Affective memory rehearsal with temporal sequences in amygdala neurons

Affective Memory Rehearsal with Temporal Sequences in Amygdala Neurons

Tamar Reitich-Stolero and Rony Paz

Dept. of Neurobiology, Weizmann Institute of Science, Rehovot, Israel.

Correspondence should be addressed to: rony.paz@weizmann.ac.il

Abstract

Affective learning and memory are essential for daily behavior, with both adaptive and maladaptive learning depending on stimulus-evoked activity in amygdala circuitry. Behavioral studies further suggest that post-association offline processing also contributes to memory formation. Here, we investigated spike-sequences across simultaneously recorded neurons while monkeys learned to discriminate between aversive and pleasant tone-odor associations. We show that triplets of neurons exhibit consistent temporal sequences of spiking activity that differed from firing patterns of individual neurons and pairwise correlations. These sequences occurred throughout the long post-trial period, contained valence-related information, declined as learning progressed, and were selectively present during the recent CS-US evoked activity. Our findings reveal that temporal sequences across neurons in the primate amygdala serve as a coding mechanism, and might aid memory formation by rehearsal of the recently experienced association.

Introduction

The role of the amygdala in learning with aversive and pleasant outcomes is well established[1-11], and impaired processing can result in maladaptive expression and discrimination of fear from safety[12-15]. During learning, an initially neutral conditioned-stimulus (CS) is paired with an unconditioned-stimulus (US) to produce plasticity that enables memory formation[6, 16-21], leading most studies to focus on stimuli-evoked responses. Nevertheless, behavioral studies have shown that the length of the post-trial interval contributes to the acquisition rate[22, 23] and that learning is diminished when introducing a novel association during that time[24]. Together, these results suggest a memory rehearsal mechanism[24, 25] that is also in-line with an amygdala-dependent fast consolidation process[26-28]. Despite evidence for affective-state-specific tonic responses that continue in the absence of external stimuli[29, 30], it remains unclear how amygdala circuitry process specific associations after the stimuli terminated to aid memory formation. Here, we demonstrate that amygdala ensembles carry such information in timing and order of spiking activity during the post-trial offline period. The focus on the order of spikes across several neurons allows examination of a lower dimension compared to that of all possible spatiotemporal patterns, and is therefore computationally tractable.

We recorded neurons in the amygdala of two monkeys acquiring pleasant and aversive tone-odor conditioning on a daily basis (Fig.1A, Supplementary Fig.1A, n=119 neurons). A discriminatory conditioned-response (CR, higher inhale volume in response to the pleasant- or aversive- associated tone) occurred in 74% of the days (n=31/42, 2-way ANOVA, $p<0.05$ for main effect of valence). The discriminatory CR developed after the first trial and progressed along the acquisition session (Fig.1B).

To establish the role of temporal-sequences as a coding and rehearsal mechanism, we examined all groups of three simultaneously recorded units (n=355) and tested for the following criteria: First, that the structure of the sequence occurs beyond what is expected from single-neuron activity; Second, that sequences are consistent across time; Third, that sequences are valence-specific, namely they are consistent within valence, hold information about the trial-type, and allow decoding of the trial valence; Fourth, that sequences contain more information early in learning than when memory formation is complete; Fifth, that the sequences that occur during the post-trial period also occur during the stimuli-pairing evoked activity. Finally, we used several shuffling approaches to validate that sequences occur beyond independent changes in firing rates, beyond single pairwise correlations, and used maximum-entropy models to demonstrate 3rd order correlations. Together, fulfilling these criteria would constitute evidence for the use of spatiotemporal spiking sequences for memory rehearsal in the post-trial offline epoch during learning.

Results

Structure in amygdala spike sequences

For an unbiased selection of time window, we quantified the number of spike-*triplets* (Fig.1C; 3-spikes across 3-neurons) in sliding windows of different sizes, and a-priori chose 150ms because it captured the majority of sequences (Fig.1D). To identify triplets with

spatiotemporal patterns that do not result from single-unit firing-rate (FR) modulations, we compared sequences to circular shuffling of the entire spiking pattern of the neurons in a random bounded duration (between ±150-300ms, Extended Data Fig.1A,B). This shuffle preserves single neuron activity and destroys inter-neuron correlations, and we therefore term triplets with distinctive activity '*Structured triplets*'. Other shuffling approaches (methods, e.g. trial-shuffle) produced stronger-better results, indicating that the circular shuffle is indeed the most stringent approach.

We first examined whether sequences existed in amygdala triplets by comparing the actual distribution to that expected from single-neuron firing rates (Fig.1E, left and middle triplets vs. rightmost, Extended Data Fig.1B). During pre-task activity, a notable proportion of amygdala triplets exhibited significant structure (Fig.1F, 49% of n=355, Monte Carlo [MC] p-values with Benjamini-Hochberg [BH] correction for multiple-comparisons). In comparison, independent neurons recorded on different days (across-days-triplets) and independent activity of the neurons from within-day did not show structured sequences (trial shuffle, Fig.1F). A similar somewhat weaker effect (20% of n=104, $\chi^2_{df=1} = 27.7, p < 10^{-6}$) was found using only neurons recorded on different electrodes (Fig.1F, inset). This difference could partly stem from the proximity between the neurons, as physical distance between electrodes was smaller for triplets exhibiting structure compared to triplets with no structure (Extended Data Fig.2A). Moreover, the magnitude of sequence structure (structure score) was higher than that of across-days-triplets (Fig.1G; unpaired t test; all triplets: $t_{699} = 12.85, p < 10^{-35}, d = 0.97$; different electrode triplets: $t_{201} = 2.16, p = 0.016, d = 0.3$).

Importantly, we found that a large proportion of the structured triplets exhibited sequences that could not be explained even when taking into account the pairwise correlated activity of either pair (Fig.1H, Supplementary Fig.2). In addition, although we initially selected an unbiased temporal window of 150ms, we further quantified the proportion of structured triplets for different durations and found that sequences occurred even in shorter time scales (Fig.1I, 25-50ms).

Taken together, these results demonstrate that triplets of neurons in the amygdala exhibit sequence structures that are different than expected from single neuron as well as pairwise activity.

Amygdala sequences are consistent across time

If sequences are indeed used as a coding mechanism in the amygdala, they should consistently and repeatedly occur across time. To evaluate this, we compared the dissimilarity between two different time segments (Extended Data Fig.3), and identified *consistent triplets* that are more similar across time (Fig.2A, left and middle triplets vs. rightmost). A large proportion of triplets exhibited consistency (Fig.2B, all triplets: 33%; Different electrodes: 11%, MC p-value BH corrected), whereas independent across-days-triplets and within-day trial-shuffle did not show consistent sequences (Fig.2B). In accordance, the distribution of consistency scores was positively skewed (Fig.2C) and higher than that of independent across-days-triplets (unpaired t test; all triplets: $t_{694} = 10.2, p <$

$10^{-22}, d = 0.77$; Different electrode triplets: $t_{198} = 2.66, p = 0.004, d = 0.38$). Here again, many triplets exhibited consistency exceeding that expected from the pairwise correlations of either pair (Fig.2D). Similarly, sequences were consistent also in shorter time scales (Fig.2E) and the physical distance was smaller for triplets exhibiting consistent activity (Extended Data Fig.2B). Finally, there was a large overlap between structured and consistent triplets (Fig.2F). These results show that spike sequences in the amygdala are consistent throughout time.

Sequences are more abundant in the Amygdala than in the dACC

Next, we examined if spike-sequences during valence discriminatory learning occur more in the amygdala than in another region. We obtained recordings in the dorsal-anterior-cingulate-cortex (dACC, n=228; simultaneously recorded triplets: n=564; Supplementary Fig.1A), and repeated the same analyses. We found that a larger proportion of amygdala triplets were significantly structured compared to dACC triplets (Extended Data Fig.4A), with higher mean structure score (Extended Data Fig.4B). Similarly, the proportion of consistent triplets in the amygdala was larger than in the dACC (Extended Data Fig.4C), with higher mean consistency score (Extended Data Fig.4D). Because the BLA is smaller in size, we validated the analysis on triplets with similar anatomical distance in the dACC as in the amygdala and found similar results. We note that a similar proportion of dACC and Amygdala neurons exhibit FR response to the aversive and pleasant CS or US (aversive: amygdala: 11%, dACC: 8%, $\chi_1^2 = 1$, $p = 0.32$; pleasant: amygdala: 30%, dACC: 29%, $\chi_1^2 = 0.0008$, $p = 0.97$) and differentiate between aversive and pleasant CS or US (amygdala: 22%, dACC: 25%, $\chi_1^2 = 0.44$, $p = 0.5$).

This strengthens the finding that Amygdala triplets exhibit spike-sequences in this context of discriminatory affective learning compared to another region, the dACC, that is also involved in affective learning and shows similar stimulus-evoked responses. The spike-sequences might also underlie previous findings of more synchronized activity in the amygdala[31].

Amygdala sequences are consistent within valence

After having established the existence of temporal spike sequences across neurons, we sought to examine if they code for valence during the learning of affective-associations. To examine this, we sampled the distribution of sequences during the offline post-trial epoch and identified triplets that exhibited different distributions of spike-sequence for the pleasant versus the aversive trials (Fig.2G, trials sorted by type for presentation only). To confirm such valence-specific sequences, we examined whether the similarity between sequences following trials with similar valence, is higher than that following trials of different valence.

For all available triplets, the mean dissimilarity between aversive-related sequences was significantly lower than the mean dissimilarity between sequences of different valence (Fig.2H), and this difference was larger than the difference in independent across-days-triplets (Fig.2H middle-left). To further control for single-neuron activity and correlations, we

compared consistency scores and found that their mean within valence was significantly higher than the mean score between valence, and larger than the difference in across-days-triplets (Fig.2H middle-right). Similar results were found when comparing to independent trial-shuffle triplets (Supplementary Fig.3A). As the lower dissimilarity and higher scores imply higher similarity within the aversive post-trial comparisons, these results suggest that a significant subset of the triplets exhibit aversive specific sequences.

Similarly, the mean dissimilarity between distributions of pleasant-related sequences was lower than the mean dissimilarity between sequences of different valence (Fig.2I), also compared to independent across-days-triplets (Fig.2I middle-left), and the mean consistency score was higher within the pleasant post-trial epochs than between sequences of different valence, also compared with across-days-triplets (Fig.2I middle-right). Here again, similar results were found for trial-shuffle control (Supplementary Fig.3B). As for aversive, these results suggest that a significant subset of the triplets exhibit pleasant specific sequences.

Therefore, a significant proportion of triplets of amygdala neurons produce sequences that are specific to the valence of the recently presented (learned) association.

Sequences hold information about recent valence associations

To test if sequences hold information about aversive vs. pleasant associations of the recent trial, we examined the difference between decoding of valence using the sequences and decoding based on independent neurons (Supplementary Fig.4). We found that 20% of amygdala triplets were able to decode the valence of the previous trial above chance level (Extended Data Fig.5D, binomial test for each triplet, BH corrected). This proportion of correctly classifying triplets (20%) was higher than the proportion in independent across-days-triplets (Fig.3A, 6.7%), higher than in trial-shuffle data (Extended Data Fig.5A, MC p-value, all triplets: $p = 0.024$) and higher compared to the dACC (0.5%, $\chi^2_{df=1} = 111, p < 10^{-10}$). Similarly, the mean decoding hit rate was higher than in trial-shuffle triplets (Extended Data Fig.5B). We also found that these sequences differ from pre-task sequences, namely before associative-learning started (Extended Data fig.5C), because decoding based on valence-related triplets allowed correct discrimination between post-trial activity and pre-task activity (n = 71, BH corrected [FDR<0.05], aversive: 90%, pleasant: 45%). Note that there was no stereotypic or preparatory inhale behavior during this post-trial period (Supplementary Fig.5).

Notably, the discrimination was achieved using sequences that occur long after the stimuli terminated (2-12sec after the CS and the US). Moreover, a high proportion of decoding triplets exhibited stable decoding for more than 25 seconds after US offset (Fig.3B, ranging from 15-35%, p<0.05, one-tailed χ^2 test) and this decoding was enabled by similar sequences across different times (Supplementary Fig.6).

Interestingly, highly discriminating triplets achieved better decoding than inter-spike-intervals (ISI) or firing-rates, as the mean hit rate was higher based on sequences than on ISI (Extended Data Fig.5D, n=76, 130 respectively, one tailed independent t-test: $t_{204} = 1.69$, $p = 0.046, d = 0.24$). Similarly, sequence-based decoding achieved higher hit rates than ISI or FR based decoding in up to 10% of the significant triplets (Fig.3C, 28% of sequence-

significant triplets, one-tailed Wilcoxon rank-sum test, p<0.05). This benefit was not observed in independent across-days-triplets or in trial-shuffle triplets (Fig.3C, insets).

These results show that a significant proportion of the triplets hold more information than independent firing patterns and that this information is available long after the stimulus has terminated.

Sequences hold more information in early than in late learning

If temporal sequences are used to strengthen the learning of a recent association, their information should fade as learning progresses and the memory strengthens, in a teaching-signal like manner. Indeed, repeating the decoding with ten trials extracted from different phases of the learning (Fig.3D,E), we found a higher hit rate in the initial phase of acquisition (trials 1-10) compared to the intermediate phase (trials 11-20, Bonferroni corrected sign-rank test, $Z = 5.85, p < 10^{-8}$), and compared to the final phase (Fig. 3D,E, trials 21-30, $Z = 6.47, p < 10^{-9}$). This decline in decoding performance was not due to changes in FR or ISI distributions (Supplementary Fig.7), or a result of changes in local-field-potential (LFP) that could point to a different overall brain-state (Supplementary Fig.8). Furthermore, we quantified trial-by-trial decoding performance (proportion of correct classification across triplets) and found a negative correlation with the mean conditioned response (CR, Fig.3F, rank-order correlation $r = -0.41$, resampling p-value: $p = 0.016$, n=29; when removing the first trial- bottom dot- as outlier: $r = -0.44,\ p = 0.008$).

To further demonstrate that this reduction occurs also in information in addition to the decoding approach, we calculated the mutual information (MI) between sequence activity and recent trial valence (29% of n=328 triplets contain significant information about valence, BH corrected permutation test). Here also, we found that the proportion of triplets with significant MI during the initial phase of learning (46% of n=258) was larger compared to the final stage (Fig.3G-I, 26% of n=269, χ^2 test for independence, $\chi^2_{df=1} = 22.3, p < 10^{-5}$), as well as a significant reduction in MI between the initial and later phases (Fig.3G-I, initial vs. intermediate: n=249, Bonferroni corrected sign-rank test, $Z = 6.53, p < 10^{-9}$; initial vs. final: n=243, $Z = 3.74, p < 0.001$; intermediate vs. final: n=254, $Z = -2.03, p = 0.064$). These results were specific to the task-related information, as the overall number of sequences did not decrease along the learning (Fig.3D-top inset).

We conclude that sequences hold information in the post-trial epoch when the association is still being acquired and this information decreases as learning progresses, a characteristic of a memory-rehearsal process.

Trial-specific sequences are repeated in the post-trial epoch

Finally, if the sequences indeed serve as a post-trial rehearsal mechanism, then we can hypothesize that the same sequences should be present also in evoked responses during the CS-US presentation (Fig.4A). We therefore examined whether post-trial valence-specific sequences occurred also during the preceding CS-US presentation. We repeated the decoding approach that was trained on post-trial sequences only, but this time tested the

performance on activity during the CS-US presentation. Post-trial triplets that significantly decoded preceding trial valence ($p < 0.05$, n=101), displayed an average hit rate significantly above chance also when tested on CS-US evoked activity (Fig.4B, inset). This hit rate was also higher than the hit rate in triplets with no post-trial decoding (Fig.4B, inset, n=254). Accordingly, there was a positive correlation between the post-trial decoding and the hit rate based on activity during the CS-US presentation (Fig.4B, Spearman rank-order correlation: $r = 0.28, p < 10^{-6}$). This suggests that some of the sequences that occur in the stimulus-evoked activity are later repeated during post-trial activity.

To demonstrate this more directly, we examined the occurrence of aversive or pleasant -specific post-trial sequences during CS-US related activity. For each triplet, we identified aversive-/pleasant- specific sequences in post-trial epochs (Fig.4C) and quantified their presence in evoked activity during the CS-US presentation. As expected, aversive-specific post-trial sequences were more abundant in aversive CS-US activity (Fig.4D, one tailed sign-rank test: $Z = 2.7, p = 0.004$) whereas pleasant-specific post-trial sequences were more frequent in pleasant CS-US activity (Fig.4D, $Z = 3.89, p < 10^{-4}$). This rehearsal activity did not exhibit itself in pre-task activity (Supplementary Fig.9).

Together, these results suggest that a portion of post-trial sequences are repetitions of the activity that occurs during the acquisition trial, implying a rehearsal mechanism for the recent association in post-trial activity.

Maximum entropy (ME) models validate the role of sequences

To further demonstrate sequence activity in triplets, we fitted two types of ME models[32-35] to amygdala activity. We first implemented the standard spatial model fitted on simultaneously recorded quadruplets of neurons in order to quantify the gain obtained by using triplets compared to pairwise (Extended Data Fig.6A, 'spatial-ME', n=358). In addition, because the spatial model does not consider the order of spikes in a triplet, we further developed a novel ME model to examine the sequential activity of triplets (Extended Data Fig.6B, 'Sequence-ME', n=291). For both models, we re-tested structure, consistency, decoding, and CS-US rehearsal.

