine screw was fixed to the skull in order to anchor the dental cement. For behavior only experiments, the skull surface was then immediately covered in a thin layer of cement (MetaBond, Parkell Inc). A titanium head plate was lowered onto the skull, centered at lambda and secured with additional applications of cement.

For DREADD inactivation and calcium imaging experiments, following attachment of the machine screw, a craniotomy was performed using a microdrill to access the brain for viral injection. All injections were performed using a microinjection pipette, Nanoject II (Drummond 6563A01). For inactivation experiments, we made bilateral injections of 200-400 nl of hSyn-hM4Di-tdTomato (Addgene, 50475-AAV2) into the BLA (AP:+/-3.30 , ML: -1.55, DV: -4.95, -4.85, relative to bregma). For calcium imaging experiments, we made a unilateral injection of 600nl CAMKII-GCaMP6f (Addgene, 100834-AAV1) in the left hemisphere using the same coordinates. Viral injections were followed either by headpost attachment or both headpost attachment and GRIN lens implantation. Mice were given 2 weeks to recover following headpost only surgery and 4 weeks to allow for adequate viral expression following viral injections before starting behavioral experiments.

Implantation of GRIN lens

In the group of mice that underwent two-photon imaging, a 0.85mm microendoscope probe (Grintech GmBH) was stereotactically implanted in the BLA. Brain tissue was first aspirated to a depth of ~1mm below skull surface to reduce intracranial pressure from the implant. The lens implant was lowered to DV -4.8mm relative to bregma and fixed in place using VetBond adhesive (3M) followed by a layer of dental cement and head post attachment. A layer of Kwik-cast (World Precision Instruments) was applied to the GRIN lens to protect it from damage between imaging sessions. After 4-6 weeks, the level of GcaMP6f expression was assessed using the two-photon

microscope. Imaging experiments began when viral expression was sufficiently bright, as judged by visual inspection.

Implantation of thermocouple

To measure respiration, one mouse was implanted with a custom-made thermocouple. Thermocouple wires were inserted in the intranasal cavity and affixed to the skull using dental cement.

Virtual Burrow Assay

Detailed methodology of the virtual burrow assay design and operation is described in the original publication[15]. Briefly, the virtual burrow is composed of a 3D-printed tube enclosure (virtual burrow). A trimmed, absorbent underpad is attached to the floor of the burrow, underneath the mouse, and a wooden stick is adhered to the front of the tube to serve as a handlebar for mice to grip with their forepaws. The virtual burrow is mounted on near-frictionless air bushings set at 15 psi (New Way Air Bearings). These air bushings slide on two parallel rails to constrain movement in one dimension. The animal is head-fixed using two custom-machined clamps fitted to the titanium headplate cemented to the mouse skull. The burrow is tethered to a linear actuator with a fishing line. The actuator is employed to place the mouse in a neutral position within the burrow during intertrial periods. The linear displacement of the mouse is measured using a laser displacement sensor focused on a 3D-printed flag that is affixed to the side of the tube enclosure. The laser readout provides a continuous, time-dependent, one-dimensional variable of position. All analog voltage signals from the laser displacement sensor are digitized at 1kHz using a National Instruments NI-DAQ for behavior-only experiments, or at 10kHz for calcium imaging experiments using a Blackrock Microsystems DAQ.

Odor delivery

Odorants acetophenone, octanal, pinene, and limonene (Sigma Aldrich) were mixed at a 1% dilution in mineral oil before the start of each experiment. Odorants were delivered through a custom-designed olfactometer. Custom-written Arduino code was used to structure the trials, which included communication between the virtual burrow assay and the olfactometer. Odor stimulus onsets were logged on the same DAQ as incoming burrow displacement data. Odor concentrations during each trial were recorded using a photoionization detector (PID) as voltage signals also collected on the DAQ.

Aversive conditioning paradigm

During pre-test, mice in the virtual burrow assay were presented with odor stimuli in 10 blocks of trials. Each block consisted of the 4 odors delivered in a random order for a total of 40 trials. Odor stimulus duration was 8 seconds and odor identity was varied for different mice (e.g., acetophenone was a CS+ for one mouse and a CS- for another mouse). The next day, mice were placed in a Med Associates chamber fitted with a grid floor in order to deliver foot shocks. During conditioning, each odor was delivered 8 times, with two odors paired with a 0.7mA footshock while the other two remained unpaired. Odor delivery was controlled by an Arduino and footshock

triggered with TTL pulses to Med Associates shock scrambler. The following day, mice were tested in the virtual burrow assay using the same block structure as pre-test day.

Two-photon imaging

All images were acquired using a multi-photon microscopy system (Prairie Technologies, now Bruker) with a 20x objective (0.5 NA, Nikon) and an ultra-fast pulsed laser beam (920-nm wavelength, Coherent). Time series were collected at 6.5 Hz and 128x256 pixels per frame. We continuously acquired fluorescence with a photomultiplier tube (GaAsP PMT) operated by PrairieView software. Each mouse received two imaging sessions, during pretest and test days (all data presented is from test day).

Extraction of calcium traces

Calcium imaging time-series were motion corrected using a planar hidden Markov model (SIMA)[64]. Data were then imported to MATLAB. Cell regions of interest (ROIs) were identified using the MATLAB CaImAn package[65]; spatial footprints and deconvolved signal for ROIs were extracted using CNMF.

Neural network classification of behavior and respiration

We trained a neural network to distinguish between periods of stationary displacement, tremble displacement, and directed motion. The model was designed to perform the classification task of distinguishing between the three movement states based on displacement and spectrograms of the displacement. First, we randomly sampled subset of 100 trials, with equal contributions from each of the 24 mice and with both CS+ and CS- trials present in the subset. For these trials, we labeled segments of the displacement as either “stationary”, “tremble”, or “motion”. The labeled data was resampled to produce equal numbers of each of the three movement categories. These labels then served as the ground truth for training and evaluating the neural network model. As features for the model, we input the displacement, as well as the displacement convolved with kernels of varying sizes to capture past and future displacement, and lastly, values from each trial’s spectrogram computed over the displacement using MATLAB’s cwt (continuous wavelet transform) function. A two-layer feed-forward neural network was trained using MATLAB’s patternnet function. During training, the network learns to map the engineered features to the correct class labels. Categorical cross-entropy is used as the loss function. A subset of test data from a held-out set of trials is taken to validate the model. Model performance is evaluated using metrics including accuracy, precision, recall, and F1-score on the testing dataset. The trained neural network was then used to label every time point of every trial’s spectrogram and displacement data into “stationary”, “tremble”, or “motion”.

Behavior and respiration quantification

The continuous wavelet transform was computed from the displacement and respiration data using MATLAB’s cwt function on a per-trial basis, with parameters set to 10 octaves and 48 voices per octave. Additionally, for each trial, “ingress onset” was defined as the first timepoint when displacement exceeded 1 mm. The “ingress magnitude” was calculated as a ratio of the maximum displacement attained following an ingress divided by the maximum achievable within the virtual

burrow assay for the session. The “tremble duration” was computed as the total time from all the tremble segments in a trial.

Prediction of ingress using tremble features

To explore a possible relationship between tremble and ingress, the trials were separated into ingress and no-ingress trials. Three logistic regression models were trained and tested using sklearn’s LogisticRegression module within Python with balancing of class weights. The three models were based on (1) tremble timing features including the duration of a segment, the time that segment began, and the total time spent trembling in that trial; (2) the tremble’s mean spectral power between 4-8 Hz and the tremble’s maximum frequency; and (3) the trial number on which a tremble occurred. The cross-validated accuracy was computed using 150 cross-validations for each model.

