Reviewed Preprint
v1 • May 15, 2026
Not revised

Reviewed Preprint
v2 • September 18, 2026
Revised by authors

Complementary stable and dynamic prelimbic ensembles encode learned threat value underlying generalization and discrimination

Marc E Normandin[1,^], Pedro M Ogallar[1,^], Matthew R Lopez[1,^], Manuel M Ramos-Alvarez[2], Isabel A Muzzio[1]

[1]Department of Psychological & Brain Sciences, University of Iowa, Iowa City, United States • [2]Psychology Department, University of Jaen, Jaén, Spain

For correspondence: isabel-muzzio@uiowa.edu

[^] Authors contributed equally

Competing interests: No competing interests declared

Funding: See page 51

Reviewing editor: Joshua Johansen, RIKEN Center for Brain Science, Japan

eLife Assessment

This **valuable** study examines how the rodent prelimbic cortex represents learned and generalized threat over time and identifies distinct stable and dynamic neuronal populations that contribute to these representations. The evidence is **convincing**, supported by longitudinal calcium imaging, appropriate control groups, and sophisticated analyses showing that the relevant neural signals cannot be explained simply by freezing behavior. The work provides a conceptual framework for understanding how stable threat-related representations supporting memory generalization and discrimination can be maintained despite ongoing changes in neuronal ensemble composition.

https://doi.org/10.7554/eLife.111177.2.sa4

Abstract

Adaptive behavior requires assigning emotional value to sensory cues and generalizing to similar stimuli. The prelimbic cortex (PL) is critical for threat expression and discrimination, yet its neuronal ensembles undergo substantial turnover over time. How stable population codes emerge despite this reorganization remains unknown. Using longitudinal calcium imaging in freely moving mice, we tracked PL activity for 30 days during auditory discriminative fear learning and retrieval. Despite substantial ensemble turnover, stable graded representations persisted in PL populations. A generalized linear model showed that learned threat value explained graded PL activity better than freezing, indicating that PL primarily encodes stimulus significance. Population similarity across stimulus presentations was highest for shock-associated and highly generalized tones, suggesting a stable population code underlying threat generalization. Stable graded neurons preserved learned threat value across sessions, whereas learning modulated the activity of dynamic frequency-selective neurons. Together, these populations enabled stable threat-value coding despite ongoing ensemble reorganization.

Introduction

Generalization allows learned responses to extend from a conditioned stimulus to similar cues, whereas discrimination enables animals to distinguish threat from safety. Adaptive behavior requires a balance between these processes, permitting defensive responses to potential danger while minimizing these responses to safe stimuli. Despite the importance of this balance for survival, how neural populations simultaneously support generalization and discrimination remains poorly understood. Two major theoretical frameworks have been proposed to explain these processes. Sensory similarity models posit that generalization arises primarily from overlap in sensory representations (Shepard, 1987). More recent work, however, suggests that sensory similarity alone cannot fully account for generalization. Instead, the brain is proposed to construct

representations of learned threat-value, whereby neural activity reflects the estimated probability or magnitude of danger associated with different stimuli (Dunsmoor & LaBar, 2013; Laufer et al., 2016; Verra et al., 2026). Under this framework, sensory similarity provides the basis for evaluating novel stimuli, whereas associative learning determines their predicted threat-value. These models therefore make distinct predictions about neural activity. Sensory similarity models predict that neural activity is organized according to perceptual similarity, with generalized responses arising from overlapping neural representations and successful discrimination emerging from increasingly distinct representations, sometimes involving separate neuronal populations that encode threat and safety (Corches et al., 2019; Grosso et al., 2018). In contrast, threat-value models propose that neural activity reflects the learned threat value assigned to each stimulus, with generalization and discrimination represented along a continuum of expected danger rather than being determined solely by perceptual similarity (Levy & Schiller, 2021). In this study, we investigated how the prelimbic cortex (PL), a region implicated in fear expression and threat generalization (Burgos-Robles et al., 2009; Rosas-Vidal et al., 2025; Sierra-Mercado et al., 2011; Sotres-Bayon & Quirk, 2010; Stujenske et al., 2022), supports generalization and discrimination.

Previous studies have established that the prelimbic cortex is required for discriminating between threat and safety cues. During discriminative fear conditioning, transient inhibition of the PL causes animals to freeze similarly to conditioned threat and safe stimuli, resulting in overgeneralization (Likhtik et al., 2014). Likewise, disruption of endocannabinoid signaling in the PL abolishes the neuronal distinction between threat and neutral cues (Rosas-Vidal et al., 2025). Together, these findings indicate that the PL plays a critical role in organizing neural representations that support appropriate discrimination while limiting threat generalization. However, how individual PL neurons encode these computations remains unclear. Human neuroimaging studies have shown that medial prefrontal activity scales with gradients of learned threat, consistent with value-based representations (Lissek et al., 2014). In contrast, other studies suggest that activity in this region primarily reflects the representational similarity of affective or valence information (Chavez & Heatherton, 2015). Adding further complexity, PL neuronal activity is also strongly correlated with freezing behavior—a hallmark defensive response during fear retrieval—regardless of the specific cue predicting an aversive outcome (Casanova et al., 2024; Kyriazi et al., 2020). Consequently, it remains unclear whether PL population activity primarily reflects sensory similarity, learned threat value, or defensive behavior, and how these factors together shape threat generalization and discrimination.

An additional challenge in identifying the cortical neural codes underlying threat generalization and discrimination is that both episodic memory ensembles (DeNardo et al., 2019; Kitamura et al., 2017) and procedural memory ensembles (Do-Monte et al., 2015; Iqbal et al., 2026) undergo substantial reorganization across days—a hallmark of systems consolidation—even while learned behaviors remain remarkably stable. This dissociation between neuronal instability and behavioral persistence suggests that stable behavior can be supported by population-level representations despite continuous changes in their neuronal composition (Deitch et al., 2021; Gallego et al., 2020). Because threat generalization requires evaluating novel stimuli in relation to previously learned threat-value representations, often long after the original threat association has been acquired, it may be particularly sensitive to ongoing ensemble reorganization. These observations raise several important questions. Can stable population codes supporting threat generalization emerge despite continuous changes in their neural composition? Do distinct neuronal subpopulations differ in their long-term stability? If so, do stable and dynamic neuronal subpopulations differentially represent learned threat value, sensory similarity, or the expression of defensive behavior?

We hypothesized that the PL generates population-level representations of learned threat value that support threat generalization and discrimination, and that these representations are maintained through the complementary contributions of stable and dynamic neuronal ensembles despite ongoing ensemble reorganization. Furthermore, although freezing behavior was expected to account for a portion of the variance in neuronal activity, we predicted that learned threat

value would explain substantially more of the neural response. To test these hypotheses, we combined longitudinal calcium imaging, computational clustering analyses, and a generalized linear model (GLMs) to characterize PL ensemble dynamics during auditory discriminative fear learning and memory retrieval in freely moving mice. Our results show that learned threat value is represented as graded neural responses at both the population and single-neuron levels despite extensive ensemble turnover. Moreover, population similarity across stimuli closely tracked behavioral generalization, suggesting that sensory similarity determines the overlap between population neuronal responses, whereas associative learning organizes these responses according to learned threat value. Finally, activity across cell types was better explained by learned threat value than by freezing behavior. Together, these findings provide a neural framework for understanding how the PL supports adaptive threat generalization and discrimination.

Results

Logarithmic tone separation predicts freezing level

To assess memory retrieval and generalization, GRIN-lens–implanted and non-implanted mice were trained in a differential auditory fear-conditioning paradigm (Fig. 1a). One tone (CS^+; 80 dB) was paired with a mild foot shock (0.5 mA, 0.5 s), whereas a second tone (CS^-; 80 dB) was never paired with shock. In one group, the CS^+ was 15 kHz and the CS^- was 3 kHz (CS^+15; N = 27; 7 implanted, 20 non-implanted); in a second group, contingencies were reversed (CS^+3; N = 22; 5 implanted, 17 non-implanted). A no-shock control group was exposed to the same tones without shock (N = 13; 6 implanted, 7 non-implanted).

Retrieval was assessed 1, 15, and 30 days after conditioning to probe recent, long-term, and remote memory (Bontempi et al., 1996). During each retrieval session, mice were tested in a novel context with the CS^+, the CS^-, and two intermediate frequencies (7 and 11 kHz), presented in a semi-random order with each tone repeated three times (Fig. 1a). The intermediate frequencies were selected to evaluate whether perceptual similarity predicts threat generalization. Although the separation between the CS^+ and CS^- was identical in both groups (2.32 octaves), the spacing between the CS^+ and the nearest intermediate frequency differed (0.45 octaves for 11–15 kHz versus 1.22 octaves for 3–7 kHz). Thus, if threat generalization is determined primarily by sensory similarity, freezing should generalize more strongly to 11 kHz in CS^+15 mice than to 7 kHz in CS^+3 mice.

In the CS^+15 group, mice consistently discriminated between the CS^+ and CS^-. Although generalization to intermediate frequencies was broad on day 1, it became progressively restricted to 11 kHz—the frequency adjacent to the CS^+—during long-term (day 15) and remote (day 30) retrieval ($p < 0.05$; Fig. 1b, left). CS^+3 mice also discriminated between the CS^+ and CS^-; however, these animals exhibited greater freezing to the CS^+ than to all intermediate frequencies across testing days ($p < 0.05$; Fig. 1b, center), indicating reduced generalization to the adjacent tone when the logarithmic separation was larger. No-shock control mice showed no significant differences in freezing across frequencies on any testing day ($p > 0.05$; Fig. 1b, right), confirming that the graded freezing patterns resulted from associative learning rather than the acoustic properties of the tones. Together, these results indicate that logarithmic spacing (i.e., frequency similarity), rather than simple proximity to the CS^+, was the primary determinant of generalization.

Because our experimental design involved repeated retrieval testing over the course of a month, we examined whether repeated tone presentations produced extinction. We quantified stimulus discrimination using discrimination ratios (DRs), which measure discrimination between each tone and the CS+. As shown in Fig. S1, DRs increased over successive retrieval sessions in both conditioned groups. In CS+15 mice, significant increases were observed for the 3 and 7 kHz tones ($p < 0.05$), whereas persistent generalization remained only for the 11 kHz tone. Similarly, CS+3 mice exhibited significant increases in DRs for the CS− and all intermediate frequencies ($p < 0.05$, Fig. S1). Thus, repeated testing progressively sharpened discrimination rather than reducing conditioned responding. This pattern is inconsistent with extinction, which would be expected to

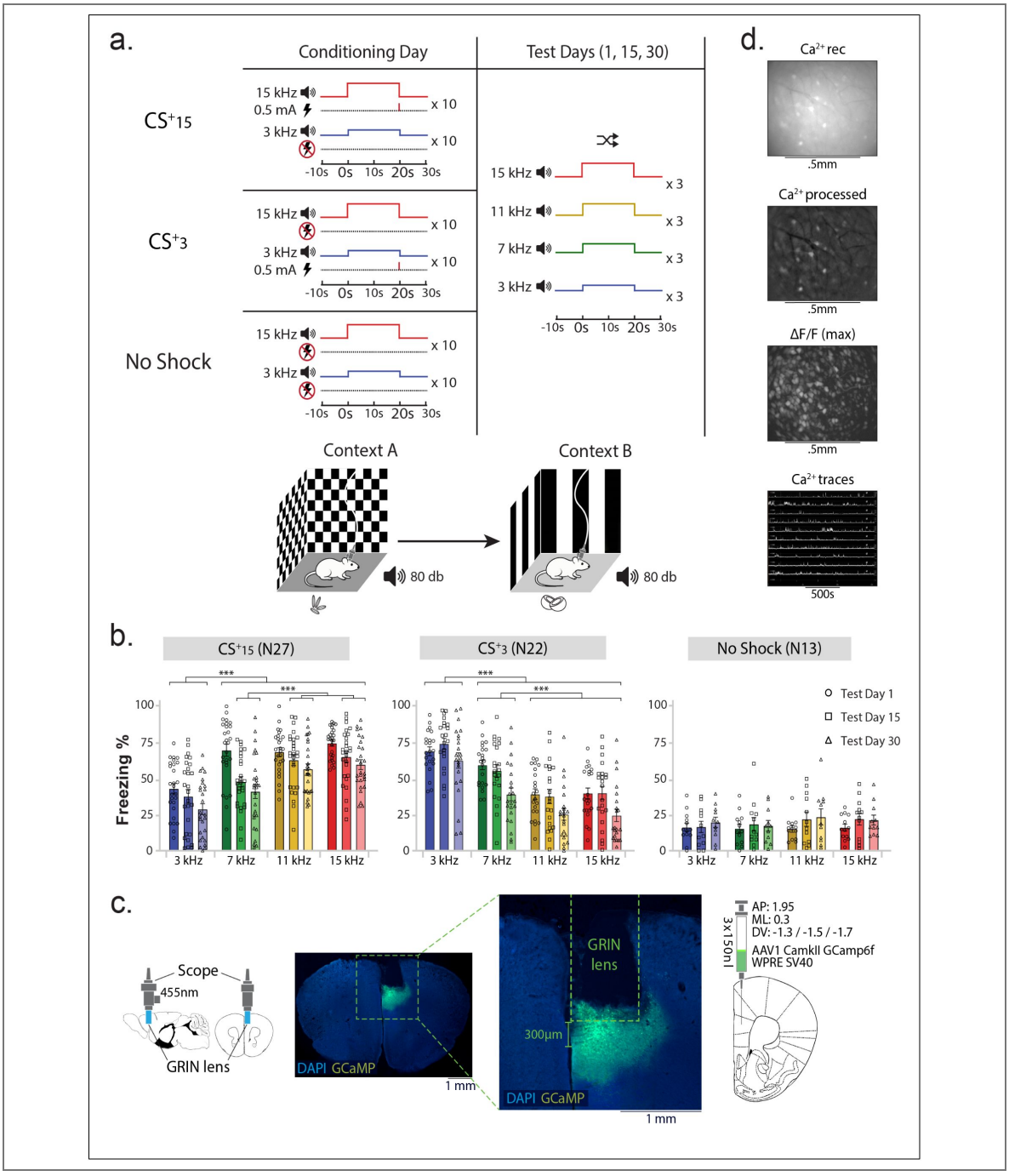

Figure 1. Experimental design, behavior, histology and calcium imaging.

a. Experimental design illustrating discriminative fear conditioning with either a 15 kHz CS$^+$ or a 3 kHz CS$^+$ (experimental groups), and no-shock controls tested at identical time points but never exposed to footshock. All groups were tested with 3 and 15 kHz tones and two intermediate frequencies (7 and 11 kHz) on days 1, 15 and 30 after conditioning. **b.** Behavioral responses across testing days (CS$^+$15 kHz: n = 27 mice; CS$^+$3 kHz: n = 22 mice; no-shock controls: n = 13 mice). Two-way repeated-measures ANOVA: CS$^+$15 : effect of day $F(2,52) = 13.28$, $p < 0.001$, frequency: $F(3,78) = 85.09$, $p < 0.001$, day × frequency interaction: $F(6,156) = 3.79$, $p = 0.002$; CS$^+$3: effects of day $F(2,42) = 14.42$, $p < 0.001$, frequency: $F(3,63) = 58.81$, $p < 0.001$, day × frequency interaction: $F(6,126) = 1.41$, $p = 0.217$; no-shock controls: effects of day: $F(2,20) = 1.38$, $p = 0.276$, frequency: $F(3,30) = 1.43$, $p = 0.253$; day × frequency interaction: $F(6,60) = 0.335$, $p = 0.916$. Significant Tukey multiple comparisons are denoted by asterisks, *** $p < 0.001$. **c.** Representative histological section showing GCaMP6f expression. Imaging depth did not exceed 300 µm, restricting recordings to PL. **d.** Example calcium imaging data showing raw fluorescence signals, deconvolved activity, thresholded events and corresponding activity traces.

flatten discrimination gradients, and instead agrees with previous studies showing that discrimination training narrows behavioral generalization gradients (Dunsmoor & LaBar, 2013; Herzog et al., 2021; Jenkins & Harrison, 1960; Lommen et al., 2017).

No sex differences were observed in CS$^+$15 mice (11 females, 16 males) or (6 females, 7 males; $p > 0.05$). In CS$^+$3 mice (11 females, 11 males), minor and inconsistent sex-related effects were detected on days 1 and 15, but these effects were absent by day 30 (Fig. S2). Accordingly, sex was not included as a factor in subsequent neural analyses. Behavioral performance did not differ between implanted and non-implanted mice on days 1 and 15 ($p > 0.05$). On day 30, a modest main effect of group was observed in the CS$^+$3: group, reflecting higher overall freezing levels in implanted animals compared with non-implanted mice ($p < 0.03$). Importantly, both experimental groups exhibited robust frequency effects across days ($p < 0.001$), with no group × frequency interactions ($p > 0.05$), indicating preserved threat discrimination and generalization profiles.

PL sound-responsive ensembles exhibit dynamic properties over time

To examine the temporal evolution of PL neuronal responses, we performed longitudinal calcium imaging in CS$^+$15 (N = 7), CS$^+$3 (N = 5), and no-shock control mice (N = 6). Across groups, neurons were tracked during conditioning and all retrieval sessions. To assess registration quality, we used the conditioning session as a reference and measured the centroid shifts between matched neurons across retrieval sessions. For each animal, we generated histograms of these shifts (μm) and compared them with the median equivalent circular footprint diameter (Fig. S3). Across animals, 99–100% of registered neurons exhibited centroid shifts smaller than the median footprint diameter, indicating highly reliable cell registration. The proportion of registered neurons was equivalent across groups (Fig. S3). Fig. 1c illustrates the GCaMP6f viral injection strategy and representative histological sections showing the pattern of viral expression. Fig. 1d shows a representative raw calcium video frame, processed frames, and activity traces.

Sound-evoked responses were visualized using activity plots with calcium activity ranked across simultaneously recorded neurons (Fig. 2a). Across animals, we visualized distinct neuronal subpopulations showing positive modulation, negative modulation, mixed responses, or no consistent response to sound in each session (Fig. 2b). Sound-responsive neurons were identified using a sound responder statistical test that detected activity modulation relative to baseline, allowing reliable identification of sound responses while remaining robust to noise. This approach was effective for neurons exhibiting either brief calcium transients that rose and decayed rapidly or sustained activity throughout the tone presentation. Fig. 2c summarizes the proportions of response types pooled across animals and sessions.

Given the ongoing debate over whether neocortical memory ensembles stabilize or remain dynamic over time (DeNardo et al., 2019; Kitamura et al., 2017; Kupke & Oliveira, 2025; Lopez et al., 2024; Mau et al., 2020; Rao-Ruiz et al., 2021; Refaeli et al., 2023; Terranova et al., 2023; Zaki & Cai, 2024), we next assessed the temporal stability of PL ensembles by tracking the cellular footprints of sound-responsive neurons across retrieval sessions. Ensemble stability was visualized using Venn diagrams (Fig. 2c). A moderate proportion of neurons was active across all retrieval sessions, with no differences between groups ($p > 0.05$). Likewise, there were no group differences in the proportion of neurons overlapping across only two sessions ($p > 0.05$). For neurons active in only a single session, control mice exhibited a higher proportion on day 30 compared with CS15+ mice ($p < 0.05$, Table S1), with no other differences observed. Overall, these data indicate that the majority of neurons overlapped across only a limited number of sessions or were transiently active, reflecting pronounced population-level dynamism over time.

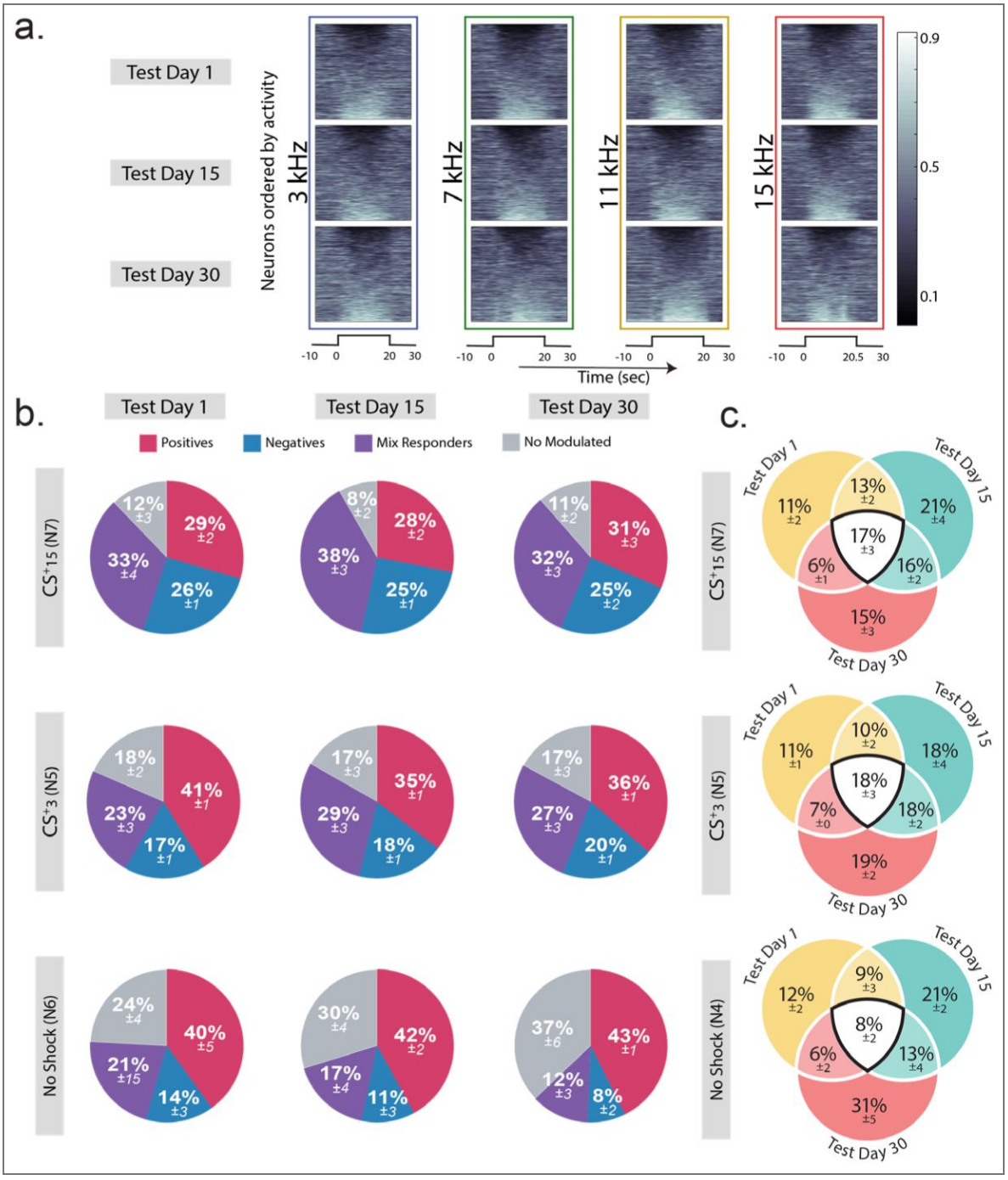

Figure 2. Distribution of stable and dynamic cell types across sessions.

a. Activity plots from an animal trained with a 15 kHz CS+, ordered by activity level. Activity plots illustrate positive sound-responsive neurons (bottom), negative sound-responsive neurons (top), and mixed or non-responsive neurons (middle) before and after sound presentation. **b.** Proportions of neuronal response types across experimental groups and testing days. Number of cells: CS+15: conditioning: 3,834; day 1: 3,177; day 15: 4,186; day 30: 3,693, CS+3 conditioning: 1,628; day 1: 1,381; day 15: 1,881; day 30: 1,833, control: conditioning: 1,118; day 1: 986; day 15: 1,386; day 30: 563 (2 mice only yielded data up to day 15).**c.** Venn diagrams showing the proportions of consistently active neurons (neurons active across all retrieval sessions), partially active neurons (active in two sessions), and transiently active neurons (active in a single session). The proportions of cells in the different categories were similar across groups, with the exception of temporary cells on Day 30, which were more abundant in the control group. However, this difference reached significance only for the comparison between the control and CS+15 groups ($p < 0.05$; Table S1). Diagrams and analyses for control mice include only mice recorded for 30 days.

Sound-modulated PL population responses encode generalization gradients

To assess population-level representations in response to conditioned and novel tones, calcium activity was averaged across three presentations for each frequency during a 10-s baseline, 20-s tone presentation, and 10-s post-stimulus interval. In CS$^+$15 mice, positively modulated sound-responsive neurons exhibited graded tone activity reflecting learned contingency value across testing days (Fig. 3a, top). Area-under-the-curve (AUC) analyses obtained by averaging recorded cells per animal revealed the highest responses at 15 and 11 kHz, intermediate responses at 7 kHz, and the lowest at 3 kHz; with no difference between 11 and 15 kHz ($p > 0.05$) across testing days. This population level gradient mirrored behavioral generalization (Fig. 1b, left). By contrast, negatively modulated neurons showed overlapping responses across frequencies with no significant differences at any time point ($p > 0.05$; Fig. 3a, bottom). A complementary pattern was observed in CS$^+$3 mice, with graded responses that peaked at 3 kHz (Fig. 3b). As in CS$^+$15 mice, negatively modulated neurons exhibited overlapping responses across frequencies and days ($p > 0.05$). In no-shock controls, although both positive and negative responses were present, population activity was not modulated by tone frequency ($p > 0.05$; Fig. 3c, bottom panel), indicating that neuronal graded responses require associative learning.

To determine how AUC varied across groups over time, we performed a three-way mixed-effects ANOVA with group (CS+15, CS+3, and no shock), frequency (3, 7, 11, and 15 kHz), and time (test days 1, 15, and 30) as factors, with repeated measures on frequency and time. For positive responder neurons, the analysis revealed significant main effects of group ($p < 0.001$) and time ($p < 0.05$), as well as a significant group × frequency interaction ($p < 0.001$), whereas the time × frequency and group × time × frequency interactions were not significant ($p > 0.05$; Table S2a). Tukey-corrected post hoc comparisons showed that, in the CS+15 group, AUC differed between all frequency pairs except 11 and 15 kHz ($p < 0.05$). In the CS+3 group, the AUC at 3 kHz differed from those at 7, 11, and 15 kHz ($p < 0.05$), whereas no significant frequency differences were observed in the no-shock controls ($p > 0.05$). For negative responder neurons, the only significant effect was a time × frequency interaction ($p < 0.01$). Tukey-corrected simple-effects analyses revealed that, on day 30, the AUC at 15 kHz differed from those at 3, 7, and 11 kHz ($p < 0.05$; Table S2b). Because this pattern was observed across all experimental groups, including the no-shock controls, it is unlikely to reflect associative learning. These results indicate that although the AUC exhibited modest changes over time, these changes were not group-specific and therefore do not support learning-dependent alterations in neuronal responses. Together, these results show that despite substantial neuronal turnover, PL population responses encode generalization gradients, closely matching behavioral expression.

Consistently active neurons preserve stable generalization gradients, whereas newly recruited neurons refine them over time

Longitudinal tracking of individual neurons revealed how subpopulations with distinct stability profiles shape the evolution of representations that parallel behavioral expression. We compared population responses across three stability types: consistently active neurons (active during conditioning and all retrieval sessions), emerging–retained neurons (recruited after conditioning and persisting through day 30), and transiently active neurons (only active during a single retrieval session). Consistently active, positively modulated neurons exhibited graded population responses reflecting a generalization gradient from day 1 onward in both experimental groups, which was quantified by calculating AUC per animal ($p < 0.05$; Fig. S4a–b, Table S3). In contrast, negatively responding neurons showed no graded tuning across days ($p > 0.05$), except for a day 1 difference between 15 and 3 kHz in CS$^+$15 mice ($p < 0.05$). In control mice, positively and negatively modulated neurons displayed variable and inconsistent activity across frequencies in all cell categories (consistently active, emerging–retained, and transiently active). As a result, these neurons were not included in further cell-type analyses.

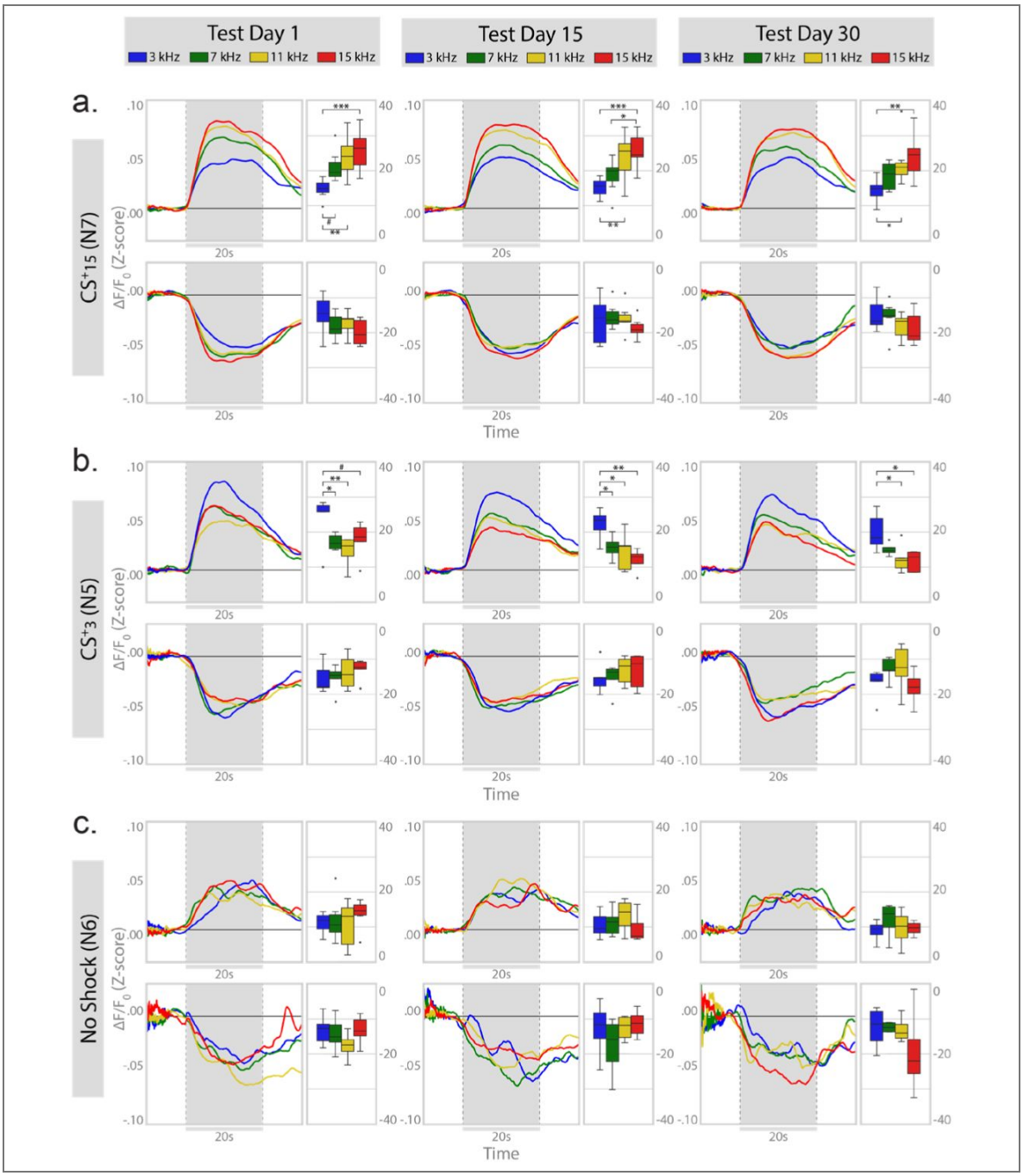

Figure 3. Population activity of positive sound responders shows emotional graded threat-value patterns of activity in response to tones.

a-c. Population responses in animals trained with a 15 kHz CS+ (a), a 3 kHz CS+ (b), and no-shock controls (c). In all groups, upper panels show positively responsive neurons and lower panels negatively responsive neurons. Boxplots show the median (center line), interquartile range (box), and whiskers extending to ±1.5× the interquartile range; points outside the whiskers represent individual observations beyond this range. CS+15: positively modulated: day 1: $F(3,18) = 9.963$, $p < 0.001$; day 15: $F(3,18) = 9.973$, $p < 0.001$; day 30: $F(3,18) = 6.627$, $p = 0.003$; negatively modulated: day 1: $F(3,18) = 2.483$, $p = 0.094$; day 15: $F(3,18) = 1.877$, $p = 0.178$; day 30: $F(3,18) = 2.753$, $p = 0.073$). CS+3: positively modulated: day 1: $F(3,12) = 6.899$, $p = 0.006$; day 15: $F(3,12) = 9.247$, $p = 0.002$; day 30: $F(3,12) = 6.123$, $p = 0.009$; negatively modulated: (day 1: $F(3,12) = 0.87$, $p = 0.484$; day 15: $F(3,12) = 0.512$, $p = 0.641$; Day 30: $F(3,12) = 1.448$, $p = 0.278$); no shock control: positively modulated: day 1: $F(3,15) = 0.527$, $p = 0.670$; day 15: $F(3,15) = 1.852$, $p = 0.181$; day 30: $F(3,9) = 1.046$, $p = 0.418$; negatively modulated: day 1: $F(3,13) = 1.205$, $p = 0.347$; day 15: $F(3,13) = 1.375$, $p = 0.294$; day 30: $F(3,9) = 0.95$, $p = 0.457$. Significant Tukey multiple comparisons are denoted by asterisks, *p < 0.05, **p < 0.01, ***p < 0.001.