We computed the independent, pairwise and triple-wise models on the data of individual trials (Fig.5A,B). We then quantified the reduction in total entropy due to the pairwise and triple-wise correlations, reflecting the contribution of these interactions to the overall activity. There was a significant contribution of triple-wise interactions to the reduction in entropy, beyond that expected from pairwise activity (Fig.5C,D, comparing to surrogate data sampled from the pairwise ME model, and see Supplementary Fig.10A,B for pairwise vs. independent). These results strengthen the conclusions of sequence-structure in triplets, demonstrating a triple-wise interaction in the sequences.

To assess consistency, we compared the JSD dissimilarity between the model in one time-segment and the data in another time-segment. We first identified groups with pairwise consistent activity (see methods), and found that in 15.3% of these quadruplets (Spatial-ME, n=13/85) and in 26.5% of these triplets (Sequence-ME, n=81/307), the dissimilarity in the

triple-wise model was smaller than in the pairwise model (BH corrected, FDR≤0.05). These results further demonstrate consistency in triplets of neurons.

For valence decoding from post-trial activity, we found better performance of the triple-wise compared to the pairwise model in the spatial-ME (Fig.5E). Similarly, there was a trend in the sequence-ME for higher performance of the triple-wise compared to the pairwise model (Fig.5F, and see Supplementary Fig.10C,D for pairwise vs. independent).

Finally, we compared the decoding of CS-US activity from the model trained on post-trial activity (as in Fig.4B). In the spatial-ME, the hit rate for decoding CS-US activity from the triple-wise and pairwise models were higher than the independent model (Extended Data Fig.7A). In the sequence-ME, the hit rate of the triple-wise model was higher than that of the independent model (Extended Data Fig.7B), and higher than the pairwise model.

These results further support the findings of the shuffle approach, and hence the notion that sequence activity during CS-US presentation is repeated during the post-trial period.

We also validated that the main findings are not different between putative excitatory projection cells and interneurons (Supplementary Fig.11), and further cannot be explained by unit-isolation (Supplementary Fig.12), non-stationarity of firing-rates (Extended Data Fig.8), or short phasic FR modulations and correlations (Supplementary Fig.13).

Discussion

Overall, our findings show that temporal sequences across multiple amygdala neurons maintain information about discriminatory valence associations. We find that specific sequences exist at baseline, as structure and consistency of triplet sequences were identified during pre-task activity and beyond pairwise and independent (firing-rate) patterns. In addition, sequences further develop according to trial valence when conditioning begins, suggesting a coding mechanism. Because these sequences were identified during the long post-trial periods, diminished as learning progressed (similar to a teaching signal), and were repetitions of CS-US evoked sequences, they likely serve as a rehearsal mechanism of the recently acquired association. This is a first demonstration of post-trial rehearsal during learning in amygdala neurons, and of coding with temporal sequences across several neurons in this circuitry. It suggests that the affective association is repeated to enhance synaptic plasticity[20, 21, 36, 37], and moreover, the short time-scales of sequences compared to the CS-US gap might reconcile previous debates about plasticity constraints during the pairing itself[21].

Although it is reminiscent of offline replay in the hippocampus[38, 39], there is a major difference between the findings. In the hippocampus, specific cells increase firing rates at specific spatial locations along the behavioral trajectory[40], so that the ordering of single-cell activity is behaviorally imposed and a time compressed sequence is repeated offline[38]. Affective conditioning does not impose external ordering, implying a different rehearsal mechanism and further introducing a technical difficulty to detect these sequences.

Our findings in triplets that exceed pairwise-correlations therefore point to a spatiotemporal code[41-45] and a first demonstration for its role during affective learning in the primate.

Therefore, the results suggest that associations are not encoded solely by firing rate (FR) changes, but also by sequences of spikes that are rehearsed offline to enhance learning. Although circuit mechanisms that can generate such reliable sequences and their readout are yet to be demonstrated conclusively, such ordinal activity as we identify here can result from the sparse sampling of three neurons (as the case in extracellular recordings) from three different yet connected sub-populations. This is in line with the varying and relatively long temporal lags of dozens of ms we observed between the spikes. In such a case, our findings are consistent with many studies showing phasic changes in FR synchrony across subpopulations of neurons[46]. Together with our findings that the reported activity exceeds short time-scale FR modulations, we argue that spike-sequences are the best explanation for the results presented here.

The sequence code and rehearsal, as well as the large proportion of triplets, suggest that they are part of a larger memory-coding ensemble in the amygdala[11, 47-50]. It remains to be seen how such larger ensembles are activated during learning and how they are enhanced or constrained by temporal patterns as shown here. Overall, we conclude that temporal-sequences in primate amygdala neurons replay recent affective associations between trials to aid memory formation.

Acknowledgments

We thank Y. Kfir, A. Taub, U. Livneh, Y. Cohen and K. Aberg, as well as E. Schneidman, and E. Karpas for scientific consult. We thank Y. Shohat for animal training, experiments and welfare; E. Kahana for medical and surgical procedures; E. Furman-Haran and F. Attar for MRI procedures. This work was supported by ISF #2352/19 and ERC-2016-CoG #724910 grants to R. Paz.

Author Contributions

T.R.S and R.P conceived and designed the experiments; T.R.S. planned and performed the analyses; T.R.S and R.P. wrote the manuscript.

Competing interests

The authors declare no competing interests.

References

1. Baxter, M.G. & Murray, E.A. The amygdala and reward. *Nat Rev Neurosci* **3**, 563-573 (2002).
2. Janak, P.H. & Tye, K.M. From circuits to behaviour in the amygdala. *Nature* **517**, 284-292 (2015).
3. Salzman, C.D., Paton, J.J., Belova, M.A. & Morrison, S.E. Flexible neural representations of value in the primate brain. *Ann N Y Acad Sci* **1121**, 336-354 (2007).
4. Sugase-Miyamoto, Y. & Richmond, B.J. Neuronal signals in the monkey basolateral amygdala during reward schedules. *J Neurosci* **25**, 11071-11083 (2005).
5. Herry, C. & Johansen, J.P. Encoding of fear learning and memory in distributed neuronal circuits. *Nat Neurosci* **17**, 1644-1654 (2014).
6. Maren, S. & Quirk, G.J. Neuronal signalling of fear memory. *Nat Rev Neurosci* **5**, 844-852 (2004).
7. Krabbe, S., Grundemann, J. & Luthi, A. Amygdala Inhibitory Circuits Regulate Associative Fear Conditioning. *Biol Psychiatry* **83**, 800-809 (2018).
8. Duvarci, S. & Pare, D. Amygdala microcircuits controlling learned fear. *Neuron* **82**, 966-980 (2014).
9. Namburi, P*., et al.* A circuit mechanism for differentiating positive and negative associations. *Nature* **520**, 675-U208 (2015).
10. Yu, K*., et al.* The central amygdala controls learning in the lateral amygdala. *Nat Neurosci* **20**, 1680-1685 (2017).
11. Josselyn, S.A., Kohler, S. & Frankland, P.W. Finding the engram. *Nat Rev Neurosci* **16**, 521-534 (2015).
12. Likhtik, E. & Paz, R. Amygdala-prefrontal interactions in (mal)adaptive learning. *Trends Neurosci* **38**, 158-166 (2015).
13. Averbeck, B.B. & Chafee, M.V. Using model systems to understand errant plasticity mechanisms in psychiatric disorders. *Nat Neurosci* **19**, 1418-1425 (2016).
14. Delgado, M.R., Olsson, A. & Phelps, E.A. Extending animal models of fear conditioning to humans. *Biol Psychol* **73**, 39-48 (2006).
15. Milad, M.R. & Quirk, G.J. Fear extinction as a model for translational neuroscience: ten years of progress. *Annu Rev Psychol* **63**, 129-151 (2012).
16. Johansen, J.P*., et al.* Optical activation of lateral amygdala pyramidal cells instructs associative fear learning. *Proc Natl Acad Sci U S A* **107**, 12692-12697 (2010).
17. Quirk, G.J., Repa, C. & LeDoux, J.E. Fear conditioning enhances short-latency auditory responses of lateral amygdala neurons: parallel recordings in the freely behaving rat. *Neuron* **15**, 1029-1039 (1995).
18. Herry, C*., et al.* Switching on and off fear by distinct neuronal circuits. *Nature* **454**, 600-606 (2008).
19. Johansen, J.P., Cain, C.K., Ostroff, L.E. & LeDoux, J.E. Molecular mechanisms of fear learning and memory. *Cell* **147**, 509-524 (2011).
20. Pape, H.C. & Pare, D. Plastic synaptic networks of the amygdala for the acquisition, expression, and extinction of conditioned fear. *Physiol Rev* **90**, 419-463 (2010).
21. Sah, P., Westbrook, R.F. & Luthi, A. Fear conditioning and long-term potentiation in the amygdala: what really is the connection? *Ann N Y Acad Sci* **1129**, 88-95 (2008).
22. Gibbon, J., Baldock, M.D., Locurto, C., Gold, L. & Terrace, H.S. Trial and intertrial durations in autoshaping. *J. Exp. Psychol. Anim. Behav. Process.* **3**, 264-284 (1977).
23. Lattal, K.M. Trial and intertrial durations in Pavlovian conditioning: issues of learning and performance. *J Exp Psychol Anim Behav Process* **25**, 433-450 (1999).
24. Wagner, A.R., Rudy, J.W. & Whitlow, J.W. Rehearsal in animal conditioning. *J Exp Psychol* **97**, 407-426 (1973).

25. Wagner, A.R. SOP: A Model of Automatic Memory Processing in Animal Behavior. *Information Processing in Animals, Memory Mechanisms*, 5-47 (1981).
26. Nader, K., Schafe, G.E. & LeDoux, J.E. The labile nature of consolidation theory. *Nat Rev Neurosci* **1**, 216-219 (2000).
27. Holland, P.C. & Schiffino, F.L. Mini-review: Prediction errors, attention and associative learning. *Neurobiol Learn Mem* **131**, 207-215 (2016).
28. McIntyre, C.K., Power, A.E., Roozendaal, B. & McGaugh, J.L. Role of the basolateral amygdala in memory consolidation. *Ann N Y Acad Sci* **985**, 273-293 (2003).
29. Lee, S.C., Amir, A., Haufler, D. & Pare, D. Differential Recruitment of Competing Valence-Related Amygdala Networks during Anxiety. *Neuron* **96**, 81-88 e85 (2017).
30. Belova, M.A., Paton, J.J. & Salzman, C.D. Moment-to-moment tracking of state value in the amygdala. *J Neurosci* **28**, 10023-10030 (2008).
31. Pryluk, R., Kfir, Y., Gelbard-Sagiv, H., Fried, I. & Paz, R. A Tradeoff in the Neural Code across Regions and Species. *Cell* **176**, 597-609 e518 (2019).
32. Schneidman, E., Berry, M.J., 2nd, Segev, R. & Bialek, W. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature* **440**, 1007-1012 (2006).
33. Martignon, L*., et al.* Neural coding: higher-order temporal patterns in the neurostatistics of cell assemblies. *Neural Comput* **12**, 2621-2653 (2000).
34. Nakahara, H. & Amari, S. Information-geometric measure for neural spikes. *Neural Comput* **14**, 2269-2316 (2002).
35. Panzeri, S. & Schultz, S.R. A unified approach to the study of temporal, correlational, and rate coding. *Neural Comput* **13**, 1311-1349 (2001).
36. Girardeau, G., Inema, I. & Buzsaki, G. Reactivations of emotional memory in the hippocampus-amygdala system during sleep. *Nat Neurosci* **20**, 1634-1642 (2017).
37. Feldman, D.E. The spike-timing dependence of plasticity. *Neuron* **75**, 556-571 (2012).
38. Carr, M.F., Jadhav, S.P. & Frank, L.M. Hippocampal replay in the awake state: a potential substrate for memory consolidation and retrieval. *Nat Neurosci* **14**, 147-153 (2011).
39. Buzsaki, G. & Llinas, R. Space and time in the brain. *Science* **358**, 482-485 (2017).
40. O'Keefe, J. & Dostrovsky, J. The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Res* **34**, 171-175 (1971).
41. Schnitzer, M.J. & Meister, M. Multineuronal firing patterns in the signal from eye to brain. *Neuron* **37**, 499-511 (2003).
42. Ikegaya, Y*., et al.* Synfire chains and cortical songs: temporal modules of cortical activity. *Science* **304**, 559-564 (2004).
43. Pillow, J.W*., et al.* Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature* **454**, 995-999 (2008).
44. Ganmor, E., Segev, R. & Schneidman, E. Sparse low-order interaction network underlies a highly correlated and learnable neural population code. *Proc Natl Acad Sci U S A* **108**, 9679-9684 (2011).
45. Oram, M.W., Wiener, M.C., Lestienne, R. & Richmond, B.J. Stochastic nature of precisely timed spike patterns in visual system neuronal responses. *J Neurophysiol* **81**, 3021-3033 (1999).
46. Buzsaki, G. Neural syntax: cell assemblies, synapsembles, and readers. *Neuron* **68**, 362-385 (2010).
47. Grewe, B.F*., et al.* Neural ensemble dynamics underlying a long-term associative memory. *Nature* **543**, 670-675 (2017).
48. Reijmers, L.G., Perkins, B.L., Matsuo, N. & Mayford, M. Localization of a stable neural correlate of associative memory. *Science* **317**, 1230-1233 (2007).

49. Rashid, A.J.*, et al.* Competition between engrams influences fear memory formation and recall. *Science* **353**, 383-387 (2016).
50. Grundemann, J. & Luthi, A. Ensemble coding in amygdala circuits for associative learning. *Curr Opin Neurobiol* **35**, 200-206 (2015).

Figure 1. Experimental setup and Structure of spatiotemporal sequences in the amygdala

(A) Each trial began with a pure tone, followed by an aversive (Propionic acid) or pleasant (banana and melon organic extract) odor. Analyses were performed prior to any stimuli ('baseline activity') and during post-trial epochs, starting 2 seconds after the termination of odor delivery. Shown also is an example raster plot of a single amygdala neuron during 5 seconds of the post-trial epoch without any external stimuli.

(B) The mean conditioned response (CR, measured as difference in full width at half maximum [FWHM] of inhale duration, see methods) showed fast initial learning (bottom inset), and progression along the session (main panel, discriminatory days, n=31, trials 1-10 vs. trials 11-20, one tailed paired t-test, $t_{df=30} = -2.13$, p=0.021, d=0.21; trials 1-10 vs. 21-30, $t_{df=30} = -2.2$, p=0.018, d=0.25; and trials 11-20 vs. 21-30, p=0.27). Top inset: single inhalation example (CR) with shorter inhale duration upon presentation of the pleasant (purple) compared to the aversive (red) conditioned stimuli (CS). FWHM are marked by corresponding dashed lines. ΔFWHM score takes the absolute value of the changes, so inhale volume can change in either direction (see methods).

(C) Estimating the probability distribution of three spike-sequences of three neurons. Left: Surrogate example of voltage traces from three neurons. The boxes symbolize a running window that starts with a spike in any of the neurons. A sequence is counted if three spikes occurred within the time window. Right: the estimated sequence probability distribution.

(D) Proportion of three spikes sequences within a time duration for amygdala triplets during the post-trial epoch. Dashed line: the unbiased a-priori chosen time duration used throughout the study unless specifically mentioned otherwise (150ms).

(E) Examples of structured (two left examples) and non-structured triplets (right). Top: mean data (blue) and shuffled data (green) sequence probability distributions, sorted by the shuffled distribution (log scale). The data and shuffled distributions are different in the structured triplets (p=0.002, right tailed Monte Carlo) and similar in the non-structured triplet (p=0.8). The shaded areas represent standard error of the mean (SEM) over 10s time segments (n=30), averaged over shuffled instances. Bottom: In the structured triplets, the mean Jensen-Shannon-divergence (JSD) dissimilarity between shuffled data sequences ($\overline{D}_{1,2\ldots500}$, green histogram) is smaller than the mean dissimilarity between the data and the shuffled sequences ($\overline{D}_{\mathrm{data}}$, , blue line).

(F) Distribution of p-values (right tailed Monte Carlo, as in E) for all simultaneously recorded triplets (blue, n=355), independent across-days-triplets (gray, n=355) and independent trial-shuffle control (turquoise, n=355). Many simultaneously triplets showed significant structure ($p \leq 0.05$). Inset: triplets from different recording electrodes (n=104).

(G) Frequency of scores for simultaneously recorded triplets (blue), independent across-days-triplets (gray) and independent trial-shuffle control (turquoise). The right tail of the simultaneously recorded distribution suggests that many triplets exhibit structure that is highly different from single neurons. Inset: triplets from different recording electrodes.

(H) Proportion of significantly structured triplets beyond either of the three pairwise activities (i.e. compared to all three single unit shuffles, right tailed Monte Carlo, as in E, p<0.05 for all three, n=195). The proportion was significantly higher than chance (dashed black) for all triplets (38%, 75/195, χ^2 test for goodness of fit for p=0.05 $\chi^2_{\mathrm{df}=1} = 459, \mathrm{p} < 10^{-20}$) as well as for triplets recorded on different electrodes (14%, $\chi^2_{\mathrm{df}=1} = 4.72, \mathrm{p} = 0.03$). Note that these triplets are structured beyond pairwise activity of single pairs (third order structure is demonstrated in Fig.5D).

(I) Proportion of structured triplets as a function of maximal sequence durations (n=355).

Error bars: standard error of the mean (SEM).