Neural decoding analysis

Neural decoding analysis was performed using the Decodanda package in Python[66] using a linear classifier based on a support vector machine architecture available within the scikit-learn package[67]. The following describes the fundamental steps implemented by Decodanda for the decoding analysis of neural activity. For details on the implementation, please refer to the package’s documentation: https://github.com/lposani/decodanda.

Data labeling

For all decoding analyses, we used inferred spiking activity as computed by CNMF. Each neuron’s spiking activity was z-scored based on the mean and standard deviation of its baseline activity from a 15-minute period prior to any CS presentation. Trials were obtained by taking time window of at least 500ms for tremble and stationary and windows of 2 seconds following ingress and were labeled according to the observed behavior and the presented stimulus (+1, +2, -1, -2). The average activity of each neuron within each trial was then computed to obtain a single vector of activity for each trial.

Pseudo-population sampling and cross-validation

The decoding performance was computed by a 20-fold cross-validation procedure. For each cross-validation fold, the trial activity vectors for each subject were randomly divided into training and testing sets (2/3 for training and 1/3 for testing) and used to construct two populations of training and testing pseudo-simultaneous (PS) activity vectors by pooling together all the recorded subjects that passed a minimum inclusion criterion of at least 4 trials per condition. To build a single PS activity vector, each subject was sampled q times to obtain vectors of length qN, where N is the total number of recorded neurons. In this work, we used $q = 5$. These pseudo-population sets were then used to train and test a linear support vector machine (SVM) to obtain a decoding accuracy value. The whole procedure, from training-testing separation to PS sampling and SVM accuracy testing, was repeated k=20 times, and the mean SVM decoding accuracy was taken as the measure of decoding performance.

Balanced sampling

Crucially, the same number of PS vectors was sampled for all conditions (combinations of values of the decoded variables) throughout the decoding analysis. Say we are decoding two variables from the same neural data: *valence* (CS+ vs. CS-) and *behavior* (tremble vs. stationary). Since, in general, tremble is more frequent during CS+ presentation than during CS-, *valence* could act as a confound when decoding *behavior*, and vice-versa. This balanced sampling ensures that the four conditions (tremble & CS+, tremble & CS-, stationary & CS+, stationary & CS-) are equally represented in the training set. Thus, variables are disentangled, and their individual contribution to neural activity is analyzed independently of each other. Unless differently noted, throughout this work, we sampled 200 PS vectors for each condition.

Decoding with a subsampled population of neurons

To analyze how decoding performance depends on the number of neurons used in the analysis, we repeated the decoding analysis, as described above, using a random subsample of the neural population with a given size n. We varied n in increments of 5% of the total number of neurons for a total of 19 values (5 to 100%). We repeated the random selection 10 times to obtain a population of decoding performances for each value of n.

Cross-condition generalization performance

Cross-condition generalization performance (CCGP) was computed as described in Bernardi et al., 2020[18]. We first constructed pseudo-simultaneous activity vectors as described above, except we did not group data from pairs of conditions with the same decoding variable. Rather, trials used for training a given classification came from one of the pairs of conditions that both contained the decoding dichotomy for a given classification while sharing the same non-decoding variable. The corresponding testing set consisted of data from the other pair of conditions that shared the non-decoding variable. For example, when decoding *valence* across *behavior,* one training set consisted of data during interactions with CS+ versus CS- during *behavior*=tremble, and the testing set consisted of data from CS+ versus CS- during *behavior*=stationary. The decoding for a given dichotomy was then repeated, swapping the classes of trials used for the training and testing data. CCGP was obtained from the mean decoding performance from the two pairs of training and testing conditions.

Null model and p-value

We tested the decoding performance and CCGP obtained by the procedures described above against a null model where the labels of trials were randomly shuffled. After each shuffling, the same procedures were repeated, obtaining a null-model value for decoding performance. We repeated the shuffling 10 times to obtain a distribution of null model performance values. The p-value for decoding performance or CCGP was then derived from the z-score of the performance computed on data compared to the distribution of null-model values.

Multi-selectivity analysis

We performed the following analysis to assess whether cells were specialized for one of two variables (hereby called variable A and variable B) or whether they encode the variables with mixed selectivity. For a given variable X and each cell i we identified the coding importance of

each individual cell, defined as w_i^X, as the absolute value of its average decoding weight normalized by the standard deviation over *k* cross-validation folds:

$$w_i^X := \left| \frac{\mathrm{E}\left[w_{i,n}^X\right]_n}{\sigma^2\left[w_{i,n}^X\right]_n} \right|$$

Where $w_{i,n}^X$ is the SVM decoding weight of cell *i* in the *n*-th cross-validation fold. We obtained a vector of all the values across the recorded population of cells:

$$\vec{W}^X = (w_1^X, w_2^X, \dots, w_N^X)$$

We denoted these vectors as $\vec{W}^A$, for variable A and $\vec{W}^B$ for variable B. If the recorded population is specialized, neurons that encode A will not encode B, and vice-versa. Therefore, a population of specialized neurons will be characterized by anti-correlated values of $\vec{W}^A$ and $\vec{W}^B$ (Methods Figure 1a). On the other hand, if neurons are not specialized (mixed selectivity), we expect no relationship between A and B coding, resulting in a null correlation between $\vec{W}^A$ and $\vec{W}^B$ (Methods Figure 1b). A third possibility is that neurons are not specialized, but information is unevenly distributed across the population. In this case, neurons will typically encode neither or both variables, resulting in a positive correlation between $\vec{W}^A$ and $\vec{W}^B$ (Methods Figure 1c).

To assess whether the recorded population was specialized, we computed the Spearman correlation between the two coding importance vectors, denoted as $\rho(\vec{W}^A, \vec{W}^B)$. We then compared the value of $\rho(\vec{W}^A, \vec{W}^B)$ with those obtained by a null model where mixed selectivity is implemented by performing a solid random rotation of the coding weights of the two variables in the neural activity space. In this null model, the two coding vectors $\vec{W}^A$ and $\vec{W}^B$ have no relationship with each other. For each null model iteration k we sample a random rotation matrix R_k and use it to rotate the weights vectors of A and B decoding before taking the average across cross-validations. This procedure was repeated 100 times, each time with a new rotation matrix, to obtain a population of null values. The recorded population of cells was then classified as either mixed or selective depending on the significance of the correlation, computed using the z-score of the Spearman correlation compared to the null model:

- $\rho(\vec{W}^A, \vec{W}^B)$ < null model: selective population (Methods Figure 1a)
- $\rho(\vec{W}^A, \vec{W}^B)$ ~ null model: mixed selectivity (Methods Figure 1b)
- $\rho(\vec{W}^A, \vec{W}^B)$ > null model: mixed selectivity (uneven distribution of coding importance, Methods Figure 1c)

L-score

In order to quantify the degree of specialization of a population of neurons for two given variables (hereby called variable A and variable B), we devised a measure of how much neurons lie on the axis in the weight-weight selectivity plot, which we called “L-score”. This measure is computed by measuring the overlap between the top *p* values for variable A and the

top *p* values for variable B, normalized so that if the two sets are the same neurons the overlap is 1:

$$O_p(A,B) := \frac{A_p \cap B_p}{A_p \cup B_p} ,$$

where A_p is the set of the top *p* selective neurons for variable A and B_p is the set of the top *p* selective neurons for variable B. By varying p from p=0 to p=1 (where *p* is expressed as fraction of the total number of neurons), we obtain a curve from which we compute the area under the curve, AUC(O_p). If the selectivities for A and for B are uncorrelated random variables, the expected value of the overlap is $O_p^{unc} = \frac{1}{\frac{2}{p}-1}$, and the corresponding area under the curve is:

$$AUC\left(O_p^{unc}(A,B)\right) = \int_0^1 \frac{1}{\frac{2}{p}-1}\, dp = \log(4) - 1 .$$

We then take the difference between the random uncorrelated case and the one computed from the data, AUC(O_p^{unc}) – AUC(O_p), as a measure of specialization in the neural population. To obtain a value that ranges between 0 and 1 we then normalize with the expected value of AUC in the case of a perfectly specialized population, i.e., when half of the neurons have high selectivity for variable A and low selectivity for variable B and the vice-versa for the other half. In this specialized case $O_p^{spec} = 0$ if $p \in [0, 0.5]$ and $O_p^{spec} = 2p - 1$ if $p \in [0.5, 1]$, so the corresponding area under the curve is AUC(O_p^{spec}) $= 0.25$.