Emerging-retained neurons recruited after conditioning exhibited graded population responses when recruitment occurred after day 1 (Fig. S5a–b, Table S3). This category encompassed neurons that emerged on day 1 and persisted through day 30, as well as neurons that emerged on day 15 and remained active through day 30. In both experimental groups, graded tuning was absent on day 1 ($p > 0.05$) but emerged on days 15 and 30, when population responses closely mirrored behavioral generalization, with responses to 11 and 15 kHz becoming increasingly similar regardless of CS^+ identity ($p > 0.05$). By contrast, emerging–retained negatively responding neurons showed no significant emotional tuning across days.

Finally, we examined transiently active neurons, defined as cells active only on a single retrieval session. Transiently positive responders showed no valence tuning when active on day 1 but exhibited clear graded population responses when recruited on days 15 or 30 in both experimental groups (Fig. S6a–b, Table S3). In contrast, transiently negative responders showed no significant frequency tuning at any time point ($p > 0.05$, Table S3). Together, these results indicate that consistently active neurons maintain stable generalization gradients across time, whereas neurons recruited after conditioning progressively sharpen their selectivity over time.

Tone identity is a stronger predictor of PL activity than freezing

Previous studies have shown that PL activity is associated with fear expression ((Burgos-Robles et al., 2009; Sierra-Mercado et al., 2011; Sotres-Bayon & Quirk, 2010), and more recent work has demonstrated that freezing can be decoded from PL neuronal activity (Casanova et al., 2024; Kyriazi et al., 2020). Because freezing covaries with learned threat value during fear generalization, however, it remains unclear whether PL neurons primarily encode behavioral expression or the learned threat value of auditory stimuli. Our experimental design provided a unique opportunity to dissociate these possibilities. The auditory stimuli were identical across sessions and animals, whereas freezing varied substantially across tones and retrieval sessions. Moreover, continuous acceleration extracted from the miniscope enabled probabilistic estimates of freezing without imposing an arbitrary threshold. To quantify the independent contributions of tone identity and freezing, we fitted a generalized linear model (GLM) to neuronal activity (calcium traces using tone identity and freezing probability as predictors.

We analyzed positively and negatively responsive neurons separately in the CS+15, CS+3, and control groups. For positively responsive neurons, freezing-corrected tone regression coefficients (β) exhibited opposite monotonic gradients that peaked at the CS+ in the two conditioning groups, whereas controls showed relatively flat coefficients (Fig. 4a; effect of groups: $p < 0.01$, group x frequency interaction: $p < 0.001$, Table S4a). To compare the relative contributions of tone identity and freezing, we calculated the freezing-to-tone β ratio. Across all groups, this ratio remained below 1 (Fig. 4b, Table S4b), indicating that tone identity contributed more strongly than freezing to neuronal activity. Consistent with this observation, only a modest fraction of neurons was classified as freezing-dominant, and this proportion remained stable across retrieval sessions, with no significant differences between groups ($p > 0.05$; Fig. 4c,d; Table S4c,d).

Negatively responsive neurons displayed a different pattern. Tone β coefficients showed little evidence of graded tuning ($p > 0.05$; Table S4e; Fig. 4e), consistent with the population responses shown in Fig. 3. Although tone identity remained the stronger predictor overall, freezing contributed proportionally more than in positively responsive neurons, resulting in higher freezing-to-tone β ratios and a larger fraction of freezing-dominant cells (Fig. 4f–h), with no differences between groups ($p > 0.05$, Table S4f). There were no differences in the proportion of freezing dominant negative responders between groups ($p > 0.05$) and these proportions remain stable across time ($p > 0.05$, Table S4g, h). Together, these findings demonstrate that PL population activity is better explained by learned threat value than by freezing behavior, particularly among positively responsive neurons.

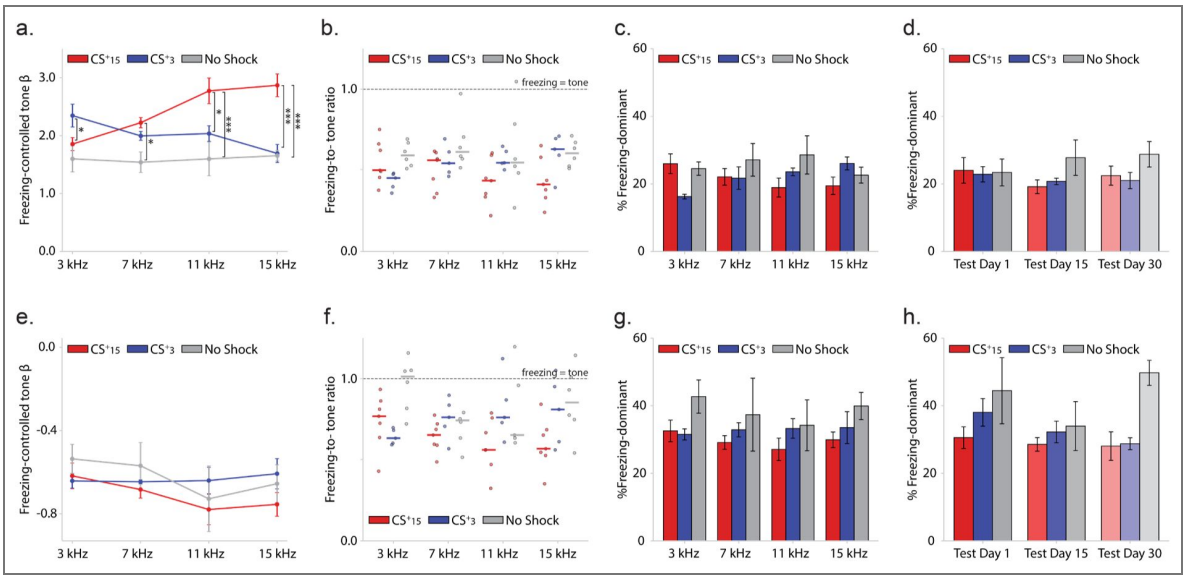

Figure 4. Generalized linear model (GLM) analysis of the relative contributions of tone identity and freezing behavior to neuronal activity.

a. Regression coefficients (β) for positive tone-selective neurons in conditioned mice exhibit opposite monotonic gradients across tones after controlling for freezing. This graded pattern is absent in non-shock controls. **b.** The freezing-to-tone ratio remains below 1 for every tone in all groups, indicating that tone identity explains more variance than freezing in positive tone-selective neurons. **c.** The proportion of freezing-dominant neurons is low across all tones, indicating that freezing does not dominate neuronal responses. **d.** When pooled across tones, the proportion of freezing-dominant neurons remains stable across retrieval sessions. **e.** Negative tone-selective neurons exhibit suppressive tone coefficients across all frequencies with only weak frequency dependence, indicating the absence of pronounced graded responses. **f.** Freezing-to-tone ratios are higher than in positive responders but remain below 1 across tones in conditioned mice. In non-shock controls, the ratio approaches 1, indicating that freezing contributes similarly to tone identity. **g.** The proportion of freezing-dominant neurons is higher than in positive responders but remains a minority across all tones and groups. **h.** The proportion of freezing-dominant neurons remains relatively stable across retrieval sessions for each group. Controls displayed a lightly higher proportion of freezing-dominant neurons but this difference was not significant. Significant Tukey multiple comparisons are denoted by asterisks, *$p < 0.05$, **p < 0.01, ***p < 0.001.

Population vector similarity at stimulus onset determines degree of generalization

The preceding results indicate that PL population activity is organized along generalization gradients, such that perceptually similar tones (i.e., adjacent frequencies) elicit similar population responses. However, the neural mechanisms that calibrate the similarity relationships remain unclear. Previous studies have shown that consistent neuronal responses across repeated stimulus presentations support stable population representations that facilitate generalization to similar cues (Hoshi et al., 2023). To assess population similarity within each retrieval session, we compared population activity across the three repeated presentations of each tone. Population activity was estimated using CASCADE, a validated spike-inference neural network (Rupprecht et al., 2021), and activity vectors were constructed from simultaneously recorded neurons. Response similarity between tone pairs was then quantified during stimulus presentation to assess the consistency of population activity across repeated trials.

Rate-based population similarity analysis revealed reliable, temporally structured responses for 3/3 kHz tone pairs in CS$^+$3 mice and for 15/15 and 15/11 kHz tone pairs in CS$^+$15 mice. In control mice high population similarity was absent for any frequency pair (Fig. 5a-d). Because population similarity peaked shortly after stimulus onset, we quantified similarity during the first 5 s after tone onset relative to the CS$^+$. In CS$^+$15 mice, early population similarity was highest for 15/15 and 15/11 tone pairs ($p < 0.001$), with no difference between them ($p > 0.05$), and was significantly greater than for 15/3 comparisons ($p < 0.05$, Fig. 5e). In CS$^+$3 mice, early similarity was highest for 3/3 tone pairs ($p < 0.004$) and significantly lower for 11/3 and 15/3 comparisons (Fig. 5f; $p < 0.05$). The 3/3 and 3/7 comparisons showed a non-significant trend ($p = 0.08$). No differences were observed among 7/15, 11/15, and 15/15 tone pairs in either group ($p > 0.05$). These findings indicate that population-level similarity at stimulus onset scales with behavioral threat generalization. In summary, PL population responses form generalization gradients that correlate with behavioral expression and encode learned threat value, while perceptual similarity across repeated presentations of threat-associated stimuli promotes similar population responses, providing a potential neural mechanism for recognizing dangerous stimuli.

Different subpopulations encode acoustic versus learned properties of sound association

Our previous analyses demonstrated that threat-value generalization gradients are represented at the population level. However, these findings do not reveal how these representations arise. Specifically, the observed population gradients could emerge either from the pooled activity of frequency-selective neurons that respond to individual tones or from neuronal subpopulations that integrate information across tones to encode their learned threat-value. To investigate these possibilities, we developed a clustering approach based on mutual information (MI), which captures both linear and nonlinear relationships between neuronal response profiles (Quian Quiroga & Panzeri, 2009). MI was used to group neurons with related activity patterns into functional subpopulations. Because MI does not preserve response polarity, neurons were subsequently classified by the sign of their pairwise correlations and re-clustered, consistently revealing two major response classes: positively and negatively modulated sound responders (Fig. 6a). Clustering was performed across all animals to derive global cluster identities and within each experimental and control group to identify learning-specific subpopulations (Figs. S7–S9).

Clustering quality was assessed by sorting signed MI values by global cluster identity and comparing within-versus across-cluster correlations (Fig. 6b). The resulting box-diagonal structure indicated robust clustering, with comparable quality indices across groups (CS$^+$15: 0.28; CS$^+$3: 0.27; controls: 0.30). In all cases, within-cluster correlations exceeded across-cluster correlations, supporting the reliability of the approach (CS$^+$15: 0.30 vs. −0.03; CS$^+$3: 0.27 vs. −0.02; controls: 0.28 vs. −0.04).

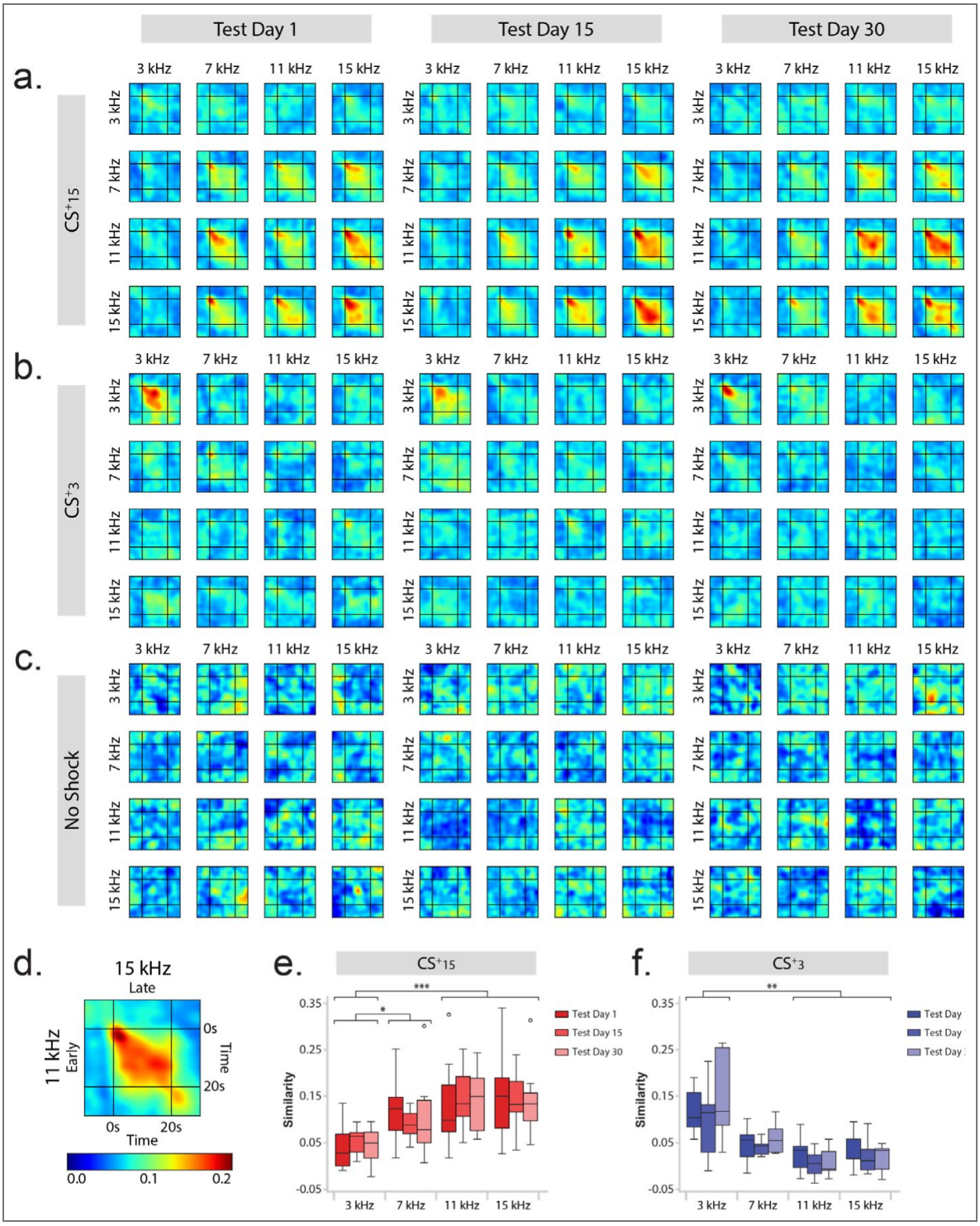

Figure 5. Population vector similarity across tones and time.

a-c. Population similarity maps for all tone pairs across the time course of sound presentation in the CS+15 (a), CS+3 (b), and no-shock control (c) groups. **d.** Schematic illustrating the color scale and map orientation; the y-axis denotes the earlier tone in the comparison, and the x-axis denotes the later tone. **e-f.** Box plots showing population similarity during the first 5 s following tone onset, quantified relative to the CS+ for the CS+15 (e, $F(3,36) = 12.025$, $p < 0.001$) and CS+3 (f, $F(3,24) = 7.435$, $p = 0.004$) groups. Significant Tukey multiple comparisons are denoted by asterisks, *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

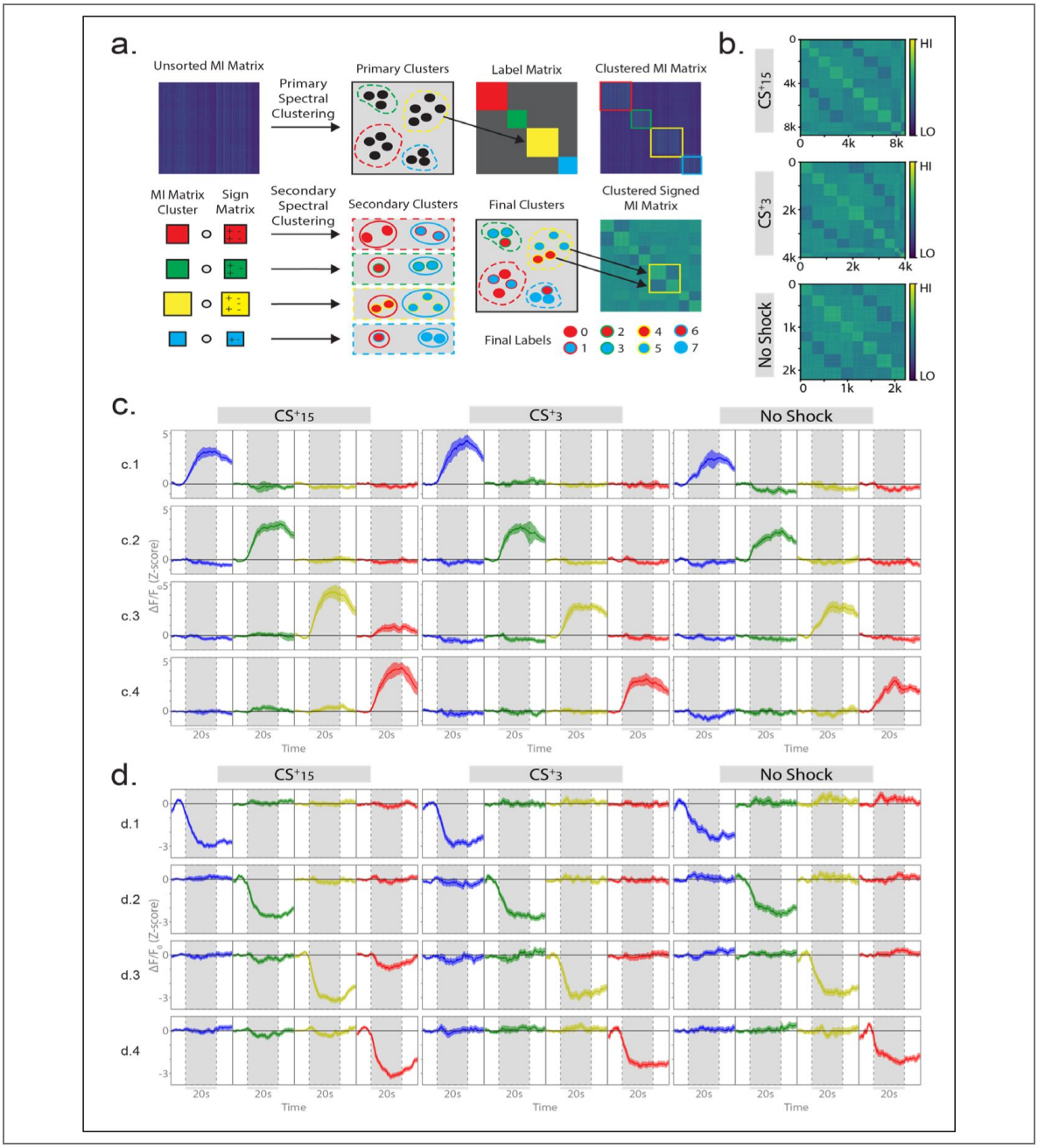

Figure 6. Clustering of PL subpopulations based on mutual information.

a. Schematic of the mutual information (MI)–based clustering pipeline. An unsorted MI matrix computed from simultaneously recorded prelimbic (PL) neurons was first subjected to spectral clustering to identify primary clusters based on shared information structure, independent of response sign. The MI matrix was then reordered according to these primary cluster assignments. Within each primary cluster, MI values were combined element-wise with a corresponding sign matrix encoding the direction of correlation between cell pairs, yielding a signed MI matrix. Spectral clustering was subsequently applied independently within each primary cluster to identify secondary subclusters with distinct signed interaction patterns. Each subcluster was assigned a unique label, and all labels were combined to generate the final clustered signed MI matrix, enabling separation of positively and negatively modulated sound-responsive neurons. **b.** Final clustered signed MI matrices for each experimental and control groups. Matrices are sorted by cluster labels for the CS+15 group (top), CS+3 group (middle), and No Shock group (bottom). Color scale indicates signed MI strength (HI to LO). **c.** Average stimulus-aligned population responses for clusters showing strong positive modulation to individual tones. C.1–C.4 show primary responses to 3 kHz, 7 kHz, 11 kHz, and 15 kHz tones, respectively, across groups. **d.** Average stimulus-aligned population responses for clusters showing strong negative modulation to individual tones. D.1–D.4 show primary responses to 3 kHz, 7 kHz, 11 kHz, and 15 kHz tones, respectively, across groups.

Across all groups, we identified positively and negatively modulated clusters selective for individual frequencies, indicating that frequency-specific representations are present in PL independent of learning (Fig. 6c–d; Figs. S7–S9), consistent with a role in auditory processing (Hockley & Malmierca, 2024; Zikopoulos & Barbas, 2006). In contrast, clusters encoding graded patterns—characterized by neurons responding to all frequencies in a graded manner—were observed exclusively in trained animals (Fig. 7; Figs. S7–S8). Graded neurons peaking at 15 kHz were specific to the CS$^+$15 group (Fig. 7a–b; Fig. S7), whereas neurons peaking at 3 kHz were unique to the CS$^+$3 group (Fig. 7c–e; Fig. S8).

If neurons encoding graded responses represent the learned contingencies, they should exhibit stability over time. To test this hypothesis, we quantified the proportion of registered neurons that retained their cluster identity across at least two retrieval sessions and compared these values to a shuffled null distribution (10,000 iterations), with multiple comparisons controlled using the Benjamini–Hochberg procedure. Only the positively modulated graded clusters showed significant identity stability across all retrieval intervals (days 1–15, 15–30, and 1–30; Fig. 7a-c).

In contrast, negatively modulated graded neurons and a smaller positively modulated cluster in the CS$^+$3 group showed stability only between adjacent sessions (Fig. 7b, d, e, Table S5). These data demonstrate that graded clusters remain consistently active, retaining their neuronal profiles, at levels exceeding chance, providing a stable representation of learned contingencies and generalization gradients.

Activity changes in frequency selective neurons encode learned value

Graded clusters encode learned contingencies but represent only a subset of the active neuronal population. Nevertheless, population-level representations, which incorporate all active neurons, robustly preserve these gradients. This led us to hypothesize that dynamically recruited, frequency-selective neurons also contribute to threat-value representations through changes in response magnitude. Consistent with hippocampal studies showing that spatially selective neurons encode task contingencies via firing-rate modulation (Gagliardi et al., 2024; Huxter et al., 2003; Sanders et al., 2019), we tested whether dynamic, frequency-selective clusters exhibited firing-rate changes proportional to learned threat value. To do so, we quantified the normalized baseline-to-stimulus response ratio (BSR) of positively responding neurons within each cluster using CASCADE-inferred spikes (Rupprecht et al., 2021).

We found that although tone-selective clusters emerged independently of learning (i.e., they were present in both conditioned and control mice), their firing rates were shaped by associative learning. Specifically, 3-kHz-selective neurons exhibited higher BSR in both conditioned groups than in controls ($p < 0.05$), regardless of whether 3 kHz served as the CS$^+$ or CS$^-$. In contrast, 11-and 15-kHz-selective neurons displayed significantly higher BSR in the CS$^+$15 group than in the CS$^+$3 or control groups on days 15 and 30 ($p < 0.05$; Fig. S10, Table S6). Notably, in the CS$^+$15 group, where 11 kHz elicited robust threat generalization, the 11- and 15-kHz-selective clusters exhibited similarly high BSRs to both tones. The higher activity rates observed for tones associated with threat or inducing generalization suggests that these neurons are shaped by the animal's associative experience, including both conditioning and retrieval-dependent updating.

As expected, positively modulated graded neurons closely mirrored population responses, exhibiting the highest BSRs to the CS$^+$ and strongly generalized tones and lower BSRs to discriminated tones ($p < 0.05$; Fig. 7h–i, Table S6). Together, these findings indicate that firing-rate modulation enables both graded and frequency-selective PL neurons to represent the learned contingencies.

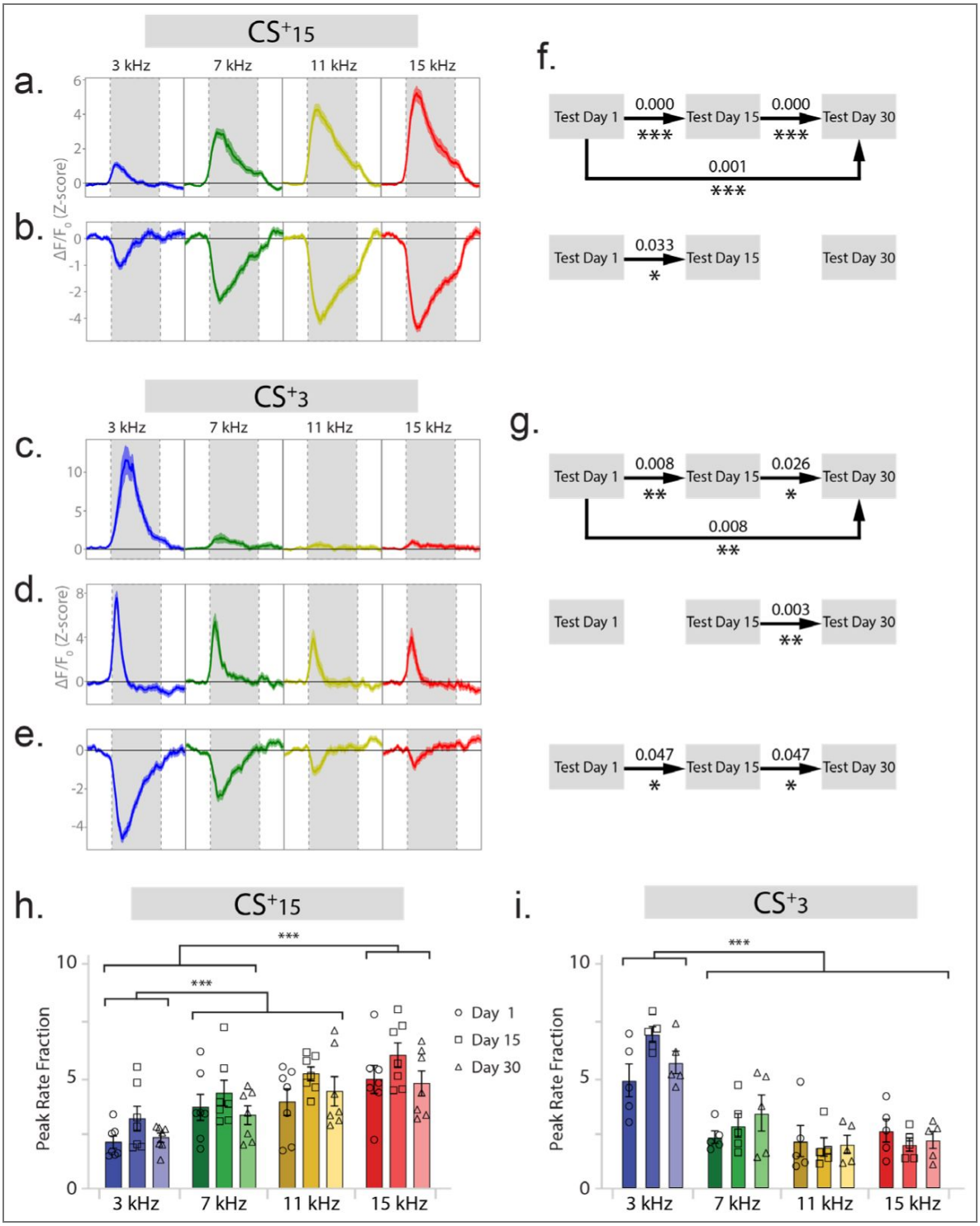

Figure 7. Graded-tone responsive neurons are only present in experimental groups.

a-b. Average stimulus-aligned population responses for clusters showing graded tone responses in animals trained with a 15 kHz CS+, showing positive (a) and negative (b) response patterns. **c-e.** Average stimulus-aligned population responses for clusters showing graded tone responses in animals trained with a 3 kHz CS+, showing positive (c-d) and negative (e) response patterns. **f-g.** Analysis of stability of neuronal identity within graded tone clusters across retrieval sessions. Benjamini–Hochberg corrected significance denoted by asterisks. **h-i.** Baseline/stimulus firing rate ratio (BSR) showing changes in activity of graded clusters only present in the experimental CS+15 (h) and CS+3 (i) groups across days. Significant Tukey multiple comparisons are denoted by asterisks, *p < 0.05, **p < 0.01, ***p < 0.001.

Stable and dynamic neuronal subpopulations encode learned threat value beyond freezing

Our earlier GLM analysis established that tone identity explains neuronal activity more strongly than freezing in the overall PL population, we next asked whether this relationship differed among the neuronal subpopulations identified by our clustering analysis. We first examined frequency-selective (single-tone) neurons, assigning each neuron's preferred tone according to its cluster identity. Summary statistics were calculated using animals as the experimental unit (CS+15: N = 7; CS+3: N = 5; controls: N = 6).

Among positively modulated single-tone neurons, freezing-corrected tone β coefficients increased monotonically toward the CS+ in opposite directions for the CS+15 and CS+3 groups, whereas coefficients remained relatively flat in controls (Fig. 8a; effect of groups: $p < 0.001$; group x frequency interaction: $p < 0.002$, Table S7a). Thus, although these neurons responded preferentially to a single frequency, their β coefficient magnitude scaled with the learned threat value of that tone rather than its absolute frequency.

Consistent with the population GLM, freezing-to-tone β ratios remained well below 1 (median ≈ 0.4; Fig. 8b), indicating that tone β coefficients were approximately two-to threefold larger than freezing β coefficients. This effect, however, was larger in experimental groups (effect of groups: $p < 0.001$, Tukey comparisons: CS+15 vs control: $p < 0.008$, CS+3 vs control: $p = 0.08$, Table S7b). Likewise, within-cell comparisons showed larger median tone than freezing β coefficients across all groups for each animal (Fig. 8c; effect of group: $p < 0.02$; effect of state condition (tone vs freezing): $p < 0.001$; interaction: $p < 0.002$, Tukey multiple comparisons: Tone vs. freezing $p < 0.001$ for all groups; β coefficients for tone: CS+15 vs controls: $p < 0.00$, CS+15 vs CS+3: $p < 0.03$, Table S7c), and only a small, stable fraction of neurons was classified as freezing-dominant throughout retrieval, with no significant differences between groups across time ($p > 0.005$; Table S7d; Fig. 8d). Together, these findings indicate that positively modulated frequency-selective neurons primarily encode learned threat value despite their narrow frequency selectivity.

Negatively modulated single-tone neurons exhibited suppression at their preferred frequency but little evidence of systematic scaling with learned threat value (Fig. 8e, effect of groups: $p < 0.001$, interaction frequency x group: $p > 0.05$; Tukey multiple comparisons: CS+15 vs control: $p < 0.001$; Table S7e). Tone identity remained the stronger predictor of activity across groups, but freezing-to-tone β ratios were lower in the experimental groups than the control (Fig. 8f, effect of group: $p < 0.001$, Tukey multiple comparisons: CS+15 vs control: $p < 0.001$, CS+3 vs control: $p < 0.006$; Table S7f). Within-cell comparisons of tone and freezing β coefficients showed a difference in the experimental groups but not the controls (Fig. 8g, effect of groups: $p < 0.05$, condition (tone vs freezing): $p < 0.001$, interaction: $p < 0.001$, Tukey multiple comparisons: tone vs. freezing: CS+15: $p < 0.001$; CS+3: $p < 0.01$; control: $p=0.977$, Table S7g). Finally, the proportion of freezing-dominant neurons was highest in controls, although this difference reached significance only relative to the CS+15 group (Fig. 8h, effect of group: $p < 0.005$, Tukey tests: CS+15 vs control: $p < 0.005$, Table S7h). These findings suggest that negatively responsive frequency-selective neurons multiplex stimulus identity and behavioral state to a greater extent than positively responsive neurons.

We next applied the same analysis to neurons classified as graded (Fig. 7). Because graded clusters were absent in controls, this analysis included only the CS+15 and CS+3 groups. For each graded cell we obtained normalized freezing-corrected β coefficients by dividing each tone beta by the largest beta out of the 4. The normalized freezing-corrected β coefficients formed steep opposing monotonic gradients that tracked learned threat value in both conditioning groups (Fig. 8i; effect of groups: $p < 0.001$, interaction group x frequency: $p < 0.001$; Tukey multiple comparisons: CS+15 vs CS+3 for 3, 11, and 15 kHz: $p < 0.001$; CS+15 vs CS+3 for 7 kHz: $p < 0.02$, Table S7i). All freezing-to-tone β ratios remained well below 1 across time (Fig. 8j), with the ratios for the CS+15 being even lower than for CS+3 on days 2 and 30 ($p < 0.05$, Table S7j). The within-cell median β coefficients for freezing and tone were different in both experimental groups ($p < 0.001$; Fig. 8k), with the tone coefficients being higher in the CS+15 group ($p < 0.003$, Table S7k). Only 4–7% of neurons were classified as freezing-dominant across retrieval sessions (Fig.