In all panels error bars mark the standard error of the mean (SEM); $* \mathrm{p} < 0.05$; $** \mathrm{p} < 0.01$; $*** \mathrm{p} < 0.001$

Figure 2. Consistent amygdala sequences throughout time and within valence.

(A) Consistent (two left examples) and non-consistent (right) triplets. Top: data (blue) and shuffled data (green) sequence probability distribution of the two subdivisions (as exemplified in the left blue bar, n=15 for each). In the two left examples the similarity between the solid blue and dashed line shows that the data sequences are similar to each other. The difference between these blue lines and the green lines shows that the data sequences are different from the shuffled sequences (p=0.002, left tailed Monte Carlo). In the right example the data sequences are similar to the shuffled sequences (p = 0.83). Bottom: histogram of mean JSD dissimilarity between the data and shuffled sequences ($\bar{C}_{1,2\ldots500}$, green) and a line indicating the dissimilarity between the data sequences ($\bar{C}_{\mathrm{data}}$, blue). The higher similarity between the data sequences suggest that they are consistent.

(B) Distribution of p-values for all possible simultaneously recorded triplets (blue, n=355), independent across-days-triplets (gray, n=355) and independent trial-shuffle control (turquoise, n=355). Many simultaneously recorded triplets showed significant consistency ($p \leq 0.05$, right tailed Monte Carlo, as in A). Inset: triplets from different recording electrodes (n=104).

(C) Frequency of consistency scores of simultaneously recorded triplets (blue), independent across-days-triplets (gray) and independent trial-shuffle control (turquoise). Inset: triplets from different recording electrodes. Scores are larger for simultaneously recorded triplets, indicating consistent sequences.

(D) Proportion of triplets significantly consistent beyond expected from either of the three pairwise activities (i.e. compared to all three single unit shuffles, right tailed Monte Carlo, as in A, p<0.05). The proportion was significantly larger than chance level (dashed black) for all triplets (20%, 28/139, χ^2 test for goodness of fit for p=0.05, $\chi^2_{\mathrm{df}=1} = 67, \mathrm{p} < 10^{-15}$) and for triplets from different recording electrodes (15%, $\chi^2_{\mathrm{df}=1} = 4.21, \mathrm{p} = 0.04$).

(E) Proportion of consistent triplets as a function of maximal sequence durations (n=355).

(F) The number and overlap between structured (pink) and consistent (purple) - triplets.

(G) Two examples of amygdala triplets with different sequence distributions in aversive (red) and pleasant (purple) post-trial epochs. Upper sections: sequence probability distributions averaged over all trials (mean and SEM). Lower sections: color maps of sequence probability distributions of single trials in the pleasant (top half) and aversive (lower half). Note that pleasant-aversive separation is only for presentation purposes; trials were interleaved.

(H) Comparison of JSD dissimilarity and consistency scores between sequence probability distributions estimated in the post-trial of two halves of the aversive trials ('aversive', n=15 vs. 15) and between the sequence probability distributions estimated in post-trial epoch of half of the aversive trials and half of the pleasant trials ('between').

Top: Single triplets' JSD (mean and SEM over subdivisions) between aversive-related sequences (x axis) and between aversive and pleasant related sequences (y axis). The JSD of many triplets is above the black identity line, implying higher similarity between aversive-

related sequences compared to the similarity across valence. Right top corner: histogram of differences between the two JSD.

middle-left: The mean JSD over all triplets between aversive-related sequences ('within day', red) was smaller than the mean JSD between aversive and pleasant related sequences ('within day', pink, one tailed paired t-test: $t_{df=354} = -3.53, p < 10^{-3}, d_{Cohen} = 0.12$), beyond the difference in the across-days-triplets control ('across days', red and pink, 2X2 mixed model ANOVA within stimulus valence and between triplet type; interaction $F_{df=1} = 13.7, p < 10^{-3}$). Shaded area: triplets from different recording electrodes.

Bottom-left: Violin plot of the difference between the JSD of aversive related sequences (corresponding to the red bar in the middle-left plot) and the JSD between aversive and pleasant related sequences (pink bar) for individual triplets (black dots, n=355). The colored surface marks the kernel density estimate of the corresponding probability distribution, the thick gray line marks the interquartile range and the black dashed line marks mean difference.

middle-right: The mean consistency score over all triplets between aversive-related sequences ('within day', red) was larger than the mean consistency score between aversive and pleasant related sequences ('within day', pink, one tailed paired t-test: $t_{df=354} = 2.71, p < 0.01, d_{Cohen} = 0.1$), beyond the difference in the independent across-days-triplets ('across days', red and pink, 2X2 mixed model ANOVA within stimulus valence and between triplet type; interaction $F_{df=1} = 4.26, p = 0.04$). Shaded area: triplets from different recording electrodes.

Bottom-right: Violin plot of the difference between the consistency score of aversive related sequences (corresponding to the red bar in the middle-right plot) and the consistency score of aversive and pleasant related sequences (pink bar) for individual triplets (black dots, n=355). Violin elements are as in the bottom left panel.

(I) Arranged as (H) for the pleasant trials.

middle-left: within day t-test: $t_{df=354} = -4.37, p < 10^{-5}, d_{Cohen} = 0.18$; interaction: $F_{df=1} = 5.7, p < 0.05$.

middle-right: within day t-test: $t_{df=354} = 4.32, p < 10^{-4}$, $d_{Cohen} = 0.17$; interaction $F_{df=1} = 10.1, p < 0.01$,

In all panels error bars and shaded area mark SEM, $* p < 0.05$; $** p < 0.01$; $*** p < 0.001$

Figure 3. Sequence based decoding and information in the post-trial epoch.

(A) Proportion of triplets with higher-than-chance hit rate (binomial test for each triplet, BH corrected, false discovery rate [FDR]≤0.05) was larger for within day triplets than independent across-day-triplets (χ^2 test for independence: $\chi^2_{\mathrm{df}=1} = 26.8, \mathrm{p} < 10^{-6}$). Inset: triplets from different recording electrodes ($\chi^2_{\mathrm{df}=1} = 5.15, \mathrm{p} = 0.023$).

(B) Proportion of triplets with significant decoding performance (χ^2 test) as a function of time from stimulus offset, calculated on 5 seconds running window (with 4 seconds overlap, n = 355). At all times, the proportion was significantly higher than chance (dashed line).

(C) Mean decoding hit rate as a function of the proportion of triplets included. Triplets (n=193) are sorted in a descending manner based on hit rates of: sequences distribution (blue), ISI distribution (green, solid) and FR distributions (green, dotted). The hit rate of high performance triplets was higher based on sequences compared to ISI and FR (significance marked by black dots). Top left inset: across-days control (n=201). Bottom right inset: trial-shuffle control (mean over n=250 repetitions).

(D) Mean decoding hit rate as a function of acquisition trials in a session (n=355). The hit rate was significantly higher in the first 10 trials of learning. Left bottom inset: boxplot of the hit rates as a function of acquisition trials, normalized (Z-score) for each triplet along the acquisition trials. Right top inset: the overall sequence rate averaged over all days.

(E) Hit rate of individual triplets in the first vs. last 10 trials (x and y axes, respectively). Blue: all triplets; purple: triplets with significant decoding performance for the entire day. Most triplets are below the black dashed identity line, suggesting higher hit rate early in learning.

(F) Trial by trial CR is negatively correlated with the proportion of classifying triplets in each trial (2-trials smoothing, n=29). Dashed line: linear regression ($\mathrm{r} = -0.44,\ \mathrm{p} = 0.008$).

(G) Mutual information in triplets (MI, mean and SEM, left y axis, blue, n=328) and proportion of triplets with significant MI (right y axis, pink) as a function of learning trials. Inset: boxplot of MI as a function of acquisition trials, normalized (Z-score) for each triplet along the acquisition trials.

(H) Same as (G) for MI rate (mean and SEM), i.e. bits per second.

(I) MI of individual triplets in the first vs. last 10 trials (x and y axes, respectively). Blue: all triplets (n=243); pink, dark purple, light purple: triplets with significant information for trials 1-10, 21-30 and both phases, respectively. Most triplets are below the black dashed identity line, suggesting higher information early in learning.

In all box plots, boxes include 25 to 75 percentile with the median marked by the middle line, whiskers mark the last data point within 1.5 interquartile range from the median.

In all panels error bars and shaded area mark SEM, $* \mathrm{p} < 0.05$; $** \mathrm{p} < 0.01$; $*** \mathrm{p} < 0.001$

Figure 4. Valence-specific post-trial sequences are repetitions of sequences that occurred during CS-US presentations.

(A) Example of CS-US evoked firing rate response and post-trial activity. Top panels: Raster plot and PSTH of a single amygdala neuron in response to pleasant (purple) and aversive (red) CS (top left panel), US (top right panel) and post-trial activity (bottom panel).

(B) Valence-decoding from trial (CS-US) sequences can be achieved based on valence-specific post-trial sequences. Main panel: decoding hit rate tested on CS-US sequences (but trained on post-trial-sequences; y-axis) is positively correlated with post-trial (train and test) decoding hit rate (x-axis, n=355). Purple and blue: triplets with significant/non-significant post-trial decoding, and a significant linear regression line (black dashed). Inset: mean hit rate for decoding CS-US valence (from post-trial training) for significant post-trial triplets (purple, n=101) is significantly higher than chance level (gray, one sample t-test, $t_{df=100} = 4.9, p < 10^{-5}, d = 0.43$) and higher than post-trial non-significant triplets (blue, n=254, independent samples t-test, $t_{df=353} = 5.48, p < 10^{-7}, d = 0.45$). Notice that this analysis does not require cross-validation as the training sequences are taken from post-trial activity and the test sequences are taken from trial (CS-US) activity.

(C) Two single triplet examples of aversive and pleasant-specific post-trial sequences. Top part: sequence probability distribution for aversive (red) and pleasant (purple). Bottom part: sequence probability ratio ($P(seq|aversive)/P(seq|pleasant)$). Differentiating sequences for aversive-specific (red rectangle) and pleasant specific (purple rectangle) were selected for each triplet. The sum of proportions of these example sequences in CS-US activity is marked by full (top example) and dashed (bottom example) gray squares in (D), where the proportions of pleasant specific sequences are marked by purple dots and aversive by red dots.

(D) For each post-trial decoding triplet (n=101), the sum of proportion of sequences that are associated with aversive (red) or pleasant (purple) post-trial activity out of all sequences present during aversive (x-axis) and pleasant (y-axis) CS-US pairings. For example, the sequences [231, 133, 112], were aversive-specific in post-trial activity (bottom example in [C]). The summed proportion of these sequences in aversive CS-US activity (0.34) was higher than the summed proportion in pleasant CS-US activity (0.27). Across all post-trial decoding triplet, aversive post-trial sequences were more frequent during aversive CS-US pairings (below the dashed black identity line) whereas pleasant post-trial sequences were more frequent in pleasant CS-US pairings (above the identity line). Main panel: Using three aversive-specific and three pleasant-specific post-trial sequences. Inset: histogram of differences between the two proportions. Bottom left/right: using 2/4 valence-specific sequences, respectively.

In all panels error bars and shaded area mark SEM, $* p < 0.05$; $** p < 0.01$; $*** p < 0.001$

Figure 5. Maximum Entropy (ME) models support structure, consistency, coding and rehearsal in triplets.

(A) Spatial-ME model. Left: a quadruplet with triple-wise correlations. Right: a quadruplet with pairwise but not triple-wise correlation. The probability of each word (Extended Data Fig.6) in each time segment (n=30) is plotted for the independent (blue), pairwise (orange) and triple-wise (yellow) models as a function of the probabilities in the real data of the quadruplet. In the left panel, the triple-wise model probabilities are proximate to the black dashed identity line while the others are scattered, indicating that only the triple-wise model is a good predictor of the data. Accordingly, the proportion of reduction of entropy due to the triple-wise interactions ($I_{(3)}/I_N$) is high. In the right panel, the independent model probabilities are scattered while the pairwise and triple-wise are proximate to the identity line, as both are good predictors of the data. Accordingly, the proportion of reduction of entropy due to the triple-wise interactions ($I_{(3)}/I_N$) is low. Insets: JSD dissimilarity between the probability distributions of the data and the distributions of each model for each time segment. The reduction in entropy is calculated as $I_{(3)} = H_2 - H_3$ and the multi information, $I_3 = H_1 - H_3$, where H_k is the entropy of the k'th order of the model.

(B) Sequence-ME model in triplets. Same presentation as in (A).

(C) Spatial-ME model (n=358). Proportion of reduction in entropy due to the triple-wise correlations (I_3/I_N) for the real data (x-axis) and for surrogate data sampled from the pairwise ME distribution (pairwise-surrogate control, y-axis). This surrogate data preserves pairwise correlations, as it is sampled from the pairwise ME model, but any third order correlations are random. Therefore, I_3/I_N in the pairwise surrogate is the reduction in entropy expected by chance. The reduction in entropy due to the triple-wise correlations is larger for the real data (below the black dashed identity line, paired t test between medians across trials: $t_{357} = 25.93, p < 10^{-20}, d = 0.78$), indicating that triple-wise correlations explain the variability beyond expected from pairwise correlations. Inset: means and SEM over all quadruplets. $*** p < 0.001$.

(D) Sequence-ME model in triplets (n=291). Same presentation as in C (Paired t test between medians across trials: $t_{290} = 14.85, p < 10^{-20}, d = 0.62$).

(E) Spatial-ME model. Decoding hit rate for single quadruplets based on the pairwise model (x-axis) and based on the triple-wise model (y-axis), with the histogram of the ratios between the hit rate of the triple-wise and pairwise models (n=119, paired one tailed t-test, $t_{118} = 2.69, p < 0.005, d = 0.25$). The higher hit rate based on the triple-wise model suggest coding in triple-wise correlations.

(F) Sequence-ME model in triplets. Same presentation as in (E) (n=150, paired one tailed t-test, $t_{149} = 1.5, p = 0.07$; Pink: triplets with significant sequence-decoding taken from Fig.3A).

Methods

Behavioral paradigm and Electrophysiological recordings

Two male macaca fascicularis (4 years old) were implanted with a recording chamber above the right amygdala and the dACC, and an MRI scan was performed to assess chamber position over dACC and amygdala (Supplementary Fig.1). Images were acquired on a 3T Trio (Siemens) Scanner, equipped with a 12 channels head matrix coil combined with a knee coil (Siemens), the primate was lying in prone position. 3D T1 weighted magnetization prepared rapid acquisition gradient-echo (MPRAGE) pulse sequence was acquired, Cartesian acquisition, field of view 160 × 130 mm, 192 x 156 matrix and 0.83 mm^3 slice thickness, resolution tilted from the sagittal plane. TE/TR/TI = 3.36ms/2500ms/1100ms, 8° flip angle, 2 averages. All surgical and experimental procedures were approved and conducted in accordance with the regulations of the Weizmann Institute Animal Care and Use Committee (IACUC), following NIH regulations and with AAALAC accreditation. Food, water, and enrichments (e.g., fruits and play instruments) were available ad libitum during the whole period, except before medical procedures.

In the behavioral paradigm, primates were seated in a dark room and engaged in a classical conditioning task in which tones (conditioned stimulus, CS) were coupled with odors (unconditioned stimulus, US)[51, 52]. Each recording day was initiated with a habituation phase of ten presentations of two conditioned stimuli (CS), pure (sinus wave) tones chosen randomly in the range between 1000-2500 Hz to induce new learning in each session. The acquisition session that followed included 30 intermixed presentations of the two CS tones paired with an aversive (Propionic acid) or pleasant (a mixture of banana and melon organic extract) odor. Odor presentation was locked to the first breath after the CS tone, but not less than 1 second (s) after tone onset.

Each day, 3–4 microelectrodes were lowered inside a metal guide into the brain using a head-tower and electrode-positioning-system (Alpha Omega). The electrodes were then moved independently further into the amygdala and dACC. Electrode signals were pre amplified, 0.3 Hz-6 KHz band-pass filtered and sampled at 25 KHz. At the end of the recording period, off-line spike sorting was performed (offline sorter, Plexon Inc).

Number of monkeys, number of recording days (sessions), and overall number of recorded neurons is similar to those reported in previous publications and as customary in the field[51, 52].

Data analysis

Behavioral conditioned response

Breath duration was quantified as full width at half maximum (FWHM) of inhale pressure. Conditioned response (CR) was quantified as inhale FWHM following the CS, normalized by the inhale FWHM in the 3 baseline breathes prior to CS:

$$\mathrm{CR} = \frac{\mathrm{FWHM}_{\mathrm{CS}} - \overline{\mathrm{FWHM}}_{\mathrm{baseline}}}{\mathrm{FWHM}_{\mathrm{CS}} + \overline{\mathrm{FWHM}}_{\mathrm{baseline}}}$$

,where $\overline{x} = \sum_i \frac{x_i}{N}$.

To examine the change of the CR along the day, the difference between each CR and the CR of the first trial (prior to any feedback) was evaluated. This response was quantified only in days with reliable pressure measurement and inhale onset detection (requiring peak amplitude > 0, time to peak < 500ms and FWHM < 800ms but > 50ms) in at least 2/3 of the trials (n=42).

Differential aversive and pleasant CR was identified by performing 2-way ANOVA (valence X trials) for each day, taking days with significance effect of valence. For these days, the difference between CRs was quantified, taking $\Delta CR = CR_{pl} - CR_{av}$ for days with $\overline{CR}_{pl} > \overline{CR}_{av}$ (n=16) and $\Delta CR = CR_{av} - CR_{pl}$ for days with $\overline{CR}_{pl} < \overline{CR}_{av}$ (n=15). The development of this response along the day, namely learning, was verified by testing for the difference between the ΔCR in the initial stage of learning (trials 1-10) and later stages (trials 11-20, 21-30).