Finally, the L-score is defined as:

$$L(A,B) := \frac{AUC\left(O_p^{unc}(A,B)\right) - AUC\left(O_p(A,B)\right)}{AUC\left(O_p^{unc}(A,B)\right) - AUC\left(O_p^{spec}(A,B)\right)} = \frac{\log(4)-1-AUC\left(O_p(A,B)\right)}{\log(4)-0.75} .$$

Ablation analysis to assess decoding importance of selective cells

To assess the relative importance of specialized and mixed neurons in the decoding performance for a given variable, we performed an "ablation" analysis where each class of cells (specialized and mixed) is selectively excluded from the data used for the decoding analysis. Say we are testing the decoding importance of two variables, A and B. First, we identified the n_A specialized cells for variable A and the n_B specialized cells for variable B as described in the multi-selectivity analysis methods. Those neurons that are not specialized for either A or B were labeled as "mixed." We then excluded all the $n_A + n_B$neurons that were identified as specialized and performed the decoding analysis for variables A and B (as described in the decoding methods) to get the decoding performances DP_A^{sel} and DP_B^{sel}. We then randomly selected the same number of mixed neurons by choosing a random value of selectivity lower than the selectivity threshold used to identify specialized neurons, which is equivalent to a random angle in the W_A vs. W_B graph (see Fig 3e and multi-selectivity methods), and choosing the $n_A + n_B$ neurons that have the selectivity closest to this random value (pink dots in Fig. 3e). We excluded these neurons and

repeated the decoding analysis to obtain the decoding performances DP_A^{mix} and DP_B^{mix}. By repeating this random choice procedure 20 times, we obtained a population of decoding values for both variables when excluding mixed selectivity neurons from the decoding analysis (pink error bars in Fig. 3i). Finally, the decoding performances DP_A^{sel} and DP_B^{sel} were compared to the population of DP_A^{mix} and DP_B^{mix} values to assess whether excluding specialized neurons affected decoding performance differently than excluding mixed-selective neurons. Statistical significance for each variable was assessed by computing the z-score of DP^{sel} compared to the corresponding DP^{mix} population.

Visualization of neural patterns in a reduced-dimensionality space

To visualize the geometry of conditions in the neural activity space, we performed a dimensionality reduction of the recorded neural patterns using a PCA approach. Say we have C conditions. These could be, for example, the 2^v combinations of v experimental binary variables or multiple values of a single variable. First, for each condition, we sampled N_c pseudo-population patterns from the activity of the recorded subjects as described in the decoding analysis methods. We then used PCA to obtain the first 3 principal components of the resulting $C \cdot N_c$ sampled patterns. We then projected the $C \cdot N_c$ patterns on the 3 principal components to obtain C collections of N_c vectors in the three-dimensional PCA space. We then computed the centroid of each condition by taking the median position of its N_c vectors in the PCA space. Using the median instead of the mean ensures that outliers that could be due to artifacts will not dominate the position of the centroid. Finally, we visualized the resulting C centroids in the three-dimensional PCA space.

Statistical Analysis

When paired comparisons were possible, paired t-tests were used to compute statistical significance. For comparing distributions, the Kolmogorov-Smirnov test was used to compute statistical significance. The alpha value for computing significance was 0.05. For all boxplots, the top and bottom edges of the box indicate the 75th and 25th percentile of the data. The whiskers extend to the most extreme data points not considered outliers. For decoder performance graphs showing shaded errors, the 95% confidence intervals obtained from the cross-validation runs are used for the upper and lower extents of the shading. Statistical analyses were performed using MATLAB (MathWorks) and Python.

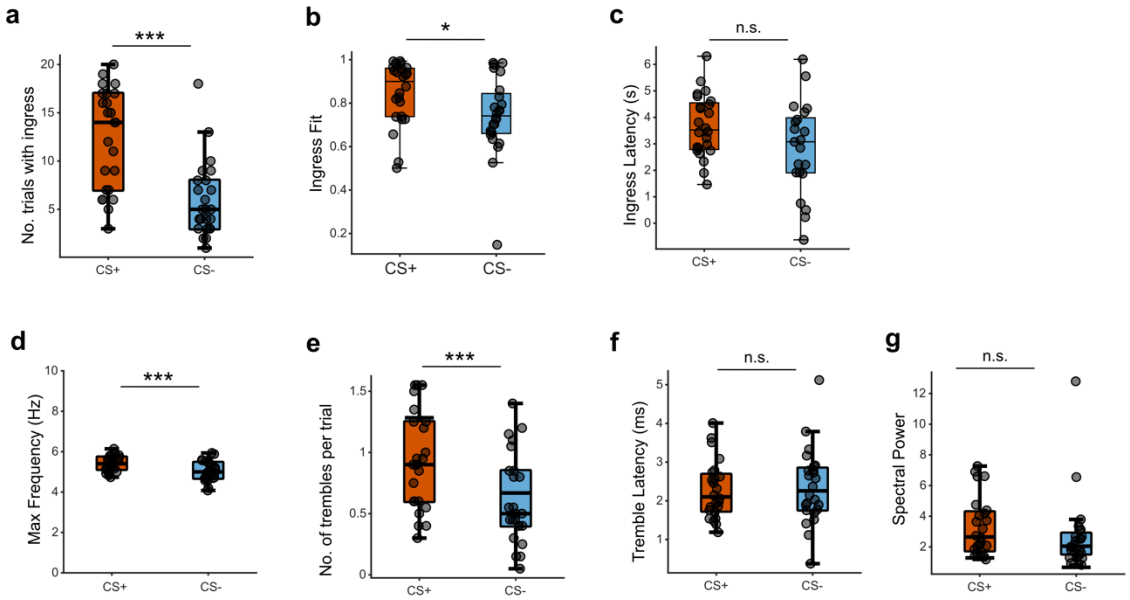
1. Blanchard, D. C. & Blanchard, R. J. Ethoexperimental approaches to the biology of emotion. *Annual review of psychology* vol. 39 Preprint at https://doi.org/10.1146/annurev.ps.39.020188.000355 (1988).
2. Blanchard, D. C., Griebel, G., Pobbe, R. & Blanchard, R. J. Risk assessment as an evolved threat detection and analysis process. *Neuroscience and Biobehavioral Reviews* vol. 35 Preprint at https://doi.org/10.1016/j.neubiorev.2010.10.016 (2011).
3. Anderson, D. J. & Adolphs, R. A framework for studying emotions across species. *Cell* Preprint at https://doi.org/10.1016/j.cell.2014.03.003 (2014).
4. LeDoux, J. E., Iwata, J., Cicchetti, P. & Reis, D. J. Different projections of the central amygdaloid nucleus mediate autonomic and behavioral correlates of conditioned fear. *Journal of Neuroscience* **8**, (1988).
5. Hitchcock, J. M., Sananes, C. B. & Davis, M. Sensitization of the Startle Reflex by Footshock: Blockade by Lesions of the Central Nucleus of the Amygdala or Its Efferent Pathway to the Brainstem. *Behavioral Neuroscience* **103**, (1989).
6. Kong, M. S. & Zweifel, L. S. Central amygdala circuits in valence and salience processing. *Behavioural Brain Research* **410**, (2021).
7. Yu, K., da Silva, P. G., Albeanu, D. F. & Li, B. Central amygdala somatostatin neurons gate passive and active defensive behaviors. *Journal of Neuroscience* (2016) doi:10.1523/JNEUROSCI.4419-15.2016.
8. Fadok, J. P. *et al.* A competitive inhibitory circuit for selection of active and passive fear responses. *Nature* **542**, (2017).
9. Moscarello, J. M. & Penzo, M. A. The central nucleus of the amygdala and the construction of defensive modes across the threat-imminence continuum. *Nat Neurosci* **25**, 999–1008 (2022).
10. Hartley, N. D. *et al.* Dynamic remodeling of a basolateral-to-central amygdala glutamatergic circuit across fear states. *Nat Neurosci* **22**, (2019).
11. Sah, P., Faber, E. S. L., De Armentia, M. L. & Power, J. The amygdaloid complex: Anatomy and physiology. *Physiological Reviews* Preprint at https://doi.org/10.1152/physrev.00002.2003 (2003).
12. Namburi, P. *et al.* A circuit mechanism for differentiating positive and negative associations. *Nature* (2015) doi:10.1038/nature14366.
13. Beyeler, A. *et al.* Organization of Valence-Encoding and Projection-Defined Neurons in the Basolateral Amygdala. *Cell Rep* **22**, (2018).
14. Kim, J., Pignatelli, M., Xu, S., Itohara, S. & Tonegawa, S. Antagonistic negative and positive neurons of the basolateral amygdala. *Nat Neurosci* (2016) doi:10.1038/nn.4414.
15. Fink, A. J., Axel, R. & Schoonover, C. E. A virtual burrow assay for head-fixed mice measures habituation, discrimination, exploration and avoidance without training. *Elife* (2019) doi:10.7554/eLife.45658.
16. Blanchard, R. J. *et al.* Defensive behaviors in wild and laboratory (Swiss) mice: The mouse defense test battery. *Physiol Behav* **65**, (1998).
17. Bolles, R. C. & Collier, A. C. The effect of predictive cues on freezing in rats. *Anim Learn Behav* **4**, (1976).
18. Bernardi, S. *et al.* The Geometry of Abstraction in the Hippocampus and Prefrontal Cortex. *Cell* (2020) doi:10.1016/j.cell.2020.09.031.