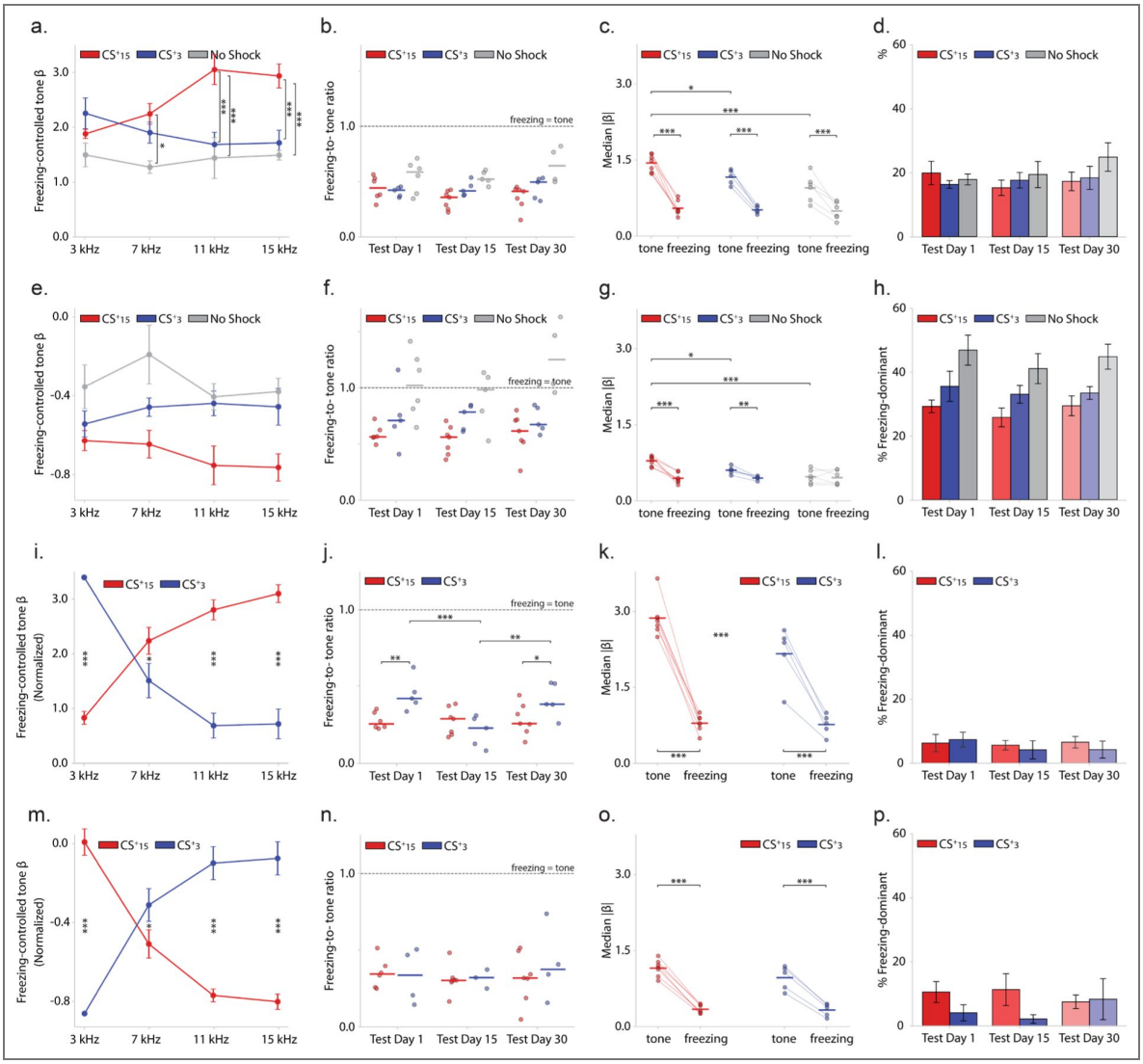

Figure 8. GLM analysis of clustered neurons.

a–h, **Positive (a–d) and negative (e–h) single tone-selective neurons. a.** Tone β coefficients in conditioned mice form opposite gradients peaking at the CS+ that are absent in non-shock controls. **b.** The freezing-to-tone ratio remains below 1 across preferred tones and retrieval sessions. **c.** Within-cell comparisons confirm that tone coefficients exceed freezing coefficients in all groups. **d.** The proportion of freezing-dominant neurons is low and stable across retrieval sessions. **e.** Negative tone-selective neurons exhibit suppressive β coefficients with weak frequency dependence and no graded responses in controls. **f.** The freezing-to-tone ratio remains below 1 but is higher than in positive responders and more variable in non-shock controls. **g.** Tone responses exceed freezing responses but by a smaller margin than in positive responders. **h.** The proportion of freezing-dominant neurons is higher than in positive responders but remains stable across retrieval sessions. **i–p. Positive (i–l) and negative (m–p) graded neurons. i.** Freezing-controlled tone β coefficients exhibit strong CS+-centered gradients. **j.** The freezing-to-tone ratio is well below 1, indicating that tone identity explains substantially more variance than freezing. **k.** Within-cell comparisons confirm that tone coefficients greatly exceed freezing coefficients. **l.** Only a small, stable fraction of positive graded neurons is freezing-dominant. **m.** Negative graded neurons exhibit strong suppressive gradients. **n**. The freezing-to-tone ratio remains below 1. **o.** Tone responses exceed freezing responses by approximately threefold. **p.** The proportion of freezing-dominant neurons remains low across retrieval sessions. Significant Tukey multiple comparisons are denoted by asterisks, *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

8l, Table S7l), a lower percentage than that observed in single-frequency neurons. Similar patterns were observed among negatively modulated graded neurons, although the difference between tone and freezing coefficients was smaller (Fig. 8m–p, Table 7Sm-Sp).

These analyses demonstrate that both dynamically recruited frequency-selective neurons and stable graded neurons encode learned threat value beyond freezing behavior. Graded neurons represented threat value with minimal influence of behavioral state, whereas frequency-selective neurons also encoded threat value, although their responses were more strongly modulated by freezing. Together, these complementary coding strategies likely enable PL populations to maintain stable representations of learned threat value despite ongoing changes in ensemble composition and behavioral state.

Discussion

Our findings suggest that sensory similarity and learned threat value are not competing explanations of threat generalization but complementary components of the neural code underlying this process. Associative learning generated graded population representations of learned contingencies and intermediate responses to novel tones that persisted despite substantial ensemble reorganization. Perceptually similar tones elicited similar population responses with the greatest consistency across stimulus presentations, indicating that sensory similarity promotes stable population representations of behaviorally significant stimuli. Importantly, these representations were not simply a reflection of freezing behavior. Although freezing contributed modestly to PL activity, GLM analyses showed that learned threat value associated with each tone explained substantially more variance in neuronal responses than freezing, with graded tone β coefficients persisting after accounting for freezing. This stable graded population code emerged through the complementary contributions of two neuronal populations: longitudinally stable graded neurons that consistently encoded learned threat value and dynamically recruited, frequency-selective neurons whose response magnitudes were modified by learning. Together, these findings suggest that the PL integrates sensory similarity with learned threat value to generate stable representations that support adaptive generalization and discrimination.

Generalization has traditionally been explained by perceptual similarity (Shepard, 1987), whereby stimuli resembling a conditioned cue recruit overlapping sensory representations and evoke similar behavioral responses (Corches et al., 2019; Grosso et al., 2018). Although perceptual similarity clearly influences the extent of generalization, accumulating evidence indicates that it cannot fully account for generalized responding (Verra et al., 2026). More recent frameworks propose that associative learning assigns learned value to novel stimuli by integrating their sensory similarity with previous experience, allowing behavior to scale according to predicted biological significance (Verra et al., 2026; Zaman et al., 2023). Our findings provide a neural framework consistent with these ideas. Sensory similarity promoted consistent neuronal population responses across tones, whereas associative learning organized these responses into graded representations that tracked learned threat value across the stimulus continuum. Thus, sensory similarity appears to define the neuronal substrate upon which associative learning constructs value-based representations that support graded behavioral generalization.

The ensemble turnover observed here is consistent with previous studies demonstrating dynamic reorganization of cortical activity patterns over time (DeNardo et al., 2019; Gallego et al., 2020; Kitamura et al., 2017; Tome et al., 2024). Such reorganization has been proposed to provide flexibility by allowing new information to be incorporated into existing cortical representations while preserving stable behavioral performance (Mau et al., 2020; Zaki & Cai, 2024). Several mechanisms could contribute to this turnover, including systems consolidation, retrieval-induced reconsolidation or memory updating, and repeated nonreinforced stimulus exposure (Lacagnina et al., 2019; Mau et al., 2020; Sangha, 2015; Zaki & Cai, 2024). Although our experiments cannot distinguish between the first two possibilities, the behavioral data argue against extinction as the primary explanation. Extinction is generally associated with the formation of new CS+-safety associations (Bouton et al., 2021), whereas discrimination ratios increased across retrieval sessions, indicating that animals progressively improved their

discrimination between threat-associated and safe stimuli rather than acquiring generalized safety responses. This pattern is consistent with previous work showing that discrimination learning sharpens stimulus representations and narrows behavioral generalization gradients (Dunsmoor & LaBar, 2013; Herzog et al., 2021; Jenkins & Harrison, 1960; Lommen et al., 2017). Importantly, turnover was not uniform across the population. Graded neurons retained remarkably consistent response profiles across retrieval sessions, and their activity remained more strongly associated with learned threat value than with freezing behavior. These observations indicate that stable components of the population code can coexist with extensive reorganization of surrounding neuronal ensembles.

These findings also inform current views of systems consolidation and cortical plasticity (Frankland & Bontempi, 2005; Lopez et al., 2024). Classical models propose that memory initially depends on subcortical structures before gradually becoming stabilized within neocortical circuits (Moscovitch & Nadel, 1998; Nadel et al., 2000). In contrast, multiple-trace and related frameworks emphasize the rapid formation of cortical memory traces and continuous cortical reorganization throughout the lifetime of a memory (Moscovitch & Nadel, 1998; Nadel et al., 2000; Tonegawa et al., 2018). Our findings are consistent with elements of both views. Representations of learned threat value emerged shortly after conditioning and remained stable within a subset of neurons despite substantial reorganization of the broader ensemble. At the same time, narrower generalization gradients and improved discrimination across retrieval sessions suggests ongoing memory updating. These observations are consistent with contemporary theories proposing that systems consolidation and retrieval-dependent updating are complementary processes through which memories continue to evolve after learning (Mau et al., 2020; Tome et al., 2024; Zaki & Cai, 2024). Our results therefore suggest a mechanism by which stable cortical representations can be preserved despite continuous neuronal reorganization, providing both the stability required for long-term memory and the flexibility needed to generalize and discriminate newly encountered stimuli.

The persistence of graded neuronal responses despite extensive ensemble turnover also has implications for understanding long-term cortical coding. Although our experiments do not establish that the stable graded neurons are memory engrams or are causally required for memory expression, they demonstrate that stable representations of learned contingencies can coexist with pronounced changes in ensemble membership. Furthermore, the fact that the GLM analysis indicates that these neurons reflect learned threat value more than freezing behavior, suggests that they encode an abstract property of the learned stimulus rather than simply mirroring behavioral output. Future studies combining longitudinal imaging with selective manipulation of stable graded neurons will be needed to confirm whether these neurons are causally required for memory storage or retrieval and whether they constitute a persistent cortical memory trace.

Notably, frequency-selective neurons were present in both conditioned and control animals, indicating that these response properties likely reflect pre-existing sensory tuning within the PL rather than neuronal populations generated through learning (Hockley & Malmierca, 2024; Zikopoulos & Barbas, 2006). Associative learning instead appeared to modify the gain of these sensory-responsive neurons, producing progressively stronger responses to threat-associated than to safety-associated tones while preserving their underlying frequency selectivity. This suggests that the inferred firing rate of these neurons is shaped by the animal's associative experience, including both conditioning and retrieval-dependent updating. This interaction between stable graded neurons and dynamically modulated sensory-responsive neurons provides a potential mechanism by which novel stimuli can be evaluated according to both their sensory similarity and their learned significance, thereby supporting flexible generalization and discrimination despite ongoing population reorganization.

In the auditory cortex, sound-evoked suppressive ("negative") responses contribute to lateral inhibition by sharpening frequency tuning and improving sensory discrimination (Kato et al., 2017; Wehr & Zador, 2003). Although negative responders did not exhibit robust graded responses at the population level, we identified negatively graded clusters whose response profiles

mirrored those of positively graded clusters across the stimulus continuum. These observations raise the possibility that suppressive neuronal populations provide complementary inhibitory contrast that enhances discrimination while preserving graded representations of learned threat value, although testing this hypothesis will require future causal experiments.

Together, our findings support a model in which sensory similarity and associative learning make complementary contributions to threat generalization. Sensory similarity determines the consistency between population responses produced by the same stimuli, whereas associative learning transforms these responses into graded representations of learned threat value that remain stable despite substantial ensemble turnover through the complementary contributions of stable graded and dynamic frequency-selective neurons. These findings provide a neural framework for understanding how the PL supports adaptive generalization and discrimination.

Methods

Subjects

Female and male C57BL/6J mice (IMSR_JAX:000664; Jackson Laboratory, Bar Harbor, ME), aged 8–10 weeks, were housed on a 12 h light/dark cycle. All experiments were conducted during the light phase of the cycle. Animal housing and care were consistent with standards set by the Association for Assessment and Accreditation of Laboratory Animal Care (AAALAC). All experimental procedures were approved by the University of Iowa Institutional Animal Care and Use Committee.

Discriminatory fear conditioning task

The training day, mice were placed in a square conditioning chamber (20 × 20 cm; Context A, Fig. 1a). The first 3 minutes served as a habituation period. Following habituation, experimental mice received presentations of a conditioned stimulus (CS^+) tone (15 or 3 kHz, 80 dB, 20.5 s) that co-terminated with a mild foot shock (unconditioned stimulus, US; 0.5 s, 0.5 mA). Between CS^+ presentations, animals were presented with a non-conditioned stimulus (CS^-) tone (3 or 15 kHz, 80 dB, 20.5 s) that was never paired with shock. Mice received 10 CS+ tone–shock pairings and 10 CS− tone presentations. Intertrial intervals between CS^+ and CS^- presentations alternated pseudo-randomly between 2, 4, and 6 minutes. Following the final CS^- presentation, mice remained in the chamber for an additional 60 s before being returned to their home cages. Control mice were exposed to the same auditory stimuli but did not receive foot shocks. On day 1, mice were placed in a novel context (Context B; 20 × 20 cm; see Fig. 1a) that differed from the training context in wall color, floor texture, and background scent. Animals were allowed to freely explore the chamber for 3 minutes, after which they were presented with auditory stimuli consisting of the conditioned tone (CS^+) the non-conditioned tone (CS^-), and two intermediate frequencies (7 and 11 kHz) spanning the same spectral range. These intermediate frequencies were spaced linearly rather than logarithmically. Tones were presented in a semi-random order, with three presentations per frequency and 3-min intertrial intervals. Following the final tone presentation, mice remained in the chamber for an additional 60 s period before being returned to their home cages.

Behavioral analysis

All behavioral sessions were video recorded using a mounted camera. Freezing behavior was defined as complete immobility except for respiratory movements and was quantified as a percentage of total trial time. Freezing was measured using FreezeFrame 5 software (Actimetrics, *RRID:SCR_014429*) for non-implanted mice and inertial measurement unit (IMU) data from the Inscopix miniscope system for implanted mice. This approach quantifies movement across three axes and circumvents visual occlusion introduced by the miniscope tether. To validate automated scoring, a subset of sessions from implanted and non-implanted mice were randomly selected and manually scored by an experimenter blind to experimental condition, following established criteria (Phillips & LeDoux, 1992; Wang et al., 2012). Automated and manual measurements

were highly correlated (Pearson's r = 0.98 and 0.96), confirming strong inter-method reliability. Freezing responses were analyzed across three retrieval sessions (days 1, 15, and 30). Two animals in the non-shock control group (JAM061 and JAM062) were excluded from repeated-measures analyses due to missing data on day 30 but were retained for day-specific comparisons.

Viral construct stereotaxic surgery

Mice were anesthetized with isoflurane (3% in oxygen at 1 L/min) and placed in a stereotaxic apparatus (David Kopf Instruments, Tujunga, CA). During surgery, anesthesia was maintained at 1.5%. The head was positioned to ensure a flat skull in both the anterior–posterior (AP) and medial–lateral (ML) planes. Rimadyl (5 mg/kg, s.c.) was administered preoperatively and for two consecutive days postoperatively for analgesia. For calcium imaging experiments, three injections (150 nL each) of AAV1-CaMKIIα-GCaMP6f-WPRE-SV40 (Addgene cat. # 100834) were delivered into the PL at the following stereotaxic coordinates relative to bregma: AP +1.95 mm, ML ±0.3 mm, and DV −1.7, −1.5, and −1.3 mm. Injections were performed at a rate of 50 nL/min. After each injection, the syringe was left in place for 5 minutes to allow for viral diffusion and to minimize backflow. Mice were allowed to recover for four weeks prior to implantation of the GRIN lens and baseplate assembly for in vivo imaging.

Miniendoscope baseplate placement

Following AAV-GCaMP6f injections, a gradient-index GRIN lens (Inscopix, diameter: 1.0 mm, length: 4.0 mm, numerical aperture: 0.5, model: 1050-004637) and baseplate assembly (Inscopix, Mountain View, CA) were implanted above the PL. Stereotaxic coordinates relative to bregma were AP +1.94 mm, ML ±0.3 mm, and DV −1.5 mm. The cortical surface was identified and set as the zero reference point. To create a tract for lens placement, a 26-gauge syringe needle was lowered to approximately two-thirds of the final depth (DV −1.0 mm) and then slowly retracted. The GRIN lens was subsequently lowered to the final DV coordinate (−1.5 mm) and secured in place with a thin layer of cyanoacrylate adhesive applied around the lens perimeter and over the exposed skull. Dental cement (Lang Dental) was used to further stabilize the baseplate and anchor it to the skull. Mice were allowed to recover for at least two weeks before the onset of behavioral training.

Calcium imaging analysis

Calcium imaging data acquired using the Inscopix nVoke 2.0 miniscope system were processed using Inscopix Data Processing Software (IDPS; Inscopix, Mountain View, CA). Raw recordings were first preprocessed using the standard IDPS pipeline, which included spatial downsampling, background subtraction, spatial filtering, and rigid motion correction to compensate for movement artifacts. Fluorescence signals were then normalized to ΔF/F, defined as the change in fluorescence relative to a baseline fluorescence level computed for each pixel, to quantify activity-dependent calcium dynamics. Following preprocessing, neuronal signals were extracted using a constrained non-negative matrix factorization approach optimized for one-photon calcium imaging (CNMF-E implemented by IDPS). This method decomposes the imaging data into spatial components corresponding to putative neurons and their associated temporal activity traces while accounting for background and neuropil contamination. Extracted components were manually curated to exclude non-neuronal signals and artifacts based on established morphological and activity criteria (Fig. 1d).

Neuronal activity was aligned to tone onset for all analyses, whereas signal representation, normalization, baseline correction, and temporal binning were adapted to the analytical objective. Sound responder classification, average stimulus-aligned traces, population trace analyses, stimulus response vectors, and the GLM were performed on ΔF/F fluorescence signals, whereas population similarity and baseline-to-stimulus ratio (BSR) analyses used CASCADE-deconvolved activity. Depending on the analysis, normalization consisted of either trial-wise or session-wise z-scoring, with baseline correction subsequently applied when quantifying stimulus-evoked responses. Analysis-specific processing steps are summarized in the table below and details provided in the relevant method subsections.

Analysis	Input	Normalization	Temporal processing	
Sound responder classification	CNMF-e Traces	Trial-wise z-score	Sliding windows	
Population traces	CNMF-e Traces	Session z-score	Averaged SAT (average-stimulus aligned procedure)	
Stimulus response vectors	CNMF-e Traces	Session z-score	Concatenated responses	
GLM	CNMF-e Traces	Session z-score	Frame-by-frame	**Sound responder**
Population similarity	CASCADE	Session z-score	1-s windows	
BSR	CASCADE	Session z-score	Session average	

classification

Sound responder classification

To quantify stimulus-evoked modulation of neuronal activity, we implemented a window-based sound response test (SRT) that detects changes in calcium activity relative to the pre-stimulus baseline while preserving the temporal structure of the response.

Data alignment and normalization

For each neuron, ΔF/F traces were aligned to tone onset and resampled onto a common relative time axis spanning a pre-stimulus baseline (10s before tone onset) and the 20-s stimulus period. All time points within that trial were normalized by subtracting the baseline mean and dividing by the standard deviation of that trial's baseline across time. A small constant was added to the denominator to prevent division by zero. This within-trial normalization places all trials on a common scale prior to being combined and is distinct from the between-trial term described below, which standardizes the reliability of the baseline across trials.

Sliding window construction

To capture temporally localized stimulus responses, normalized traces were segmented into overlapping sliding windows spanning the stimulus period. Windows were defined by a fixed duration (1 s) and stride (0.1 s), with window boundaries determined directly from the sampling interval of the data. Only windows fully contained within the stimulus epoch were included. This approach enabled detection of responses with variable onset latencies and durations without assuming a fixed response time.

Window-wise response quantification

For each window, the median normalized activity was computed and compared with the corresponding trial baseline median. Baseline-subtracted responses were then averaged across trials to obtain a mean response magnitude for each window. Separately, to down-weight windows in cells whose baseline was unreliable across repetitions, a between-trial standardization was computed for each window as the mean baseline-subtracted response divided by the standard deviation of the per-trial baseline medians across trials. This computation indexes trial-to-trial baseline reliability rather than within-trial fluctuation, and was used only to gate which windows entered the significance test (see below); the response magnitude used to classify sound responders is the trial-averaged, baseline-relative response itself, already expressed in units of each trial's baseline standard deviation. For single-trial cases, where a between-trial standard deviation is undefined, the trial-averaged baseline-relative response was used directly.

Statistical comparison to baseline

For each window, normalized activity values were compared with the pooled baseline distribution across trials using a two-sample Kolmogorov–Smirnov test. This nonparametric test was selected because it detects differences in both central tendency and distribution without assuming normality.

Identification of significant response epochs

A window was classified as significant if it met two criteria: (i) a p-value below a predefined threshold (typically $p < 0.05$) and (ii) a between-trial standardized response of at least one baseline standard deviation. Significant windows were classified as positive or negative according to the sign of the response, and contiguous windows of the same sign were merged into response epochs, allowing gaps of one window or less to account for noise. For each neuron and response direction, the longest response epoch was retained.

Summary response metrics

For each response epoch, we extracted the trial-averaged baseline-relative response and associated duration. Based on these metrics, responses were classified as positively modulated (trial-averaged baseline-relative response ≥ 3 and significant for at least 1 s) or negatively modulated (baseline-relative response ≤ −1.5 and significant for at least 1 s), where the response is expressed in units of each trial's baseline standard deviation.

Mapping active ensembles across time

For longitudinal analyses, neuronal identity across imaging sessions was tracked using spatial footprint registration to identify putatively identical neurons across days. Spatial footprints of cells detected during conditioning were registered longitudinally using a probabilistic model implemented in CellReg [Inscopix, (Evangelidis & Psarakis, 2008)]. This algorithm aligns neurons based on the similarity of their spatial footprints, enabling consistent tracking of neuronal identity across imaging sessions. To evaluate registration quality, we used the conditioning session as the reference and measured the centroid distances between each neuron and its corresponding match in each retrieval session. For each animal, we generated histograms of these centroid shifts (expressed in micrometers) and compared their distributions with the median equivalent footprint diameter of the neurons. To estimate this value, we thresholded each neuronal footprint at half of its maximum intensity and calculated the diameter of an equivalent circle having the same area. Cells classified as *consistently active* were those reliably detected across all phases of conditioning and retrieval, including sessions on days 1, 15, and 30.

Average stimulus-aligned trace procedure

For each neuron and tone frequency, session z-scored ΔF/F traces were aligned to tone onset, and 40-s epochs (10 s pre-stimulus, 20 s stimulus, 10 s post-stimulus) were extracted. Epochs were linearly interpolated to a common time base (to allow for later averaging), after which the mean activity during the pre-stimulus period was subtracted to baseline-correct each epoch. Baseline-corrected epochs were then averaged across repetitions to generate a single average stimulus-aligned trace (ASAT) for each neuron and tone.

Generation of Population Sound-Response Curves

Population sound-response curves were generated by first averaging ASATs across neurons within each animal and condition and then averaging these animal-level traces across animals within each experimental group, representing the population response to each auditory stimulus.

Population similarity across repeated tone presentations

CASCADE-inferred firing rates were Gaussian-smoothed (2-s kernel), z-scored within each session, and averaged within overlapping 1-s windows (0.5-s overlap) to generate population activity vectors. Population similarity matrices were computed by correlating these vectors across independent presentations of the same tone (within-tone comparisons) and different tones (cross-tone comparisons). Similarity matrices were not symmetrized, preserving asymmetries arising from stimulus order (e.g., 11 vs. 3 kHz or 3 vs. 11 kHz), tone identity, and stimulus-evoked population dynamics. Matrix axes represent time relative to tone onset (−10 to 30 s), with tone presentation occurring from 0 to 20 s. The earlier tone presentation is plotted on the y-axis and the later presentation on the x-axis. To quantify similarity during stimulus onset, correlation values

from the first 5 s of tone presentation were averaged for each animal and compared across tone pairs relative to the CS+ in each experimental group.al, focusing on tone pairs relative to the CS+ in each experimental group.

Classification of subpopulation tone-response patterns

Stimulus response vectors

For each neuron, baseline-corrected ΔF/F responses were extracted from 40-s stimulus-aligned epochs (10 s pre-stimulus, 20 s stimulus, 10 s post-stimulus), interpolated to a common time base, and averaged across repetitions for each tone frequency. Responses to the four frequencies were then concatenated in ascending order to generate a composite response vector for each neuron.

Mutual information matrix

For each pair of cells, we computed the mutual information (MI) between their stimulus response vectors to quantify the statistical dependence between their responses. MI measures the reduction in uncertainty about one variable given knowledge of another. In this context, low MI values indicate that the stimulus response vectors of two cells are largely independent, whereas high MI values indicate shared information or coordinated response structure. The MI values for all cell pairs were assembled into a symmetric matrix (n_cells × n_cells), with each entry representing the information shared between a given pair of cells. Separate MI matrices were computed for each experimental group (non-shock, CS+3, CS+15), pooling cells across recording days within each group. In what follows, “MI matrix” refers to one of these matrices.

Adjacency matrix construction

To transform the MI matrix into a network graph representation, we applied a summed top-k adjacency procedure. For each cell (matrix row), connections to other cells were ranked by MI strength, and only the top-k values were retained. Each retained connection was weighted by the reciprocal of its rank (1 / rank), favoring the strongest associations. The resulting matrix was symmetrized to represent an undirected network of functional connectivity (if cell A is functionally connected to cell B then this enforces that B is also functionally connected to A).

Spectral clustering

The adjacency matrix was analyzed using spectral clustering to identify functional subpopulations. Spectral clustering converts the adjacency matrix into a graph Laplacian, computes its eigenvalues and eigenvectors, and projects the data into a low-dimensional space in which clusters are more separable. K-means clustering was then applied to the eigenvector representations to delineate groups of highly connected neurons. This approach is well suited for detecting non-linear or irregular structures in neuronal co-activity subpopulations.

Hyperparameter selection

Two hyperparameters were determined: the top-k value and the number of clusters. We evaluated a range of top-k values (2 to 100) and, for each, computed the first 30 eigenvalues of the normalized graph Laplacian obtained during spectral clustering. Eigenvalues were averaged across top-k values to yield a single representative spectrum. The optimal cluster number was then determined using the eigengap method, defined as the largest gap between consecutive eigenvalues. The top-k values associated with this solution were averaged to obtain the final k.

Positively and negatively modulated subpopulations

Because MI is non-directional, it cannot distinguish between positively and negatively correlated responses. To separate these responses, Spearman’s correlation was computed for all pairs of cells within the cluster, and only the sign of each correlation was retained, resulting in a sign matrix of the same dimensions as the MI matrix. A signed MI matrix was then created by element-wise multiplication of the sign matrix by the MI matrix. The signed MI matrix was then used as input for another round of clustering, as previously described. Briefly, this involved generating a top-k adjacency matrix, computing eigenvalues from the normalized Laplacian, determining the optimal

number of clusters using the eigengap method, and assigning cluster labels. Based on this procedure, one cluster was separated into three sub-clusters, while all other clusters were separated into two sub-clusters.

Global labels

Since our clustering method begins with a MI matrix, and each group has its own MI matrix and associated subclusters and subcluster labels, we performed a final cluster labeling step that allowed for comparison across groups. The results are such that, if a set of cells in group A have been assigned cluster label L, then the set of cells in group B having similar response patterns will be assigned the same cluster label L, allowing for comparison across groups.

Cluster Stability

To assess the stability and reorganization of functional clusters across days, we tracked registered cells that were detected on at least two of the three imaging sessions (day 1, day 15, and day 30). For each pair of days (A, B) and each cluster label L on day A, we quantified label stability ("percent_same") — the percentage of cells that had label L on day A and retained the same label on day B:

$$\text{percent\_same}\ (L, A, B) = \frac{n_{\text{same}}}{n_L} \times 100$$

where n_L is the number of cells assigned to label L on day A and n_{same} is the number of those cells that were assigned the same label on day B. Higher percent_same values indicate that a larger fraction of cells maintained the same cluster identity across days, suggesting that the corresponding functional cell assemblies were stable over time. Conversely, lower percent_same values reflect cluster reorganization, where individual cells changed their cluster membership between sessions.

Baseline-to-stimulus firing rate ratio (BSR)

For each neuron, the mean CASCADE-inferred firing rate during the 20-s stimulus period was divided by the mean firing rate during the 10-s prestimulus baseline. Ratios were then averaged across cells within each animal, cell type, and recording day.

Generalized linear model (GLM)

Freezing estimate

Movement was quantified using the inertial measurement unit (IMU) integrated into the head-mounted miniscope. Because the IMU provides band-limited body acceleration rather than locomotion speed, we fitted a two-component Gaussian mixture model to the distribution of log-transformed body-acceleration magnitude. The posterior probability that each imaging frame belonged to the low-movement (freezing) component, p_freeze(t) ∈ [0,1], was used as a continuous freezing regressor. Samples containing invalid or missing IMU values were identified using a validity mask and excluded from the analysis rather than being classified as freezing.

We used a continuous estimate of freezing instead of applying a binary threshold because it captures graded variations in immobility and uncertainty in behavioral state. Consequently, the GLM removes more behavior-related variance from the neural signal than would be expected with a binary freezing classifier. This provides a conservative test of tone encoding: if neuronal activity were primarily explained by freezing, the continuous regressor would account for more of that variance, making it more difficult—not easier—to detect an independent contribution of tone identity.

Freezing-controlled time-series GLM

We fitted a generalized linear model (GLM) to the z-scored ΔF/F time series of each neuron to estimate the independent contributions of tone identity and freezing behavior. The design matrix included four tone regressors (3, 7, 11, and 15 kHz), a continuous freezing regressor derived from

IMU measurements (see previous subsection), and a trial-number regressor to account for within-session drift. The model was fitted without an intercept so that each tone coefficient represented activity relative to the inter-tone baseline. Standard errors were estimated using heteroskedasticity- and autocorrelation-consistent (HAC; Newey–West (Newly, 1987)) estimators because calcium activity exhibits substantial temporal autocorrelation, making conventional ordinary least squares standard errors anti-conservative. Autocorrelation was estimated within, but not across, trials by resetting the Bartlett kernel at trial boundaries, and the truncation lag was selected using the Newey–West plug-in rule, capped at the duration of a single trial. Statistical significance was assessed using two-sided *t*-tests based on the HAC standard errors.

To ensure that the effects of tone identity and freezing could be estimated independently, we assessed collinearity using the variance inflation factor (VIF). The freezing regressor exhibited minimal collinearity with the tone regressors (median VIF ≈ 1.1), confirming that the independent contributions of tone identity and freezing could be reliably estimated within the same model.

Derived per-cell metrics

From each fitted model we extracted the preferred-tone coefficient (β_pref), the freezing coefficient (β_freeze), the freezing-to-tone ratio (|β_freeze|/|β_pref|), and a binary freezing-dominant classification (|β_freeze|>|β_pref|, HAC p<0.05). A neuron was classified as freezing-dominant when the freezing coefficient was significant (HAC $p < 0.05$) and its absolute magnitude exceeded that of the corresponding tone coefficient.

For the general population GLM, in which neurons were not assigned to functional clusters, we used the same cells as used for the population curves and their associated beta tone values. i.e. If a cell contributed to a population response curve of a given tone, then we also used the same cell and its associated beta tone. For frequency-selective neurons, the preferred tone corresponded to the neuron's preferred tone as determined by the mutual-information (MI) clustering analysis. For the graded neurons, each neuron generated four β tone values that were normalized by dividing by the maximum of the four.

Statistics

Behavioral and neural data were analyzed using one-or two-way analyses of variance (ANOVAs) with repeated measures. For one-way ANOVAs, tone frequency was treated as the repeated factor; for two-way ANOVAs, a between-subjects group factor (e.g., experimental vs. control) was additionally included. To evaluate the effects of frequency over time across all groups, we performed a three-way mixed-effects ANOVA with group (CS+15, CS+3, and control) as the between-subjects factor and frequency and time as within-subjects (repeated-measures) factors. When ANOVAs revealed significant main effects or interactions, post hoc comparisons were performed using Tukey multiple-comparisons test. Data normality was assessed using the Shapiro–Wilk test and homogeneity of variances with the Brown–Forsythe test prior to statistical analysis. When assumptions were violated, Friedman tests were used for repeated-measures analyses, and Kruskal–Wallis tests were used for comparisons without repeated measures. Rank-based tests were followed by Dunn's post hoc multiple-comparison tests. Statistical significance was evaluated using a two-sided alpha level of 0.05. For analyses involving multiple cell-identity comparisons, p values were adjusted using the Benjamini–Hochberg false discovery rate procedure.

Histology

At the conclusion of experiments, mice were deeply anesthetized with isoflurane and transcardially perfused with phosphate-buffered saline (PBS), followed by 4% paraformaldehyde (PFA). Extracted brains were post-fixed in 4% PFA for 24 hours and subsequently transferred to a 20% sucrose solution containing sodium azide at 4 °C for an additional 24 hours. Frozen brains were coronally sectioned at 50 μm using a cryostat, and sections were mounted on Superfrost Plus microscope slides (Fisher Scientific). Histological verification was performed to confirm GCaMP expression and accurate lens placement within the PL.