Whereas the unconditioned-response (UR, the response to the odor) shows the expected lower-shorter inhale for aversive odor and higher-longer inhale for pleasant odor[52, 53], the conditioned-responses (CR) reflects a coping strategy and varies between animals and sessions. One can observe the two typical behaviors described in classical conditioning literature: either the CR and UR are in the same direction, as in early classical-conditioning theories, or they have opposite direction, as can be expected from 'naïve' reasoning (a longer inhale for the CS to prepare for the shorter inhale for the aversive odor), or as observed in electric-shock studies that show opposite direction between CR and UR of evoked autonomic measures. To measure learning and the development of the CR independent of this and in-line with our previous studies that found different strategies between animals and sessions, we tested for a difference in the half-width as long as it is consistent within a session.

Neuronal analyses

Baseline activity was taken from a 30 segments X 10s time period prior to any paradigm-related stimulus. Post-trial activity was taken as 30 trials X 10s periods starting 2s after US offset.

Sequence distributions (below) were estimated for all possible triplets of neurons that were recorded simultaneously. In addition, the results were compared to triplets based on independent neurons that were recorded in different days (across-days-triplets), preserving independent neurons activity and the dynamic of each neuron along time. The results were also compared to shuffling of two of the neurons across trials (trial-shuffle), preserving independent neurons activity and single neuron identity.

Estimation of sequence probability distributions

Sequences were defined as a sequence of three spikes from the activity of three neurons that occurred within a time lag (10-250ms). Sequences were counted by using an overlapping running window and calculating the probability of each sequence.

$$\widehat{\mathrm{P}}^{\mathrm{data}}(\mathrm{seq_i}) = \frac{\#\,(\mathrm{seq}\colon \mathrm{seq} = \mathrm{seq_i})}{\mathrm{N}}$$

Where $\mathrm{seq_i}$ is a specific sequence of three spikes (from any of the three neurons) and N is the total number of recorded sequences.

Shuffling methods

To test for differences between the spatiotemporal-structured triplet and that expected from firing-rate (FR) correlations and single-neuron firing patterns (FP) of the same three neurons, shuffled data sets (n=500) were created by circularly shuffling the entire spiking patterns of two neurons in a random duration between $\pm$150-300ms (Extended Data Fig.1A). Circular shuffling was performed on each time segment separately, i.e. on 10s time epochs during baseline activity, or on individual 10s post-trial activity.

Analyses were repeated with three additional shuffles that were applied on two of the neurons: unbounded circular shift (rather than 150-300ms), shuffling across trials, and Poisson shuffle. In trial shuffle, the order of the 30 time segments (post-trial or time segments of baseline activity) was randomly shuffled. In Poisson shuffle the number of spikes within a predetermined non-overlapping time window (150-500ms) was counted and randomly assigned back [54]. The Poisson shuffle was highly sensitive to single neuron firing patterns such that the structure and consistency analyses appeared significant even for independent across-days-triplet. These shuffles showed to be generally less stringent and are therefore not reported in the main text.

For each shuffled data instance the sequence probability distributions were estimated as:

$$\widehat{\mathrm{P}_{\mathrm{j}}^{\mathrm{Sh}}}(\mathrm{seq_i}) = \frac{\#\,(\mathrm{seq}\colon \mathrm{seq} = \mathrm{seq_i})}{\mathrm{N}}$$

Where $\widehat{\mathrm{P}}^{\mathrm{sh}}$ is the estimated sequence distribution of a shuffled data set, $\mathrm{seq_i}$ is a specific sequence of three spikes (from any of the three neurons), N is the total number of recorded sequences, j is shuffle index.

Shuffling to test for pairwise activity for all 3 pairs

To control for pairwise activity, the same shuffling method was performed only on one of the neurons, thereby preserving the joint activity of the unshuffled pair and destroying the relation to the shuffled neuron (Supplementary Fig.2A). Thus, each triplet was tested against three shuffled data sets (n=200 instances for each shuffled neuron). Tests for pairwise activity were performed for all three shuffled data sets and determined significant for $\mathrm{p} \leq 0.05$ for all three tests. This tests if the sequence activity is different from expected from either pairwise activity separately. Since all three tests are required to ascribe significance, the probability of type-1 error for all three tests is bounded by $\alpha = 0.05$:

If $\mathrm{H_0}$ is true for all three tests: $\alpha_{\mathrm{group}} = \alpha^3$

If $\mathrm{H_0}$ is true only for the first test: $\alpha_{\mathrm{group}} = \alpha * (1-\beta_2)(1-\beta_3) \leq \alpha$, where $(1-\beta_{\mathrm{i}})$ is the power of the test for the i'th shuffled unit.

Jensen–Shannon divergence (JSD) - probability distribution dissimilarity measure

JSD was used as a measure of dissimilarity between probability distribution. JSD is symmetric and bounded in the range $[0,1]$.

$$JSD(P||Q) = \frac{D_{KL}(P||M) + D_{KL}(Q||M)}{2}, \quad M = \frac{P+Q}{2}, \qquad D_{KL}(P||M) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

Benjamini–Hochberg (BH) correction for false discovery rate (FDR) in multiple comparisons

The largest p-value for which $P_i < \frac{i}{m} * \alpha$ was detected, where m is the total number of comparisons and $\alpha = 0.05$ is the maximal expected proportion of errors. The critical p-value was set as P_i, guarantying $FDR \leq \alpha$.

Structure analysis scheme

The probability of each sequence was estimated for the shuffled data sets and for the real data using the entire 300s time period or 30 acquisition post-trial epochs (30 trials X 10s).

The mean JSD between the shuffled sequence distribution and the individual sequence distribution was estimated as a measure of dissimilarity for both the data and the shuffled data sets:

$$\overline{D}_{data} = \frac{\sum_{j=1}^{n} JSD(\widehat{P_j^{Sh}}||\widehat{P}^{data})}{n}$$

$$\overline{D}_{shuffle}^{k} = \frac{\sum_{j=1}^{n} JSD(\widehat{P_j^{Sh}}||\widehat{P_k^{Sh}})}{n}$$

, where n is the number of shuffled data instances.

Large dissimilarity between data and shuffled data would suggest a structured probability distribution (Extended Data Fig.1B), so a right tailed Monte Carlo p-value for the structure measure and a structure score ($\overline{D}_{idx}$) were estimated based on shuffled data instances:

$$\widehat{P_{val}}(\overline{D}_{data}) = \frac{1 + \sum_{k=1}^{n} I\{\overline{D}_{shuffle}^{k} \geq \overline{D}_{data}\}}{1+n}$$

$$\overline{D}_{idx} = \frac{\overline{D}_{data} - \overline{\overline{D}_{sh}^{1,2,..,n}}}{\overline{D}_{data} + \overline{\overline{D}_{sh}^{1,2,..,n}}}$$

, where n is the number of shuffled data instances and $\overline{x} = \sum_i \frac{X_i}{N}$

Consistency analysis scheme

The probability of each sequence was estimated on L=100 semi-randomized subdivisions of the 30 time segments into two groups (15 segments of 10s each). Sets of subdivisions were randomly selected 1000 times and the chosen set was the one that maximized the Hamming distance between the different subdivisions. The JSD between the probability distributions of the two data segments was averaged over subdivisions ($\overline{C}_{data}$) and compared to the average JSD between one data segment and one shuffled data segment ($\overline{C}_{shuffle}^{k}$).

$$\bar{C}_{data} = \frac{\sum_{l=1}^{L} JSD(\widehat{P_{l,1}}^{data} || \widehat{P_{l,2}}^{data})}{L}$$

$$\bar{C}_{shuffle}^{k} = \frac{\sum_{l=1}^{L} \frac{JSD(\widehat{P_{l,1}}_{k}^{sh} || \widehat{P_{l,2}}^{data}) + JSD(\widehat{P_{l,2}}_{k}^{sh} || \widehat{P_{l,1}}^{data})}{2}}{L}$$

, where $\widehat{P_{l,1}}^{data}$ is the sequence probability distribution of the first group of the *l*'th subdivision of the data and $\widehat{P_{l,1}}_{k}^{sh}$ is the sequence probability distribution of the first group of the *l*'th subdivision of the *k*'th shuffle.

Large similarity between data segments would suggest a consistent distribution (Extended Data Fig.3), so left-tailed Monte Carlo p-values and consistency scores ($\bar{C}_{idx}$) were estimated:

$$\widehat{P_{val}}(\bar{C}_{data}) = \frac{1 + \sum I\{\bar{C}_{shuffle}^{k} \leq \bar{C}_{data}\}}{1 + n}$$

$$\bar{C}_{idx} = \frac{\overline{\bar{C}_{shuffle}^{1,\ldots,n}} - \bar{C}_{data}}{\overline{\bar{C}_{shuffle}^{1,\ldots,n}} + \bar{C}_{data}}$$

, where n is the number of shuffled data instances.

Consistency within versus across comparisons

For each triplet, we found the sequence duration that produced the maximal mean consistency score for between and within stimulus valence ($Seq_{lag}^{*}(pl), Seq_{lag}^{*}(av)$):

$$Seq_{lag}^{*}(pl) = \underset{Seq_{lag} \in \{10,25,50,100,150,200,250\}}{argmax} \bar{C}_{idx}^{pl-pl} + \bar{C}_{idx}^{pl-av}$$

$$Seq_{lag}^{*}(av) = \underset{Seq_{lag} \in \{10,25,50,100,150,200,250\}}{argmax} \bar{C}_{idx}^{av-av} + \bar{C}_{idx}^{pl-av}$$

, where $\bar{C}_{idx}^{pl-pl}, \bar{C}_{idx}^{av-av}$ are calculated as $\bar{C}_{idx}$, where $\widehat{P_{l,1}}$ and $\widehat{P_{l,2}}$ are estimated from pleasant or aversive post-trial epoch, respectively. $\bar{C}_{idx}^{pl-av}$ is calculated as $\bar{C}_{idx}$ where $\widehat{P_{l,1}}$ is estimated from pleasant and $\widehat{P_{l,2}}$ is estimated from aversive post-trial epoch.

Taking the relevant sequence duration for each triplet, the JSD between the post-trial sequence distributions of the same stimulus was estimated and averaged over 100 subdivisions (as in Consistency analysis scheme):

$$JSD(pl||pl) = \frac{\sum_{l=1}^{L} JSD(\widehat{P_{l,1}}^{pl} || \widehat{P_{l,2}}^{pl})}{L}$$

$$JSD(av||av) = \frac{\sum_{l=1}^{L} JSD(\widehat{P_{l,1}}^{av} || \widehat{P_{l,2}}^{av})}{L}$$

These were compared to the JSD between the post-trial sequence distributions of the different stimuli, averaged over the two possibilities:

$$\mathrm{JSD}(\mathrm{av}||\mathrm{pl}) = \frac{\sum_{l=1}^{L} \mathrm{JSD}(\widehat{P_{l,1}}^{\mathrm{av}}||\widehat{P_{l,2}}^{\mathrm{pl}}) + \mathrm{JSD}(\widehat{P_{l,1}}^{\mathrm{pl}}||\widehat{P_{l,2}}^{\mathrm{av}})}{2L}$$

Next, consistency scores were evaluated for post-trial sequence distributions of the same stimulus, $\bar{C}_{idx}^{pl-pl}$ and $\bar{C}_{idx}^{av-av}$ and compared to the consistency score between the post-trial sequence distributions of the different stimuli, $\bar{C}_{idx}^{pl-av}$.

Inter-spike interval (ISI) distribution estimation

First, a naïve estimation of the inter spike interval (ISI) distribution was estimated

$$P(ISI = x) = \frac{\#(isi: isi = x)}{N}$$

Where N is the total number of counted inter spike interval. This distribution was smoothed using Kernel density estimation with a normal kernel evaluated at 100 equally spaced points.

Firing rate (FR) distribution estimation

Firing rates (FR) were counted on non-overlapping 250ms time bins and the probability distribution was estimated naïvely, without accounting for the timing of the FR.

Likelihood ratio decoding from post-trial activity

According to Neyman–Pearson lemma, the log-likelihood ratio is the most powerful test to discriminate between two hypotheses. Therefore, it can be used to test how well a readout mechanism can discriminate between the previously presented stimulus and the current stimuli.

$$L_{\{s1\}}(r) = \log\frac{P(s_1|r_1,\ldots,r_n)}{p(s_2|r_1,\ldots,r_n)} = \log\frac{P(r_1,\ldots,r_n|s_1)*\frac{p(s_1)}{p(r_1,\ldots,r_n)}}{P(r_1,\ldots,r_n|s_2)*\frac{p(s_2)}{p(r_1,\ldots,r_n)}} = \cdots$$

$$= \log\frac{P(r_1,\ldots,r_n|s_1)}{P(r_1,\ldots,r_n|s_2)} + \log\frac{p(s_1)}{p(s_2)} = \sum \log\frac{P(r_i|s_1)}{P(r_i|s_2)}$$

Where r is the neural response (sequences, ISI or FR), s1 and s2 are the pleasant and aversive stimuli. The last equality holds for balanced stimulus presentation ($p(s_1) = p(s_2)$) and independent responses.

The conditioned probability distributions, $P(r = r_i|s)$, were estimated in the post-trial activity consecutive to the stimulus s (pleasant or aversive) of all acquisition trials except the j'th trial and $r_1, \ldots r_n$ are the responses in the post-trial of trial j (*Leave one out cross validation*).

If the log likelihood ratio of the test set was smaller than zero, the decoder classified the stimulus as s_2 and vice versa. Hit rates were calculated as $\frac{\#\text{correct classification}}{\#\text{trials}}$. Significance level of decoding performance of single triplets was tested by a binomial or χ^2 test under the null hypothesis that $p(\text{correct}) = p(\text{error}) = 0.5$.

Decoding performance as a function of time in the post-trial was assessed by decoding on 5s running window with 4s overlap, starting from US onset (-3s).

For the ISI and FR based decoding independence between the three neurons was assumed so likelihood ratios were summed over all three neurons and classified:

$L_{\{s1\}}(r) = \log\frac{P(s_1|r_{11},\ldots,r_{n1},r_{12},\ldots,r_{n2},r_{13},\ldots,r_{n3})}{p(s_2|r_{11},\ldots,r_{n1},r_{12},\ldots,r_{n2},r_{13},\ldots,r_{n3})} = \cdots = \sum_{j=1}^{3}\sum \log\frac{P(r_{ij}|s_1)}{P(r_{ij}|s_2)}$, where $r_{ij}$is the i′th response of neuron j.

As not all triplets work together to produce sequences and the ISI/FR distributions hold more information on single units, the average decoding performance of all possible triplets was expected to be smaller for sequences. To compare decoding performance on putative sequence-coding triplets, the hit rate of the best performing triplets was evaluated as a function of proportion of triplets included, taking triplets from best to worst performance. To avoid selection bias, best performing triplets were taken separately for each method, enabling an unbiased comparison between sequence-best triplets and ISI/FR-best triplets.

Likelihood ratio decoding between post-trial and pre-task activity

To verify that valence-specific sequences did not exist in pre-task activity, post-trial pleasant and aversive activity was decoded from pre-task activity. To this end, pre-task activity was divided into 30 segments of 10 second each (matched to the post-trial activity) and likelihood ratio decoding was performed between pre-task and aversive, as well as between pre-task and pleasant, post-trial activity using *Leave one out cross validation.*

Correlation between trial-by-trial decoding performance and CR

Trial by trial decoding performance was assessed by quantifying the proportion of triplets that correctly classified the i′th pleasant and aversive trials:

$$\text{proportion}(\text{trial i}) = \frac{\sum_{j=1}^{n_{triplets}} correct_{av}(i)+correct_{pl}(i)}{2*n_{triplets}}$$

Conditioned responses were estimated as ΔCR above. As these measures are noisy, we used two trial temporal smoothing (two trials running window with 1 trial overlap). The correlation was tested by resampling procedure, where trials were first shuffled, then smoothed (as the original data) and correlated. This was repeated $n = 10{,}000$ times to get:

$$p_{resampling} = \frac{1+\sum I\{r_{data} \le r_{resampled}\}}{1+n}$$

This was further multiplied by 2 to account for the comparisons with no smoothing.

FR response

CS and US FR were evaluated in a 1 sec time window after stimulus onset and baseline activity was evaluated in a 1 sec time window prior to CS onset. For each neuron and each valence (pleasant or aversive), a paired two tailed t-test was performed on the FR response across 30 trials comparing baseline activity to CS response and baseline activity to US response. In addition, differential FR response was evaluated by comparing (paired two tailed t-test) pleasant and aversive responses to the CS or the US, normalized by baseline activity ($\frac{FR_{Stimulus}-FR_{baseline}}{FR_{Stimulus}+FR_{baseline}}$).

Local field potential (LFP)

LFP signals were sampled at 781.25Hz, filtered with high-pass Butterworth filter with a cutoff of 3 Hz and a low-pass Butterworth filter with a cutoff at 90 Hz. After filtering, individual

electrodes were Z-scored and power spectral analysis and spike triggered average of the LFP signal were computed on the normalized signals[55].