19. Rigotti, M. *et al.* The importance of mixed selectivity in complex cognitive tasks. *Nature* (2013) doi:10.1038/nature12160.
20. Fusi, S., Miller, E. K. & Rigotti, M. Why neurons mix: High dimensionality for higher cognition. *Current Opinion in Neurobiology* Preprint at https://doi.org/10.1016/j.conb.2016.01.010 (2016).
21. Kriegeskorte, N. & Wei, X. X. Neural tuning and representational geometry. *Nature Reviews Neuroscience* vol. 22 Preprint at https://doi.org/10.1038/s41583-021-00502-3 (2021).
22. Chung, S. Y. & Abbott, L. F. Neural population geometry: An approach for understanding biological and artificial neural networks. *Current Opinion in Neurobiology* vol. 70 Preprint at https://doi.org/10.1016/j.conb.2021.10.010 (2021).
23. Boyle, L. M., Posani, L., Irfan, S., Siegelbaum, S. A. & Fusi, S. Tuned geometries of hippocampal representations meet the demands of social memory. *bioRxiv* (2023).
24. Kyriazi, P., Headley, D. B. B. & Pare, D. Multi-dimensional Coding by Basolateral Amygdala Neurons. *Neuron* **99**, (2018).
25. Davis, M. The Role Of The Amygdala In Fear And Anxiety. *Annu Rev Neurosci* (1992) doi:10.1146/annurev.neuro.15.1.353.
26. Bagur, S. *et al.* Breathing-driven prefrontal oscillations regulate maintenance of conditioned-fear evoked freezing independently of initiation. *Nat Commun* **12**, (2021).
27. Moberly, A. H. *et al.* Olfactory inputs modulate respiration-related rhythmic activity in the prefrontal cortex and freezing behavior. *Nat Commun* **9**, (2018).
28. Schenberg, L. C., Corral Vasquez, E. & Barcellos da Costa, M. Cardiac baroreflex dynamics during the defence reaction in freely moving rats. *Brain Res* **621**, (1993).
29. Hofer, M. A. Cardiac and respiratory function during sudden prolonged immobility in wild rodents. *Psychosom Med* **32**, (1970).
30. Gore, F. *et al.* Neural Representations of Unconditioned Stimuli in Basolateral Amygdala Mediate Innate and Learned Responses. *Cell* (2015) doi:10.1016/j.cell.2015.06.027.
31. Bolles, R. C. Species-specific defense reactions and avoidance learning. *Psychol Rev* **77**, (1970).
32. Fanselow, M. S. The role of learning in threat imminence and defensive behaviors. *Current Opinion in Behavioral Sciences* vol. 24 Preprint at https://doi.org/10.1016/j.cobeha.2018.03.003 (2018).
33. Mobbs, D., Headley, D. B., Ding, W. & Dayan, P. Space, Time, and Fear: Survival Computations along Defensive Circuits. *Trends in Cognitive Sciences* Preprint at https://doi.org/10.1016/j.tics.2019.12.016 (2020).
34. Borkar, C. D. & Fadok, J. P. A novel pavlovian fear conditioning paradigm to study freezing and flight behavior. *Journal of Visualized Experiments* **2021**, (2021).
35. Jercog, D. *et al.* Dynamical prefrontal population coding during defensive behaviours. *Nature* **595**, (2021).
36. Duvarci, S. & Pare, D. Amygdala microcircuits controlling learned fear. *Neuron* vol. 82 Preprint at https://doi.org/10.1016/j.neuron.2014.04.042 (2014).
37. Bravo-Rivera, C., Roman-Ortiz, C., Brignoni-Perez, E., Sotres-Bayon, F. & Quirk, G. J. Neural structures mediating expression and extinction of platform-mediated avoidance. *Journal of Neuroscience* (2014) doi:10.1523/JNEUROSCI.0191-14.2014.