Data availability

Datasets available at: *https://data.mendeley.com/datasets/9yyn63g346/1*

Supplementary figures and tables

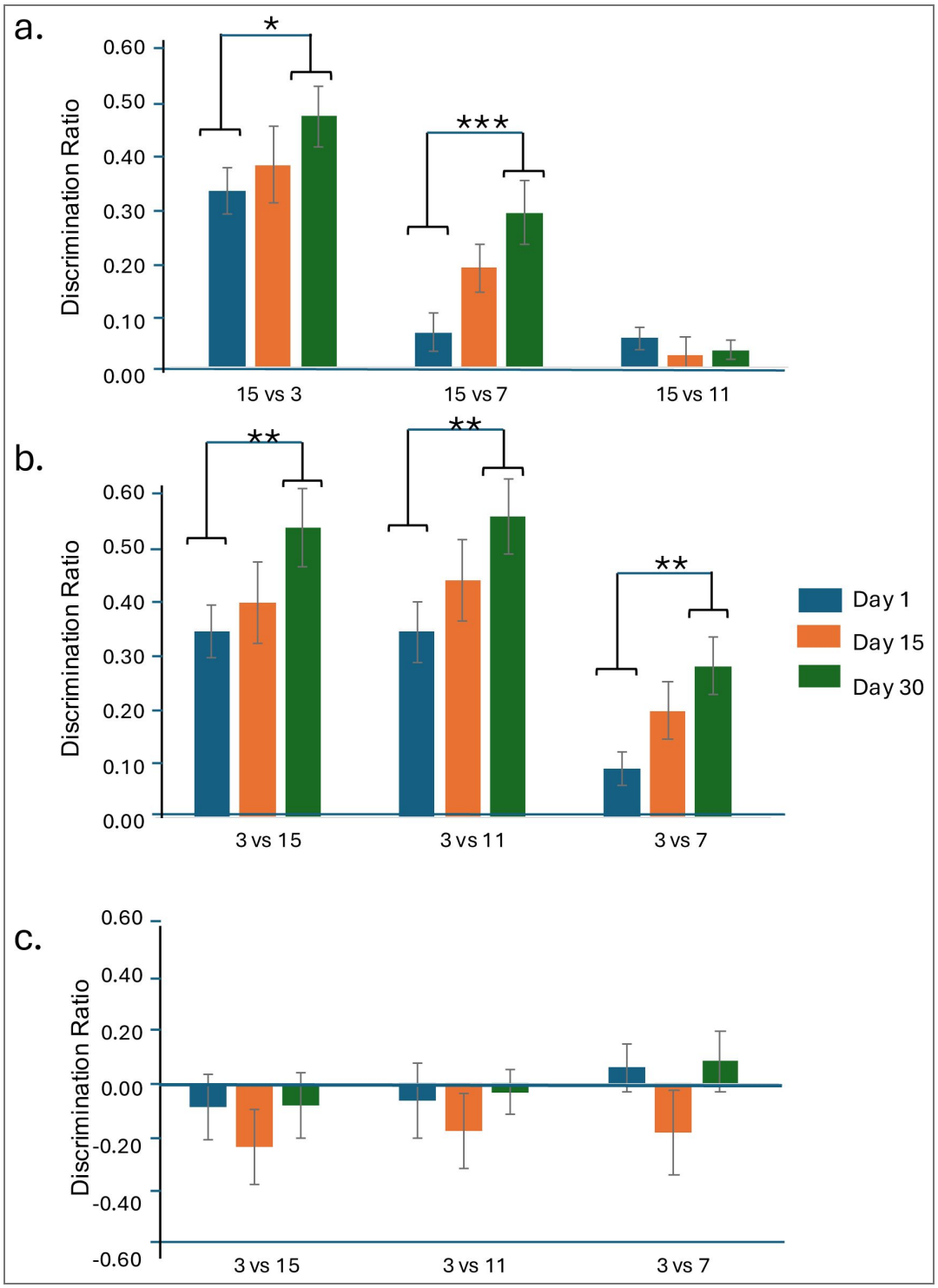

Figure S1. (a–c) Discrimination ratios (DRs) were used to quantify discrimination between each frequency and the CS+ across testing days in the CS+15 group (a), CS+3 group (b), and control animals (c). In both experimental groups, discrimination increased over time, particularly between the CS+ and the CS− and between the CS+ and the frequency adjacent to the CS−. Discrimination between the CS+ and its adjacent frequency also increased

across testing days in the CS+3 group but not in the CS+15 group (Tukey's multiple comparisons). In contrast, control animals showed no consistent changes in discrimination over time. Repeated-measures ANOVAs with frequency comparison and testing day as factors revealed significant effects of frequency comparison, testing day, and their interaction in the CS+15 group (frequency: $F(2,52) = 44.362$, $p < 0.001$; day: $F(2,52) = 4.822$, $p = 0.012$; interaction: $F(4,104) = 3.421$, $p = 0.011$). In the CS+3 group, significant effects of frequency comparison and testing day were observed, but not their interaction (frequency: $F(2,42) = 24.415$, $p < 0.001$; day: $F(2,42) = 5.499$, $p = 0.008$; interaction: $F(4,84) = 0.23$, $p = 0.921$). No significant effects were detected in control animals (frequency: $F(2,20) = 1.21$, $p = 0.319$; day: $F(2,20) = 0.767$, $p = 0.477$; interaction: $F(4,40) = 0.154$, $p = 0.960$). Asterisks denote Tukey's multiple comparisons: $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

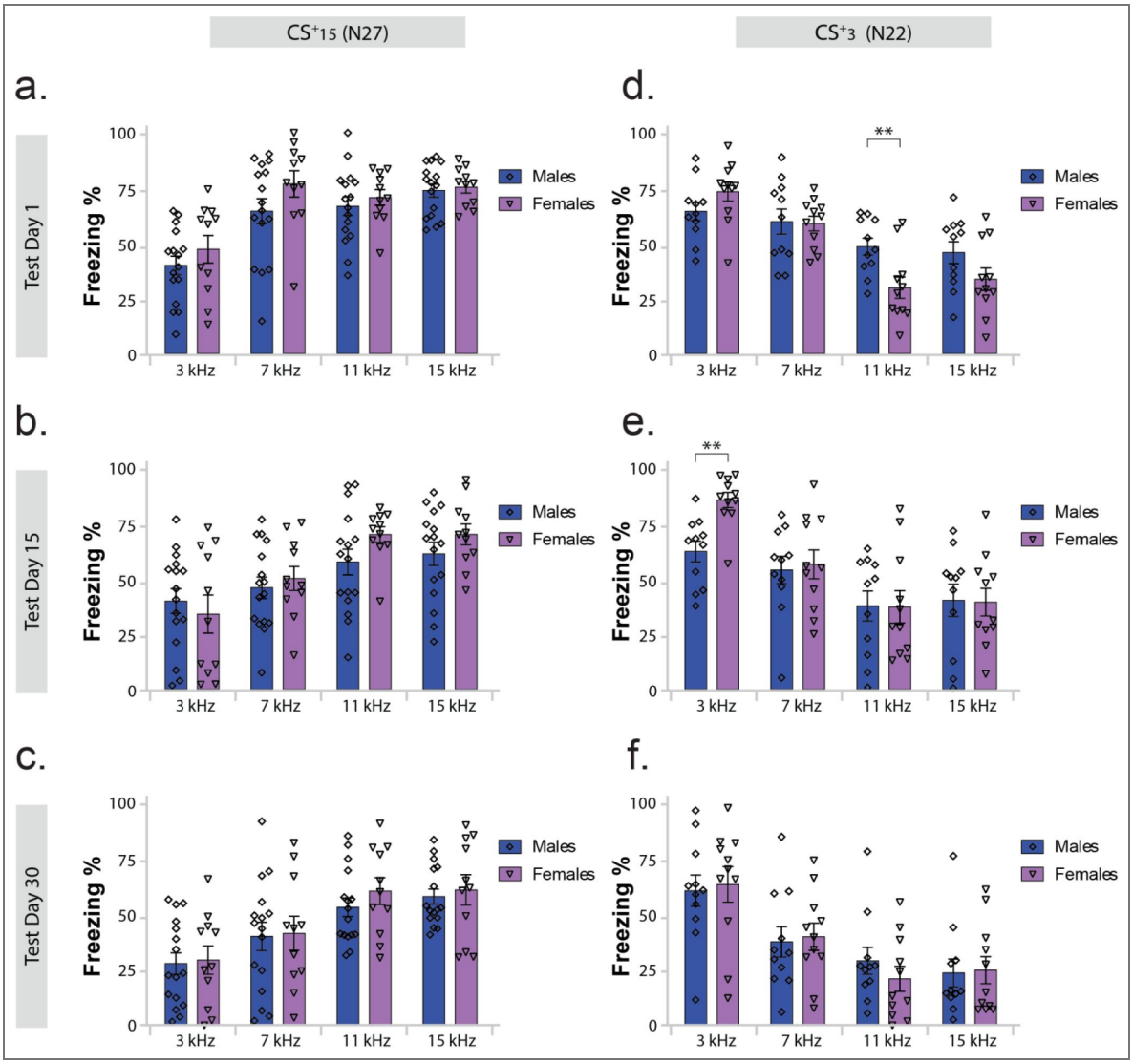

Figure S2. Sex differences in behavior.(a-c) Sex differences in CS+15-trained mice.

No sex differences were observed on any testing day (effect of sex: day 1, $F(1,25) = 1.52$, $p = 0.23$; day 15, $F(1,25) = 0.55$, $p = 0.467$; day 30, $F(1,25) = 0.17$, $p = 0.681$). On all days, there was a significant main effect of frequency, indicating successful learning in both males and females (day 1, $F(3,75) = 33.46$, $p < 0.001$; day 15, $F(3,75) = 26.32$, $p < 0.001$; day 30, $F(3,75) = 51.47$, $p < 0.001$). No sex × frequency interactions were detected on any testing day ($p > 0.05$). Post hoc comparisons showed that on day 1, mice discriminated 3 kHz from 7, 11 and 15 kHz ($p < 0.05$), but generalized among the higher frequencies ($p > 0.05$). On days 15 and 30, mice discriminated 3 kHz from 7, 11 and 15 kHz ($p < 0.05$) and 7 kHz from 11 and 15 kHz ($p < 0.05$), while generalization persisted between 11 and 15 kHz ($p > 0.05$). **(d-f) Sex differences in CS+3-trained mice.** No main effect of sex was observed on any testing day (day 1, $F(1,20) = 1.24$, $p = 0.279$; day 15, $F(1,20) = 0.78$, $p = 0.388$; day 30, $F(1,20) = 0.003$, $p = 0.956$). A significant main effect of frequency was present across all days, indicating learning in both sexes (day 1, $F(3,60) = 40.65$, $p < 0.001$; day 15, $F(3,60) = 30.17$, $p < 0.001$; day 30, $F(3,60) = 35.19$, $p < 0.001$). Post hoc comparisons indicated discrimination among all frequencies across days ($p < 0.05$), except between 11 and 15 kHz ($p > 0.05$). Significant sex × frequency interactions were observed on day 1 ($F(3,60) = 6.91$, $p < 0.001$) and day 15 ($F(3,60) = 3.55$, $p < 0.02$), reflecting sex-specific differences at 11 kHz on day 1 and at 3 kHz on day 15. These interaction effects were not present on day 30 ($F(3,60) = 0.76$, $p = 0.522$) and likely reflect increased behavioral variability rather than stable sex differences.

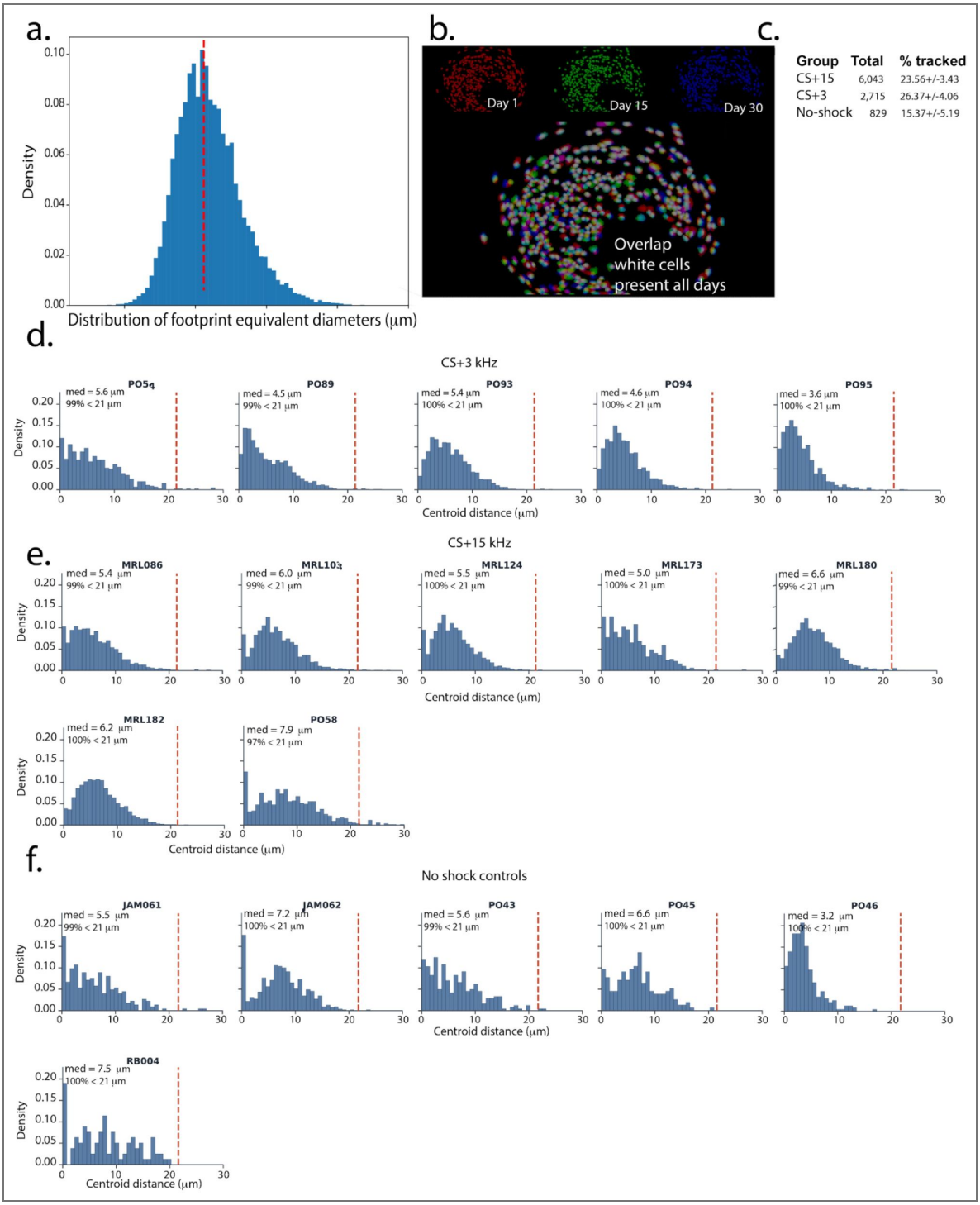

Figure S3. Evaluation of cell registration quality.

(a) Distribution of equivalent neuronal footprint diameters (μm) for all registered neurons. **(b)** Representative spatial footprints recorded on Day 1 (red), Day 15 (green), and Day 30 (blue), along with the merged image showing neurons tracked throughout the experiment (white). **(c)** Total number of neurons recorded in each experimental group and the percentage of neurons successfully registered across all sessions. Note that two control animals were excluded because they lost the GRIN lens after Day 15. **(d–f)** Distributions of centroid shifts between matched neurons in the retrieval sessions relative to the conditioning session for the CS+3 **(d),** CS+15 **(e),** and control **(f)** groups. The red line indicates the median equivalent neuronal footprint diameter (21.4 μm). The upper left corner of each panel shows the median centroid shift and the percentage of registered neurons with centroid shifts smaller than the median footprint diameter.

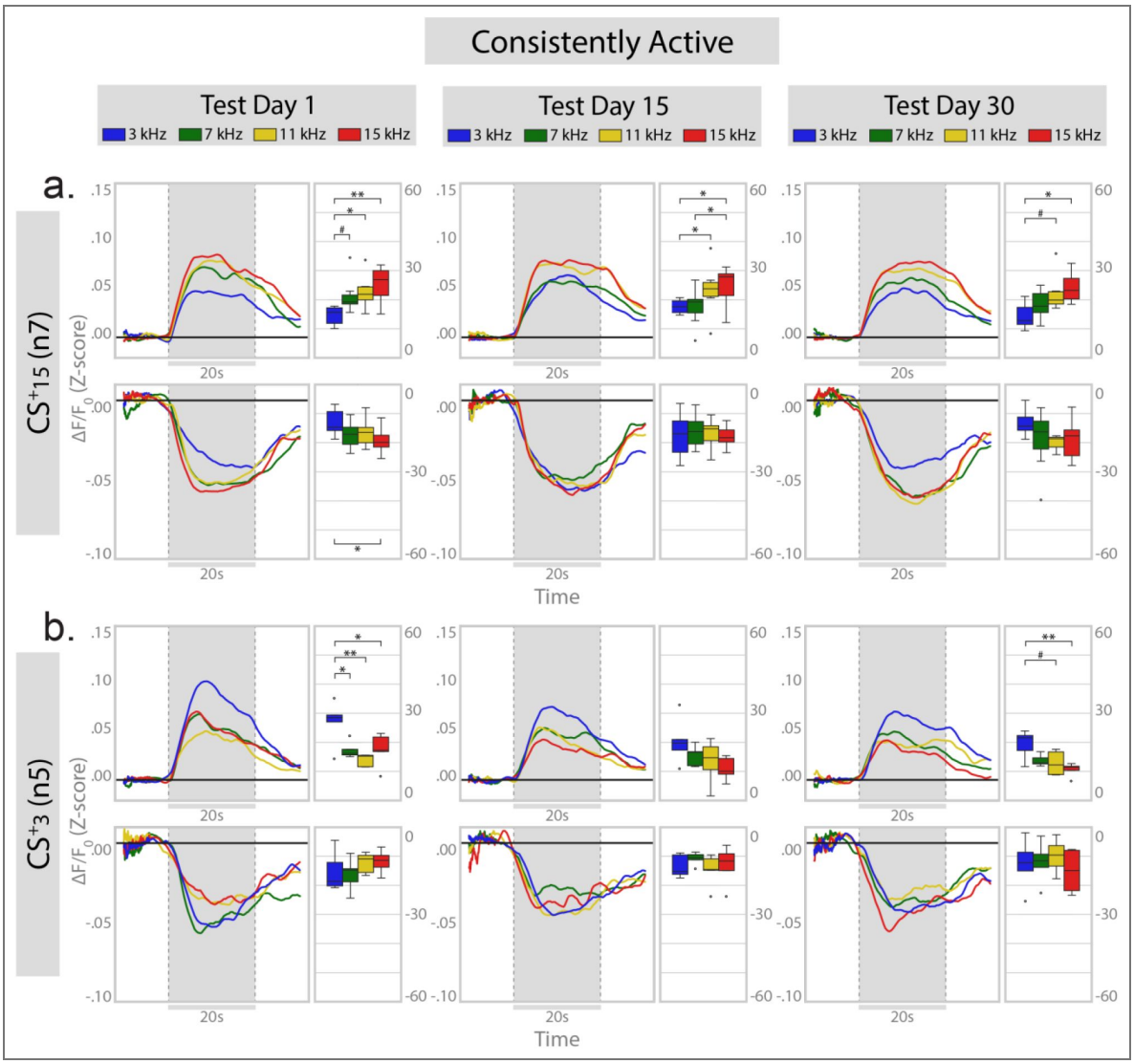

Figure S4. Population activity of consistently active neurons.

(a-b) Population activity of consistently active neurons from animals trained with a 15 kHz CS+ **(a)** or a 3 kHz CS+ **(b).** Upper panels show positively tone-responsive neurons and lower panels show negatively tone-responsive neurons. Adjacent boxplots display areas under the population response curves (AUC). Boxplots show the median (center line), interquartile range (box), and whiskers extending to ±1.5× the interquartile range; points outside the whiskers represent individual observations beyond this range. Consistently active positive responders exhibit graded responses across test days (day 1 to day 30). *p < 0.05; **p < 0.01.

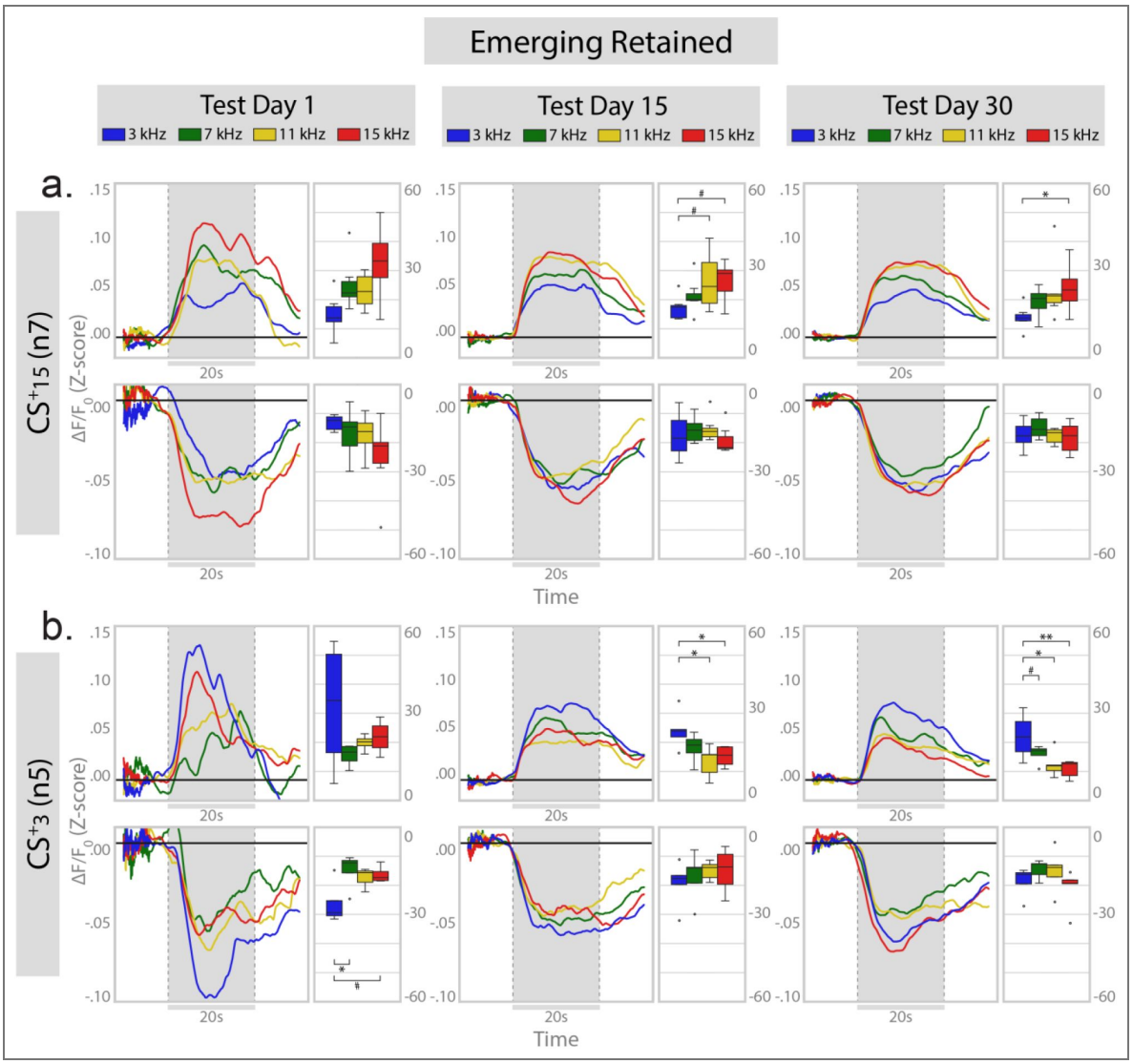

Figure S5. Population activity of emerging-retained neurons.

Emerging-retained neurons become part of the active ensemble after conditioning and remaining active after that. **(a-b)** Population activity of emerging-retained neurons from animals trained with a 15 kHz CS+ **(a)** or a 3 kHz CS+ **(b).** Upper panels show positively tone-responsive neurons and lower panels show negatively tone-responsive neurons. Adjacent boxplots display areas under the population response curves. Emerging-retained neurons show greater variability on day 1 compared with later test days. Only positive responders show significant graded valence. *p < 0.05; **p < 0.01.

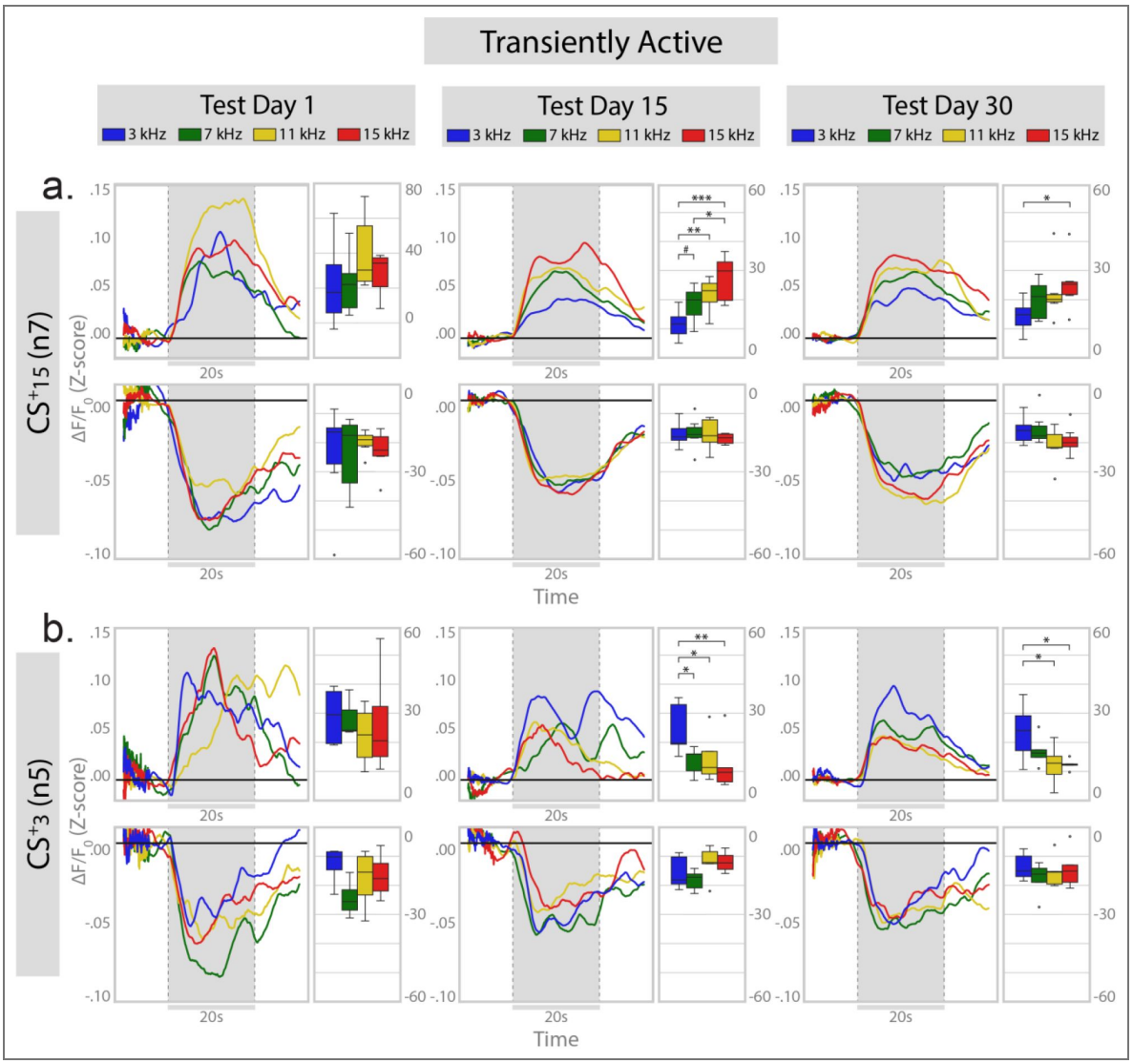

Figure S6. Population activity of transiently active neurons.

Transiently active neurons were active on only a single test day. **(a-b)** Population activity of transiently active neurons from animals trained with a 15 kHz CS+ **(a)** or a 3 kHz CS+ **(b).** Upper panels show positively tone-responsive neurons and lower panels show negatively tone-responsive neurons. Adjacent boxplots display areas under the population response curves. Transiently active neurons did not show graded population responses on day 1; graded responses emerged by day 15 and were maintained through day 30 in positive sound responder neurons. *p < 0.05; **p < 0.01.

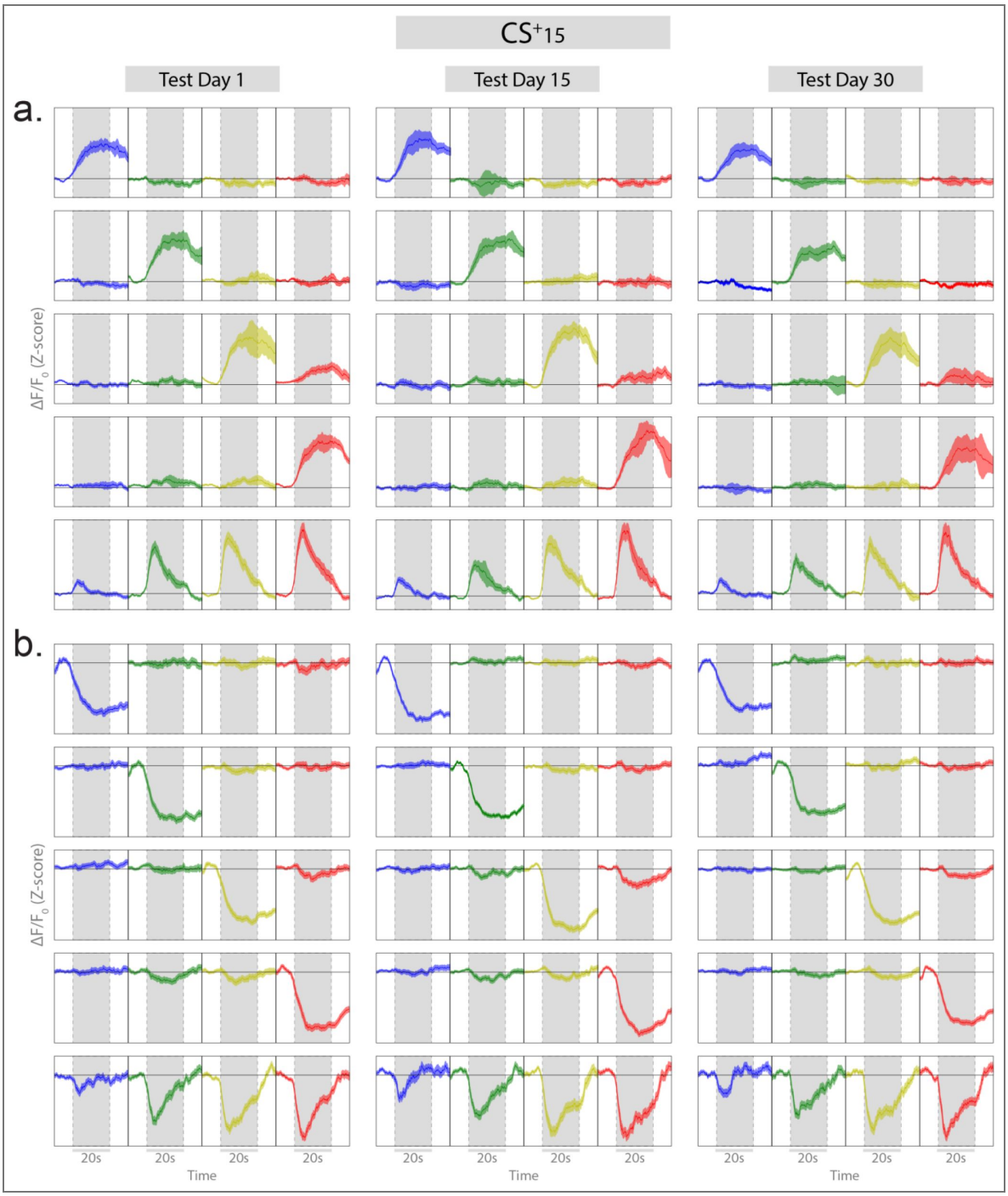

Figure S7. Clustering of PL subnetworks based on signed mutual information for the CS+ 15 kHz group per testing session.