Mutual information (MI)

Mutual information (MI) between sequences of individual triplets and stimulus valence (pleasant vs. aversive) was calculated by sampling the sequence distribution for all 30 trials of each valence or for ten trials along learning.

$$\mathrm{MI}_{\mathrm{naive}}(\mathrm{R}||\mathrm{S}) = \mathrm{H}(\mathrm{R}) - \mathrm{H}(\mathrm{R}|\mathrm{S}) = \mathrm{H}(\mathrm{sequences}) - \mathrm{H}(\mathrm{sequences}|\mathrm{valence})$$

, where H(R) is the entropy and H(R|S) is the conditioned entropy.

To decrease under-sampling bias, MI was calculated only for sequence distributions with sufficient sampling, taking a sampling criterion: $\frac{\mathrm{N_s}}{\mathrm{R}} \geq 12$, where $\mathrm{N_s}$ is the total number of observed sequences in all pleasant and in all aversive trials and R is the size of the sampled space of sequences in either stimuli (≤ 27) [56]. The under-sampling bias [56] was estimated by:

$\mathrm{bias}[\mathrm{MI}(\mathrm{R}||\mathrm{S})] = \frac{1}{2\mathrm{N}\ln(2)}\{\sum_{\mathrm{s=pl,av}}[\mathrm{R_s} - 1] - [\mathrm{R} - 1]\}$, where N is the total number of sequences and $\mathrm{R_s}$ is the size of the sampled space of sequences for the pleasant or aversive stimulus.

The presented MI are corrected such that:

$$\mathrm{MI} = \mathrm{MI}_{\mathrm{naive}}(\mathrm{R}||\mathrm{S}) + \mathrm{bias}[\mathrm{MI}(\mathrm{R}||\mathrm{S})]$$

The MI estimates the average information (i.e. reduction in uncertainty) between sequences and valence in a single event, namely a single sequence. To estimate the average information transmitted by sequences in one second, we multiplied the MI of individual triplets in each time segment by the sequence rate in that time segment.

To test the significance of the MI we performed a 1000-iterations permutation test where post-trial activity segments were randomly assigned (without replacement) to pleasant or aversive groups and the same sufficient sampling criterion and bias correction were applied.

CS-US by post-trial Likelihood ratio decoding

Valence (pleasant vs. aversive) was decoded from CS-US activity based on post-trial probability distributions. CS-US sequences were counted in a 2s window starting from CS onset, where US onset was set to the next breath onset (≥ 1s and < 3s after CS onset).

$$\mathrm{L}_{\mathrm{CS-US\ \{s1\}}}(\mathrm{r}) = \sum \log \frac{\mathrm{P}(\mathrm{r_i}|\mathrm{s_1})}{\mathrm{P}(\mathrm{r_i}|\mathrm{s_2})}$$

The decoder was trained on post-trial epochs of all trials (estimating the conditioned distribution $\mathrm{P}(\mathrm{r} = \mathrm{r_i}|\mathrm{s})$). To ensure proper sampling of CS-US activity, the decoder was tested on 30 sets of CS-US sequences from 15 randomly chosen trials J = $\{\mathrm{j}_1, \ldots \mathrm{j}_{15}\}$ (summing over all sequences in CS-US responses of all trials in J).

Proportion of post-trial valence-specific sequences in CS-US evoked activity

Valence-specific sequences were categorized by examining the ratio $\frac{\mathrm{P}(\mathrm{seq_i}|\mathrm{av})}{\mathrm{P}(\mathrm{seq_i}|\mathrm{pl})}$, evaluated from post-trial epoch of all trials. Aversive/pleasant sp\ecific sequences were taken as m

sequences with maximal/minimal ratio, respectively, while ignoring single neuron sequences (e.g. [1,1,1]). The proportion of valence-specific sequences was evaluated during aversive and pleasant CS-US activity: $\frac{\#\text{ aversive specific sequences}}{\#\text{ sequences}}$ and $\frac{\#\text{ pleasant specific sequences}}{\#\text{ sequences}}$, for $m = 2,3,4$.

Pleasant rehearsing triplet were defined as triplets with a larger proportion of pleasant specific sequences in the pleasant CS-US activity than the proportion in aversive CS-US activity:

$$\frac{\#\text{ pleasant specific sequences (plesant)}}{\#\text{ sequences (pleasant)}} > \frac{\#\text{ pleasant specific sequences (aversive)}}{\#\text{ sequences (aversive)}}$$

Aversive rehearsing triplets were defined as triplets with a larger proportion of aversive specific sequences in aversive CS-US activity:

$$\frac{\#\text{ aversive specific sequences (aversive)}}{\#\text{ sequences (aversive)}} > \frac{\#\text{ aversive specific sequences (pleasant)}}{\#\text{ sequences (pleasant)}}$$

To test if pleasant- and aversive-rehearsed sequences were present in pre-task activity, the proportion of valence-specific sequences was compared between CS-US response and pre-task activity.

Maximum entropy (ME) models

The Maximum Entropy Toolbox for MATLAB, version 1.0.2. 2017[57] was used to fit exact solutions to the models described below.

Unless stated otherwise, models were fit with a threshold of $th = 10^{-4}$ standard deviation of the expected measurement noise.

Spatial-ME model: The ME model for triple-wise spatial connection is of the form: $P(x) = \frac{1}{Z}\exp(\sum_{i=1}^{N} h_i x_i + \sum_{i<j} j_{ij} x_i x_j + \sum_{i<j<k} m_{ijk} x_i x_j x_k)$, where z is a scaling factor, $N = 4$ is the number of neurons in each group and i, j, k are indexes for neurons.

It is fitted to the data based on three groups of constraints:

(independent spike rate) $< \theta_i > \; = \frac{1}{T}\sum_{t=1}^{T} \theta_i(t)$

(pairwise correlations) $< \theta_{ij} > \; = \frac{1}{T}\sum_{t=1}^{T} \theta_i(t)\theta_j(t)$

(triple-wise correlations) $< \theta_{ijk} > \; = \frac{1}{T}\sum_{t=1}^{T} \theta_i(t)\theta_j(t)\theta_k(t)$

The pairwise model is only constrained by the independent and pairwise constraints and takes $m_{ijk} = 0$, and the independent model is only constrained by the independent constraint and takes also $j_{ij} = 0$. This model was fitted to all groups of 4 neurons, binned into 50ms binary words (Extended Data Fig.6, cases where $n_{spikes} > 1$ were taken as $n_{spikes} = 1$). The 50ms was taken due to the sequences structure found in this time duration (Fig.1I).

Sequence-ME model: To capture the temporal characteristics of the sequences in three neurons while using the ME model, a reduced data set was generated with 1ms bins, neglecting all time bins where none of the neurons spiked or more than one neuron spiked (Extended Data Fig.6). Triplets with time segments of less than 20 samples were disqualified.

The ME model for three steps spatiotemporal model is of the form $P(x_{T,T+1,T+2}) = \frac{1}{Z}\exp(\sum_{i=1}^{N}\sum_{t=T}^{T+2} h_{i(t)}\, x_i(t) + \sum_{i<j}\sum_{t=T}^{T+2} h_{ij(t)}\, x_i(t)x_j(t) + \sum_{i,j}\sum_{t=T}^{T+1} j_{i(t)j(t+1)}\, x_i(t)x_j(t+1) + \sum_{i,j,k} m_{i(t)j(t+1)k(t+2)}\, x_{i,t}x_{j,t+1}x_{k,t+2})$, where z is a scaling factor, $N = 3$ is the number of neurons in each group, i, j, k are indexes for neurons and *t* is time index.

It was fitted based on three groups of constraints:

(independent) $< \theta_i > \; = \frac{1}{T}\sum_{t=1}^{T} \theta_i(t)$; $< \theta_{ij} > \; = \frac{1}{T}\sum_{t=1}^{T} \theta_i(t)\theta_j(t)$

(pairwise spatiotemporal correlations) $< \theta_{i(t)j(t+1)} > \; = \frac{1}{T}\sum_{t=1}^{T} \theta_i(t)\theta_j(t+1)$

(triple-wise spatiotemporal correlations) $< \theta_{i(t)j(t+1)k(t+2)} > \; = \frac{1}{T}\sum_{t=1}^{T} \theta_i(t)\theta_j(t+1)\theta_k(t+2)$

The pairwise model is only constrained by the independent and pairwise constraints and takes $m_{i(t)j(t+1)k(t+2)} = 0$, while the independent model is only constrained by the independent constraints and takes also $j_{i(t)j(t+1)} = 0$.

Notice that the independent constraint in this model includes pairwise spatial correlations (but not temporal), as these are bound from model construction (simultaneous spikes from two neurons were not allowed) and tends to be severely overestimated. Namely, when the sparse 1ms spike matrix is taken without no-spikes time bins, it becomes very abundant in spikes, but there are no events where two neurons spike simultaneously. This is very unpredictable based on the rates of the neurons, as many simultaneous spiking events are expected, such that it creates biased probability distributions compared to the data. This bias is fixed by the learning of pairwise connections, as a spike of one neuron predicts that there is no spike of the others and low co-firing is predicted.

Testing structure using the ME model

For each group the spatial-ME and sequence-ME models were fitted to 30 time segments of 10sec each from the pre-task data.

To quantify the contribution of the pairwise and triple-wise correlation to the uncertainty in the data (i.e. pairwise and triple-wise structure), the proportion of reduction in entropy by each order was calculated as the ratio between $I_{(k)} = H_{k-1} - H_k$ and the multi information, $I_N = H_1 - H_N$, where H_k is the entropy of the model with k'th order correlations and H_N is the entropy of the data[32].

Since by definition the data is better explained by higher order models, this measure was compared to a surrogate data set (matching in the number of samples to the real data), sampled from the independent model (p_1) or from the pairwise model (p_2). New models were fitted to these generated data sets and the same measures were calculated. These comparison guarantees that the contribution of the pairwise and triple-wise correlations is not a result of chance or overfitting the model to the data.

Consistency account using the ME model

For each group of neurons, the spatial-ME and sequence-ME models were fitted to 30 time segments of 10sec each from the pre-task data. 200 sets of train-test subdivisions were created, with 90% train segments ($n_{train} = 27$ segments of 10sec each) and 10% test segments ($n_{test} = 3$ segments). For each train-test subdivision the probabilities of the model

fitted to individual trials were averaged and the JSD between the model train distribution (P_i^{train}) and the data test distribution (p_{data}^{test}) were calculated:

$$JSD_{\text{model order}} = JSD(P_i^{train} || p_{data}^{test})$$

Where p_i ($i = 1,2,3$) is the probability distribution corresponding to the model order.

For each triplet/quadruplet the JSD of low and high model orders were compared using a paired t-test. Pairwise consistent triplets/quadruplets had significantly lower JSD in the pairwise compared to the independent model. Triple-wise consistent triplets/quadruplets had significantly lower JSD in the triple-wise compared to the pairwise model.

Likelihood ratio decoding from ME models

For each group of neurons, the spatial-ME and sequence-ME model ($th = 0.1$ standard deviations, to reduce over-fitting) were fitted to post-trial activity of all pleasant and aversive trials (30 trials of 10sec from each stimulus). For each order of the ME model, the ME probability distribution were used to train the decoder and it was tested on the real data using *Leave one out cross validation*:

$L_{\{s1\}}(r) = \sum \log \frac{P_{ME}(r_i|s_1)}{P_{ME}(r_i|s_2)}$, where $P_{ME}(r|s)$ is taken as the average probability of the maximum entropy model of all trial but trial j, and r are taken from the data of trial j. For the sequence-ME, triplets were included if the ME model was valid in at least 75% of the trials.

To avoid overfitting in the case of triplets and quadruplets that only code the stimulus independently, the comparison between the pairwise and triple-wise models were done only on groups that were not clearly coding independently:

$$\text{hit rate}_{\text{pairwise}} > \text{hit rate}_{\text{independent}} \cup \text{hit rate}_{\text{triple-wise}} > \text{hit rate}_{\text{independent}}$$

These preconditions do not create selection bias, as they are symmetric with respect to the pairwise and triple-wise orders.

To test CS-US decoding from post-trial activity, the decoder was trained on post-trial epochs of all trials but trial j. The sequence-ME was tested on 30 sets of CS-US sequences from 15 randomly chosen trials $J = \{j_1, \dots j_{15}\}$, to ensure sufficient sampling (as the number of sample was dependent on the activity). The spatial-ME model was tested only on trial j (as the number of samples was fixed, n=40).

Putative interneurons and projection cells

Spike durations were measured on unfiltered voltage traces and defined as the interval between trough and peak for the negative spikes and the interval between peak and trough for positive spikes. To minimize misclassification, we applied two criteria for putative interneurons: $FR \geq 7\text{Hz}$ and spike duration $\leq 0.5\text{ms}$, and two criteria for putative projection cells: $FR \leq 1\text{Hz}$ and spike duration $\geq 0.7\text{ms}$. Since the number of neurons that were classified using this method was low, the firing rates and spike duration of all neurons pertaining to significant and non-significant triplets in different criteria were also examined (Supplementary Fig.11).

Since each neuron can take part in more than one triplet, comparison of the groups was done by permutation tests that preserve neurons identity. Thus, the FR and spike durations

were shuffled across the neurons, but the composition of triplets from neurons was preserved, thereby preserving dependencies between triplets. This shuffling approach was repeated 10000 times, and MC p-value was extracted by comparing the mean difference in FR or spike durations in the real data to the mean difference in FR or spike durations in the shuffled data.

Similarly, the difference in the probabilities of putative interneurons and projection cells to have significant structure, decoding and rehearsal were tested by shuffling the identity of the interneurons and projection cells across all classified neurons. Here again, the shuffling approach was repeated 10000 times, and MC p-value was extracted by comparing the mean difference in probabilities between interneurons and projection cells based on the real classification of the neurons to that of the shuffled classification.

FR Stationarity

Two tests for stationarity were employed: 1. Two tailed t-test comparing the average firing rate in the first and last 150 seconds of the pre-task activity (FR t-test). 2. Runs-test examining if inter-spike-intervals (ISI) along the pre-task activity were drawn randomly from a single distribution. The proportion of structured and consistent triplets in stationary and non-stationary triplets was compared and the reduction in entropy analysis that demonstrated 3-wise sequence activity was repeated.

Isolation score (unit isolation)

Isolation scores[58] were calculated as the isolation between unit_1 and unit_2:

$$\mathrm{P_X(Y)} = \frac{\exp\left(-\mathrm{d(X,Y)}\left(\frac{\lambda}{\mathrm{d_0}}\right)\right)}{\sum_{\mathrm{Z}\neq\mathrm{X}}\exp\left(-\mathrm{d(X,Z)}\left(\frac{\lambda}{\mathrm{d_0}}\right)\right)}$$

Where $\mathrm{d(X,Y)}$ is the Euclidian distance, $\mathrm{X,Y,Z}$ are spike shapes of unit_1 and unit_2, $\lambda = 10$ is a scaling factor and $\mathrm{d_0}$ is the average Euclidian distance between all spike shapes of the two units.

$$\mathrm{P(X)} = \sum_{\mathrm{Y}\in\mathrm{unit}_1} \mathrm{P_X(Y)}$$

$$\text{Isolation score}\ (\mathrm{unit}_1) = \frac{1}{|\mathrm{unit}_1|}\sum_{\mathrm{X}\in\mathrm{unit}_1} \mathrm{P(X)}$$

Where $|\mathrm{unit}_1|$ is the number of spike shapes in unit_1 cluster.

This quantifies a measure of similarity between each spike shape and all other spike shapes, normalized as probability of similarity in the two units, and summed over shapes within the same unit. This measure can be intuitively viewed as the average probability that an event that was classified as a spike belongs to the neuron it was classified to and not to the other neurons from the same electrode[58].

Error bars

All error bars represent standard error of the mean (SEM), unless specifically stated otherwise.

Effect size

Cohen's d was calculated as: $\mathrm{d} = \frac{\overline{X_1} - \overline{X_2}}{S_{pooled}}$ for two samples ; $\mathrm{d} = \frac{\overline{X_1} - \mu_0}{S_1}$ for one sample.

r_{rb} is the rank-biserial correlation coefficient.

Statistical tests

Statistical testing was done using t-test, ANOVA, Wilcoxon rank-sum, sign-rank test, permutation testing and Monte-Carlo p-value with resampling procedures. Significance level was set to $p<0.05$ unless otherwise mentioned. Correction for multiple comparison was done using Tukey correction for family wise error or using Benjamini–Hochberg (BH) correction for FDR.

All statistical tests were two sided, unless specifically stated otherwise.

In some statistical tests, data distributions were assumed to be normal and/or with equal variances but this was not formally tested.

Randomization

Pleasant and aversive trials were pseudorandomly presented to the monkeys but equalized in total number. Tones were randomly selected daily for pleasant and aversive CS. As randomization is irrelevant to triplets of neurons (all simultaneously recorded triplets were analyzed in this study), randomization was achieved by randomizing the control groups. Thus, shuffling lags were randomly chosen to the shuffled data sets and trials were randomly matched for the trial shuffle controls.

Blinding

Blinding is done as spike sorting is blind to the timing of the stimuli.

Data exclusions

Data was not excluded from the analysis.

Reporting Summary

Further information is available in the Nature Research Life Sciences Reporting Summary linked to this article.

Code availability

Custom code for behavioral and electrophysiological tests is available from the corresponding author upon reasonable request.

Data availability

All data supporting the findings of this study are available from the corresponding author upon reasonable request.