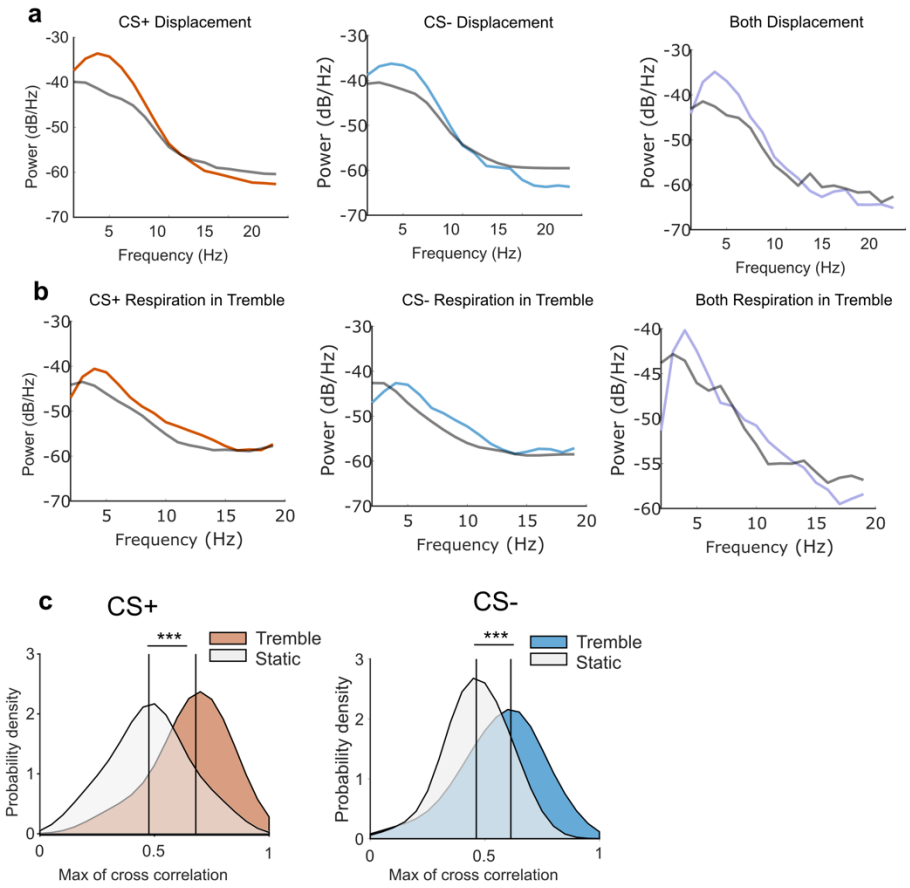
38. Paton, J. J., Belova, M. A., Morrison, S. E. & Salzman, C. D. The primate amygdala represents the positive and negative value of visual stimuli during learning. *Nature* (2006) doi:10.1038/nature04490.
39. Schoenbaum, G., Chiba, A. A. & Gallagher, M. Neural Encoding in Orbitofrontal Cortex and Basolateral Amygdala during Olfactory Discrimination Learning. *Journal of Neuroscience* (1999) doi:10.1523/JNEUROSCI.19-05-01876.1999.
40. Uwano, T., Nishijo, H., Ono, T. & Tamura, R. Neuronal responsiveness to various sensory stimuli, and associative learning in the rat amygdala. *Neuroscience* (1995) doi:10.1016/0306-4522(95)00125-3.
41. Grewe, B. F. *et al.* Neural ensemble dynamics underlying a long-term associative memory. *Nature* **543**, 670–675 (2017).
42. Belova, M. A., Paton, J. J. & Salzman, C. D. Moment-to-Moment Tracking of State Value in the Amygdala. *Journal of Neuroscience* (2008) doi:10.1523/jneurosci.1400-08.2008.
43. Morrison, S. E., Saez, A., Lau, B. & Salzman, C. D. Different Time Courses for Learning-Related Changes in Amygdala and Orbitofrontal Cortex. *Neuron* (2011) doi:10.1016/j.neuron.2011.07.016.
44. Saez, A., Rigotti, M., Ostojic, S., Fusi, S. & Salzman, C. D. Abstract Context Representations in Primate Amygdala and Prefrontal Cortex. *Neuron* **87**, 869–881 (2015).
45. Fustiñana, M. S., Eichlisberger, T., Bouwmeester, T., Bitterman, Y. & Lüthi, A. State-dependent encoding of exploratory behaviour in the amygdala. *Nature* **592**, (2021).
46. Gründemann, J. *et al.* Amygdala ensembles encode behavioral states. *Science (1979)* (2019) doi:10.1126/science.aav8736.
47. Beyeler, A. *et al.* Divergent Routing of Positive and Negative Information from the Amygdala during Memory Retrieval. *Neuron* **90**, 348–361 (2016).
48. Zhang, W. *et al.* Functional circuits and anatomical distribution of response properties in the primate amygdala. *Journal of Neuroscience* **33**, (2013).
49. Zhang, X. *et al.* Genetically identified amygdala–striatal circuits for valence-specific behaviors. *Nat Neurosci* **24**, (2021).
50. Burgos-Robles, A. *et al.* Amygdala inputs to prefrontal cortex guide behavior amid conflicting cues of reward and punishment. *Nat Neurosci* (2017) doi:10.1038/nn.4553.
51. Stuber, G. D. *et al.* Excitatory transmission from the amygdala to nucleus accumbens facilitates reward seeking. *Nature* (2011) doi:10.1038/nature10194.
52. Massi, L. *et al.* Disynaptic specificity of serial information flow for conditioned fear. *Sci Adv* **9**, eabq1637 (2023).
53. Piantadosi, S. C. *et al.* Holographic stimulation of opposing amygdala ensembles bidirectionally modulates valence-specific behavior. *bioRxiv* (2022).
54. Hagihara, K. M. *et al.* Intercalated amygdala clusters orchestrate a switch in fear state. *Nature* **594**, (2021).
55. Lonsdorf, T. B. & Kalisch, R. A review on experimental and clinical genetic associations studies on fear conditioning, extinction and cognitive-behavioral treatment. *Translational Psychiatry* vol. 1 Preprint at https://doi.org/10.1038/tp.2011.36 (2011).
56. Pitman, R. K. *et al.* Biological studies of post-traumatic stress disorder. *Nature Reviews Neuroscience* vol. 13 Preprint at https://doi.org/10.1038/nrn3339 (2012).

57. Lissek, S. Toward an account of clinical anxiety predicated on basic, neurally mapped mechanisms of Pavlovian fear-learning: The case for conditioned overgeneralization. in *Depression and Anxiety* vol. 29 (2012).
58. Lissek, S. *et al.* Classical fear conditioning in the anxiety disorders: A meta-analysis. *Behaviour Research and Therapy* **43**, (2005).
59. Jovanovic, T. & Ressler, K. J. How the neurocircuitry and genetics of fear inhibition may inform our understanding of PTSD. *American Journal of Psychiatry* vol. 167 Preprint at https://doi.org/10.1176/appi.ajp.2009.09071074 (2010).
60. Resnik, J. & Paz, R. Fear generalization in the primate amygdala. *Nat Neurosci* **18**, (2015).
61. Shaban, H. *et al.* Generalization of amygdala LTP and conditioned fear in the absence of presynaptic inhibition. *Nat Neurosci* **9**, (2006).
62. Ponsen, M., Taylor, M. E. & Tuyls, K. Abstraction and generalization in reinforcement learning: A summary and framework. in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* vol. 5924 LNAI (2010).
63. Barto, A. G. & Mahadevan, S. Recent Advances in Hierarchical Reinforcement Learning. *Discrete Event Dynamic Systems: Theory and Applications* vol. 13 Preprint at https://doi.org/10.1023/A:1022140919877 (2003).
64. Kaifosh, P., Zaremba, J. D., Danielson, N. B. & Losonczy, A. SIMA: Python software for analysis of dynamic fluorescence imaging data. *Front Neuroinform* **8**, 1–10 (2014).
65. Giovannucci, A. *et al.* CaImAn: An open source tool for scalable Calcium Imaging data Analysis. *bioRxiv* (2018) doi:10.1101/339564.
66. Van Rossum, G. & L. Drake, F. *Python 3 Reference Manual, Scotts Valley*. *Scotts Valley, CA* (2009).
67. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python Pedregosa, Varoquaux, Gramfort et al. *Journal of Machine Learning Research* **12**, (2011).