(a-b) Average stimulus-aligned population responses for clusters showing positive **(a)** or negative **(b)** modulation to individual tones (3, 7, 11, or 15 kHz, upper panels) or graded emotional tuning **(a-b, bottom panels).**

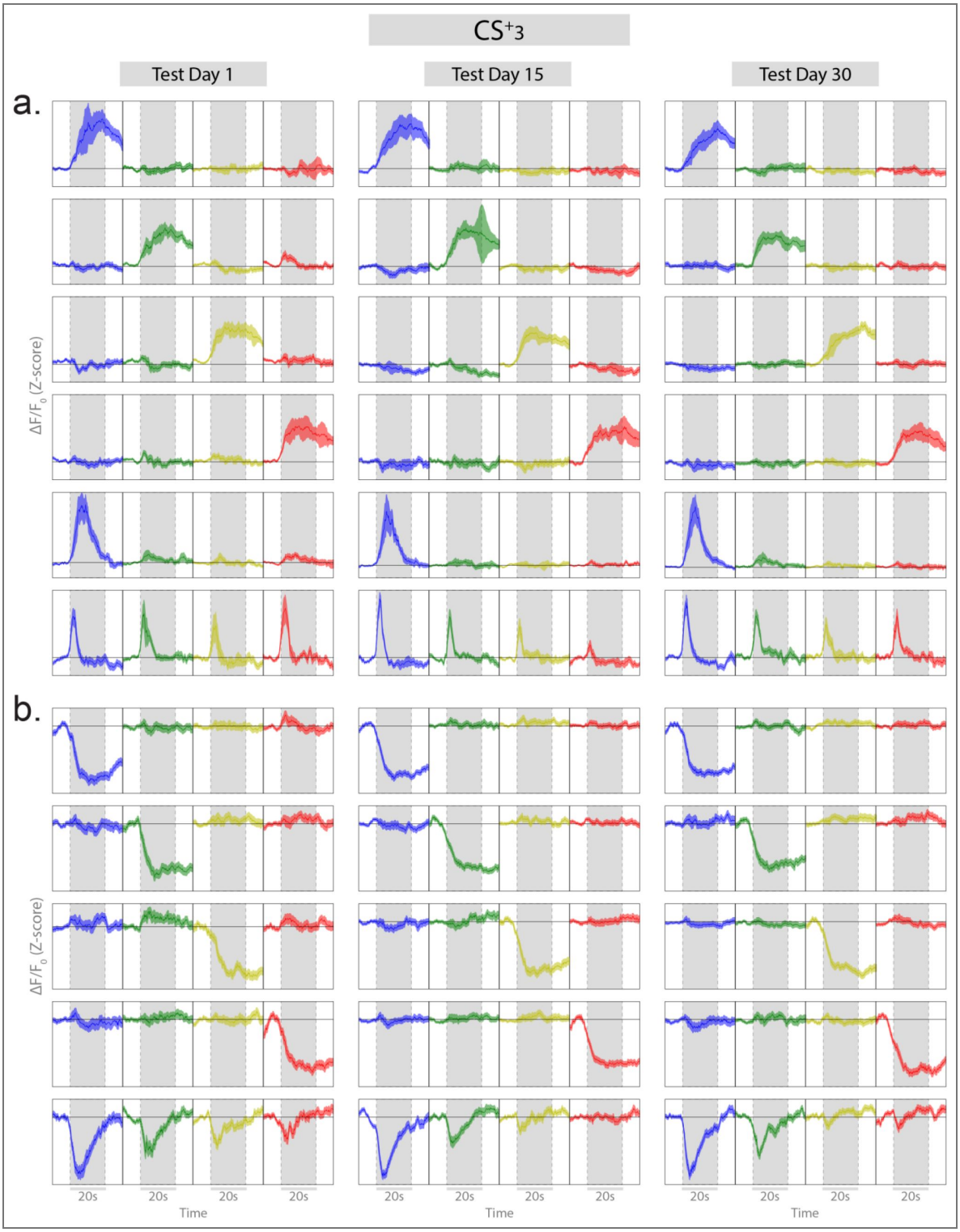

Figure S8. Clustering of PL subnetworks based on signed mutual information for the CS+ 3 kHz group per testing session.

(a-b) Average stimulus-aligned population responses for clusters showing positive **(a)** or negative **(b)** modulation to individual tones (3, 7, 11, or 15 kHz, upper panels) or graded emotional tuning **(a-b, bottom panels).**

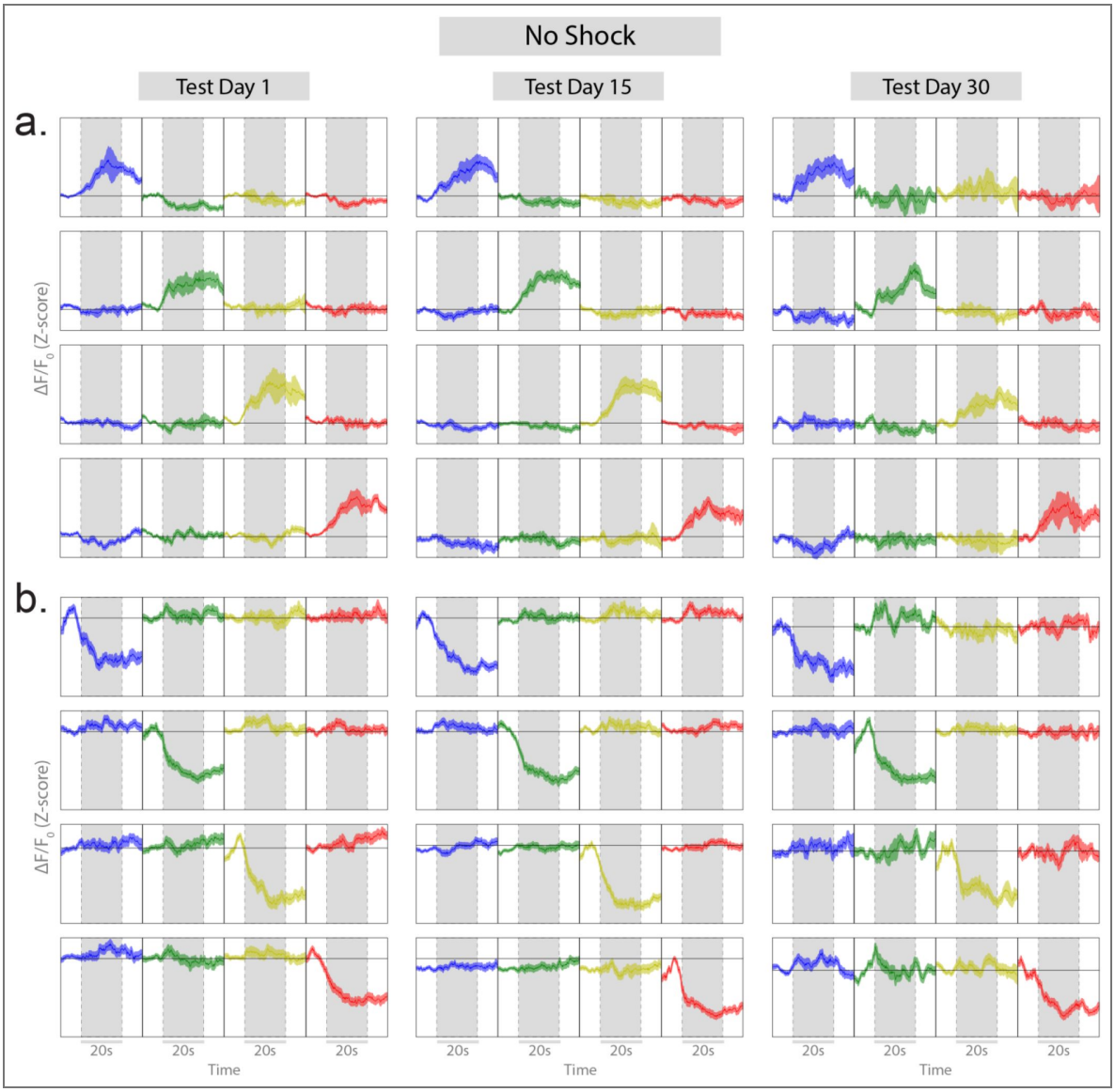

Figure S9. Clustering of PL subnetworks based on signed mutual information for the no shock control per testing session.

(a-b) Average stimulus-aligned population responses for clusters showing positive **(a)** or negative **(b)** modulation to individual tones (3, 7, 11, or 15 kHz).

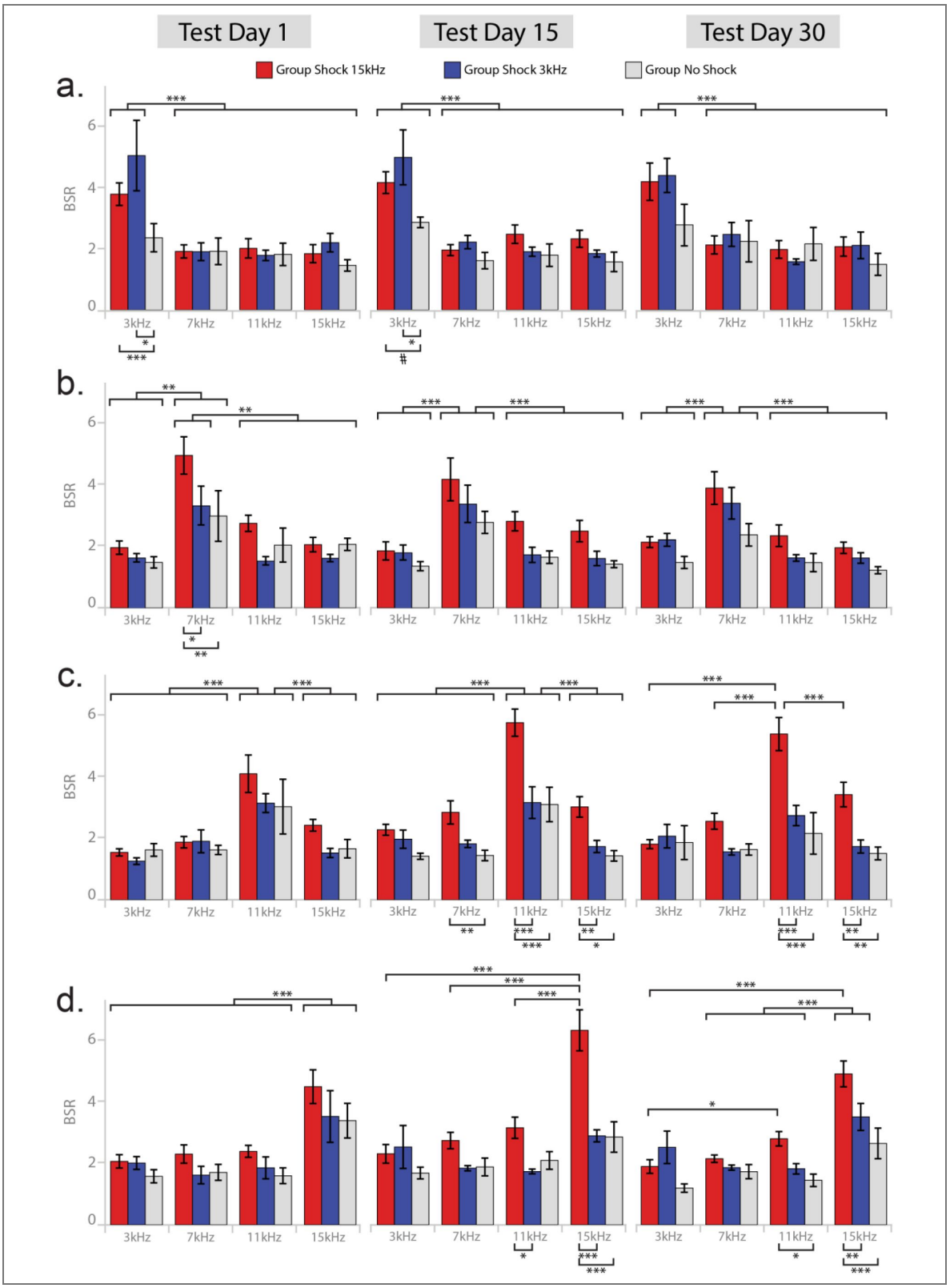

Figure S10.

a–d, Baseline-to-stimulus firing rate ratio (BSR) illustrating changes in positively responding neurons within tone-specific clusters shown in Fig. 5, for the CS+15 (red), CS+3 (blue), and control groups. **(a)** BSR of neurons in clusters primarily responsive to 3 kHz (Fig. 5c.1), **(b)** 7 kHz (Fig. 5c.2), **(c)** 11 kHz (Fig. 5c.3), and **(d)** 15 kHz (Fig. 5c.4). ANOVA results are reported in Table S3; Tukey's multiple-comparison tests indicate significance ($p < 0.05$; $\boldsymbol{p} < 0.01$; $p < 0.001$).

Fig. 2.

Analysis	Cell class/session	Variable	Sum Sq.	Mean Sq.	F	df	p	ηp²
One-way ANOVA	Consistently active neurons	Group	0.0261	0.0130	2.704	2, 13	0.104	0.29
One-way ANOVA	Emerged-retained neurons: Day 1 and 15	Group	0.0065	0.00325	1.445	2, 13	0.271	0.18
One-way ANOVA	Emerged-retained neurons: Day 15 and 30	Group	0.00697	0.00349	1.018	2, 13	0.388	0.14
One-way ANOVA	Emerged-retained neurons: Day 1 and 30	Group	0.0084	0.0042	0.332	2, 13	0.723	0.05
One-way ANOVA	Transiently active neurons: Day 1	Group	0.000194	0.000097	0.0598	2, 13	0.942	0.01
Kruskal-Wallis	Transiently active neurons: Day 15	Group	NA	NA	1.901	2	0.387	ε²=0.00
Kruskal-Wallis	Transiently active neurons: Day 30	Group	NA	NA	7.515	2	0.023*	ε²=0.42

Pairwise multiple comparisons (Dunn's method)

Comparison	Diff. of ranks	Q	p
No shock vs. CS+15	8.179	2.741	0.018*
No shock vs. CS+3	5.350	1.675	0.282
CS+15 vs. CS+3	2.829	1.015	0.931

Table S1. Statistical analyses corresponding to the Venn diagrams shown in Fig. 2.

One-way ANOVAs were used when the assumptions of normality and homogeneity of variance were met; otherwise, Kruskal–Wallis tests were performed. Significant Kruskal–Wallis tests were followed by Dunn's multiple-comparisons tests. Effect sizes are reported as partial eta squared (ηp^2) for ANOVAs and epsilon squared (ε^2) for Kruskal– Wallis tests.

Fig. 3. 3-way mixed model ANOVA

a. Positive modulated

Modulation	Variable	Sum Sq.	Mean Sq.	F	df	p	ηp²
Positive	Group	41614	20807.1	16.803	2, 17.907	< 0.001***	0.65
	Time	11118	5558.8	4.489	2, 191.314	< 0.02*	0.04
	Frequency	3480	1160.1	0.937	3, 189.815	0.424	0.01
	Group x Time	3617	904.3	0.730	4, 191.152	0.572	0.02
	Group x Frequency	189326	31554.4	25.482	6, 189.815	< 0.001***	0.45
	Time x Frequency	9127	1521.1	1.228	6, 189.815	0.293	0.04
	Group x Time x Frequency	7936	661.3	0.534	12, 189.815	0.891	0.03

Simple effects analysis conducted for Group x Frequency interaction

Group = CS+15

Contrast	Estimate	SE	df	t.ratio	p
3 vs 7	-45.634	11.9	230	-3.820	< 0.0005***
3 vs 11	-87.231	11.9	230	-7.302	< 0.0001***
3 vs 15	-103.651	11.9	230	-8.677	< 0.0001***
7 vs 11	-41.598	11.9	230	-3.482	< 0.0012**
7 vs 15	-58.018	11.9	230	-4.857	< 0.0001***
11 vs 15	-16.420	11.9	230	-1.375	0.1706

Group = CS+3

Contrast	Estimate	SE	df	t.ratio	p
3 vs 7	59.286	14.1	230	4.194	< 0.0002***
3 vs 11	81.162	14.1	230	5.742	< 0.0001***
3 vs 15	80.288	14.1	230	5.680	< 0.0001***
7 vs 11	21.876	14.1	230	1.548	0.3692
7 vs 15	21.002	14.1	230	1.486	0.3692
11 vs 15	-0.874	14.1	230	-0.062	0.9508

Group = No shock control

Contrast	Estimate	SE	df	t.ratio	p
3 vs 7	-15.841	13.9	230	-1.137	1.0000
3 vs 11	-10.626	13.9	230	-0.762	1.0000
3 vs 15	-7.660	13.9	230	-0.550	1.0000
7 vs 11	5.216	13.9	230	0.374	1.0000
7 vs 15	8.181	13.9	230	0.587	1.0000
11 vs 15	2.966	13.9	230	0.213	1.0000

b. Negative modulated

Modulation	Variable	Sum Sq.	Mean Sq.	F	df	p	ηp²
Negative	Group	5701.6	2850.8	2.2670	2, 18.142	0.1322	0.200
	Time	963.6	481.8	0.3831	2, 189.038	0.6822	0.004
	Frequency	6148.4	2049.5	1.6298	3, 186.216	0.1840	0.030
	Group x Time	1204.1	301.0	0.2394	4, 188.686	0.9158	0.005
	Group x Frequency	14708.2	2451.4	1.9494	6, 186.203	0.0750	0.059
	Time x Frequency	23070.6	3845.1	3.0577	6, 186.257	< 0.0071**	0.090
	Group x Time x Frequency	23011.2	1917.6	1.5249	12, 186.238	0.1182	0.089

Table S2. Statistics comparing areas under the curve for the CS+15, CS+3, and No shock control groups across all frequencies and testing sessions.

(a-b) Three-way mixed model ANOVAs using Group (CS+15, CS+3, No shock), Frequency (3, 7, 11, 15 kHz), and Time (days 1, 15, 30) as variables for positive (a) and negative (b) cell responders. Simple effects were calculated using Holm correction.

Simple effect analysis conducted for Time x Frequency interaction

Time: Day 1

Contrast	Estimate	SE	df	t.ratio	p
3 vs 7	14.787	13.2	226	-1.124	1.0000
3 vs 11	-15.219	13.7	227	-1.109	1.0000
3 vs 15	-8.945	13.2	226	-0.680	1.0000
7 vs 11	-0.433	13.7	227	-0.032	1.0000
7 vs 15	5.842	13.2	226	0.444	1.0000
11 vs 15	-6.274	13.7	227	0.457	1.0000

Time: Day 15

Contrast	Estimate	SE	df	t.ratio	p
3 vs 7	-14.825	13.7	227	-1.080	1.0000
3 vs 11	9.416	13.2	226	0.716	1.0000
3 vs 15	3.458	13.2	226	0.263	1.0000
7 vs 11	24.241	13.7	227	1.766	0.4726
7 vs 15	18.283	13.7	227	1.332	0.9212
11 vs 15	5.958	13.2	226	-0.453	1.0000

Time: Day 30

Contrast	Estimate	SE	df	t.ratio	p
3 vs 7	-15.841	13.9	230	-1.137	1.0000
3 vs 11	-10.626	13.9	230	-0.762	1.0000
3 vs 15	-7.660	13.9	230	-0.550	1.0000
7 vs 11	5.216	13.9	230	0.374	1.0000
7 vs 15	8.181	13.9	230	0.587	1.0000
11 vs 15	2.966	13.9	230	0.213	1.0000

Table S2. (continued)

Fig. S4, S5, S6

Group	Type	Modulation	Test	Statistic	df	p	ηp²
CS+15	Consistent	Positive	Test Day 1	F = 6.11	3, 18	0.005 **	0.51
			Test Day 15	F = 6.29	3, 18	0.004**	0.51
			Test Day 30	F = 4.43	3, 18	0.017*	0.43
		Negative	Test Day 1	F = 3.76	3, 18	0.030*	0.39
			Test Day 15	F = 0.31	3, 18	0.817	0.05
			Test Day 30	F = 1.80	3, 18	0.183	0.23
	Emerged Retained	Positive	Test Day 1	F = 3.18	3, 13	0.060	0.42
			Test Day 15	F = 3.24	3, 18	0.046 *	0.35
			Test Day 30	χ^2 = 8.66	3	0.0.4*	NA
		Negative	Test Day 1	F = 1.89	3, 17	0.170	0.25
			Test Day 15	F = 1.45	3, 18	0.263	0.20
			Test Day 30	F = 2.63	3, 18	0.308	0.31
	Transiently Active	Positive	Test Day 1	F = 1.18	3, 17	0.332	0.17
			Test Day 15	F = 11.17	3, 18	<0.001***	0.65
			Test Day 30	F = 3.87	3, 18	0.030 *	0.39
		Negative	Test Day 1	F = 0.39	3, 17	0.764	0.06
			Test Day 15	F = 0.30	3, 18	0.827	0.05
			Test Day 30	F = 2.08	3, 18	0.139	0.26
CS+3	Consistent	Positive	Test Day 1	F = 9.32	3, 12	0.002 **	0.70
			Test Day 15	F = 3.18	3, 12	0.063	0.44
			Test Day 30	F = 5.39	3, 12	0.014 *	0.57
		Negative	Test Day 1	F = 1.28	3, 12	0.326	0.24
			Test Day 15	F = 0.53	3, 12	0.668	0.12
			Test Day 30	F = 0.47	3, 12	0.709	0.11
	Emerged Retained	Positive	Test Day 1	F = 1.49	3, 6	0.309	0.43
			Test Day 15	F = 6.18	3, 9	0.014 *	0.67
			Test Day 30	F = 9.51	3, 9	0.004 **	0.76
		Negative	Test Day 1	F = 4.75	3, 8	0.035	0.64
			Test Day 15	F = 0.65	3, 12	0.595	0.14
			Test Day 30	χ^2 = 1.39	3	0.356	NA
	Transiently Active	Positive	Test Day 1	F = 0.28	3, 8	0.839	0.10
			Test Day 15	F = 6.45	3, 12	0.008 **	0.62
			Test Day 30	F = 5.49	3, 12	0.013 *	0.58
		Negative	Test Day 1	F = 1.12	3, 9	0.398	0.27
			Test Day 15	F = 1.57	3, 11	0.253	0.30
			Test Day 30	F = 0.426	3, 12	0.738	0.10

Table S3. Statistics corresponding to consistently active, emerging-retained, and transiently active cells.

One-way ANOVAs with repeated measures were used when normality and equal variance tests were passed to test the effect of frequency. Friedman tests were used (χ^2 with Kendall's W) when assumptions were violated, Friedman tests were used (χ^2 with Kendall's W).

Fig. 4

Positive modulated cells

Fig. 4a

Modulation	Variable	Sum Sq.	Mean Sq.	F	df	p	ηp²
Positive	Group	8959	4.48	9.14	2, 15	0.003**	0.55
	Frequency	7451	0.49	2.12	3, 45	0.111	0.12
	Group x Frequency	4859	0.81	8.69	6, 45	< 0.001***	0.54

Tukey post hoc comparisons

Tone	Comparison	Diff. of means	df	q	p
3 kHz	CS+3 vs. No shock	0.747	3	3.975	0.022*
	CS+3 vs. CS+15	0.491	3	2.706	0.151
	CS+15 vs. No shock	0.255	3	1.479	0.554
7 kHz	CS+15 vs. No shock	0.684	3	3.963	0.022*
	CS+15 vs. CS+3	0.229	3	1.259	0.650
	CS+3 vs. No shock	0.455	3	2.423	0.215
11 kHz	CS+15 vs. No shock	1.174	3	6.805	< 0.001***
	CS+15 vs. CS+3	0.737	3	4.058	0.019*
	CS+3 vs. No shock	0.437	3	2.328	0.241
15 kHz	CS+15 vs. No shock	1.215	3	7.040	< 0.001***
	CS+15 vs. CS+3	1.176	3	6.476	< 0.001***
	CS+3 vs. No shock	0.0387	3	0.206	0.988

Fig. 4b

Modulation	Variable	Sum Sq.	Mean Sq.	F	df	p	ηp²
positive	Group	0.26	0.13	2.79	2, 15	0.094	0.27
	Day	0.02	0.007	0.806	3, 45	0.497	0.05
	Group x Day	0.154	0.025	2.84	6, 45	0.020*	0.27

Tukey post hoc comparisons

Tone	Comparison	Diff. of means	df	q	p
3 kHz	No shock vs. CS+3	0.164	3	2.818	0.130
	No shock vs. CS+15	0.0454	3	0.849	0.821
	CS+15 vs. CS+3	0.119	3	2.107	0.308
7 kHz	No shock vs. CS+15	0.166	3	3.099	0.087
	No shock vs. CS+3	0.105	3	1.800	0.420
	CS+3 vs. CS+15	0.061	3	1.084	0.726
11 kHz	No shock vs. CS+15	0.185	3	3.455	0.051
	No shock vs. CS+3	0.0418	3	0.718	0.868
	CS+3 vs. CS+15	0.143	3	2.539	0.186
15 kHz	No shock vs. CS+15	0.172	3	3.207	0.074
	No shock vs. CS+3	0.000427	3	0.00734	1.000
	CS+3 vs. CS+15	0.171	3	3.040	0.095

Fig. 4c

Modulation	Variable	Sum Sq.	Mean Sq.	F	df	p	ηp²
Positive	Group	255.23	127.62	1.148	2, 15	0.344	0.13
	Frequency	26.999	9	0.221	3, 45	0.881	0.01
	Group x Frequency	587.08	97.846	2.402	6, 45	0.042*	0.24

Tukey post hoc comparisons

Tone	Comparison	Diff. of means	df	q	p
3 kHz	CS+15 vs. CS+3	9.714	3	3.072	0.087
	CS+15 vs. No shock	1.450	3	0.482	0.938
	No shock vs. CS+3	8.265	3	2.527	0.185
7 kHz	No shock vs. CS+3	5.421	3	1.658	0.475
	No shock vs. CS+15	5.057	3	1.683	0.465

Table S4. Statistical analysis of GLM-derived measures for cells used in the population curves shown in Fig. 4.

Two-way ANOVAs with repeated measures were conducted to evaluate effects of group (CS+15, CS+3, or No shock), frequency (3, 7, 11, 15 kHz) or day (days 1, 15, 30), and their interactions. Tukey post hoc comparisons are shown where simple effects were tested.

	CS+15 vs. CS+3	0.364	3	0.115	0.996
11 kHz	No shock vs. CS+15	9.676	3	3.221	0.069
	No shock vs. CS+3	5.016	3	1.534	0.528
	CS+3 vs. CS+15	4.660	3	1.474	0.554
15 kHz	CS+3 vs. CS+15	6.622	3	2.094	0.309
	CS+3 vs. No shock	3.461	3	1.058	0.736
	No shock vs. CS+15	3.161	3	1.052	0.739

Fig. 4d

Modulation	Variable	Sum Sq.	Mean Sq.	F	df	p	ηp²
Positive	Group	156.357	78.178	1.166	2, 15	0.336	0.13
	Day	17.999	68.24	0.174	2, 26	0.842	0.01
	Group x Day	106.236	26.559	0.512	4, 26	0.727	0.07

Negative modulated cells

Fig. 4e

Modulation	Variable	Sum Sq.	Mean Sq.	F	df	p	ηp²
Negative	Group	0.135	0.0677	0.968	2, 15	0.402	0.11
	Frequency	0.101	0.0336	1.453	3, 44	0.240	0.09
	Group x Frequency	0.072	0.012	0.519	6, 44	0.791	0.07

Fig. 4f

Modulation	Variable	Sum Sq.	Mean Sq.	F	df	p	ηp²
Negative	Group	1.038	0.519	3.066	2, 15	0.076	0.29
	Frequency	0.141	0.0471	0.399	3, 44	0.754	0.03
	Group x Frequency	0.468	0.0779	0.661	6, 44	0.681	0.08

Fig. 4g

Modulation	Variable	Sum Sq.	Mean Sq.	F	df	p	ηp²
Negative	Group	864.929	432.464	2.238	2, 15	0.141	0.23
	Day	218.567	72.856	0.644	3, 44	0.591	0.04
	Group x Day	217.491	36.248	0.320	6, 44	0.923	0.04

Fig. 4h

Modulation	Variable	Sum Sq.	Mean Sq.	F	df	p	ηp²
Negative	Group	886.884	443.442	2.756	2, 15	0.092	0.27
	Day	400.272	200.136	1.500	2, 26	0.242	0.10
	Group x Day	468.189	117.047	0.877	4, 26	0.491	0.12

Table S4. (continued)

Fig. 6 and 7

Group	Cluster	Mod.	% Stable D1-D15	P (D1-D15)	% Stable D15-D30	P (D15-D30)	% Stable D1-D30	P (D1-D30)
CS+15 Fig. 6	Single tone c1	+	13.5	0.391	5.0	1.000	9.3	0.870
	Single tone c2	+	3.4	1.000	7.1	1.000	8.2	1.000
	Single tone c3	+	11.1	0.576	13.4	0.391	19.2	0.124
	Single tone c4	+	15.5	0.391	15.1	0.391	20.9	0.121
	Single tone d1	-	19.3	0.079	11.9	0.391	12.9	0.576
	Single tone d2	-	10.4	0.765	10.0	0.661	10.3	0.576
	Single tone d3	-	11.9	0.697	20.2	0.334	17.8	0.697
	Single tone d4	-	20.2	0.011*	11.8	1.000	10.3	1.000
Fig. 7	Graded a		28.2	0.000***	22.8	0.000***	21.3	0.001***
	Graded b		11.1	0.034*	5.5	0.576	5.9	0.576
CS+3 Fig. 6	Single tone c1	+	23.9	0.014*	16.5	0.460	19.0	0.476
	Single tone c2	+	13.4	0.708	6.5	0.982	11.9	0.792
	Single tone c3	+	10.8	0.852	9.8	0.939	21.1	0.185
	Single tone c4	+	10.4	0.756	7.7	0.982	19.6	0.521
	Single tone d1	-	20.9	0.230	20.6	0.074	13.0	0.708
	Single tone d2	-	11.5	0.714	8.0	0.579	8.6	0.592
	Single tone d3	-	9.7	0.739	5.3	0.982	10.3	0.846
	Single tone d4	-	8.9	0.872	7.6	0.972	12.8	0.592
Fig. 7	Graded c		21.2	0.008**	17.9	0.026*	25.0	0.008**
	Graded d		9.1	0.303	40.0	0.003**	14.3	0.303
	Graded e		12.9	0.047*	15.4	0.047*	8.3	0.589
No Shock Fig. 6	Single tone c1	+	7.8	1.000	8.3	1.000	22.2	1.000
	Single tone c2	+	11.5	1.000	10.0	1.000	0.0	1.000
	Single tone c3	+	21.6	1.000	16.7	1.000	40.0	1.000
	Single tone c4	+	11.5	1.000	5.9	1.000	0.0	1.000
	Single tone d1	-	10.0	1.000	7.1	1.000	0.0	1.000
	Single tone d2	-	5.5	1.000	21.4	1.000	21.4	1.000
	Single tone d3	-	12.0	1.000	4.0	1.000	33.3	1.000
	Single tone d4	-	6.4	1.000	22.2	1.000	0.0	1.000

Table S5. Statistical analyses identifying neuronal clusters that remained stable over time, defined as neurons that retained the same functional response profile on days 15 and 30 as on day 1.

These analyses correspond to the clusters shown in Figs. 6 and 7, as indicated in the table. For each cluster, the observed proportion of neurons that preserved their response profile was compared with a null distribution, with multiple comparisons controlled using the Benjamini– Hochberg procedure. Only the graded neuronal clusters in each experimental group (Fig. 7) exhibited significant stability across time. No graded clusters were identified in the no-shock control group.

Fig. S10.

Cluster	Groups	Day	Statistic	df	p	ηp^2
Statistics for clusters responding to individual frequencies (Fig. 6c)						
Fig. 6c, c.1 Fig. S7, S8, S9 Maximal response at 3 kHz	CS+15 CS+3 No shock	Test Day 1	Group: F =1.49 Frequency: F=30.51 Interaction: 4.50	2, 15 3, 6 6, 45	0.258 0.001***. 0.001***.	0.17 0.94 0.38
		Test Day 15	Group: F =4.59 Frequency: F=31.99 Interaction: 2.28	2, 15 3, 6 6,4 5	0.03* 0.001***. 0.053.	0.38 0.94 0.23
		Test Day 30	Group: F =0.36 Frequency: F=30.02 Interaction: 2.78	3, 13 3, 6 6, 39	0.705 0.001***. 0.024*.	0.08 0.94 0.30
Fig. 6c, c.2 Fig. S7, S8, S9 Maximal response at 7kHz	CS+15 CS+3 No shock	Test Day 1	Group: F =2.29 Frequency: F=30.18 Interaction: 2.62	2,15 3, 6 6, 45	0.135 0.001***. 0.029*.	0.23 0.94 0.26
		Test Day 15	Group: F =3.50 Frequency: F=27.46 Interaction: 0.99	2, 15 3, 6 6, 45	0.057 0.001***. 0.441.	0.32 0.93 0.12
		Test Day 30	Group: F =3.51 Frequency: F=24.64 Interaction: 1.057	2,13 3, 6 6, 3 9	0.061 0.001***. 0.404.	0.35 0.93 0.14
Fig. 6c, c.3 Fig. S7, S8, S9 Maximal response at 11 kHz	CS+15 CS+3 No shock	Test Day 1	Group: F =1.45 Frequency: F=20.44 Interaction: 0.912	2, 15 3, 6 6, 4 5	0.266 0.001***. 0.495.	0.16 0.91 0.11
		Test Day 15	Group: F =15.47 Frequency: F=38.53 Interaction: 3.40	2, 15 3, 6 6,4 5	0.001*** 0.001***. 0.007**.	0.67 0.95 0.31
		Test Day 30	Group: F =11.46 Frequency: F=14.20 Interaction: 6.12	2, 13 3, 6 6, 3 9	0.001*** 0.001***. 0.001***.	0.64 0.88 0.49
Fig. 6c, c.4 Fig. S7, S8, S9 Maximal response at 15 kHz	CS+15 CS+3 No shock	Test Day 1	Group: F =2.71 Frequency: F=22.28 Interaction: 0.38	3, 15 3, 6 6,4 5	0.099 0.001***. 0.89.	0.35 0.92 0.05
		Test Day 15	Group: F =7.64 Frequency: F=33.86 Interaction: 10.58	2,1 5 3, 6 6, 45	0.005** 0.001***. 0.001***.	0.51 0.94 0.59
		Test Day 30	Group: F =8.01 Frequency: F=30.05 Interaction: 4.15	2, 13 3, 6 6, 39	0.005** 0.001***. 0.003**.	0.55 0.94 0.39
Statistics for clusters showing graded valence (Fig.7)						
Graded cluster Fig. 7a Fig. S7a	Experimental CS+15	Test Day 1, 15, and 30	Day: F =2.89 Frequency: F=29.28 Interaction: 0.351	2, 12 3, 12 6,36	0.10 0.001***. 0.905.	0.33 0.88 0.06
Graded cluster Fig. 7c Fig. S8a	Experimental CS+3	Test Day 1, 15, and 30	Day F =0.467 Frequency: F=53.37 Interaction: 2.63	2, 8 3,12 6, 24	0.643 0.001***. 0.042*.	0.11 0.93 0.40

Table S6. Statistical analyses corresponding to the normalized baseline-to-stimulus firing rate (BSR) for the neuronal clusters shown in Figs. S7–S9.