Methods-only References

51. Livneh, U. & Paz, R. Amygdala-prefrontal synchronization underlies resistance to extinction of aversive memories. *Neuron* **75**, 133-142 (2012).
52. Livneh, U. & Paz, R. Aversive-bias and stage-selectivity in neurons of the primate amygdala during acquisition, extinction, and overnight retention. *J Neurosci* **32**, 8598-8610 (2012).
53. Livneh, U. & Paz, R. An implicit measure of olfactory performance for non-human primates reveals aversive and pleasant odor conditioning. *J Neurosci Methods* **192**, 90-95 (2010).
54. Harrison, M.T., Amarasingham, A. & Truccolo, W. Spatiotemporal conditional inference and hypothesis tests for neural ensemble spiking precision. *Neural Comput* **27**, 104-150 (2015).
55. Nauhaus, I., Busse, L., Carandini, M. & Ringach, D.L. Stimulus contrast modulates functional connectivity in visual cortex. *Nat Neurosci* **12**, 70-76 (2009).
56. Panzeri, S., Senatore, R., Montemurro, M.A. & Petersen, R.S. Correcting for the sampling bias problem in spike train information measures. *J Neurophysiol* **98**, 1064-1072 (2007).
57. Maoz, O. & Schneidman, E. maxent_toolbox: Maximum Entropy Toolbox for MATLAB, version 1.0.2. 2017. URL: https://orimaoz.github.io/maxent_toolbox. (2017).
58. Joshua, M., Elias, S., Levine, O. & Bergman, H. Quantifying the isolation quality of extracellularly recorded action potentials. *J Neurosci Methods* **163**, 267-282 (2007).

Figure 1

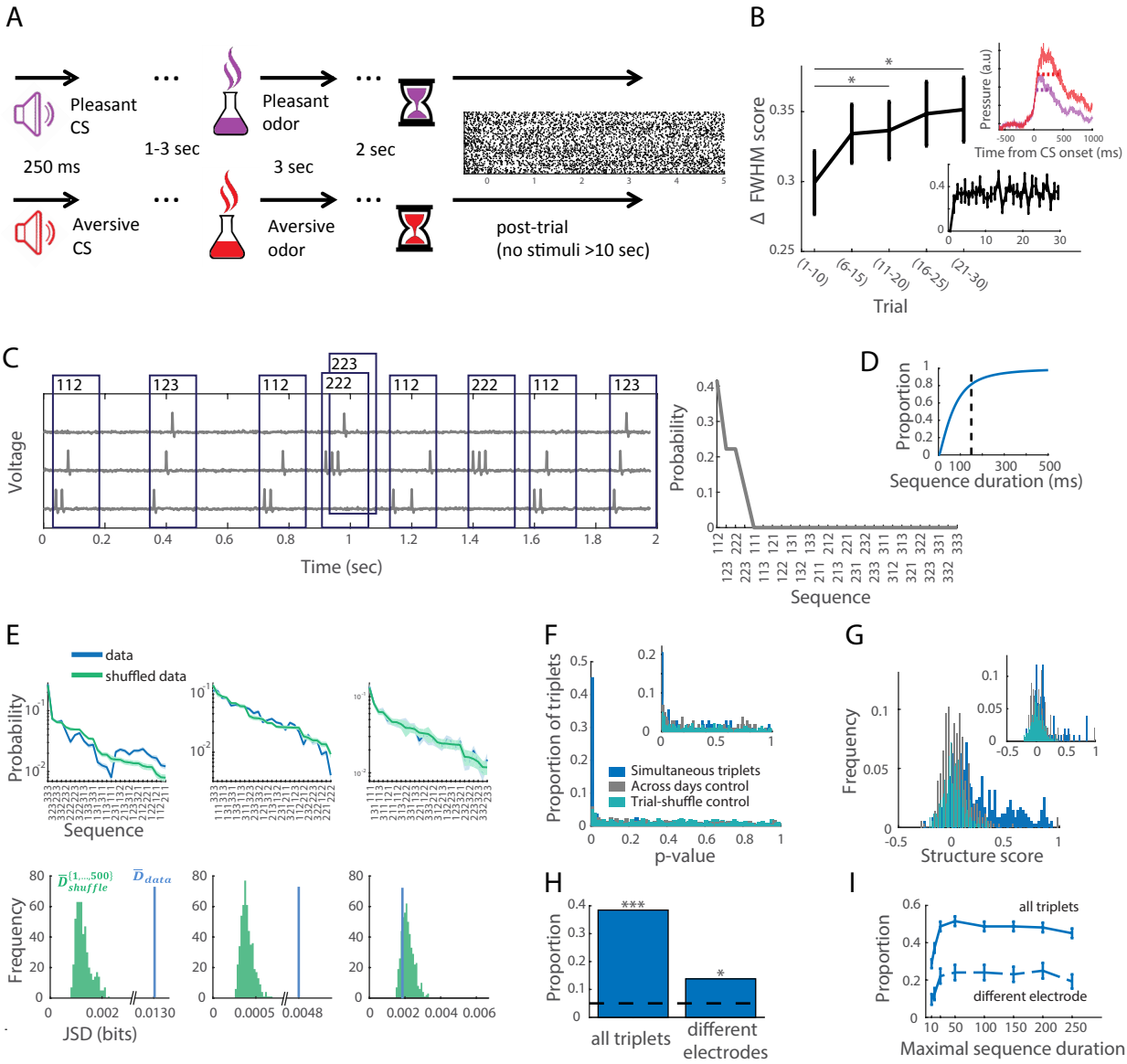

A
Pleasant CS
250 ms
1-3 sec
Pleasant odor
3 sec
2 sec
Aversive CS
Aversive odor
post-trial
(no stimuli >10 sec)
B
Δ FWHM score
Trial
(1-10)
(6-15)
(11-20)
(16-25)
(21-30)
Pressure (a.u)
Time from CS onset (ms)
C
Voltage
Time (sec)
Probability
Sequence
D
Proportion
Sequence duration (ms)
E
data
shuffled data
Probability
Sequence
Frequency
JSD (bits)
F
Proportion of triplets
p-value
Simultaneous triplets
Across days control
Trial-shuffle control
G
Frequency
Structure score
H
Proportion
all triplets
different electrodes
I
Proportion
all triplets
different electrode
Maximal sequence duration

Figure 2

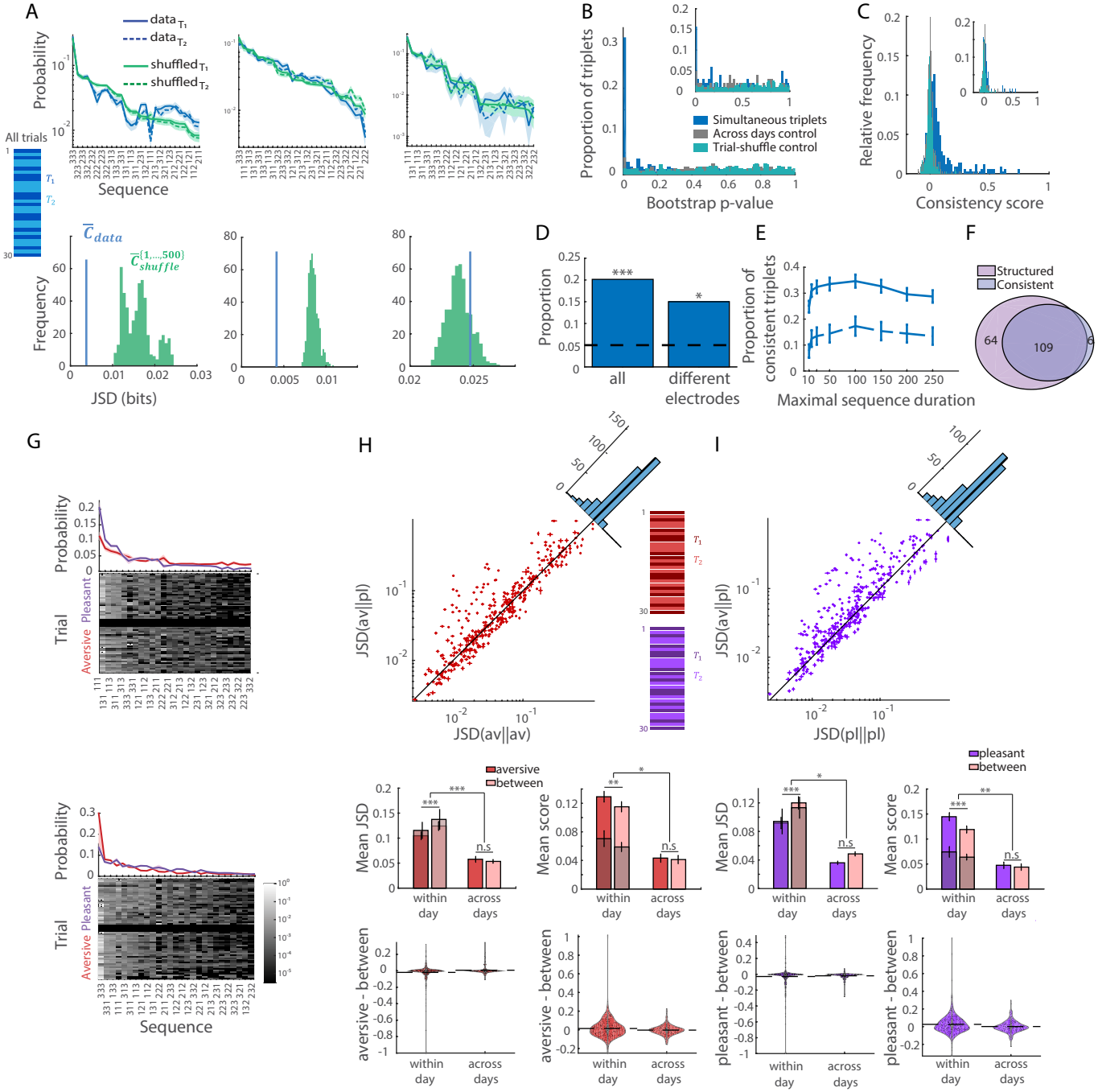

A
B
C
D
E
F
G
H
I
Probability
Sequence
All trials
Frequency
JSD (bits)
Proportion of triplets
Bootstrap p-value
Simultaneous triplets
Across days control
Trial-shuffle control
Relative frequency
Consistency score
Proportion
all
different electrodes
Proportion of consistent triplets
Maximal sequence duration
Structured
Consistent
64
109
Trial
Aversive
Pleasant
JSD(av||pl)
JSD(av||av)
JSD(pl||pl)
aversive
between
pleasant
Mean JSD
Mean score
within day
across days
n.s
aversive - between
pleasant - between

Figure 3

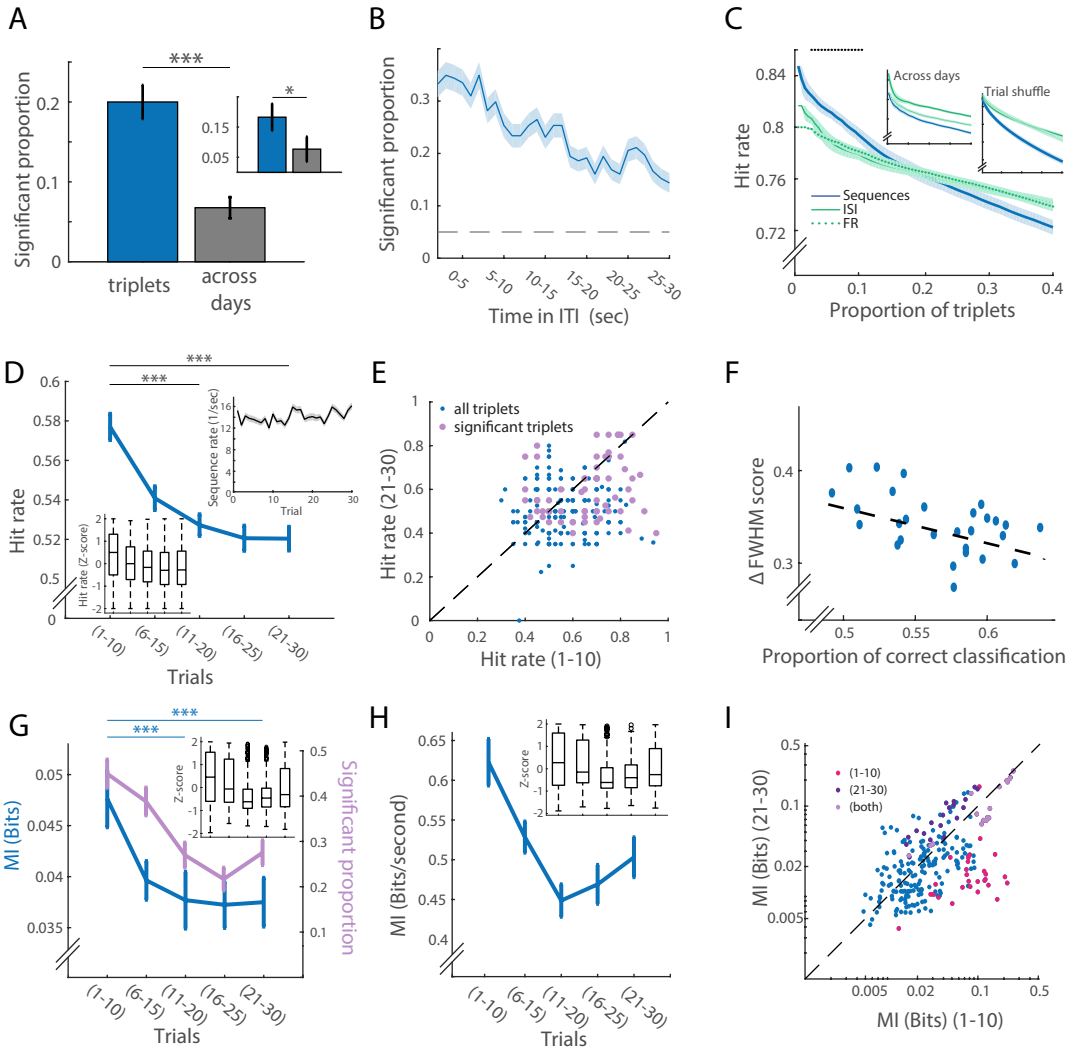

A
B
C
Significant proportion
triplets
across days
Time in ITI (sec)
Hit rate
Across days
Trial shuffle
Sequences
ISI
FR
Proportion of triplets
D
E
F
Hit rate
Trials
Sequence rate (1/sec)
Trial
Hit rate (Z-score)
all triplets
significant triplets
Hit rate (21-30)
Hit rate (1-10)
ΔFWHM score
Proportion of correct classification
G
H
I
MI (Bits)
Significant proportion
MI (Bits/second)
Z-score
(1-10)
(21-30)
(both)
MI (Bits) (21-30)
MI (Bits) (1-10)