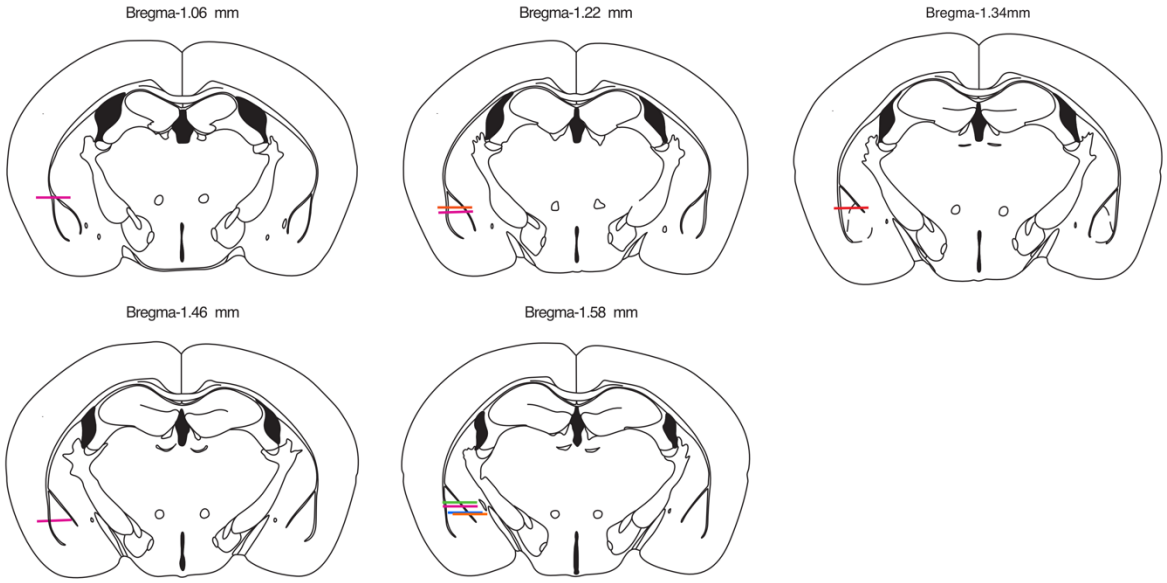
SUPPLEMENTAL FIGURES



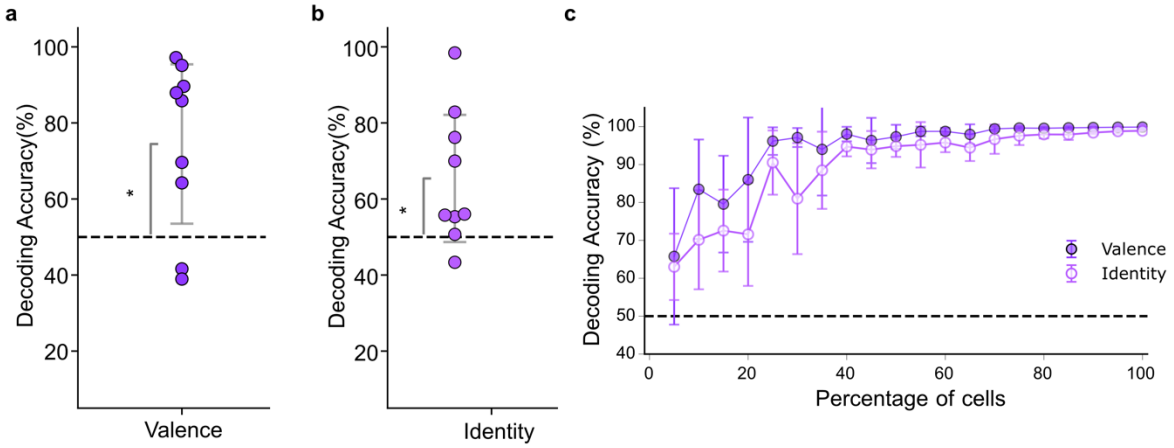
Supplemental Figure 1. Quantification of CS+/- behavior in virtual burrow assay. a, Number of trials with ingress for CS+ and CS- trials across mice. **b**, Ingress goodness of fit for CS+ and CS- across mice. **c**, Average latency to ingress across mice. **d**, Maximum frequency of a tremble bout across mice. **e**, Average number of trembles per trial across mice. **f**, Average latency to first tremble across mice. **g**, Spectral power of tremble across mice. For all plots, n = 24 mice, * $p<0.05$, ***$p<0.001$, paired t-test.



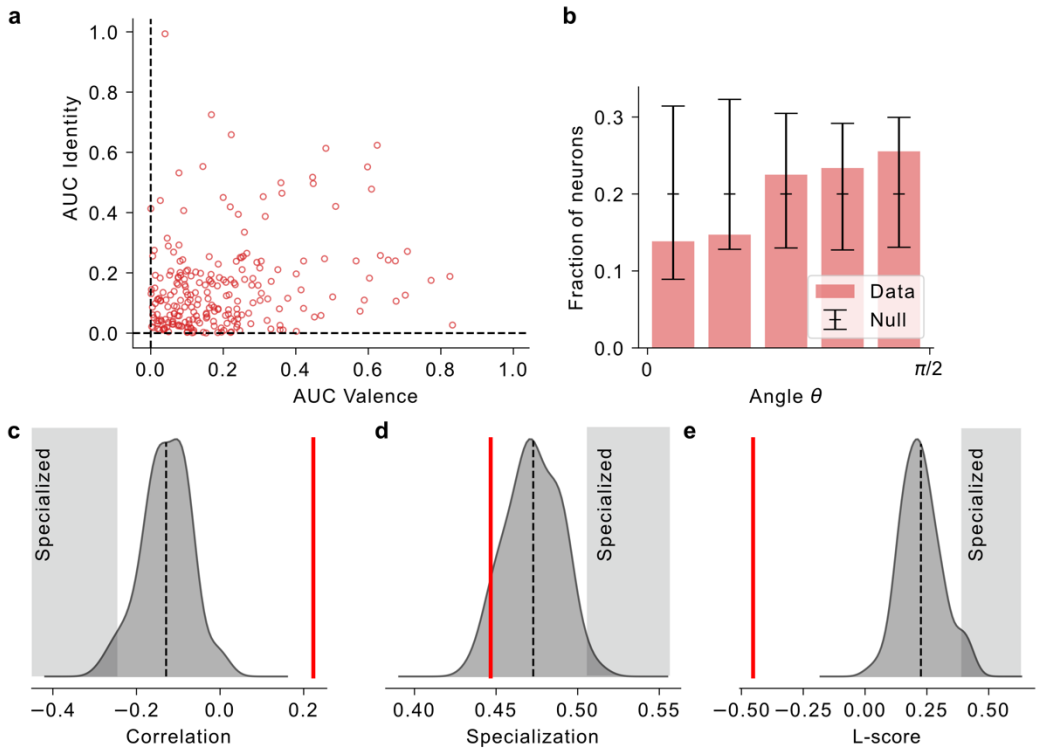
Supplemental Figure 2. Quantification of the relationship between respiration and tremble. a, Average power spectra of tremble bouts in CS+ trials, CS- trials, and CS+ and CS- trials combined. **b**, Average power spectra of respiration within tremble bouts in CS+ trials, CS- trials, and CS+ and CS- trials combined. **c**, Probability density distributions of the maximum correlation between displacement and respiration during tremble vs stationary periods from CS+ and CS- trials separately (vertical lines depict median of distributions, ***p<0.001, Kolmogorov–Smirnov test).