BSRs for the clusters shown in Fig. S7-S9 are presented in Fig. S10, whereas BSRs for the graded clusters shown in Figs. S7-S8 are presented in Figs. 7h–i. For pure tone-selective clusters, two-way repeated-measures ANOVAs were used to evaluate the effects of group (CS+15, CS+3, or No Shock), frequency (3, 7, 11, and 15 kHz), and their interaction. For graded clusters, one-way repeated-measures ANOVAs were used to evaluate the effect of time. Tukey's multiple-comparisons tests are reported in Figs. S10 and 7h–i.

Fig. 8

Positive modulated single-frequency neurons

Fig. 8a

Modulation	Variable	Sum Sq.	Mean Sq.	F	df	p	ηp²
Positive	Group	15.905	7.952	19.447	2, 15	< 0.001***	0.72
	Frequency	0.832	0.277	1.108	3, 45	0.356	0.07
	Group x Frequency	6.319	1.053	4.21	6, 45	0.002**	0.36

Tukey post hoc comparisons

Tone	Comparison	Diff. of means	df	q	p
3 kHz	CS+3 vs. No shock	0.757	3	3.284	0.061
	CS+3 vs. CS+15	0.374	3	1.677	0.467
	CS+15 vs. No shock	0.383	3	1.810	0.412
7 kHz	CS+15 vs. No shock	0.969	3	4.574	< 0.006**
	CS+15 vs. CS+3	0.341	3	1.530	0.529
	CS+3 vs. No shock	0.628	3	2.723	0.141
11 kHz	CS+15 vs. No shock	1.609	3	7.597	< 0.001***
	CS+15 vs. CS+3	1.363	3	6.116	< 0.001***
	CS+3 vs. No shock	0.246	3	1.065	0.733
15 kHz	CS+15 vs. No shock	1.443	3	6.811	< 0.001***
	CS+15 vs. CS+3	1.218	3	5.464	< 0.001***
	CS+3 vs. No shock	0.225	3	0.975	0.771

Fig. 8b

Modulation	Variable	Sum Sq.	Mean Sq.	F	df	p	ηp²
Positive	Group	0.207	0.104	6.62	2, 15	0.008**	0.47
	Day	0.0178	0.0089	1.205	2, 26	0.316	0.08
	Group x Day	0.0379	0.00948	1.281	4, 26	0.303	0.16

Tukey post hoc comparisons

Comparison	Diff. of means	df	q	p
No shock vs. CS+15	0.168	3	5.026	< 0.008**
No shock vs. CS+3	0.117	3	3.316	0.080
CS+3 vs. CS+15	0.0513	3	1.654	0.488

Fig. 8c

Modulation	Variable	Sum Sq.	Mean Sq.	F	df	p	ηp²
Positive	Group	0.49	0.245	5.277	2, 15	0.018*	0.41
	Condition	3.962	3.962	254.641	1, 15	< 0.001***	0.94
	Group x Condition	0.318	0.159	10.205	2, 15	0.002**	0.58

Tukey post hoc comparisons

Factor	Comparison	Diff. of means	df	q	p
within CS+15	Tone vs. freezing	0.899	2	19.078	< 0.001***
within CS+3	Tone vs. freezing	0.651	2	11.676	< 0.001***

Table S7. Statistical analysis of positive and negative modulated single-frequency and graded neurons shown in Fig. 8.

Two-way ANOVAs with repeated measures were conducted to evaluate effects of group (CS+15, CS+3, or No shock), frequency (3, 7, 11, 15 kHz), day, condition (tone vs. freezing), and their interactions. Tukey post hoc comparisons are shown where simple effects were test

within No shock	Tone vs. freezing	0.459	2	9.005	< 0.001***
within freezing	CS+15 vs. No shock	0.0531	3	0.767	0.851
	CS+15 vs. CS+3	0.0317	3	0.435	0.949
	CS+3 vs. No shock	0.0214	3	0.284	0.978
within tone	CS+15 vs. No shock	0.494	3	7.134	< 0.001***
	CS+15 vs. CS+3	0.280	3	3.839	0.031*
	CS+3 vs. No shock	0.214	3	2.843	0.131

Fig. 8d

Modulation	Variable	Sum Sq.	Mean Sq.	F	df	p	ηp²
Positive	Group	83.099	41.549	0.763	2, 15	0.482	0.09
	Day	62.468	31.234	0.649	2, 26	0.531	0.05
	Group x Day	130.193	32.548	0.677	4, 26	0.614	0.09

Negative modulated single-frequency neurons

Fig. 8e

Modulation	Variable	Sum Sq.	Mean Sq.	F	df	p	ηp²
Negative	Group	1.765	0.882	11.246	2, 15	0.001***	0.60
	Frequency	0.121	0.0404	1.274	3, 45	0.295	0.08
	Group x Frequency	0.166	0.0277	0.872	6, 45	0.523	0.10

Tukey post hoc comparisons

Comparison	Diff. of means	df	q	p
No shock vs. CS+15	0.365	3	6.624	< 0.001***
No shock vs. CS+3	0.142	3	2.361	0.249
CS+3 vs. CS+15	0.223	3	3.852	0.039*

Fig. 8f

Modulation	Variable	Sum Sq.	Mean Sq.	F	df	p	ηp²
Negative	Group	1.474	0.737	18.566	2, 15	< 0.001***	0.71
	Frequency	0.131	0.0656	1.509	2, 26	0.240	0.10
	Group x Frequency	0.185	0.0462	1.062	4, 26	0.395	0.14

Tukey post hoc comparisons

Comparison	Diff. of means	df	q	p
No shock vs. CS+15	0.451	3	8.650	< 0.001***
No shock vs. CS+3	0.284	3	5.176	0.006**
CS+3 vs. CS+15	0.167	3	3.450	0.067

Fig. 8g

Modulation	Variable	Sum Sq.	Mean Sq.	F	df	p	ηp²
Negative	Group	0.149	0.0746	4.47	2, 15	0.030*	0.37
	Day	0.261	0.261	39.684	1, 15	< 0.001***	0.73
	Group x Day	0.174	0.087	13.233	2, 15	< 0.001***	0.64

Tukey post hoc comparisons

Factor	Comparison	Diff. of means	df	q	p

Table S7. (continued)

within CS+15	Tone vs. freezing	0.343	2	11.207	< 0.001***
within CS+3	Tone vs. freezing	0.154	2	4.257	< 0.009**
within No shock	Tone vs. freezing	0.0178	2	0.538	0.709
within freezing	No shock vs. CS+15	0.0125	3	0.294	0.977
	No shock vs. CS+3	0.00849	3	0.184	0.991
	CS+3 vs. CS+15	0.00397	3	0.0889	0.998
within tone	CS+15 vs. No shock	0.313	3	7.379	< 0.001***
	CS+15 vs. CS+3	0.185	3	4.144	0.019*
	CS+3 vs. No shock	0.128	3	2.772	0.143

Fig. 8h

Modulation	Variable	Sum Sq.	Mean Sq.	F	df	p	ηp²
Negative	Group	1464.631	732.315	7.469	2, 15	0.005**	0.50
	Day	177.73	88.865	1.756	2, 26	0.193	0.12
	Group x Day	1316.013	23.848	0.471	4, 26	0.756	0.07

Tukey post hoc comparisons

Comparison	Diff. of means	df	q	p
No shock vs. CS+15	14.17	3	5.354	< 0.005**
No shock vs. CS+3	8.058	3	2.890	0.136
CS+3 vs. CS+15	6.112	3	2.492	0.216

Positive modulated graded neurons

Fig. 8i

Modulation	Variable	Sum Sq.	Mean Sq.	F	df	p	ηp²
Positive	Group	0.449	0.449	29.231	1, 15	< 0.001***	0.66
	Frequency	0.0709	0.977	0.977	3, 30	0.417	0.09
	Group x Frequency	3.933	1.311	54.221	3, 30	< 0.001***	0.84

Tukey post hoc comparisons

Tone	Comparison	Diff. of means	df	q	p
3 kHz	CS+3 vs. CS+15	0.757	2	12.335	< 0.001***
7 kHz	CS+15 vs. CS+3	0.215	2	3.495	0.018*
11 kHz	CS+15 vs. CS+3	0.624	2	10.173	< 0.001***
15 kHz	CS+15 vs. CS+3	0.702	2	11.446	< 0.001***

Fig. 8j

Modulation	Variable	Sum Sq.	Mean Sq.	F	df	p	ηp²
Positive	Group	0.055	0.055	4.795	1, 10	0.053	0.32
	Frequency	0.0933	0.0467	6.548	2, 19	0.007**	0.41
	Group x Frequency	0.0929	0.0464	6.518	2, 19	0.007**	0.41

Tukey post hoc comparisons

Factor	Comparison	Diff. of means	df	q	p
within CS+15	15 vs. 2	0.00771	3	0.228	0.986
	30 vs. 2	0.0194	3	0.575	0.914
	30 vs. 15	0.0117	3	0.368	0.964

Table S7. (continued)

within CS+3	2 vs. 15	0.236	3	6.256	< 0.001***
	2 vs. 30	0.0346	3	0.916	0.796
	30 vs. 15	0.202	3	5.340	0.004**
within 2	CS+3 vs. CS+15	0.180	2	4.454	0.004**
within 15	CS+15 vs. CS+3	0.0639	2	1.661	0.250
within 30	CS+3 vs. CS+15	0.126	2	3.275	0.028*

Fig. 8k

Modulation	Variable	Sum Sq.	Mean Sq.	F	df	p	ηp²
positive	Group	0.779	0.779	4.761	1, 10	0.054	0.32
	Condition	17.748	17.748	221.136	1, 23	< 0.001***	0.91
	Group x Condition	0.671	0.671	8.366	1, 23	< 0.016*	0.27

Tukey post hoc comparisons

Factor	Comparison	Diff. of means	df	q	p
within CS+15	Tone vs. freezing	2.084	2	19.458	< 0.001***
within CS+3	Tone vs. freezing	1.405	2	11.090	< 0.001***
within freezing	CS+15 vs. CS+3	0.0261	2	0.181	0.900
within tone	CS+15 vs. CS+3	0.705	2	4.874	0.003**

Fig. 8l

Modulation	Variable	Sum Sq.	Mean Sq.	F	df	p	ηp²
Positive	Group	5.092	5.092	0.105	1, 10	0.752	0.01
	Condition	17.586	8.793	0.470	2, 19	0.632	0.05
	Group x Condition	20.386	10.193	0.545	2, 19	0.588	0.05

Negative modulated graded neurons

Fig. 8m

Modulation	Variable	Sum Sq.	Mean Sq.	F	df	p	ηp²
Negative	Group	0.513	0.513	13.677	1, 10	0.004**	0.58
	Frequency	0.00725	0.00242	0.0836	3, 30	0.968	0.01
	Group x Frequency	6.404	2.135	73.84	3, 30	< 0.001***	0.88

Tukey post hoc comparisons

Tone	Comparison	Diff. of means	df	q	p
3 kHz	CS+15 vs. CS+3	1.006	2	13.781	< 0.001***
7 kHz	CS+3 vs. CS+15	0.229	2	3.131	0.033*
11 kHz	CS+3 vs. CS+15	0.776	2	10.628	< 0.001***
15 kHz	CS+3 vs. CS+15	0.841	2	11.519	< 0.001***

Fig. 8n

Modulation	Variable	Sum Sq.	Mean Sq.	F	df	p	ηp²
Negative	Group	0.00308	0.00308	0.122	1, 10	0.733	0.01
	Day	0.0437	0.0218	1.323	2, 15	0.296	0.15
	Group x Day	0.0299	0.015	0.907	2, 15	0.425	0.11

Table S7. (continued)

Fig. 8o

Modulation	Variable	Sum Sq.	Mean Sq.	F	df	p	ηp²
Negative	Group	0.0609	0.0609	1.421	1, 10	0.261	0.12
	Day	3.081	3.081	443.863	1, 10	< 0.001***	0.98
	Group x Day	0.0436	0.0436	6.285	2, 15	0.031*	0.46

Tukey post hoc comparisons

Factor	Comparison	Diff. of means	df	q	p
within CS+15	Tone vs. freezing	0.813	2	25.825	< 0.001***
within CS+3	Tone vs. freezing	0.640	2	17.184	< 0.001***
within freezing	CS+15 vs. CS+3	0.0157	2	0.240	0.868
within tone	CS+15 vs. CS+3	0.189	2	2.888	0.062

Fig. 8p

Modulation	Variable	Sum Sq.	Mean Sq.	F	df	p	ηp²
Negative	Group	174.242	174.242	2.001	1, 10	0.188	0.17
	Day	11.597	5.799	0.065	2, 19	0.937	0.01
	Group x Day	149.463	74.732	0.838	2, 19	0.448	0.08

Table S7. (continued)

Acknowledgements

This work has been funded by NSF (NSF/IOS 2303305 to IAM), NIH (R01 MH123260-01 to IAM; RISE GMO60655 to MRL).

Additional information

Author Contributions

MEN developed method to identify subpopulations, wrote code for analysis, and contributed to writing the manuscript, PMO collected data, conducted behavioral and in vivo recording experiments, and contributed to writing the manuscript, MRL contributed to experimental design, collected data, conducted behavioral and in vivo recording experiments. MMRA contributed to the statistical analysis. IAM supervised design, experiments, analysis, and writing of the manuscript.

Funding

Funder	Grant reference number	Author
National Science Foundation (NSF)	2303305	Isabel A Muzzio
National Institute of Mental Health and Neurosciences (NIMHANS)	MH123260-01	Isabel A Muzzio

Author ORCID iDs

Isabel A Muzzio: https://orcid.org/0000-0003-0156-0088

References

Bontempi B, Jaffard R, Destrade C (1996) Differential temporal evolution of post-training changes in regional brain glucose metabolism induced by repeated spatial discrimination training in mice: visualization of the memory consolidation process?. *Eur J Neurosci* **8**:2348-2360 https://doi.org/10.1111/j.1460-9568.1996.tb01198.x | PubMed

Bouton ME, Maren S, McNally GP (2021) Behavioral and Neurobiological Mechanisms of Pavlovian and Instrumental Extinction Learning. *Physiol Rev* **101**:611-681 https://doi.org/10.1152/physrev.00016.2020

Burgos-Robles A, Vidal-Gonzalez I, Quirk GJ (2009) Sustained conditioned responses in prelimbic prefrontal neurons are correlated with fear expression and extinction failure. *J Neurosci* **29**:8474-8482 https://doi.org/10.1523/JNEUROSCI.0378-09.2009 | PubMed

Casanova JP, Pouget C, Treiber N, Agarwal I, Brimble MA, Vetere G (2024) Threat-dependent scaling of prelimbic dynamics to enhance fear representation. *Neuron* **112**:2304-2314.e2306. https://doi.org/10.1016/j.neuron.2024.04.029 | PubMed

Chavez RS, Heatherton TF (2015) Representational similarity of social and valence information in the medial pFC. *J Cogn Neurosci* **27**:73-82 https://doi.org/10.1162/jocn_a_00697 | PubMed

Corches A, Hiroto A, Bailey TW, Speigel JH, Pastore J, Mayford M, Korzus E (2019) Differential fear conditioning generates prefrontal neural ensembles of safety signals. *Behav Brain Res* **360**:169-184 https://doi.org/10.1016/j.bbr.2018.11.042 | PubMed

Deitch D, Rubin A, Ziv Y (2021) Representational drift in the mouse visual cortex. *Curr Biol* **31**:4327-4339.e4326. https://doi.org/10.1016/j.cub.2021.07.062 | PubMed

DeNardo LA, Liu CD, Allen WE, Adams EL, Friedmann D, Fu L, Guenthner CJ, Tessier-Lavigne M, Luo L (2019) Temporal evolution of cortical ensembles promoting remote memory retrieval. *Nat Neurosci* **22**:460-469 https://doi.org/10.1038/s41593-018-0318-7 | PubMed

Do-Monte FH, Quinones-Laracuente K, Quirk GJ (2015) A temporal shift in the circuits mediating retrieval of fear memory. *Nature* **519**:460-463 https://doi.org/10.1038/nature14030 | PubMed

Dunsmoor JE, LaBar KS (2013) Effects of discrimination training on fear generalization gradients and perceptual classification in humans. *Behav Neurosci* **127**:350-356 https://doi.org/10.1037/a0031933 | PubMed

Evangelidis GD, Psarakis EZ (2008) Parametric image alignment using enhanced correlation coefficient maximization. *IEEE Trans Pattern Anal Mach Intell* **30**:1858-1865 https://doi.org/10.1109/TPAMI.2008.113 | PubMed

Frankland PW, Bontempi B (2005) The organization of recent and remote memories. *Nat Rev Neurosci* **6**:119-130 https://doi.org/10.1038/nrn1607 | PubMed

Gagliardi CM, Normandin ME, Keinath AT, Julian JB, Lopez MR, Ramos-Alvarez MM, Epstein RA, Muzzio IA (2024) Distinct neural mechanisms for heading retrieval and context recognition in the hippocampus during spatial reorientation. *Nat Commun* **15**:5968 https://doi.org/10.1038/s41467-024-50112-7 | PubMed

Gallego JA, Perich MG, Chowdhury RH, Solla SA, Miller LE (2020) Long-term stability of cortical population dynamics underlying consistent behavior. *Nat Neurosci* **23**:260-270 https://doi.org/10.1038/s41593-019-0555-4 | PubMed

Grosso A, Santoni G, Manassero E, Renna A, Sacchetti B (2018) A neuronal basis for fear discrimination in the lateral amygdala. *Nat Commun* **9**:1214 https://doi.org/10.1038/s41467-018-03682-2 | PubMed

Herzog K, Andreatta M, Schneider K, Schiele MA, Domschke K, Romanos M, Deckert J, Pauli P (2021) Reducing Generalization of Conditioned Fear: Beneficial Impact of Fear Relevance and Feedback in Discrimination Training. *Front Psychol* **12** https://doi.org/10.3389/fpsyg.2021.665711 | PubMed

Hockley A, Malmierca MS (2024) Auditory processing control by the medial prefrontal cortex: A review of the rodent functional organisation. *Hear Res* **443**:108954 https://doi.org/10.1016/j.heares.2024.108954 | PubMed

Hoshi A, Hirayama Y, Saito F, Ishiguro T, Suetani H, Kitajo K (2023) Spatiotemporal consistency of neural responses to repeatedly presented video stimuli accounts for population preferences. *Sci Rep* **13**:5532 https://doi.org/10.1038/s41598-023-31751-0 | PubMed

Huxter J, Burgess N, O'Keefe J (2003) Independent rate and temporal coding in hippocampal pyramidal cells. *Nature* **425**:828-832 https://doi.org/10.1038/nature02058 | PubMed

Iqbal J, Kim S, Lawal S, Shah A, Punepalle L, Sanghvi H, Gallagher A, Wilson A, Xu B, Shrestha P (2026) A prelimbic molecular clock of protein synthesis for memory persistence. *bioRxiv* https://doi.org/10.64898/2026.01.02.697403 | PubMed

Jenkins HM, Harrison RH (1960) Effect of discrimination training on auditory generalization. *J Exp Psychol* **59**:246-253 https://doi.org/10.1037/h0041661 | PubMed

Kato HK, Asinof SK, Isaacson JS (2017) Network-Level Control of Frequency Tuning in Auditory Cortex. *Neuron* **95**:412-423.e414. https://doi.org/10.1016/j.neuron.2017.06.019 | PubMed

Kitamura T, Ogawa SK, Roy DS, Okuyama T, Morrissey MD, Smith LM, Redondo RL, Tonegawa S (2017) Engrams and circuits crucial for systems consolidation of a memory. *Science* **356**:73-78 https://doi.org/10.1126/science.aam6808 | PubMed

Kupke J, Oliveira AMM (2025) The molecular and cellular basis of memory engrams: Mechanisms of synaptic and systems consolidation. *Neurobiol Learn Mem* **219**:108057 https://doi.org/10.1016/j.nlm.2025.108057 | PubMed

Kyriazi P, Headley DB, Pare D (2020) Different Multidimensional Representations across the Amygdalo-Prefrontal Network during an Approach-Avoidance Task. *Neuron* **107**:717-730.e715. https://doi.org/10.1016/j.neuron.2020.05.039 | PubMed

Lacagnina AF, Brockway ET, Crovetti CR, Shue F, McCarty MJ, Sattler KP, Lim SC, Santos SL, Denny CA, Drew MR (2019) Distinct hippocampal engrams control extinction and relapse of fear memory. *Nat Neurosci* **22**:753-761 https://doi.org/10.1038/s41593-019-0361-z | PubMed

Laufer O, Israeli D, Paz R (2016) Behavioral and Neural Mechanisms of Overgeneralization in Anxiety. *Curr Biol* **26**:713-722 https://doi.org/10.1016/j.cub.2016.01.023 | PubMed

Levy I, Schiller D (2021) Neural Computations of Threat. *Trends Cogn Sci* **25**:151-171 https://doi.org/10.1016/j.tics.2020.11.007 | PubMed

Likhtik E, Stujenske JM, Topiwala MA, Harris AZ, Gordon JA (2014) Prefrontal entrainment of amygdala activity signals safety in learned fear and innate anxiety. *Nat Neurosci* **17**:106-113 https://doi.org/10.1038/nn.3582 | PubMed

Lissek S, Bradford DE, Alvarez RP, Burton P, Espensen-Sturges T, Reynolds RC, Grillon C (2014) Neural substrates of classically conditioned fear-generalization in humans: a parametric fMRI study. *Soc Cogn Affect Neurosci* **9**:1134-1142 https://doi.org/10.1093/scan/nst096 | PubMed

Lommen MJJ, Duta M, Vanbrabant K, de Jong R, Juechems K, Ehlers A (2017) Training discrimination diminishes maladaptive avoidance of innocuous stimuli in a fear conditioning paradigm. *PLoS One* **12**:e0184485 https://doi.org/10.1371/journal.pone.0184485 | PubMed

Lopez MR, Wasberg SMH, Gagliardi CM, Normandin ME, Muzzio IA (2024) Mystery of the memory engram: History, current knowledge, and unanswered questions. *Neurosci Biobehav Rev* **159**:105574 https://doi.org/10.1016/j.neubiorev.2024.105574 | PubMed

Mau W, Hasselmo ME, Cai DJ (2020) The brain in motion: How ensemble fluidity drives memory-updating and flexibility. *eLife* **9** https://doi.org/10.7554/eLife.63550 | PubMed

Moscovitch M, Nadel L (1998) Consolidation and the hippocampal complex revisited: in defense of the multiple-trace model. *Curr Opin Neurobiol Current Opinion in Neurobiology* **8**:297-300 https://doi.org/10.1016/s0959-4388(98)80155-4 | PubMed

Nadel L, Samsonovich A, Ryan L, Moscovitch M (2000) Multiple trace theory of human memory: computational, neuroimaging, and neuropsychological results. *Hippocampus* **10**:352-368 https://doi.org/10.1002/1098-1063(2000)10:4<352::aid-hipo2>3.0.co;2-d | PubMed

Newly WK, West KD (1987) A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* **55**:703-708 https://doi.org/10.2307/1913610

Phillips RG, LeDoux JE (1992) Differential contribution of amygdala and hippocampus to cued and contextual fear conditioning. *Behav Neurosci* **106**:274-285 https://doi.org/10.1037//0735-7044.106.2.274 | PubMed

Quian Quiroga R, Panzeri S (2009) Extracting information from neuronal populations: information theory and decoding approaches. *Nat Rev Neurosci* **10**:173-185 https://doi.org/10.1038/nrn2578 | PubMed

Rao-Ruiz P, Visser E, Mitric M, Smit AB, van den Oever MC (2021) A Synaptic Framework for the Persistence of Memory Engrams. *Front Synaptic Neurosci* **13**:661476 https://doi.org/10.3389/fnsyn.2021.661476 | PubMed

Refaeli R, Kreisel T, Groysman M, Adamsky A, Goshen I (2023) Engram stability and maturation during systems consolidation. *Curr Biol* **33**:3942-3950.e3943. https://doi.org/10.1016/j.cub.2023.07.042 | PubMed

Rosas-Vidal LE, Naskar S, Mayo LM, Perini I, Masroor R, Altemus M, Ramos-Medina L, Zaidi SD, Engelbrektsson H, Jagasia P, *et al.* (2025) Prefrontal correlates of fear generalization during endocannabinoid depletion. *J Clin Invest* **135** https://doi.org/10.1172/JCI179881 | PubMed

Rupprecht P, Carta S, Hoffmann A, Echizen M, Blot A, Kwan AC, Dan Y, Hofer SB, Kitamura K, Helmchen F, *et al.* (2021) A database and deep learning toolbox for noise-optimized, generalized spike inference from calcium imaging. *Nat Neurosci* **24**:1324-1337 https://doi.org/10.1038/s41593-021-00895-5 | PubMed

Sanders H, Ji D, Sasaki T, Leutgeb JK, Wilson MA, Lisman JE (2019) Temporal coding and rate remapping: Representation of nonspatial information in the hippocampus. *Hippocampus* **29**:111-127 https://doi.org/10.1002/hipo.23020 | PubMed

Sangha S (2015) Plasticity of Fear and Safety Neurons of the Amygdala in Response to Fear Extinction. *Front Behav Neurosci* **9**:354 https://doi.org/10.3389/fnbeh.2015.00354 | PubMed

Shepard RN (1987) Toward a universal law of generalization for psychological science. *Science* **237**:1317-1323 https://doi.org/10.1126/science.3629243 | PubMed

Sierra-Mercado D, Padilla-Coreano N, Quirk GJ (2011) Dissociable roles of prelimbic and infralimbic cortices, ventral hippocampus, and basolateral amygdala in the expression and extinction of conditioned fear. *Neuropsychopharmacology* **36**:529-538 https://doi.org/10.1038/npp.2010.184 | PubMed

Sotres-Bayon F, Quirk GJ (2010) Prefrontal control of fear: more than just extinction. *Curr Opin Neurobiol* **20**:231-235 https://doi.org/10.1016/j.conb.2010.02.005 | PubMed

Stujenske JM, O'Neill PK, Fernandes-Henriques C, Nahmoud I, Goldburg SR, Singh A, Diaz L, Labkovich M, Hardin W, Bolkan SS, *et al.* (2022) Prelimbic cortex drives discrimination of non-aversion via amygdala somatostatin interneurons. *Neuron* **110**:2258-2267.e2211. https://doi.org/10.1016/j.neuron.2022.03.020 | PubMed

Terranova JI, Yokose J, Osanai H, Ogawa SK, Kitamura T (2023) Systems consolidation induces multiple memory engrams for a flexible recall strategy in observational fear memory in male mice. *Nat Commun* **14**:3976 https://doi.org/10.1038/s41467-023-39718-5 | PubMed

Tome DF, Zhang Y, Aida T, Mosto O, Lu Y, Chen M, Sadeh S, Roy DS, Clopath C (2024) Dynamic and selective engrams emerge with memory consolidation. *Nat Neurosci* **27**:561-572 https://doi.org/10.1038/s41593-023-01551-w | PubMed

Tonegawa S, Morrissey MD, Kitamura T (2018) The role of engram cells in the systems consolidation of memory. *Nat Rev Neurosci* **19**:485-498 https://doi.org/10.1038/s41583-018-0031-2 | PubMed

Verra L, Spitzer B, Schuck NW, Zika O (2026) Increased generalisation in trait anxiety is driven by aversive value transfer. *Commun Psychol* **4** https://doi.org/10.1038/s44271-026-00415-w | PubMed

Wang ME, Wann EG, Yuan RK, Ramos Alvarez MM, Stead SM, Muzzio IA (2012) Long-term stabilization of place cell remapping produced by a fearful experience. *J Neurosci* **32**:15802-15814 https://doi.org/10.1523/JNEUROSCI.0480-12.2012 | PubMed

Wehr M, Zador AM (2003) Balanced inhibition underlies tuning and sharpens spike timing in auditory cortex. *Nature* **426**:442-446 https://doi.org/10.1038/nature02116 | PubMed

Zaki Y, Cai DJ (2024) Memory engram stability and flexibility. *Neuropsychopharmacology* **50**:285-293 https://doi.org/10.1038/s41386-024-01979-z | PubMed

Zaman J, Yu K, Verheyen S (2023) The idiosyncratic nature of how individuals perceive, represent, and remember their surroundings and its impact on learning-based generalization. *J Exp Psychol Gen* **152**:2345-2358 https://doi.org/10.1037/xge0001403 | PubMed

Zikopoulos B, Barbas H (2006) Prefrontal projections to the thalamic reticular nucleus form a unique circuit for attentional mechanisms. *J Neurosci* **26**:7348-7361 https://doi.org/10.1523/JNEUROSCI.5511-05.2006 | PubMed

Peer reviews

Reviewer #1 (Public review):

Summary:

The authors combine discriminative auditory fear conditioning with longitudinal in vivo calcium imaging to ask how prelimbic (PL) representations of learned and generalized threat evolve across recent and remote memory time points. Using two different CS+ frequencies and a no-shock control group, they report that PL population activity tracks graded behavioral generalization, that population similarity is highest for tones eliciting strong threat responding, and that distinct subnetworks can be identified that appear to encode tone-specific sensory features versus learned threat-related response structure.

To my knowledge, this may be the first study to comprehensively examine neural encoding of fear generalization in prelimbic cortex (PL). The manuscript is ambitious and technically interesting, and several aspects are potentially important. In particular, the suggestion that neurons showing graded, learning-related response patterns become selectively stabilized over time is intriguing. The inclusion of two CS+ training conditions and a no-shock control also strengthens the case that at least some of the reported effects are related to associative learning rather than simple sensory differences. However, in its current form, the manuscript does not yet fully support the strength of the conceptual claims. Several issues limit confidence in the interpretation, including the possibility that repeated testing itself contributes to changes across days, uncertainty about the relationship between neural activity and freezing behavior, limited quantitative documentation of longitudinal cell registration, and a number of problems in figure clarity and statistical framing. Overall, the study contains promising observations, but the claims should be narrowed, and several analyses or controls would be needed to fully support the proposed framework.

Comments on revised version.

The authors have addressed my previous concerns well, and the revised manuscript is substantially improved. In particular, the additional analyses strengthen the conclusion that prelimbic cortical activity reflects learned threat value rather than simply freezing behavior, while the revised framing and additional controls clarify the interpretation of the longitudinal neural dynamics. This paper represents an important contribution to our understanding of the neural mechanisms supporting aversive learning, memory, and generalization.

https://doi.org/10.7554/eLife.111177.2.sa3

Reviewer #2 (Public review):

The authors have substantially revised the paper in response to the original review, which is greatly appreciated. It is clear that it will eventually make a nice contribution to the literature. This being said, the following points are somewhere between major and minor in term of their implications for interpretation of the study results. If they were to be addressed, the paper would be again improved.

There are a few remnants of the past language that are not helpful re interpretation of the study results: 1) "Specifically, the observed population gradients could emerge either from the pooled activity of frequency-selective neurons that respond to individual tones or from neuronal subpopulations that integrate information across tones to encode their learned threat-value."; and 2) "Together, these findings suggest that the PL integrates sensory similarity with learned threat value to generate stable representations that support adaptive generalization and discrimination." Neither of these statement follows what has been shown in the study, even with inclusion of the results from the GLM analysis (see point 4 below).

(1) This paragraph in the Discussion is difficult to follow: "Generalization has traditionally been explained by perceptual similarity (Shepard, 1987), whereby stimuli resembling a conditioned cue recruit overlapping sensory representations and evoke similar behavioral responses (Corches et al., 2019; Grosso et al., 2018). Although perceptual similarity clearly influences the extent of generalization, accumulating evidence indicates that it cannot fully account for generalized responding (Verra et al., 2026). More recent frameworks propose that associative learning assigns learned value to novel stimuli by integrating their sensory similarity with previous experience, allowing behavior to scale according to predicted biological significance (Verra et al., 2026; Zaman et al., 2023). Our findings provide a neural framework consistent with these ideas. Sensory similarity promoted consistent neuronal population responses across tones, whereas associative learning organized these responses

into graded representations that tracked learned threat value across the stimulus continuum. Thus, sensory similarity appears to define the neuronal substrate upon which associative learning constructs value-based representations that support graded behavioral generalization."

While the revisions have removed the many unnecessary references to inference and integration, this paragraph seems like it is adhering to the original idea of how the authors wished to present their work. If the authors wished to talk about something more than perceptual similarity in the context of generalization, they should have used a task that lends itself to a more-than-perceptual-similarity explanation. Again, the inclusion of the GLM analysis is suggestive for some of what the authors wish to say, but doesn't justify the statements that: "Sensory similarity promoted consistent neuronal population responses across tones, whereas associative learning organized these responses into graded representations that tracked learned threat value across the stimulus continuum." In short, the analysis does not substitute for the design that could have and should have been used to assess learned threat value independently of sensory similarity.