Figure 4

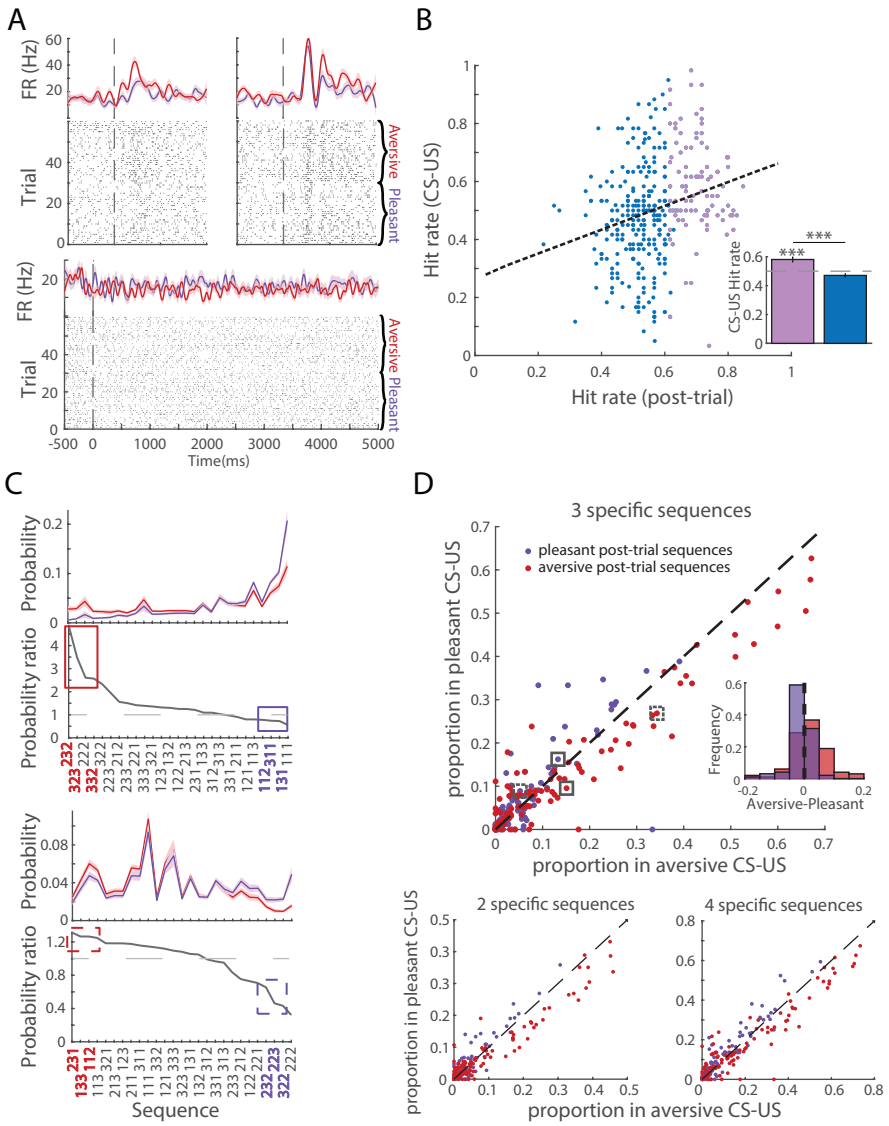
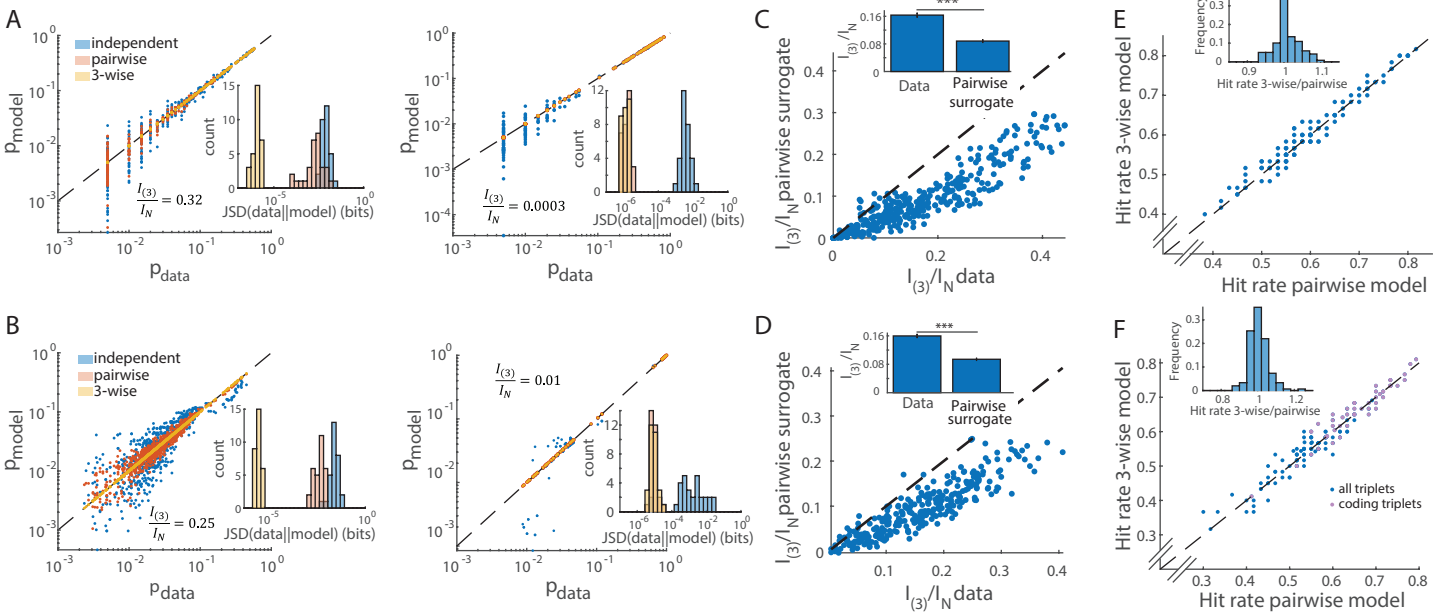
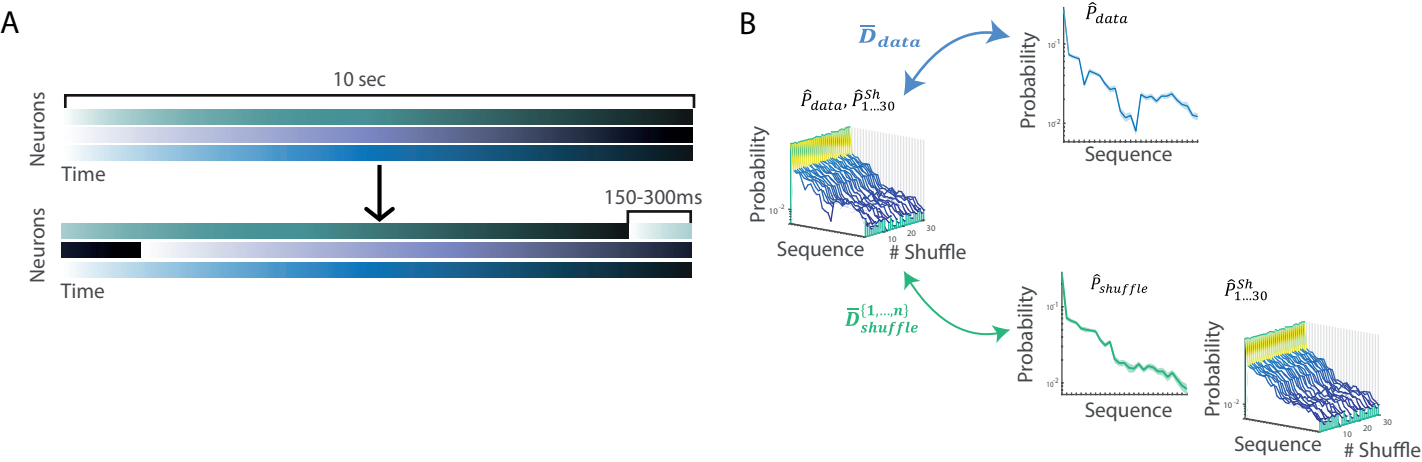
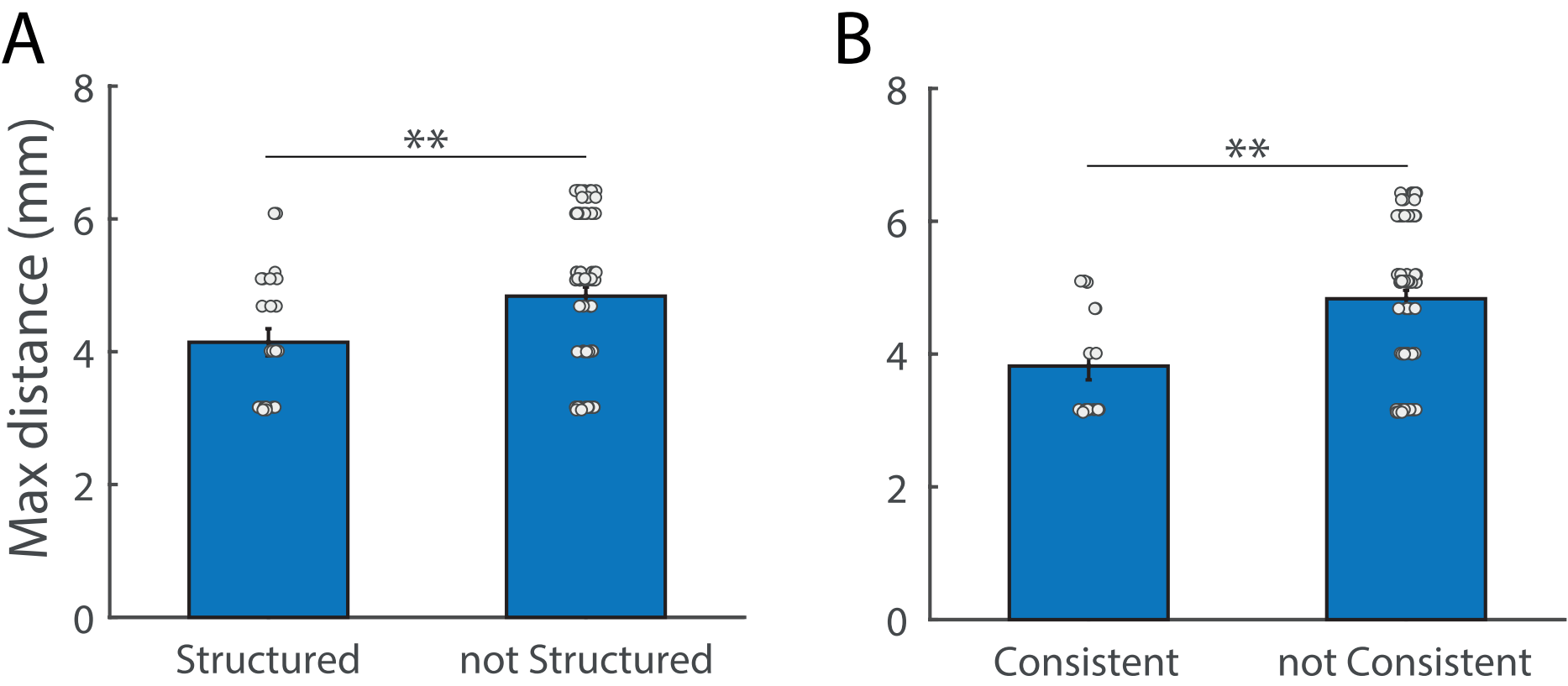

A
FR (Hz)
Trial
Aversive
Pleasant
Time(ms)
B
Hit rate (CS-US)
Hit rate (post-trial)
CS-US Hit rate
***
C
Probability
Probability ratio
Sequence
D
3 specific sequences
pleasant post-trial sequences
aversive post-trial sequences
proportion in pleasant CS-US
proportion in aversive CS-US
Frequency
Aversive-Pleasant
2 specific sequences
4 specific sequences

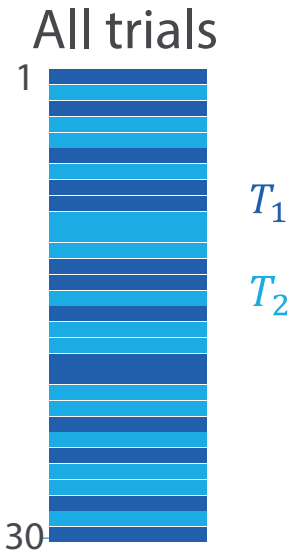
Figure 5


A
independent
pairwise
3-wise
$\frac{I_{(3)}}{I_N} = 0.32$
$\frac{I_{(3)}}{I_N} = 0.0003$
count
JSD(data||model) (bits)
p_{model}
p_{data}
B
$\frac{I_{(3)}}{I_N} = 0.25$
$\frac{I_{(3)}}{I_N} = 0.01$
C
D
***
Data
Pairwise surrogate
$I_{(3)}/I_N$ pairwise surrogate
$I_{(3)}/I_N$ data
E
F
Frequency
Hit rate 3-wise/pairwise
Hit rate 3-wise model
Hit rate pairwise model
all triplets
coding triplets


A
10 sec
Neurons
Time
150-300ms
Neurons
Time
B
$\overline{D}_{data}$
$\hat{P}_{data}$
Probability
Sequence
$\hat{P}_{data}, \hat{P}^{Sh}_{1...30}$
Probability
Sequence
Shuffle
$\overline{D}^{\{1,...,n\}}_{shuffle}$
$\hat{P}_{shuffle}$
Probability
Sequence
$\hat{P}^{Sh}_{1...30}$
Probability
Sequence
Shuffle


A
B
Max distance (mm)
8
6
4
2
0
**
**
Structured
not Structured
Consistent
not Consistent

Trials bisection example

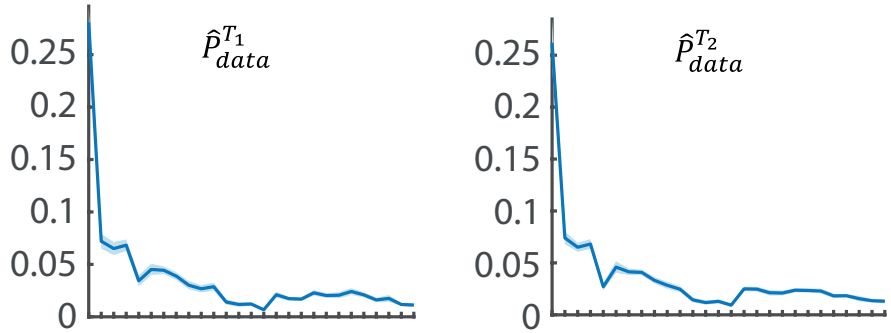


$\overline{C}_{data}$

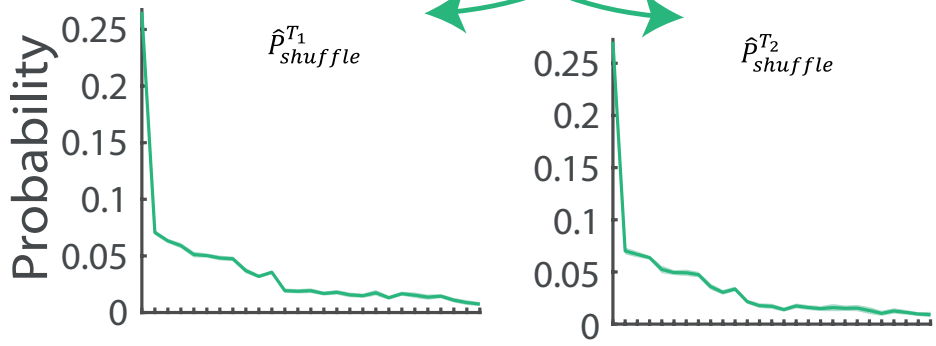
$\times 100$

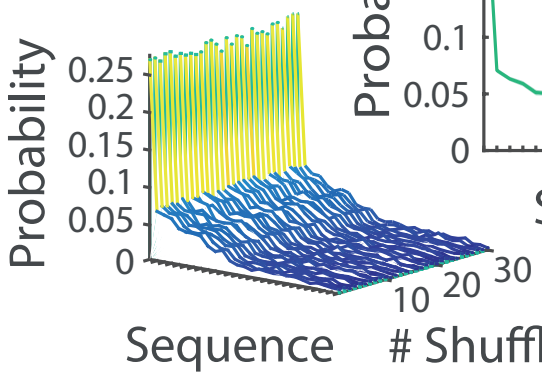
$T_1 : \{t \subset t_1, \ldots, t_{30}\}_{|T|=15}$