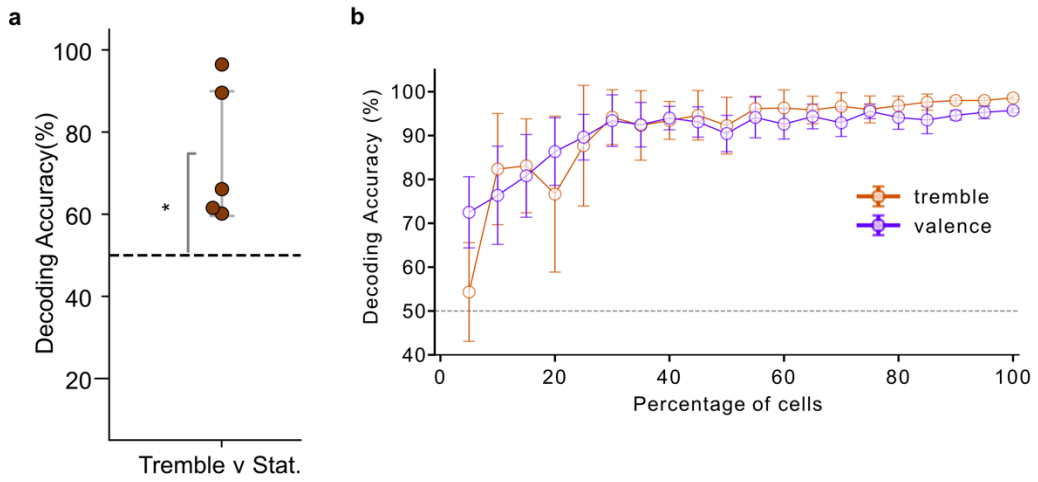
Supplemental Figure 3. GRIN lens locations for mice in two photon imaging cohort. Each line depicts location of GRIN lens implant within BLA from subsequent histological examination.



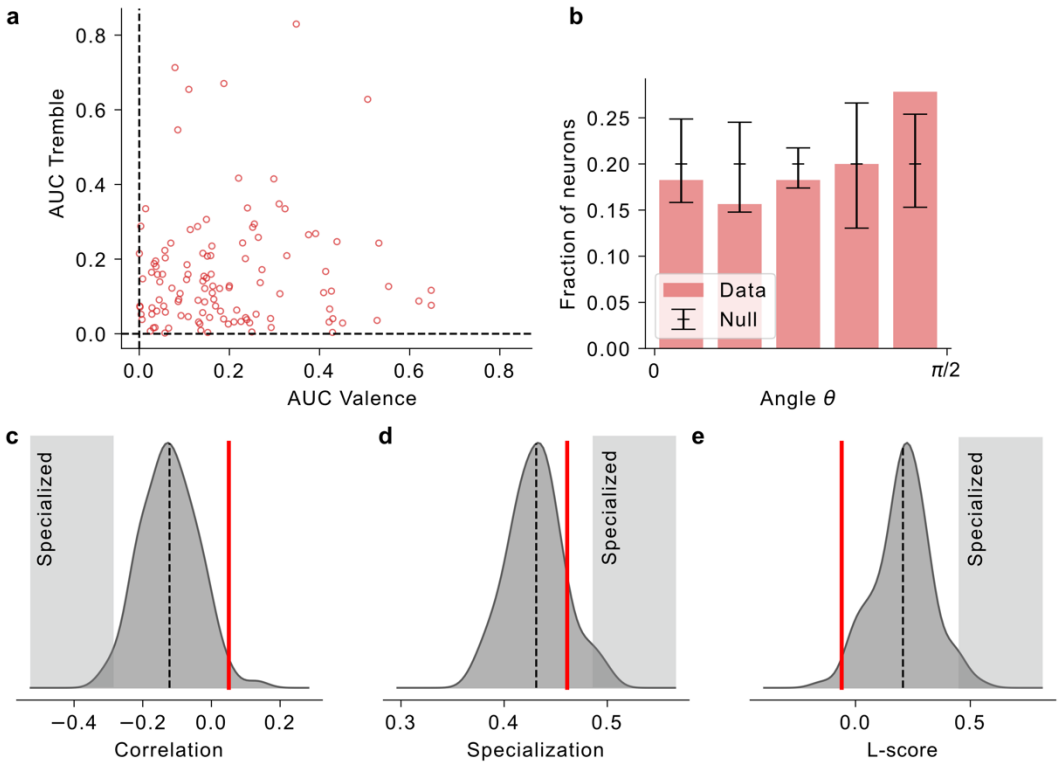
Supplemental Figure 4. Individual animal decoding and number of neurons needed for decoding valence and identity. a, Decoding accuracy for CS valence for each mouse (dots) (mean: 74% accuracy, n=9 mice, *p<0.05, t-test from 50% chance accuracy) **b**, Decoding accuracy for one of the two dichotomies related to CS identity for each mouse (dots) (mean: 65% accuracy, n=9 mice, *p<0.05) **c**, Decoding accuracy of valence and one of the two identity-related dichotomies (CS+1/CS-1 vs CS+2/CS-2) as a function of the percentage of neurons included in decoding analyses (resulting from analysis that down-samples neurons from the population).



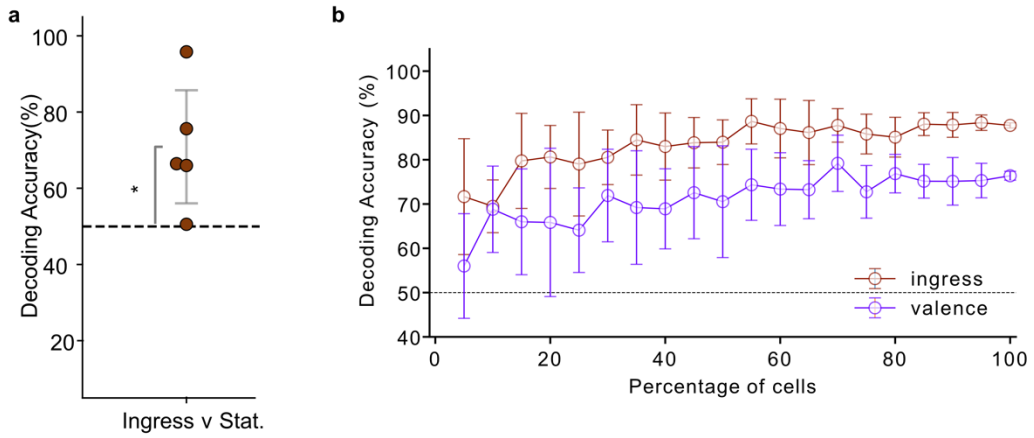
Supplemental Figure 5. BLA mixed selectivity for valence and CS identity characterized from neuron firing rates. **a**, For each neuron, the area under the curve (AUC) computed from an ROC analysis is plotted for both valence and for one of the CS identity-related dichotomies (CS+1/CS-1 vs CS+2/CS-2). **b**, Fraction of neurons contained within five polar slices of the scatterplot shows density distributed across angular slices. Error bars, 95% confidence interval from the null model. **c**, Spearman's correlation calculated between the AUC values for valence and the other dichotomy calculated from the data in (a) (red line) and from iterations of the null model (distribution) **d**, Mean specialization of neurons calculated from AUC values, as in c.. **e**, L-score calculated from AUC values as in c.