(2) The next paragraph in the Discussion is also confusing. "Such reorganization has been proposed to provide flexibility by allowing new information to be incorporated into existing cortical representations while preserving stable behavioral performance (Mau et al., 2020; Zaki & Cai, 2024). Several mechanisms could contribute to this turnover, including systems consolidation, retrieval-induced reconsolidation or memory updating, and repeated nonreinforced stimulus exposure (Lacagnina et al., 2019; Mau et al., 2020; Sangha, 2015; Zaki & Cai, 2024). Although our experiments cannot distinguish between the first two possibilities, the behavioral data argue against extinction as the primary explanation. Extinction is generally associated with the formation of new CS+-safety associations (Bouton et al., 2021), whereas discrimination ratios increased across retrieval sessions, indicating that animals progressively improved their discrimination between threat-associated and safe stimuli rather than acquiring generalized safety responses. This pattern is consistent with previous work showing that discrimination learning sharpens stimulus representations and narrows behavioral generalization gradients (Dunsmoor & LaBar, 2013; Herzog et al., 2021; Jenkins & Harrison, 1960; Lommen et al., 2017). Importantly, turnover was not uniform across the population. Graded neurons retained remarkably consistent response profiles across retrieval sessions, and their activity remained more strongly associated with learned threat value than with freezing behavior. These observations indicate that stable components of the population code can coexist with extensive reorganization of surrounding neuronal ensembles."

The issue with repeated testing is *not* caused by extinction per se. The issue is that non-reinforcement across the repeated testing should differentially affect the CS+ and CS-. Specifically, it should extinguish responding to the CS- stimulus at a rate that matches its distance from the CS+, thereby sharpening the CS+ versus CS- discrimination in precisely the ways that have been observed. Ergo, the repeated testing *is* a problem for inferences that might be drawn about the way that generalization gradients change with time; and *is* a problem for statements regarding "dynamic reorganization of cortical activity patterns over time." There is nothing in the study that allows one to comment on the reorganization of cortical activity patterns over time. The reorganization can and should be attributed to the repeated testing, which is confounded with time. Nonetheless, the reorganization must be due to the repeated testing and NOT time as the present findings are inconsistent with the well-documented broadening of generalization gradients with time.

(3) In the next paragraph, the authors state: "At the same time, narrower generalization gradients and improved discrimination across retrieval sessions suggests ongoing memory updating. These observations are consistent with contemporary theories proposing that systems consolidation and retrieval-dependent updating are complementary processes

through which memories continue to evolve after learning (Mau et al., 2020; Tome et al., 2024; Zaki & Cai, 2024)."

In general, I'm not sure why one would invoke systems consolidation or retrieval-induced reconsolidation as an explanation for any of the present findings: they are not explanations of much at all. In this specific text, the authors seem to be implying an updating process that occurs independently of what is learned across the repeated sessions of testing. Why? The changes that occur in the behaviour and neuronal representations are perfectly explicable in terms of additional learning that occurs - of the sort that I hope to have made clear in my previous comment. Why invoke more than what is needed to explain the observed pattern of results?

(4) Re the GLM analysis - The authors write that: "the fact that the GLM analysis indicates that these neurons reflect learned threat value more than freezing behavior, suggests that they encode an abstract property of the learned stimulus rather than simply mirroring behavioral output."

This is fine if freezing fully indexes the state of conditioned fear and there are no other behaviours in which animals express their fear. If, however, fear is expressed in a range of other behaviours that are likely coordinated by the PL (e.g., startle, vigilance, scanning, orienting to source of danger), this interpretation of the GLM analysis is unwarranted. This is an important point and would be worth noting somewhere in the paragraph where the statement appears.

https://doi.org/10.7554/eLife.111177.2.sa2

Reviewer #3 (Public review):

Summary:

Normandin et al. explore the coding of stimuli predicting an aversive event in the prelimbic cortex. Stimuli could either be explicitly paired, explicitly unpaired, or novel but with an inferred association with the aversive event (generalization). Long-term tracking of GCaMP positive neurons allowed them to examine how coding evolves out to a month following training. In general, they found two types of ensemble codes. One was ensembles coding for each stimulus independently, but with enhanced responding to the one eliciting a freezing response. The other was ensembles that responded to all stimuli in proportion to their similarity to the stimulus paired with the aversive event, either increasing or decreasing their activation with the degree of freezing elicited by a stimulus. Importantly, this second set of ensembles was more stable across days, potentially providing a memory trace.

Strengths:

(1) The authors track ensembles in prelimbic cortex over long time scales, providing valuable information on the consolidation of neural codes.

(2) Neural coding of generalization is examined, which is under examined in the field.

Comments on revised version.

The authors have convincingly and thoroughly addressed my concerns. I have no further issues regarding this study.

https://doi.org/10.7554/eLife.111177.2.sa1

Author response:

The following is the authors' response to the original reviews.

> ***Public review:***
>
> ***Reviewer #1 (Public review):***
>
> *Summary:*
>
> *The authors combine discriminative auditory fear conditioning with longitudinal in vivo calcium imaging to ask how prelimbic (PL) representations of learned and generalized threat evolve across recent and remote memory time points. Using two different CS+ frequencies and a no-shock control group, they report that PL population activity tracks graded behavioral generalization, that population similarity is highest for tones eliciting strong threat responding, and that distinct subnetworks can be identified that appear to encode tone-specific sensory features versus learned threat-related response structure. To my knowledge, this may be the first study to comprehensively examine neural encoding of fear generalization in prelimbic cortex (PL). The manuscript is ambitious and technically interesting, and several aspects are potentially important. In particular, the suggestion that neurons showing graded, learning-related response patterns become selectively stabilized over time is intriguing. The inclusion of two CS+ training conditions and a no-shock control also strengthens the case that at least some of the reported effects are related to associative learning rather than simple sensory differences. However, in its current form, the manuscript does not yet fully support the strength of the conceptual claims. Several issues limit confidence in the interpretation, including the possibility that repeated testing itself contributes to changes across days, uncertainty about the relationship between neural activity and freezing behavior, limited quantitative documentation of longitudinal cell registration, and a number of problems in figure clarity and statistical framing. Overall, the study contains promising observations, but the claims should be narrowed, and several analyses or controls would be needed to fully support the proposed framework.*
>
> *Detailed Comments*
>
> *(1) A general concern is that the repeated test procedure itself may contribute to extinction. Because the animals are exposed to multiple CS frequencies across multiple test days, and each tone is presented three times per session, some of the reported changes in behavior and neural activity across days could reflect extinction or repeated nonreinforced retrieval rather than the passage of time per se. This is especially relevant given that the manuscript makes claims about recent versus remote representations and representational drift over 30 days. At a minimum, the authors should discuss this limitation explicitly and temper claims about time-dependent changes. Ideally, they would include a control group in which animals are tested only once or twice (e.g., at an early and later time point with fewer CS frequencies), or a reduced-frequency testing design that minimizes extinction while still allowing evaluation of recent versus remote memory.*

We agree with the reviewer that repeated testing is an inherent limitation of longitudinal memory studies and may itself contribute to neural changes across sessions. Repeated retrieval can induce memory updating (reconsolidation) or extinction, the latter involving the formation of a new association between the CS+ and safety. Although memory updating may have contributed to the ensemble reorganization observed here, several aspects of our findings argue against extinction as the primary explanation for the observed neural changes.

First, we observed substantial neuronal ensemble turnover beginning with the first retrieval session. This early turnover is consistent with previous observations in the prefrontal cortex [1, 2] and with growing evidence that cortical memory representations remain dynamic throughout systems consolidation [3, 4]. Longitudinal studies have shown that neurons are continuously recruited into and removed from cortical memory ensembles while memory expression remains stable [1-4].

Second, we calculated discrimination ratios to quantify discrimination of each tone relative to the CS+ across retrieval sessions (Figure S1). These analyses showed that discrimination increased, rather than decreased, over successive retrieval sessions, a pattern inconsistent with the behavioral profile expected if repeated testing had induced extinction.

Finally, one of the most novel findings of our study is that ensemble turnover does not affect all neuronal populations equally. The graded neurons identified by our clustering analysis maintained their identity and functional organization across retrieval sessions, and their activity was better explained by tone threat value than by freezing behavior (Figure 8). This selective stability indicates that ensemble reorganization is not a uniform process but instead preferentially affects specific neuronal subpopulations while preserving a stable threat-value generalization gradient. Thus, although repeated retrieval may contribute to ongoing ensemble reorganization, our results demonstrate that this process is selective and largely spares the neuronal subpopulations that encode graded threat-value representations.

Accordingly, we have revised the Discussion to explicitly acknowledge these points as follows:

“The ensemble turnover observed here is consistent with previous studies demonstrating dynamic reorganization of cortical activity patterns over time [1-3, 5]. Such reorganization has been proposed to provide flexibility by allowing new information to be incorporated into existing cortical representations while preserving stable behavioral performance [4, 6]. Several mechanisms could contribute to this turnover, including systems consolidation, retrieval-induced reconsolidation or memory updating, and repeated nonreinforced stimulus exposure [4, 6-8]. Although our experiments cannot distinguish between the first two possibilities, the behavioral data argue against extinction as the primary explanation. Extinction is generally associated with the formation of new CS+-safety associations [9], whereas discrimination ratios increased across retrieval sessions, indicating that animals progressively improved their discrimination between threat-associated and safe stimuli rather than acquiring generalized safety responses. This pattern is consistent with previous work showing that discrimination learning sharpens stimulus representations and narrows behavioral generalization gradients [10-13]. Importantly, turnover was not uniform across the population. Graded neurons retained remarkably consistent response profiles across retrieval sessions, and their activity remained more strongly associated with learned threat value than with freezing behavior. These observations indicate that stable components of the population code can coexist with extensive reorganization of surrounding neuronal ensembles.” Pg. 19

> *(2) More generally, some of the reported learning-related neural differences may be driven by behavioral differences, particularly freezing, rather than by learning or generalization per se. For example, animals that freeze more to certain frequencies may show corresponding neural response differences simply because freezing alters PL activity. The authors should examine this possibility more directly. Analyses testing whether recorded cells encode freezing behavior, or whether tone frequency-related neural differences remain robust when comparing high- and low-freezing epochs, would help determine whether the reported effects reflect learned stimulus value rather than behavioral state differences.*

This is an important point, which was also highlighted by the other reviewers. To directly address this concern, we implemented the generalized linear model (GLM) analysis suggested

by Reviewer 3. We modeled the neuronal activity time series using both tone identity and freezing behavior as simultaneous predictors. Because tone identity was fixed across trials whereas freezing varied from trial to trial, the GLM allowed us to dissociate their independent contributions to neuronal activity.

As described in the original submission, freezing was estimated from the miniscope's onboard inertial measurement unit (IMU), which measures body acceleration along three axes. Rather than classifying freezing using a fixed threshold, we estimated the continuous probability of freezing from the accelerometer signal using a Gaussian mixture model. This probabilistic estimate was incorporated directly into the GLM together with tone identity, providing a conservative test of whether neuronal activity was better explained by freezing behavior or by the auditory stimulus.

We applied the GLM both to all sound-responsive neurons contributing to the population response curves (Figure 4) and to the graded and frequency-selective neuronal subpopulations identified by our clustering analysis (Figure 8). Across both experimental groups and all analyses, the median regression coefficients (β) associated with tone identity were consistently larger than those associated with freezing, indicating that tone identity contributed more strongly to neuronal activity. Moreover, tone coefficients exhibited graded monotonic profiles that closely tracked the learned threat value of each tone, with graded neurons showing the strongest gradients (Figures 4a, 8a, and 8e). Consistent with previous reports [14, 15] freezing accounted for a modest but significant component of PL activity. However, only 6–8% of graded neurons were classified as freezing-dominant, indicating that for the vast majority of these neurons, tone identity was the stronger predictor. Together, these findings demonstrate that the graded representation of learned threat value persists after accounting for freezing behavior, supporting our conclusion that PL activity reflects learned threat value rather than merely the behavioral expression of fear.

> *(3) A central feature of the manuscript is the analysis of neural response properties over an extended period of time, up to 30 days after learning. However, aside from a brief mention in the Methods that spatial registration was used, the manuscript provides very little quantitative information about this critical aspect of the study. The paper would be strengthened by including explicit metrics describing longitudinal cell tracking, such as the number and proportion of ROIs retained across all sessions, distributions of spatial-footprint correlations or centroid distances across days, and representative examples of matched imaging fields over time. Without this information, it is difficult to assess how strongly the longitudinal claims are supported.*

We thank the reviewer for this suggestion. We now include measures of registration quality in the resubmission. Specifically, we calculated shifts in centroid distances, proportion of ROIs retained across all sessions, and representative examples of matched imaging fields over time (Fig, S3).

> *(4) The text states that "Figs. 1c and 1d show GCaMP6f expression in PL, representative calcium footprints, and activity traces". However, the figure as presented does not clearly show all of these elements, at least not in a way that matches the description in the Results. The correspondence between text and figure should be corrected.*

We corrected correspondence between text and Figure.

> *(5) The labeling of Figure 2a is insufficient for interpretation. The legend states that the panel shows raster plots of sound responsiveness, but the axes and scaling are not clearly defined. It is not clear from the figure what the x-axis represents, whether the y-axis corresponds to individual neurons, where the CS period occurs, or what the activity scale at the right denotes. Also, the term 'rasters' implies that spikes were analyzed. It seems that the spike inference approach (CASCADE) was only used for later analyses.*

> *Perhaps 'heat-plot' would be more accurate here? Generally, this figure should be annotated more clearly so that the reader can understand it without referring back to the Methods.*

We clarified the labelling of the Figure 2a and call the graphs “activity-plots”.

> *(6) In relation to Figure 3, the analysis of population-averaged responses across tone frequencies is useful, but the manuscript would be stronger with additional statistical analyses across time and across groups. For example, if the authors want to argue that learning induces graded changes in neural responses and that these evolve across time, they should directly compare within-group responses across days and also compare matched frequencies between the conditioned groups and the no-shock controls. These analyses would help establish whether the observed differences are genuinely learning dependent and whether they change significantly over time.*

For Figure 3, we maintained the previous one-way ANOVAs assessing changes in AUC per day to be able to note significance on the Figure panels. However, we added a three-way mixed-effects analysis, using group (CS15, CS3, no shocks), frequency (3, 7, 11, 15), and day of testing (2, 15, 30) as variables, with frequency and day of testing as repeated measures. The results were described as follows (statistical details Table S1):

“To determine how AUC varied across groups over time, we performed a three-way mixed-effects ANOVA with group (CS+15, CS+3, and no shock), frequency (3, 7, 11, and 15 kHz), and time (test days 1, 15, and 30) as factors, with repeated measures on frequency and time. For positive responder neurons, the analysis revealed significant main effects of group ($p < 0.001$) and time ($p < 0.05$), as well as a significant group × frequency interaction ($p < 0.001$), whereas the time × frequency and group × time × frequency interactions were not significant ($p > 0.05$; Table S2a). Tukey-corrected post hoc comparisons showed that, in the CS+15 group, AUC differed between all frequency pairs except 11 and 15 kHz ($p < 0.05$). In the CS+3 group, the AUC at 3 kHz differed from those at 7, 11, and 15 kHz ($p < 0.05$), whereas no significant frequency differences were observed in the no-shock controls ($p > 0.05$). For negative responder neurons, the only significant effect was a time × frequency interaction ($p < 0.01$). Tukey-corrected simple-effects analyses revealed that, on day 30, the AUC at 15 kHz differed from those at 3, 7, and 11 kHz ($p < 0.05$; Table S2b). Because this pattern was observed across all experimental groups, including the no-shock controls, it is unlikely to reflect associative learning. These results indicate that although the AUC exhibited modest changes over time, these changes were not group-specific and therefore do not support learning-dependent alterations in neuronal responses. Together, these results show that despite substantial neuronal turnover, PL population responses encode generalization gradients, closely matching behavioral expression.” Pg. 9

> *(7) The inclusion of two different CS+ frequencies and a no-shock control is a strength of the study and substantially improves the interpretation that graded neural responses are related to learning and generalization rather than to simple sensory processing or passage of time. That said, I am not entirely comfortable with the use of the term "inference" throughout the manuscript. What is being measured here appears closer to sensory generalization than inference in a stronger cognitive sense. The current task does not clearly require that animals infer hidden structure or stimulus value through abstract reasoning; rather, the generalized stimulus may simply be treated as similar to the conditioned cue. The terminology should therefore be reconsidered or softened.*

We thank the reviewer for appreciating the strengths of the experimental design and for this thoughtful suggestion regarding terminology. We agree that the term inference may overstate the cognitive processes engaged by the current task. Accordingly, we revised the terminology throughout the manuscript to describe these effects as graded generalization of threat value across stimuli. The new GLM analyses further support this interpretation by demonstrating

that, in the conditioned groups, neuronal activity at both the population and single-neuron levels is explained substantially better by tone identity than by freezing behavior (Figures 4 and 8). We therefore retained the term threat value, as our results indicate that PL activity primarily reflects learned threat value rather than simply the expression of freezing behavior, but removed inference.

> *(8) I also found the use of the term "valence" somewhat problematic. The manuscript appears to use valence to refer to graded responding across tones with different aversive significance, but valence typically refers more broadly to distinctions between appetitive and aversive value. Here, terms such as "threat value," "aversive value," may be more precise. The authors should consider revising this language throughout.*

We corrected the language and replaced valence for “threat value”

> ***Reviewer #2 (Public review):***
>
> *Summary:*
>
> *The following points are those that occurred to me across readings of the paper. They are listed in what I take to be the order of their significance. Many of the points relate to the loose use of language and invocation of concepts that are not warranted, given the study design and results obtained.*
>
> *Major Comments:*
>
> *(1) The concept of ensemble turnover is interesting - the way it is introduced and discussed implies some type of spontaneous change in the neural underpinnings of fear discrimination and generalization in the PL. But, of course, every trial involves an opportunity to learn about the threat CS or the generalization test stimuli, and I am troubled by the thought that stability in the neural underpinnings of fear discrimination and generalization will actually reflect the level of defensive behaviours evoked on different trial types and/or the discrepancy between those behaviours and the outcome of a given trial in the generalization test. That is, stability in the neural underpinnings may be related to an animal's certainty or uncertainty in the contingency between a stimulus and danger; or, put another way, an animal's confidence that danger will or won't occur given the presence of some stimulus. This is not uninteresting. It is, however, not considered anywhere in the paper, which is overloaded with references to inferred threat values and integration of information across different types of stimuli. The protocol is not one that requires inference about anything or integration across anything.*

We thank the reviewer for this thoughtful comment. We agree that our original wording may have implied that turnover was a spontaneous process. Repeated retrieval provides opportunities for updating the learned contingencies associated with both the conditioned and generalization stimuli, and therefore changes in ensemble composition across sessions need not arise independently of experience. We also agree that the stability of graded neuronal representations may be related to the animal's certainty about the learned contingencies. However, in our data the graded neuronal population remained remarkably stable across retrieval sessions, whereas changes occurred primarily within the dynamic, frequency-selective neuronal populations. This suggests that stable ensembles preserve representations of learned threat value while updating is concentrated in a distinct neuronal subpopulation. We have now incorporated these ideas into the Discussion.

> *(2) I appreciate the link to Gu and Johansen in paragraph 3 of the Introduction, but the type of generalization under investigation here is not the same as the type of 'generalization' studied by Gu and Johansen [who used a sensory preconditioning protocol]. Nonetheless, the authors have forced the language used by Gu and Johansen*

> *into their paper, and this has created tension [at least for this reader] as the concepts introduced by Gu and Johansen [inference, integration] are simply not relevant given the generalization protocol used here. Here are a few examples of points where the tension might interfere with a reader's understanding:*

We thank the reviewer for these specific criticisms. We revised the manuscript throughout to remove or redefine terms like "inferred valence" and "integration," replacing them with clearer, more accurate descriptions of gradient generalization of threat value. Below we address each point raised by the reviewer regarding terminology clarifications.

> *(a) 'We hypothesized that generalization to novel stimuli depends on stable subnetwork organization that enables comparisons between learned and inferred valence, as well as population-level features that reduce variability across related representations.'*
>
> *I understand the words in the hypothesis, but can't form a representation of what is being said because of the reference to terms that stand in need of clarification [inferred valence, variability across related representations], but, ultimately, won't be clarified. This needs to be re-expressed so that the reader can appreciate what is being said.*

(a) We hypothesized that the PL generates representations of learned threat value that support threat generalization and discrimination, and that these representations emerge from the coordinated activity of stable and dynamic neuronal subnetworks, preserving consistent relationships among stimuli despite ongoing cellular turnover.

> *(b) 'Our results show that stable cortical subnetworks integrate the emotional "gist" of memory and inferred valence for novel cues over time, despite ongoing ensemble reorganization, and that population-level firing rate similarity across stimulus presentations determines threat generalization.'*
>
> *Again, what does this mean? How is the gist of a memory integrated with inferred valence for novel cues over time? The statement simply doesn't make sense. This needs to be rewritten for clarity.*

(b) The summary statement was rewritten: " Together, these findings provide a neural framework for understanding how the PL supports adaptive threat generalization and discrimination." pg. 4

> *(c) 'In CS^+ 15 mice, positively modulated sound-responsive neurons exhibited graded tone activity reflecting the contingency learned valence as well as the inferred valence of novel tones across testing days...'.*
>
> *Can this be rewritten as 'In CS^+15 mice, positively modulated sound-responsive neurons exhibited graded activity to the tone CS and its variants that were used to assess generalization.'? The overloading of the text with references to 'contingency learned valence' and 'inferred valence' is unnecessary and makes it much harder to understand what has been shown in the results.*

We adopted the reviewer's suggested rewording: " In CS^+ 15 mice, positively modulated sound-responsive neurons exhibited graded tone activity reflecting learned contingency value across testing days" pg. 9

We will systematically review the entire manuscript to ensure consistency with this revised framing.

> *(3) Re the same passage of text as in 2c:*
>
> *Is it the case that these neurons are simply tracking the expression of freezing to the various tones? The same question applies to the results obtained for the CS+3 mice. If*

> *this is the case, then why should the results be taken to support the banner statement that 'Sound-modulated PL population responses encode learned and inferred valence' - these analyses do not support that statement. And, as indicated, I don't believe that the language of learned and inferred valence is appropriate to such statements, given the nature of the protocol used and results obtained. It is a study looking at how populations of neurons in the PL respond during presentations of auditory stimuli that were subject to discriminative conditioning, and during tests of generalized freezing to other [intermediate] auditory stimuli.*

The reviewer is correct that the graded population responses observed in PL could reflect freezing behavior across tone frequencies rather than encoding an abstract threat-value representation. This important concern was also raised by other reviewers. To address it directly, we followed Reviewer 3's suggestion and implement a Generalized Linear Model (GLM) using the time series activity derived from the Ca2+ signals, with both tone identity and freezing behavior included as predictors. This analysis allowed us to dissociate the respective contributions of tone frequency and freezing to the graded neural responses. Based on the outcome of this analysis, we concluded that tone identity was a stronger predictor of neuronal activity than freezing. These results are summarized in Figures 4 for all cells contributing to population responses and Figure 8 for the main neuron types identified in the clustering analysis (frequency-selective and graded neurons). All details of this extensive new analysis are shown in red in the revised resubmission.

In addition, we revised the text to remove the terminology of "learned and inferred valence" throughout the manuscript.

> *(4) It is stated that:*
>
> *'In no-shock controls, although both positive and negative responses were present, population activity was not modulated by tone frequency or valence'.*
>
> *What does this mean? I can understand that population activity was not modulated by tone frequency. But what does it mean to say that it was not modulated by valence? Why should it have been when none of the tones were conditioned in this group and, hence, mice were responding to all the tones equally? And given that this is true, I don't understand the use of 'valence' here, or the subsequent statements in this paragraph that 'graded responses require associative learning' and that 'PL population responses encode graded sound-valence associations that reflect both learning and inference, closely matching behavioral generalization.' The latter statement is particularly unwarranted and, again, highlights a major issue with the paper. It could and should be rewritten as 'PL population responses reflect behavioral generalization.' There is nothing in the additional language that adds to the reader's understanding of what has been shown. The reference to 'graded sound-valence associations that reflect both learning and inference' is completely unwarranted, given the nature of this study. It is anathema to the vast literature on stimulus generalization. If the authors wished to make statements of this sort, they should have taken a different approach, perhaps using protocols like those featured in Gu and Johansen.*

We thank the reviewer for this helpful comment. We agree that our use of the term valence in describing the no-shock controls was imprecise. Because none of the tones was associated with reinforcement in this group, there was no learned valence that could modulate neuronal activity. Our intention was simply to convey that, although both positive and negative sound-responsive neurons were present, the population responses did not vary systematically across tone frequencies. We have revised this section accordingly.

We also agree that our original wording overstated the interpretation of the graded population responses. Our data do not demonstrate that associative learning is required for

sound responsiveness itself; rather, they show that associative learning is required for the emergence of graded population responses that distinguish tones according to their learned threat value. We have revised the text to make this distinction explicit.

Finally, we agree that our previous references to "learning and inference" were not justified by the behavioral paradigm. We have removed this language throughout the manuscript and now describe the findings more directly as graded representations of learned threat value that closely parallel the observed behavioral generalization gradients.

> *(5) The section titled, 'Consistently active neurons preserve valence representations as newly recruited neurons sharpen remote memory traces' ends with the following summary:*
>
> *'Together, these results indicate that consistently active neurons maintain stable representations of learned and inferred sound associations across time, whereas neurons recruited after conditioning progressively acquire graded tuning at later retrieval stages. This dynamic refinement suggests that cortical memory representations become increasingly selective during systems consolidation, while a stable neuronal subpopulation preserves the core emotional content of the memory.'*
>
> *Once again, the summary is not in keeping with the results obtained. The 'dynamic refinement' of representations is far more likely to reflect the repeated testing across days 1, 15, and 30 rather than anything to do with systems consolidation - at the very least, it is the simplest interpretation of the results. The impact of repeated testing is evident in the sharpening of generalization gradients over time, which is contrary to what is otherwise observed in the literature - the incredibly well -documented broadening of generalization gradients with time. Given this impact of repeated testing, surely the changes in the neuronal population that underlie performance are more likely to reflect the learning that occurs on days 1, 15, and 30, which is reflected in reduced freezing to the non-conditioned tones. If this is a reasonable take on the results, then I don't see the basis for invoking systems consolidation at all, and I don't see the basis for inferring a stable neuronal subpopulation that preserves the emotional content of the memory. Rather, non-reinforced presentations of 'never-reinforced' tones result in recruitment of additional neurons that result in suppression of freezing responses to those stimuli.*

We thank the reviewer for this thoughtful comment. We agree that repeated retrieval is an inherent limitation of longitudinal memory studies and that repeated non-reinforced presentations of the tones provide opportunities for memory updating. Accordingly, we have revised the Discussion to explicitly acknowledge that repeated retrieval may contribute to the ensemble reorganization observed across sessions through memory updating or reconsolidation processes (Discussion, pg. 19).

We also agree that the progressive sharpening of the behavioral generalization gradients across retrieval sessions is consistent with memory updating. Both the behavioral data (increased discrimination ratios) and the neuronal data (progressively sharper population generalization gradients among neurons active after conditioning) indicate that the memory representation became more precise over time. We now discuss this possibility explicitly in the revised Discussion. We also agree that fear generalization often broadens with time; however, this is not universal. Under discriminative conditioning paradigms, repeated retrieval can instead produce progressively narrower generalization gradients [11]. We have revised the Discussion to clarify this distinction and added the appropriate references (pg. 19).

While the reviewer's interpretation is therefore plausible, we do not believe it fully accounts for our observations. If repeated non-reinforced presentations were the sole driver of the observed neuronal changes, one might expect a more uniform reorganization across the

neuronal populations engaged by the task. Instead, the reorganization was highly selective. Neurons encoding graded threat value remained remarkably stable across retrieval sessions, whereas neuronal turnover occurred primarily within the frequency-selective subpopulations. Thus, although repeated retrieval may update the memory representation, the neuronal substrate supporting graded threat-value coding is largely preserved while refinement occurs within a distinct neuronal subpopulation.

Moreover, we observed substantial neuronal turnover beginning with the first retrieval session, consistent with previous longitudinal studies showing that cortical memory ensembles remain dynamic despite stable memory [1-4]. This early emergence of turnover suggests that repeated testing alone is unlikely to account for the continuous population dynamics observed throughout the experiment.

Rather than viewing these findings as evidence exclusively for either memory updating or systems consolidation, we believe they are more consistent with current models proposing that these processes occur in parallel. Several influential frameworks argue that memories are continuously modified through retrieval while simultaneously undergoing systems-level reorganization [4, 6, 16, 17].We have therefore revised the Discussion to interpret the longitudinal changes more conservatively as reflecting the combined influence of retrieval-dependent memory updating and systems-level reorganization.

In summary, we have revised the manuscript to better acknowledge the contribution of repeated retrieval while emphasizing what we believe is the principal finding of our study: despite substantial turnover within the overall ensemble, the neuronal population encoding graded threat value remained remarkably stable, whereas refinement occurred primarily within dynamic frequency-selective neuronal populations.

> *(6) In the section titled, 'Population vector similarity at stimulus onset determines degree of generalization', it is stated that:*
>
> *'Because population similarity peaked shortly after stimulus onset, we quantified similarity during the first 5 s after tone onset relative to the CS^+. In CS^+15 mice, population similarity was highest for 15/15 and 15/11 tone pairs with no differences between them.'*
>
> *Isn't this consistent with the view that the population response in the PL simply reflects the level of freezing? Freezing to the 15-15 and 15-11 tones is most likely to be similar on their first presentation prior to the effects of extinction on the 11 Hz tone; hence the results obtained. That is, these results appear to clearly indicate that neuronal responses in the PL reflect the degree of stimulus generalization, as evidenced in freezing behavior. Given all that we know about the involvement of the PL in expressing fear responses, it is not appropriate to claim that 'population vector similarity at stimulus onset *determines* the degree of generalization. The PL responses simply reflect the varying levels of performance displayed to the different types of tones. What have I missed that could be taken to support additional statements?*

We agree that, because population similarity is highest for the 3/3, 15/15, and 15/11 tone pairs and freezing is also greatest for these same stimuli, the neural data could, in principle, reflect a correlate of behavioral expression rather than an independent representation of learned threat value.

To directly address this possibility, we implemented a generalized linear model (GLM) to dissociate the contributions of tone identity and freezing behavior to neuronal activity. Across all analyses, tone identity consistently explained substantially more variance in neuronal activity than freezing behavior. Importantly, this finding held not only for the full population of sound-responsive neurons used to generate the population similarity analyses

(Figure 4), but also for both the stable graded neurons and the dynamic tone-selective neuronal populations identified by our clustering analysis (Figure 8). Thus, although freezing behavior contributes modestly to PL activity, it cannot account for the enhanced similarity of population vectors across stimulus presentations or the graded population responses that form the basis of our conclusions.

In addition, the temporal dynamics of the population vector similarity analysis are not entirely consistent with the interpretation that PL activity simply reflects the expression of freezing behavior. Population vector similarity peaked during the first 5 seconds following tone onset, whereas freezing occurred intermittently throughout the tone presentations. Although this temporal relationship does not establish causality, it is consistent with the interpretation that PL activity reflects the learned threat value associated with each tone rather than merely tracking the magnitude of freezing.

Finally, we have revised the manuscript to more clearly acknowledge the correlational nature of these analyses. Specifically, we now state that population vector similarity is associated with, rather than determines, the degree of threat generalization.

> *Later in the same section, it is stated that 'population-level similarity at stimulus onset scales with behavioral threat generalization and is maximal for tones associated with robust threat responses.' For simplicity and, therefore, clarity, this should be rewritten as 'population-level similarity at stimulus onset reflects behavioral threat generalization.'*

We made this correction. (“These findings indicate that population-level similarity at stimulus onset scales with behavioral threat generalization”. pg. 13)

> *(7) In the section titled, 'Different subnetworks encode acoustic versus learned properties of sound association', it is stated that:*
>
> *'Our previous analyses show that learned and inferred associations are represented at the population level. However, these results do not resolve whether graded responses arise from pooled activity of frequency-selective neurons or from subnetworks encoding integrated learned valence across tones.'*
>
> *What does it mean to say 'integrated learned valence across tones'? As it presently stands, the meaning of the phrase is unclear. It only makes sense if one supposes that generalized freezing responses to the 11 and 7 kHZ tones reflect separate associations between those tones and the aversive foot shock US. This supposition is inconsistent with the rich literature on generalization of Pavlovian conditioned fear responses. Specifically, it is inconsistent with the many theories of fear generalization, which attribute the reduction in fear as one moves away from the specific conditioned stimulus to a decrement in the ability of the test stimulus to activate the trained CS-US association. My strong impression is that the authors would do well to ground their findings in theories of stimulus/fear generalization, of which there are many. This would better serve the results obtained [and the reader's appreciation of them] - at present, the unnecessary invocation of concepts does very little to enhance the reader's appreciation or understanding of what has been found in the study.*

We agree that the phrase "integrated learned valence" is unnecessarily opaque and we replaced it with more precise language “Our previous analyses demonstrated that threat-value generalization gradients are represented at the population level. However, these findings do not reveal how these representations arise. Specifically, the observed population gradients could emerge either from the pooled activity of frequency-selective neurons that respond to individual tones or from neuronal subnetworks that integrate information across tones to encode their learned threat-value.” (Pg. 13)

> *(8) Another example of what has been a common theme in this review:*

> *'...we hypothesized that the PL active ensemble segregates into functionally distinct subnetworks: one encoding tone-specific sensory features with dynamic characteristics, and another responding to all frequencies encoding stable core memory content and inferred emotional valence.'*
>
> *What does it mean to say 'all frequencies encoding stable core memory content and inferred emotional valence'? Do the authors mean to say '...and another that tracks freezing/defensive responses regardless of whether they were elicited by the trained CS or one of the generalization test stimuli'?*

We thank the reviewer for pointing out that this section was unclear. We agree that our original wording was imprecise and could be interpreted as implying cognitive processes that were not directly tested in the present study. Accordingly, we have revised the terminology throughout the manuscript. We no longer refer to "inferred emotional valence" or "core memory content" and instead describe these neurons more specifically as exhibiting graded representations of learned threat value.