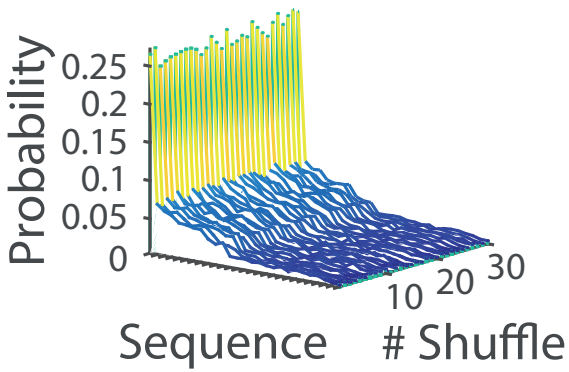
$T_2 = \overline{T}_1$

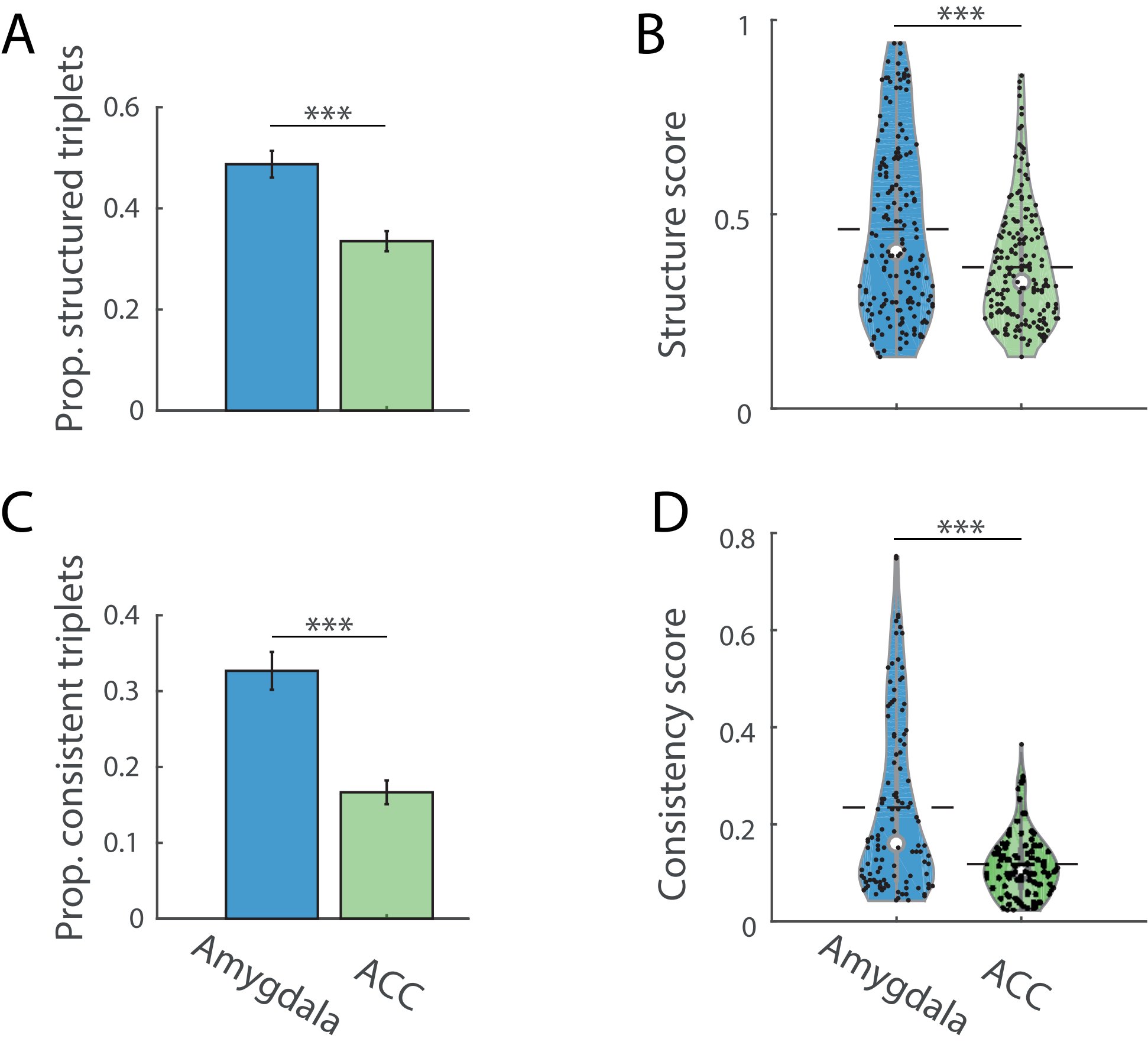


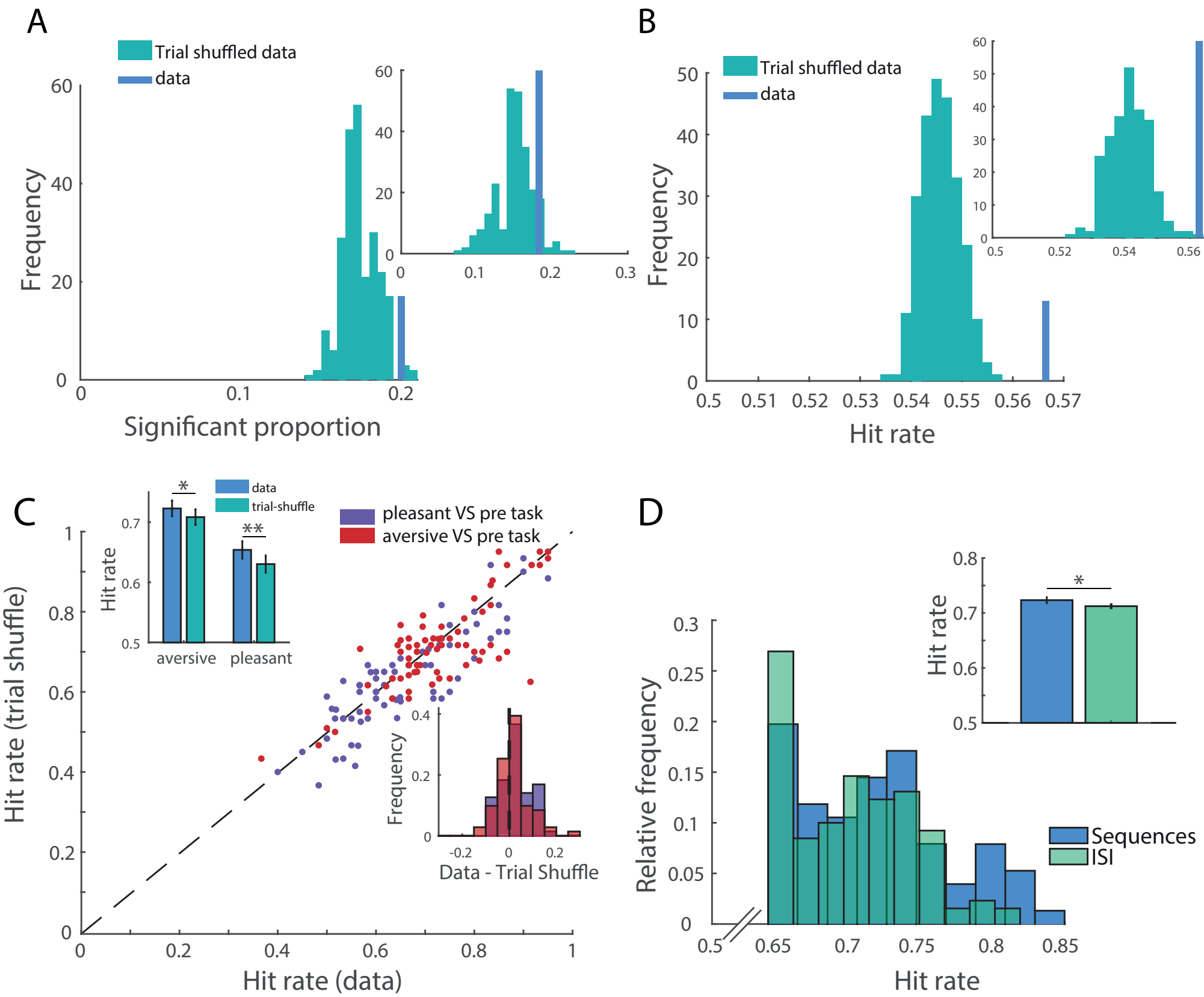
$\overline{C}^{\{1,\ldots,n\}}_{shuffle}$



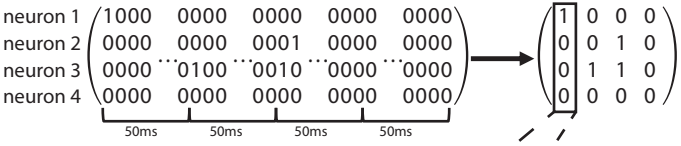





A
Prop. structured triplets
0.6
0.4
0.2
0
***
B
Structure score
1
0.5
0
***
C
Prop. consistent triplets
0.4
0.3
0.2
0.1
0
***
Amygdala
ACC
D
Consistency score
0.8
0.6
0.4
0.2
0
***
Amygdala
ACC


A
Trial shuffled data
data
Frequency
Significant proportion
B
Trial shuffled data
data
Frequency
Hit rate
C
data
trial-shuffle
Hit rate
aversive
pleasant
pleasant VS pre task
aversive VS pre task
Hit rate (trial shuffle)
Hit rate (data)
Frequency
Data - Trial Shuffle
D
Relative frequency
Hit rate
Sequences
ISI

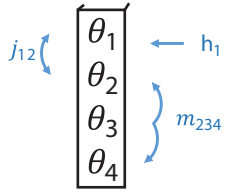
A



(independent)$< \theta_i > = \frac{1}{T}\sum_{t=1}^{T} \theta_i(t)$

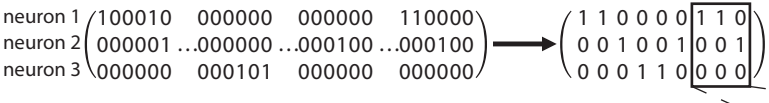
(pairwise) $< \theta_{ij} > = \frac{1}{T}\sum_{t=1}^{T} \theta_i(t)\theta_j(t)$

(triple-wise) $< \theta_{ijk} > = \frac{1}{T}\sum_{t=1}^{T} \theta_i(t)\theta_j(t)\theta_k(t)$



$P(x) = \frac{1}{Z}\exp(\sum_{i=1}^{N} h_i x_i + \sum_{i<j} j_{ij} x_i x_j + \sum_{i<j<k} m_{ijk} x_i x_j x_k)$

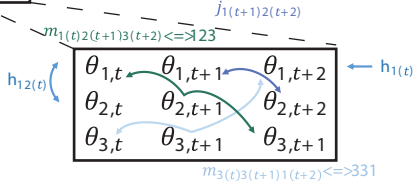
B



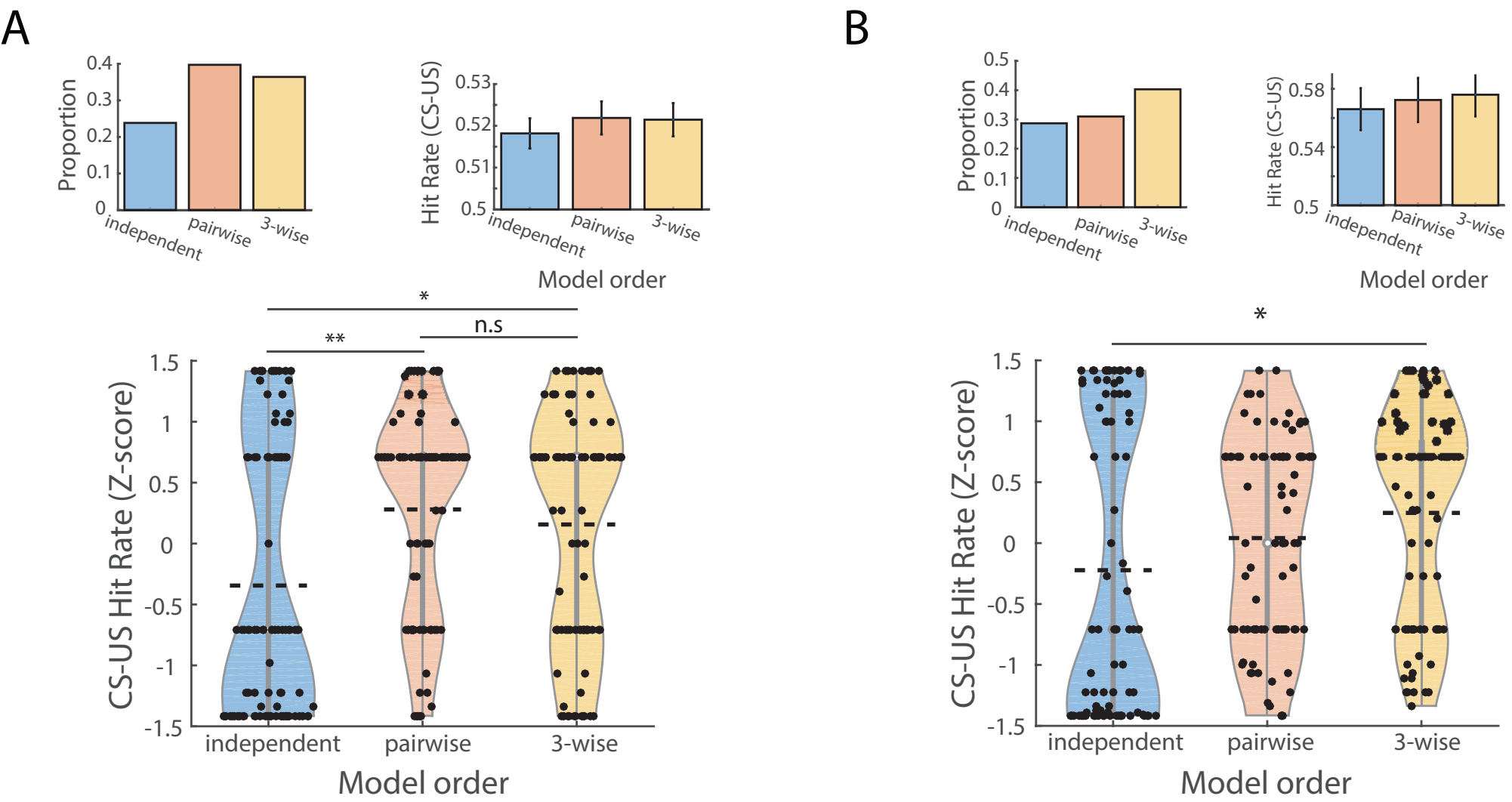
(independent)$< \theta_i > = \frac{1}{T}\sum_{t=1}^{T} \theta_i(t)$; $< \theta_{ij} > = \frac{1}{T}\sum_{t=1}^{T} \theta_i(t)\theta_j(t)$

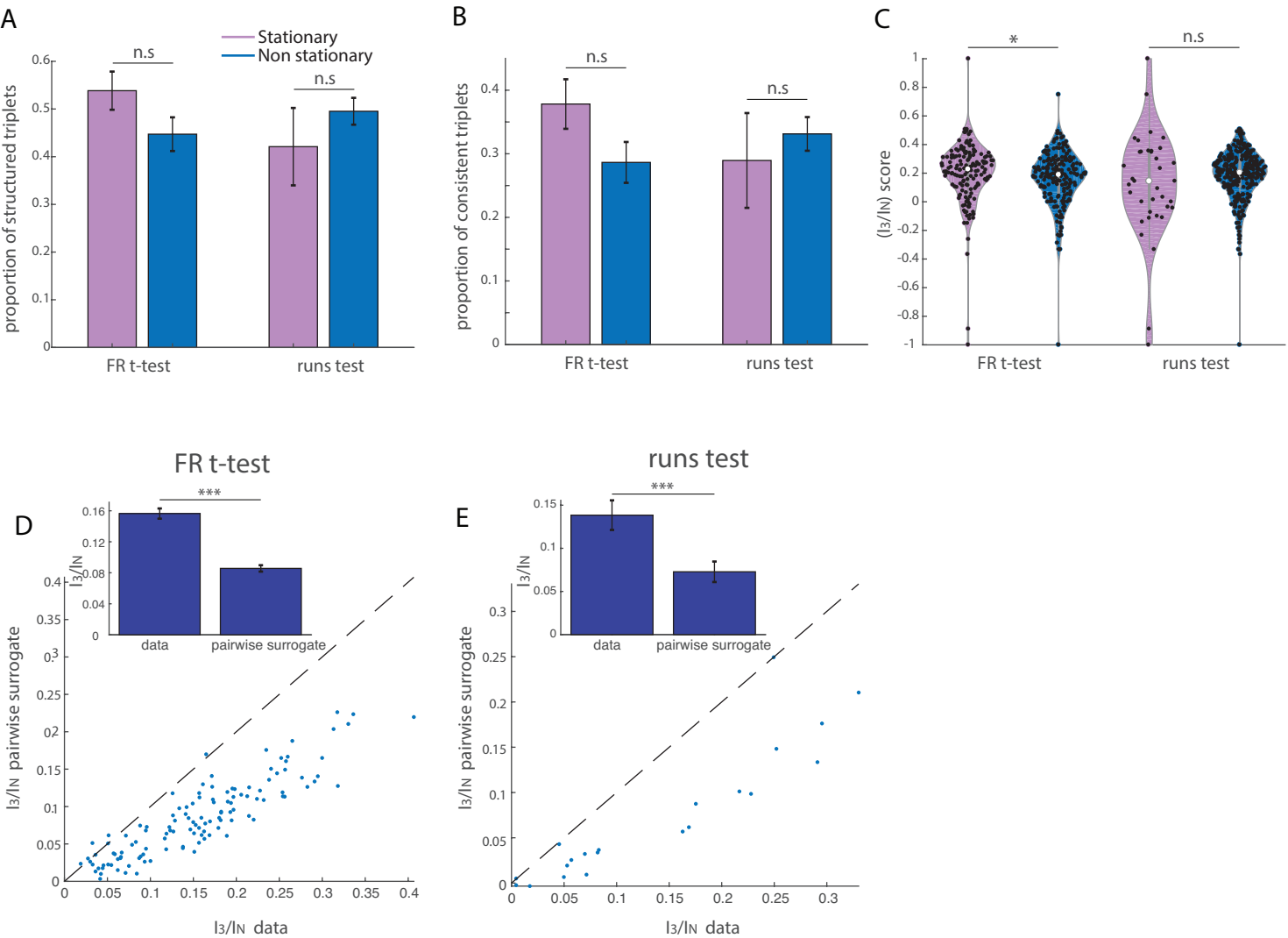
(pairwise) $< \theta_{i(t)\,j(t+1)} > = \frac{1}{T}\sum_{t=1}^{T} \theta_i(t)\theta_j(t+1)$

(triple-wise) $< \theta_{i(t)\,j(t+1)\,k(t+2)} > = \frac{1}{T}\sum_{t=1}^{T} \theta_i(t)\theta_j(t+1)\theta_k(t+2)$



$P(x_{T,T+1,T+2}) = \frac{1}{Z}\exp(\sum_{i=1}^{N}\sum_{t=T}^{T+2} h_{i(t)} x_i(t) + \sum_{i<j}\sum_{t=T}^{T+2} h_{ij(t)} x_i(t) x_j(t) + \sum_{i,j}\sum_{t=T}^{T+1} j_{i(t)j(t+1)} x_i(t) x_j(t+1) + \sum_{i,j,k} m_{i(t)j(t+1)k(t+2)} x_{i,t} x_{j,t+1} x_{k,t+2})$


A
Proportion
0.4
0.3
0.2
0.1
0
independent
pairwise
3-wise
Hit Rate (CS-US)
0.53
0.52
0.51
0.5
independent
pairwise
3-wise
Model order
*
n.s
**
CS-US Hit Rate (Z-score)
1.5
1
0.5
0
-0.5
-1
-1.5
independent
pairwise
3-wise
Model order
B
Proportion
0.5
0.4
0.3
0.2
0.1
0
independent
pairwise
3-wise
Hit Rate (CS-US)
0.58
0.54
0.5
independent
pairwise
3-wise
Model order
*
CS-US Hit Rate (Z-score)
1.5
1
0.5
0
-0.5
-1
-1.5
independent
pairwise
3-wise
Model order


A
Stationary
Non stationary
n.s
n.s
proportion of structured triplets
0.6
0.5
0.4
0.3
0.2
0.1
0
FR t-test
runs test
B
n.s
n.s
proportion of consistent triplets
0.4
0.3
0.2
0.1
0
FR t-test
runs test
C
*
n.s
(I3/IN) score
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
FR t-test
runs test
FR t-test
D
***
I3/IN
0.16
0.12
0.08
0.04
0
data
pairwise surrogate
I3/IN pairwise surrogate
I3/IN data
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
runs test
E
***
I3/IN
0.15
0.1
0.05
0
data
pairwise surrogate
I3/IN pairwise surrogate
I3/IN data
0.3
0.25
0.2
0.15
0.1
0.05
0
0.05
0.1
0.15
0.2
0.25
0.3