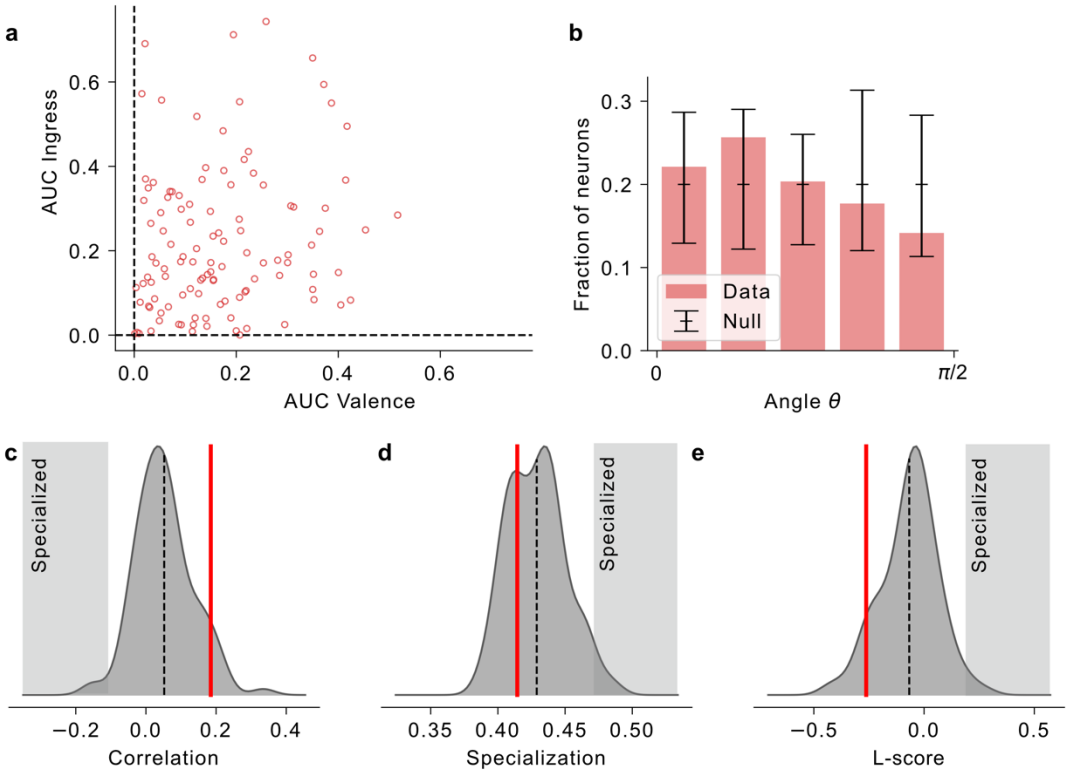
Supplemental Figure 6. Individual animal decoding and number of neurons needed for decoding valence and tremble. a, Decoding accuracy for tremble each mouse (dots) (mean: 75% accuracy, n=5 mice, *p<0.05, t-test from 50% chance accuracy). **b**, Decoding accuracy of tremble and valence plotted as a function of the percentage of neurons included in decoding analyses (resulting from analysis that down-samples neurons from the population).



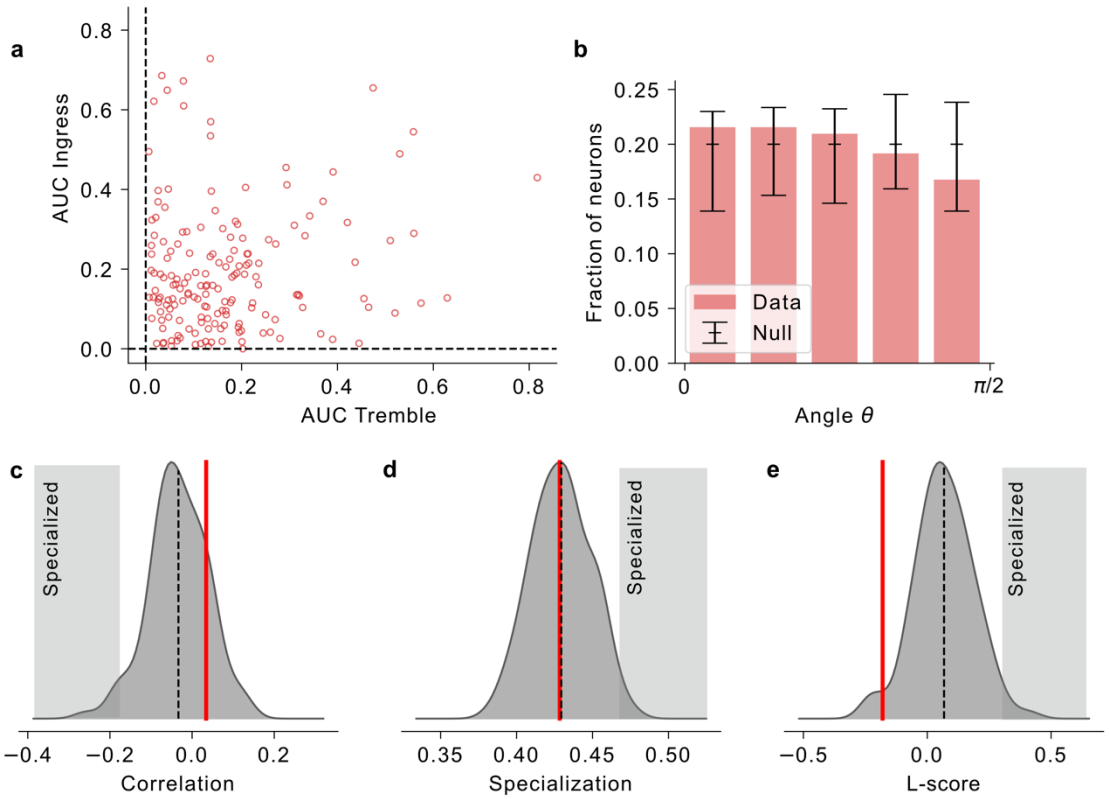
Supplemental Figure 7. BLA mixed selectivity for valence and tremble characterized from neuron firing rates. **a**, For each neuron, the area under the curve (AUC) computed from an ROC analysis is plotted for both valence and tremble. **b**, Fraction of neurons contained within five polar slices of the scatterplot shows density distributed across angular slices. Error bars, 95% confidence interval from the null model. **c**, Spearman's correlation calculated between the AUC values for valence and tremble calculated from the data in (a) (red line) and from iterations of the null model (distribution) **d**, Mean specialization of neurons calculated from AUC values, as in c. **e**, L-score calculated from AUC values as in c.



Supplemental Figure 8. Individual animal decoding and number of neurons needed for decoding valence and ingress. a, Decoding accuracy for ingress for each mouse (dots) (mean: 71% accuracy, n=5 mice, *p<0.05, t-test from 50% chance accuracy). **b**, Decoding accuracy of ingress and valence plotted as a function of the percentage of neurons included in decoding analyses (resulting from analysis that down-samples neurons from the population).



Supplemental Figure 9. BLA mixed selectivity for valence and ingress characterized from neuron firing rates. **a**, For each neuron, the area under the curve (AUC) computed from an ROC analysis is plotted for both valence and ingress. **b**, Fraction of neurons contained within five polar slices of the scatterplot shows density distributed across angular slices. Error bars, 95% confidence interval from the null model. **c**, Spearman's correlation calculated between the AUC values for valence and ingress calculated from the data in (a) (red line) and from iterations of the null model (distribution) **d**, Mean specialization of neurons calculated from AUC values, as in c. **e**, L-score calculated from AUC values as in c.



Supplemental Figure 10. BLA mixed selectivity for tremble and ingress characterized from neuron firing rates. **a**, For each neuron, the area under the curve (AUC) computed from an ROC analysis is plotted for both ingress and tremble. **b**, Fraction of neurons contained within five polar slices of the scatterplot shows density distributed across angular slices. Error bars, 95% confidence interval from the null model. **c**, Spearman's correlation calculated between the AUC values for ingress and tremble calculated from the data in (a) (red line) and from iterations of the null model (distribution) **d**, Mean specialization of neurons calculated from AUC values, as in c. **e**, L-score calculated from AUC values as in c.