This is not the interpretation we intended. To determine whether these neurons primarily reflected defensive behavior rather than learned stimulus value, we implemented a Generalized Linear Model (GLM) that dissociates the contributions of tone identity and freezing behavior to neuronal activity. Across the entire neuronal population, as well as within the stable graded and dynamic tone-selective neuronal subpopulations, tone identity consistently explained substantially more variance than freezing behavior (Figures 4 and 8). Furthermore, after accounting for freezing, the regression coefficients of the graded neurons continued to follow the learned threat value of the tones, exhibiting opposite monotonic gradients in the CS+15 and CS+3 groups. If these neurons simply tracked defensive behavior irrespective of the stimulus presented, this relationship would not be expected to persist after accounting for freezing. We therefore conclude that the activity of this stable neuronal subpopulation is better explained by graded representations of learned threat value than by defensive behavior alone, and we have revised the manuscript accordingly.

> *(9) It is stated that - 'Graded clusters encode emotional valence but constitute only a fraction of the active population; yet valence coding at the population level remains accurate and precise. This indicates that neurons newly recruited into the population-likely frequency-selective and organized within learning-independent clusters-can be shaped by associative processes through modulation of firing activity.'*
>
> *What does this mean? Are the authors trying to say that - 'Some clusters of PL neurons track freezing responses. In spite of the fact that these are only a fraction of the total active neuronal population, the population-level response of PL neurons also tracks the levels of fear to the trained tone and its variants used in the test for generalization.' If this is what one wants to say, then the final statement in the reproduced section does not follow. That is, there is no indication that 'neurons newly recruited into the population-likely frequency-selective and organized within learning-independent clusters-can be shaped by associative processes through modulation of firing activity.' As noted, the characteristics of other ensembles that become active across the repeated tests on days 1, 15, and 30 are more likely to reflect learning from non-reinforcement that occurs within and across those sessions. Perhaps this is what is meant by the phrase, 'shaped by associative processes'? If so, it should be stated explicitly instead of left to the reader to work out.*

We thank the reviewer for highlighting that this section was unclear. We agree that the original phrasing was insufficiently precise. Our intention was to convey that only a subset of PL neurons displays graded tuning that tracks behavioral generalization across tones. Nevertheless, despite constituting only a fraction of the total active population, this graded coding is also reflected at the population level. This observation led us to hypothesize that

neurons recruited into the active population after conditioning— likely dynamic, frequency-selective neurons—also contribute to these graded population responses through modulation of their firing rates.

The reviewer correctly notes that the phrase "shaped by associative processes" was too vague. By this we meant that the firing properties of these neurons are modified by the animal's associative history, including both the original conditioning experience and any retrieval-dependent updating that may occur during subsequent test sessions. We have revised the manuscript to make this interpretation explicit rather than leaving it to the reader to infer.

To test this hypothesis, the GLM analysis we implemented dissociated the contributions of tone identity and freezing behavior to neuronal activity. After accounting for freezing, tone identity (i.e., learned threat value) remained a significant predictor of neuronal responses. Importantly, this was also true for the dynamic, frequency-selective neurons (Fig. 8e–f), indicating that these neurons contribute to population-level representations of learned threat value through firing-rate modulation rather than simply reflecting defensive behavior.

To clarify our interpretation, we have rewritten the relevant section as follows:

"Graded clusters encode generalization gradients but constitute only a subset of the active neuronal population. Nevertheless, population-level representations, which incorporate all active neurons, remain robust and accurately preserve these gradients. This observation led us to hypothesize that neurons recruited over time (e.g., dynamic, frequency-selective cells) also contribute to threat-value representations. Consistent with findings in the hippocampus showing that neurons can encode task contingencies through firing-rate modulation despite responding selectively to a single location (Gagliardi et al., 2024; Huxter et al., 2003; Sanders et al., 2019), we tested whether dynamic, frequency-selective clusters exhibited firing-rate differences proportional to learned threat value." (page 15)

Regarding the reviewer's suggestion that the characteristics of the newly recruited neurons may reflect learning during repeated non-reinforced test sessions, we agree that retrieval-dependent memory updating likely contributes to the reorganization of the dynamic neuronal population, and we now explicitly acknowledge this possibility in the Discussion. However, we do not believe that our findings are fully explained by repeated non-reinforced retrieval alone. First, no-shock control animals underwent the same repeated testing but failed to develop graded neuronal representations, indicating that repeated exposure in the absence of associative learning is insufficient to account for the observed changes. Second, both behavioral discrimination and the corresponding population-level neural gradients became progressively sharper over time, consistent with refinement of learned threat representations rather than an effect of repeated testing alone, which must lead to extinction.

In summary, we thank the reviewer for highlighting both the ambiguity of our original wording and an important alternative interpretation. In response, we have clarified the text to explicitly define what we mean by associative processes, added a GLM analysis demonstrating that the newly recruited neurons encode learned threat value beyond freezing behavior, and revised the Discussion to acknowledge that retrieval-dependent memory updating likely contributes to the reorganization of the dynamic neuronal population.

> *(10) The following points all relate to the Discussion and reiterate many of the points above.*
>
> *(a) 'A subset of neurons remains consistently active across sessions, preserving core components of the memory trace and supporting inference of emotional valence for novel sounds, while neurons recruited after conditioning progressively acquire valence selectivity at remote time points.'*

> *'Inference of emotional valence' is unclear and unwarranted for all of the reasons provided above regarding the use of language.*

We modified the language as stated in the prior points.

> *(b) '...Our data reconcile these views by demonstrating that cortical representations of emotional valence emerge rapidly after learning and persist within stable subnetworks, even as the broader population undergoes substantial turnover. This architecture preserves core mnemonic content while allowing flexibility in the surrounding ensemble.'*
>
> *These statements assume that the PL neuronal responses reflect something more than the levels of freezing behavior to the different stimuli; what are the grounds for this assumption?*

We incorporated the new GLM analysis to address this point and conclusions.

> *(c) 'Importantly, these subnetworks encode both learned contingencies and the inferred valence of novel stimuli along a graded representational axis, suggesting that strong recurrent connectivity provides a stable scaffold for emotional memory representations.'*
>
> *What is a graded representational axis, and what part of the first statement suggests that 'strong recurrent connectivity provides a stable scaffold for emotional memory representations'? If the authors' goal was to make statements about emotional memory representations vis-à-vis emotional memory content, they should have used protocols that allowed them to probe such content. The auditory fear conditioning protocol used here [followed by tests for generalization to other auditory stimuli that differ in frequency from the conditioned tone] is not one that lends itself to analysis of emotional memory representations or content.*

We agree that the term "graded representational axis" was insufficiently defined and could be interpreted in multiple ways. Because this terminology was not essential to our conclusions, we have removed it and instead describe the observed phenomenon as a graded population representation of learned threat value across tone frequencies. We also removed the statement suggesting that recurrent connectivity provides a stable scaffold for these representations, as this mechanistic interpretation is not directly supported by our data.

We also agree that some sections of the manuscript overstated the scope of our conclusions and have revised the wording accordingly. Our study uses neuronal activity recorded during memory retrieval after learning, an approach widely used in studies of systems consolidation to infer how learned information is represented within neural populations. Accordingly, we have revised the manuscript to explicitly state that our findings pertain to neural representations of learned threat value during memory retrieval rather than the broader content of emotional memories.

Finally, we agree that our data are correlational and do not establish the causal role of the neuronal representations we identify. Throughout the manuscript, we now refer more precisely to population- and single-neuron correlates of learned threat value during memory retrieval following auditory fear conditioning.

> *(d) 'Dynamic tone-selective responsive neurons emerge independently of learning, as they are present in both control and experimental mice, reflecting pre-existing PL sensory-driven properties (Hockley & Malmierca, 2024; Zikopoulos & Barbas, 2006).'*
>
> *Maybe. They are also likely to have developed as a consequence of the repeated testing on days 1, 15, and 30, which involved intermixed exposures to the tones of different frequencies. That is, rather than 'pre-existing PL sensory-driven properties', the responses of these neurons might reflect the emergence of discrimination between the various*

> *tones across testing, and greater suppression of freezing to the non-trained tones compared to the trained tone across the various test intervals.*

We thank the reviewer for this thoughtful comment. Our interpretation that these neurons reflect preexisting sensory-driven properties of PL cortex is based on two observations. First, tone-selective neuronal clusters were present in both conditioned and no-shock control animals, consistent with previous reports of sensory responsiveness in PL cortex [18, 19]. Second, these responses were already present during the first retrieval session, when the intermediate frequencies were presented for the first time. Thus, they cannot be explained by repeated exposure to those tones across subsequent test sessions.

We therefore interpret the frequency-selective response properties as pre-existing features of PL circuitry that are present independently of conditioning. In contrast, associative learning modifies the firing activity of these neurons, allowing them to contribute to graded representations of learned threat value. This interpretation is supported by our GLM analysis, which showed that, after accounting for freezing, tone identity significantly predicted the activity of frequency-selective neurons in conditioned animals but not in no-shock controls. Thus, while the frequency-selective response properties are present independently of learning, associative learning modifies how these neurons encode learned threat value. We have revised the manuscript to clarify this distinction.

> ***Reviewer #3 (Public review):***
>
> *Summary:*
>
> *Normandin et al. explore the coding of stimuli predicting an aversive event in the prelimbic cortex. Stimuli could either be explicitly paired, explicitly unpaired, or novel but with an inferred association with the aversive event (generalization). Long-term tracking of GCaMP-positive neurons allowed them to examine how coding evolves out to a month following training. In general, they found two types of ensemble codes. One was ensembles coding for each stimulus independently, but with enhanced responding to the one eliciting a freezing response. The other was ensembles that responded to all stimuli in proportion to their similarity to the stimulus paired with the aversive event, either increasing or decreasing their activation with the degree of freezing elicited by a stimulus. Importantly, this second set of ensembles was more stable across days, potentially providing a memory trace.*
>
> *Strengths:*
>
> *(1) The authors track ensembles in prelimbic cortex over long time scales, providing valuable information on the consolidation of neural codes.*
>
> *(2) Neural coding of generalization is examined, which is under-examined in the field.*

We thank the reviewer for appreciating our design to track ensembles over time and the relevance of studying the neural substrates of generalization.

> *Weaknesses:*
>
> *(1) Difficult to determine if responses treated as encoding stimulus valence are driven instead by the behavior that the stimulus elicits, freezing.*

We thank the reviewer for this thoughtful and constructive comment. We agree that an alternative interpretation is that the graded neuronal responses may partially reflect freezing-related activity rather than representations of learned threat value. In the revised manuscript, we acknowledge that previous studies have identified PL neurons whose activity tracks freezing independently of stimulus identity or associative content. To directly address this possibility, we implemented the reviewer's suggestion by fitting a generalized linear

model (GLM) to the neuronal activity time series derived from the Ca^{2+} signals, using tone identity and freezing behavior as predictors. Because tone identity is fixed across trials, whereas freezing varies both during tone presentation and across trials (see below our answer to the Recommendations to Authors), this approach allowed us to dissociate their respective contributions to neuronal activity. We are grateful for this excellent suggestion, which has substantially strengthened both the manuscript and the conclusions that can be drawn from our data. The new analyses are summarized in Figures 4 and 8.

In the points below we summarize the new findings.

> *(2) The study implies that the identified ensembles are causally related to valence memory, but no experimental interventions are performed to justify this.*

We appreciate the reviewer's point. We agree that our data are correlational in nature and that establishing a causal relationship between identified ensembles and valence memory would require experimental interventions such as combinations of optogenetic and two-photon manipulations, which are beyond the scope of the present study but represent an important direction for future work.

We examined inter-individual variability in freezing relative to the proportion of graded cells but the number of mice used in this study (CS+3= 5 and CS+15=7) did not give us enough power to reach significance.

Therefore, we modified the manuscript terminology accordingly, replacing causal language with phrasing that accurately reflects the correlational nature of our conclusions.

> ***Recommendations for the authors:***
>
> ***Reviewer #2 (Recommendations for the authors):***
>
> *Many sections of the paper should be rewritten along the lines that I have suggested in my public review and below.*
>
> *Minor Comments:*
>
> *(1) INTRO - 'This broad accessibility reduces spatial specificity and increases learning variability...'.*
>
> *Broad accessibility of what, exactly? And how does the 'broad accessibility' reduce spatial specificity and increase learning variability? That is, I do not understand what the terms 'reduced spatial specificity' and 'increased learning variability' refer to at this point in the first paragraph...*

We rewrote the introduction and discussion to address the points raised by the reviewer.

> *(1) INTRO - 'The prelimbic cortex (PL) contributes to the expression (Burgos-Robles et al., 2009; SierraMercado et al., 2011; Sotres-Bayon & Quirk, 2010) and the proper discrimination and generalization of threat memories (Rosas-Vidal et al., 2025; Stujenske et al., 2022).'*
>
> *What is achieved by calling it 'proper' discrimination and generalization? Can't one simply say that the PL contributes to the expression, discrimination, and generalization of threat memories?*

This was corrected.

> *(3) INTRO - '... and that population-level firing rate similarity across stimulus presentations determines threat generalization'.*

> *Or, alternatively, that generalization of conditioned freezing responses from the tone CS to variants along the dimension of Hz values is reflected in systematic changes in the firing rate of PL neuronal ensembles; when the test stimulus is similar to the conditioned stimulus, the two elicit similar behavioural responses and evoke a similar population-level firing rate in the respective PL neuronal ensembles.*

We have revised the Introduction as stated above. However, as discussed in our detailed responses, freezing behavior cannot fully account for the observed patterns of PL activity.

> *(4) METHODS - 'Memory retrieval was tested on days 1, 15, and 30 after conditioning to probe early, long-term, and remote memory (Bontempi et al., 1996). During retrieval, mice were tested in a novel context with the CS^+, $CS1^-$, and two intermediate frequencies (7 and 11 kHz), presented in semirandom order, with each tone repeated three times (Fig. 1a).'*
>
> *Why was testing conducted in a different context than that of conditioning? This is likely to result in an underestimation of generalization to the different tones...*

In tone fear conditioning, it is always customary to test in a different context to dissociate conditioning to the context vs conditioning to the tones, which usually take place simultaneously in the same context [20]. Therefore, testing generalization in a novel context gives the correct estimate of generalization to the tones in the absence of contextual conditioning confounds. Please note that while overall freezing levels may be lower in a novel context due to the absence of contextual conditioning, the relative generalization gradient across tones — which is what your study measures — is unlikely to be systematically distorted by context change.

> *(5) RESULTS - 'No-shock control mice showed no significant differences in freezing across frequencies on any testing day ($p > 0.05$; Fig. 1b, right), confirming that freezing reflected associative learning.'*
>
> *The inference doesn't follow from the result described. Was there more freezing among animals in the shocked groups compared to those in the no-shock group? I presume so - my point is that this comparison is the one that most directly speaks to the presence or absence of associative learning.*

Experimental animals exhibited not only higher overall freezing but also graded freezing responses across tone frequencies. It is important to note that no-shock controls did not display this pattern, ruling out the possibility that the different frequencies themselves elicited graded behavioral responses. To clarify this point, we revised the sentence as follows: "No-shock control mice showed no significant differences in freezing across frequencies on any testing day ($p > 0.05$; Fig. 1b, right), confirming that the graded freezing patterns resulted from associative learning rather than the acoustic properties of the tones." (Pg. 6)

> *(6) 'Across animals and sessions, we identified distinct neuronal populations showing positive modulation, negative modulation, mixed responses, or no consistent response to sound (Fig. 2b)...'*
>
> *To be clear, do you mean to say that there were distinct neuronal populations that consistently [i.e., across all three sessions] increased their responses to the tones [positive modulation], decreased their responses to the tones [negative modulation], showed variable responses to the tones [mixed responses], and did not respond to tones [not modulated]?*

The sentence refers to neuronal populations identified within each recording session based on their responses to the tones, not to neurons that maintained the same response profile across all three sessions. We have revised the text to make this distinction explicit. The only

stable patterns across sessions were observed in graded neurons that were stable across retrieval.

“Across animals, we identified distinct neuronal subpopulations showing positive modulation, negative modulation, mixed responses, or no consistent response to sound in each session (Fig. 2b)” Pg. 7

> *(7) What does 'active' mean in relation to Figure 2? Does this refer to neurons that displayed either positive responses, negative responses, and/or mixed responses? In the text, it is stated that 'Sound responder neurons were classified using a test that detected modulation based on magnitude relative to baseline variability, allowing reliable identification of both transient and sustained responses while remaining robust to noise...'*
>
> *I can't work out if this is the same classification criteria used for the determination of positive modulation, negative modulation, and mixed responding.*

We thank the reviewer for pointing out this ambiguity. In Figure 2, the term "active" referred to neurons that exhibited significant sound-evoked modulation and were subsequently classified as showing positive, negative, or mixed responses. Thus, active and sound-responsive refer to the same population of neurons. We removed the word active to avoid confusion.

The reference to transient and sustained responses describes the temporal profile of the calcium signals rather than separate response categories. Some neurons exhibited brief calcium transients that rose and decayed rapidly, whereas others displayed sustained activity throughout the tone presentation. The sound-response detection algorithm was designed to reliably identify both temporal response profiles. We have revised the manuscript to make these definitions explicit. We modified the sentence as follows: “Sound-responsive neurons were identified using a statistical test that detected activity modulation relative to baseline variability, allowing reliable identification of responses while remaining robust to noise. This approach was effective for neurons exhibiting either brief calcium transients that rose and decayed rapidly or sustained activity throughout the tone presentation.” Pg. 7-8

> *(8) 'A moderate proportion of neurons was present across all retrieval sessions, with no differences between groups ($p > 0.05$).'*
>
> *Do you mean to say that 'A moderate proportion of neurons was ACTIVE across all retrieval sessions, with no differences between groups ($p > 0.05$)'?*

We replaced the word present and replaced it with “active”. Pg. 8

> *(9) In the section titled, 'Different subnetworks encode acoustic versus learned properties of sound association', it is stated that:*
>
> *'If neurons encoding graded responses carry core mnemonic information, they should exhibit enhanced stability over time. To test this hypothesis, we quantified the proportion of registered neurons that retained their cluster identity across at least two retrieval sessions and compared these values to a shuffled null distribution (10,000 iterations), with multiple comparisons controlled using the BenjaminiHochberg procedure.'*
>
> *What does the comparison to the shuffled null distribution tell us exactly? I accept that some neurons were stable positive responders across at least two sessions. The comparison to the shuffled null distribution creates a false impression about the robustness of this stability or the 'enhanced stability over time'.*

Our intention in comparing the observed stability to a shuffled null distribution was to evaluate whether the proportion of neurons retaining cluster identity exceeded chance levels

expected from random assignment. The shuffled distribution therefore provides a statistical baseline against which the observed degree of stability can be evaluated. We agree, however, that the wording “enhanced stability over time” may be confusing regarding this finding. We rephrased this paragraph to clarify that a subset of neurons retained cluster identity across all retrieval sessions at levels greater than expected by chance as follows:

“These data demonstrate that graded clusters remain consistently active at levels exceeding chance, preserving their cellular identity and providing a stable representation of learned contingencies and generalization gradients.” Pg. 15.

> *(10) ABSTRACT. The abstract states that, 'Stimulus-evoked population similarity scaled precisely with behavioral generalization, and consistent population states emerged only for tones associated with shock or those eliciting strong generalized freezing, indicating that population-level similarity predicts inferred threat.'*
>
> *I believe that the sentence could be rewritten as, 'Stimulus-evoked population similarity reflected the degree of generalization, and consistent population states emerged only for tones associated with shock or those eliciting strong generalized freezing.'*

We revised the text according the reviewer’s suggestion; however, we had to shorten the sentence due to word limits. “Population similarity tracked behavioral generalization, whereas consistent population states emerged only for shock-associated or highly generalized tones.” Pg. 2

> ***Reviewer #3 (Recommendations for the authors):***
>
> *Major points:*
>
> *(1) The ensembles with graded activation in proportion to stimulus valence are described at various points in the manuscript as "maintaining the emotional 'gist'", "preserving core components of the memory trace", and "preserving core components of the memory trace". This conclusion is premature because there is an alternative interpretation. The graded response ensembles would also be consistent with coding for the freezing behavior itself, irrespective of the specific memory or stimulus association that drives it. An ensemble that encodes a behavior in this way would not be considered mnemonic, just as motor neurons in the spinal cord are not, even if they may fire during a conditioned response. Indeed, previous work has identified neurons in the prelimbic cortex that encode freezing independently from the stimuli that signal an aversive outcome (e.g., Kyriazi, Headley, and Pare 2020; Casanova, Pouget, ..., Vetere 2024).*
>
> *There are two ways the authors can address this point.*
>
> *(a) Fit a generalized linear model to the time series of inferred spiking activity from the Ca2+ signal and include stimuli and freezing as predictors. Since freezing behavior is inconsistent across trials, while stimulus presence is fixed, they can be disassociated. If, after accounting for freezing, responsiveness neurons still show a graded coding of stimuli that agrees with inferred aversiveness, this would strengthen their claim that they have identified an ensemble that corresponds with mnemonic or salience aspects of the stimuli.*
>
> *(b) Conduct no further analysis but cover the issue in the discussion as a limitation to their study and to dampen some of the language throughout the manuscript that implies that a memory trace has been identified.*

We thank the reviewer for this thoughtful and constructive comment. We agree that an important alternative interpretation is that graded-response ensembles could reflect freezing-related activity rather than representations of learned threat value. To directly

address this possibility, we implemented a Generalized Linear Model (GLM) analysis, as suggested by the reviewer. The GLM was fitted to the activity of every sound-responsive neuron included in the population analyses and simultaneously incorporated tone identity and continuous freezing probability (derived probabilistically from miniscope acceleration) as predictors, allowing us to quantify their independent contributions to neuronal activity.

We want to note that freezing was quantified from the miniscope's inertial measurement unit (IMU) using a two-component Gaussian mixture model applied to the log-transformed body-acceleration signal. Rather than classifying freezing with a binary threshold, we used the posterior probability of the low-movement state as a continuous freezing regressor. This approach captures graded variations in immobility and provides a more conservative test of tone encoding, because it accounts for more behaviour-related variance than a binary classifier, making it more difficult to detect an independent contribution of tone identity.

We applied the GLM both to all sound-responsive neurons contributing to the population response curves and separately to the identified frequency-selective and graded neuronal subpopulations. Across all analyses, tone identity consistently explained neuronal activity better than freezing. Furthermore, the freezing-corrected tone β coefficients scaled with learned threat value, with graded neurons exhibiting the strongest monotonic gradients, indicating that they provide the most robust representation of learned threat value. These findings demonstrate that the graded coding of learned threat value persists after accounting for freezing behavior and therefore cannot be explained simply by the behavioral expression of fear. The new analyses are presented in Figures 4 and 8. Notably, although freezing-dominant neurons were present in both the tone-selective and graded populations, they represented only a small fraction of each group and were least prevalent among graded neurons (6–8%), further supporting the conclusion that graded neurons primarily encode learned threat value.

In addition, we revised the manuscript to avoid language implying that these neuronal populations constitute a mnemonic trace. Instead, we consistently describe them as encoding learned threat value, a more accurate interpretation that is directly supported by the new GLM analyses.

> *(2) The title makes a seemingly causal claim by using the term 'arise', "Learned and inferred valence arise from interactions between stable and dynamic subnetworks". While it is true that the authors show that both stable and dynamic ensembles encode valence, they do not demonstrate that the behavioral expression of valence depends on these codes, nor their interaction. Experimentally testing this is beyond the scope of this study (holographic two-photon stimulation of transient and stable ensembles?), but they may be able to get closer to it by examining inter-individual variability. The authors could measure the proportion of neurons in each subject that participate in the stable (graded responding) and dynamic (stimulus-specific) ensembles, and see if they predict individual differences in the expression of freezing behavior or its generalization. Indeed, this correlation may change across testing days.*

We agree that the term “arise” in the title may imply a stronger causal relationship than is directly supported by the present data. We modified the title in the resubmission as follows: “Complementary stable and dynamic prelimbic ensembles encode learned threat value underlying generalization and discrimination”

The new LGM analysis confirms that a large proportion of neural activity can be predicted by tone threat value; therefore, we think this title fully captures our findings.

We also appreciate the reviewer's suggestion to examine inter-individual variability. In the revised manuscript, we tested whether the proportion of graded neurons correlated with freezing behavior. However, the limited number of experimental animals in each

experimental group provided insufficient statistical power to reliably assess this relationship. Accordingly, we revised the manuscript to clarify that our conclusions are based on correlational observations rather than causal inferences.

In summary, we revised the title and related language throughout the manuscript to avoid implying causal mechanisms beyond the scope of the current experiments.

> *Minor points:*
>
> *(1) I was surprised by the absence of an ensemble in the No-shock group that responded uniformly to all stimuli. Can the authors confirm this?*

Yes, we confirm this finding. It was unexpected to us as well. We would like to clarify, however, that some control neurons may have responded to more than one frequency, but these responses were too infrequent or too weak to be classified as a distinct graded neuronal population by our clustering algorithm. Thus, while broadly responsive neurons may have been present in the control group, they did not form a robust, identifiable ensemble comparable to that observed after fear conditioning.

> *(2) Several different approaches were used to analyze the same Ca2+ responses to stimuli across testing days. These were the "Sound responder classification", "Average stimulus-aligned trace procedure", "Population similarity over time across tone pairs", and the construction of "Stimulus response vectors". These feature differing alignment/binning/interpolation, normalization, and response quantification procedures, and it is unclear why they cannot all be in agreement, at least when it comes to alignment and normalization.*

We thank the reviewer for this careful reading of our Methods. All analyses were performed on the same underlying calcium imaging dataset, but they were designed to address different aspects of the data and therefore required different preprocessing steps. The analyses share a common initial pipeline leading to the calcium traces (all z-scored across the session). Differences in subsequent processing (e.g., use of ΔF/F versus CASCADE-deconvolved activity, normalization, baseline correction, temporal binning, and interpolation) were introduced only when required by the specific analysis method.

To make this clearer, we have substantially revised the Methods. We added a new overview of preprocessing section that summarizes the common preprocessing pipeline and explicitly distinguishes the shared steps from those that are analysis-specific. We also included a summary table describing the input signal (ΔF/F or CASCADE-deconvolved activity), normalization procedure, and temporal processing used for each analysis. Finally, the individual Methods sections were revised to eliminate redundancies and more clearly describe the steps to avoid confusion. We hope these revisions make the rationale for the different preprocessing procedures and the overall analytical workflow more transparent (Pg. 22-23)

> *(3) In the methods section "Window-wise response quantification" the Ca2+ signal was baselinesubtracted and divided by the standard deviation in the baseline across trials, but that data was already presumably z-normalized to the baseline of each trial ("Data alignment and normalization"). This second step of normalization seems excessive. Why is it not sufficient to just take the average peri-stimulus response across the z-normalized trials from the "Data alignment and normalization" section? This is simpler and would capture the effect size of the response relative to baseline.*

We thank the reviewer for this careful observation.

The two operations are also not the same normalization applied twice; they standardize different sources of variability. During the alignment step, each trial is z-scored relative to its

own baseline by dividing by the standard deviation of that trial's baseline across time. This places all trials on a common within-trial scale before averaging. In the window-wise step, the trial-averaged, baseline-subtracted response is expressed relative to the standard deviation of the per-trial baseline levels across trials—a distinct quantity that reflects trial-to-trial baseline stability rather than within-trial fluctuations. The purpose of this second term was to down-weight windows in cells with unstable baselines across trials, and it entered the analysis only as a significance criterion; the magnitude threshold defining a sound responder was applied to the trial-averaged baseline-relative response itself. We have revised the Methods to clarify the distinct roles of these two normalization steps.

To further address this concern, we re-ran the sound-responder classification after removing the second (between-trial) normalization step, so that responder detection depended only on the per-trial baselinerelative response magnitude and its temporal persistence. Across all cells, tones, and sessions (n = 89,504 cell–tone–session classifications), the two procedures agreed on 95.3% of labels. The small fraction of cells whose labels changed were almost exclusively those lying immediately at the detection threshold: 86.6% of changes involved cells moving into or out of the "modulated" category, whereas direct reversals between excitatory and inhibitory classification occurred in only 8 of 89,504 cases (0.009%). Consistent with the between-trial standard deviation being a less stable quantity when few trials are available, label changes were approximately twice as frequent in the three-trial retrieval sessions (5.2%) as in the ten-trial conditioning sessions (2.1%). Overall responder proportions changed only minimally (positive responders +2.2%, negative responders +4.3%), and all population-level findings—including the graded threat-value gradient across tones, its absence in no-shock controls, and its persistence after controlling for freezing in the , as analysis—were unaffected. These analyses demonstrate that our conclusions are robust to this methodological choice.

> *(4) It would increase confidence in the tracking of neurons across days if the authors showed some example images of neurons tracked across days.*

We added an example in the Supplement. Additionally, we now provide measures of registration quality (Fig. S3)

> *(5) Table S2 is a bit confusing. I take it that Graded A/B were only for CS15, and Graded C/D/E were only for CS3. Also, Common B1-4 were the cells with positive responses to individual stimuli, and Common C1-4 were the cells with negative responses to stimuli. If this is the case, it should be explained in the figure legend (or even better, clusters should be named and numbered consistently in all figures.*

Thank you for pointing this out, we corrected the Table to indicate which test corresponds to which figure and cluster, specifying which ones were positive or negative modulated. Please note old Table 2 is now Table 5

> *(6) The term network and subnetworks implies some connectivity between neurons, but in this study, it is used to refer to the ensembles of cells activated in a similar manner. Since connectivity is never assessed, it would be better if the authors stuck to the terms ensembles or populations.*

We changed the wording and now use ensembles or populations

> *(7) The legend for Figure S3 has the text 'eded', which seems to be a typo.*

We corrected this typo.

References

(1) Kitamura, T., et al., Engrams and circuits crucial for systems consolidation of a memory. Science, 2017. 356(6333): p. 73–78.

(2) DeNardo, L.A., et al., Temporal evolution of cortical ensembles promoting remote memory retrieval. Nat Neurosci, 2019. 22(3): p. 460–469.

(3) Tome, D.F., et al., Dynamic and selective engrams emerge with memory consolidation. Nat Neurosci, 2024. 27(3): p. 561–572.

(4) Mau, W., M.E. Hasselmo, and D.J. Cai, The brain in motion: How ensemble fluidity drives memory-updating and flexibility. Elife, 2020. 9.

(5) Gallego, J.A., et al., Long-term stability of cortical population dynamics underlying consistent behavior. Nat Neurosci, 2020. 23(2): p. 260–270.

(6) Zaki, Y. and D.J. Cai, Memory engram stability and flexibility. Neuropsychopharmacology, 2024. 50(1): p. 285–293.

(7) Lacagnina, A.F., et al., Distinct hippocampal engrams control extinction and relapse of fear memory. Nat Neurosci, 2019. 22(5): p. 753–761.

(8) Sangha, S., Plasticity of Fear and Safety Neurons of the Amygdala in Response to Fear Extinction. Front Behav Neurosci, 2015. 9: p. 354.

(9) Bouton, M.E., S. Maren, and G.P. McNally, Behavioral and Neurobiological Mechanisms of Pavlovian and Instrumental Extinction Learning. Physiol Rev, 2021. 101(2): p. 611–681.

(10) Jenkins, H.M. and R.H. Harrison, Effect of discrimination training on auditory generalization. J Exp Psychol, 1960. 59: p. 246–53.

(11) Dunsmoor, J.E. and K.S. LaBar, Effects of discrimination training on fear generalization gradients and perceptual classification in humans. Behav Neurosci, 2013. 127(3): p. 350–6.

(12) Herzog, K., et al., Reducing Generalization of Conditioned Fear: Beneficial Impact of Fear Relevance and Feedback in Discrimination Training. Front Psychol, 2021. 12: p. 665711.

(13) Lommen, M.J.J., et al., Training discrimination diminishes maladaptive avoidance of innocuous stimuli in a fear conditioning paradigm. PLoS One, 2017. 12(10): p. e0184485.

(14) Casanova, J.P., et al., Threat-dependent scaling of prelimbic dynamics to enhance fear representation. Neuron, 2024. 112(14): p. 2304–2314 e6.

(15) Kyriazi, P., D.B. Headley, and D. Pare, Different Multidimensional Representations across the Amygdalo-Prefrontal Network during an Approach-Avoidance Task. Neuron, 2020. 107(4): p. 717–730 e5.

(16) McKenzie, S. and H. Eichenbaum, Consolidation and reconsolidation: two lives of memories? Neuron, 2011. 71(2): p. 224–33.

(17) Winocur, G. and M. Moscovitch, Memory transformation and systems consolidation. J Int Neuropsychol Soc, 2011. 17(5): p. 766–80.

(18) Hockley, A. and M.S. Malmierca, Auditory processing control by the medial prefrontal cortex: A review of the rodent functional organisation. Hear Res, 2024. 443: p. 108954.

(19) Zikopoulos, B. and H. Barbas, Prefrontal projections to the thalamic reticular nucleus form a unique circuit for attentional mechanisms. J Neurosci, 2006. 26(28): p. 7348–61.

(20) Phillips, R.G. and J.E. LeDoux, Differential contribution of amygdala and hippocampus to cued and contextual fear conditioning. Behav Neurosci, 1992. 106(2): p. 274–85.

https://doi.org/10.7554/eLife.111177.2.sa0