eLife

Reviewed Preprint
v1 • March 11, 2026
Not revised

Reviewed Preprint
v2 • May 11, 2026
Revised by authors

For correspondence:
antares0715@gmail.com
sukwoo12@snu.ac.kr

# K.S. and D.Y. contributed equally to this manuscript.

Author contributions: K.S., J.L., and S.C. designed research; and K.S., and D.Y. performed research; K.S., D.Y., J.L., and S.C. analyzed data; and K.S., D.Y., J.L., and S.C. wrote the paper.

Competing interests: No competing interests declared

Funding: See page 33

Reviewing editor: Tatyana O Sharpee, Salk Institute for Biological Studies, United States

Computational mechanisms for temporal integration in the anterior claustrum

Kuenbae Sohn#, Donghyeon Yoon#, Junghwa Lee, Sukwoo Choi
School of Biological Sciences, College of Natural Sciences, Seoul National University, Seoul, Republic of Korea

eLife Assessment

This work provides an **important** modeling-based framework for understanding the processes of temporal integration in the claustrum. These mechanisms could support a broader range of integrative brain function. The manuscript presents **solid** evidence for how claustrum may integrate temporal disparate signals via a novel computational phenomenon with neural dynamics evolving along neural trajectories as opposed to settling into fixed-point attractor states.

https://doi.org/10.7554/eLife.109539.2.sa3

Abstract

The claustrum, with its extensive reciprocal connections to nearly all cortical regions, has long been hypothesized as a key hub for integrating diverse cognitive, sensory and motor information. However, despite its anatomical connectivity, whether and how it functionally integrates different inputs to generate coherent representations has remained unclear. Here, we developed a recurrent neural network (RNN) trained via supervised learning on behavioral metrics of delayed escape—a behavioral paradigm that requires integration of temporally separated task-relevant signals. A subset of RNN neurons exhibited dynamics similar to those of anterior claustral neurons during this behavior. These neurons formed a recurrent cluster, a structure supported by in vitro stimulation experiments in claustral brain slices. We analyzed the computational properties of this claustrum-like cluster via dimensionality reduction of population activity. The network showed nonlinear integration of temporally distributed inputs and increased synergistic information. Rather than settling into attractors, integrated information was dynamically encoded along continuously evolving neural trajectories. Notably, similar trajectory patterns associated with dynamic integration were observed in claustral recordings, suggesting the model's biological plausibility. We propose that the anterior claustrum dynamically integrates task-relevant input signals over time and broadcasts the evolving representation to downstream brain regions capable of reading and interpreting it in a context-dependent manner.

Introduction

The claustrum is a thin, elongated sheet of gray matter that maintains extensive reciprocal connections with nearly all cortical areas and several subcortical regions. This dense interconnectivity has led to the longstanding hypothesis that the claustrum serves as a hub for integrating internal and external signals into a unified percept, potentially supporting conscious awareness and higher-order cognition (1). However, direct experimental evidence for this hypothesis has remained scarce. One recent study reported that individual claustral neurons can receive convergent input from multiple cortical areas, which is critical for a behavioral task requiring multimodal sensory inputs (2). In contrast, numerous studies, including those from the Mathur and Citri groups, have implicated the claustrum in diverse cognitive processes such as attention, salience detection, and cognitive control (3–14). This apparent discrepancy may reflect

differences in the specific claustral subregions examined. Although the claustrum is less differentiated in rodents than in primates, important distinctions exist even among rodent species (15). In mice, the anterior boundary of the claustrum typically does not extend beyond the rostral end of the striatum (16). In contrast, in rats, the claustrum extends further rostrally beneath the forceps minor of the corpus callosum (fmi), where claustrum-specific gene expression has been reported (17, 18).

Only a few previous studies have directly investigated this rostral segment of the rat claustrum (rostral-to-striatum, rsCla) (18–22). However, recent findings suggest that neuronal activity in this region plays a critical role in behavioral tasks in which two different kinds of information must be integrated (17, 23), indicating a potentially distinct computational role for the rsCla. Among these tasks, the newly developed delayed escape task provides a behavioral paradigm in which rats must rapidly escape to a neutral zone following a fear-inducing conditioned stimulus (CS), after a delay that separates the CS from the opening of an outlet. Crucially, the task requires no prior training and instead depends on the flexible integration of two temporally separated events that become jointly relevant for guiding escape behavior. Neural recordings from this task revealed that CS-related information was maintained through persistent activity in rsCla neurons, suggesting that these neurons may integrate the sustained internal threat signal with the subsequent outlet-opening signal to guide future escape behavior.

To investigate how the anterior claustrum integrates information at the population level, we employed a delayed escape task in Long-Evans rats. Because the task could only be performed once per animal, and only a small number of single units were recorded per subject, direct population-level analysis was severely limited. To overcome these constraints, we trained a recurrent neural network (RNN) model solely on behavioral outputs—a strategy previously shown to yield biologically plausible dynamics (24–30). We then compared the network's emergent population activity patterns with in vivo claustral recordings, and used the model to explore circuit-level mechanisms of integrating temporally separated inputs.

Results

RNNs trained on the Delayed Escape Task show claustrum-like dynamical patterns

The delayed escape task allows only a single test trial per animal, which means that opportunities to obtain neurophysiological data are extremely limited. To overcome this limitation, we used recurrent neural network (RNN) simulations, a strategy that has been widely adopted in recent neuroscience research (26, 27, 31). In the delayed escape task, when a conditioned stimulus (CS) previously associated with electric shock is presented for the first time in a novel environment, animals infer that a value-neutral alternative space is likely to be safer and therefore exhibit faster escape behavior (Fig. 1A, This figure is adapted from Fig. 1A of Han et al. (17)). Importantly, the passage to this neutral space opens after a delay following CS offset, making it essential for CS-related information to be maintained beyond the actual presence of the CS. As a control group, animals were placed in the same environment during the CS period but did not receive the CS. In previous experiments, animals that received the CS escaped to the neutral space significantly faster than controls (17). To model this behavior, we built upon the continuous-time RNN framework developed by Ehrlich et al. (2021) (24) and Ehrlich & Murray (2022) (25) (Fig. 1B).

The trained RNN reproduced the escape latency pattern reported by Han et al. (2024) (17): the CS + door-opening condition escaped faster than the control door-opening only group (Fig. 1C), reflecting the in vivo behavioral results. Furthermore, it showed similarities to the claustral network dynamics observed in that study (Fig. 1D–L, Fig. 1D, E, F and L(left) are adapted from Figs. 4B, 4D, 4E and S4E of Han et al. (17)). In the previous biological study, the responses of individual claustral units representing each behavioral epoch were subjected to dimensionality reduction and clustering (Fig. 1D). One of these clusters showed persistent firing during and

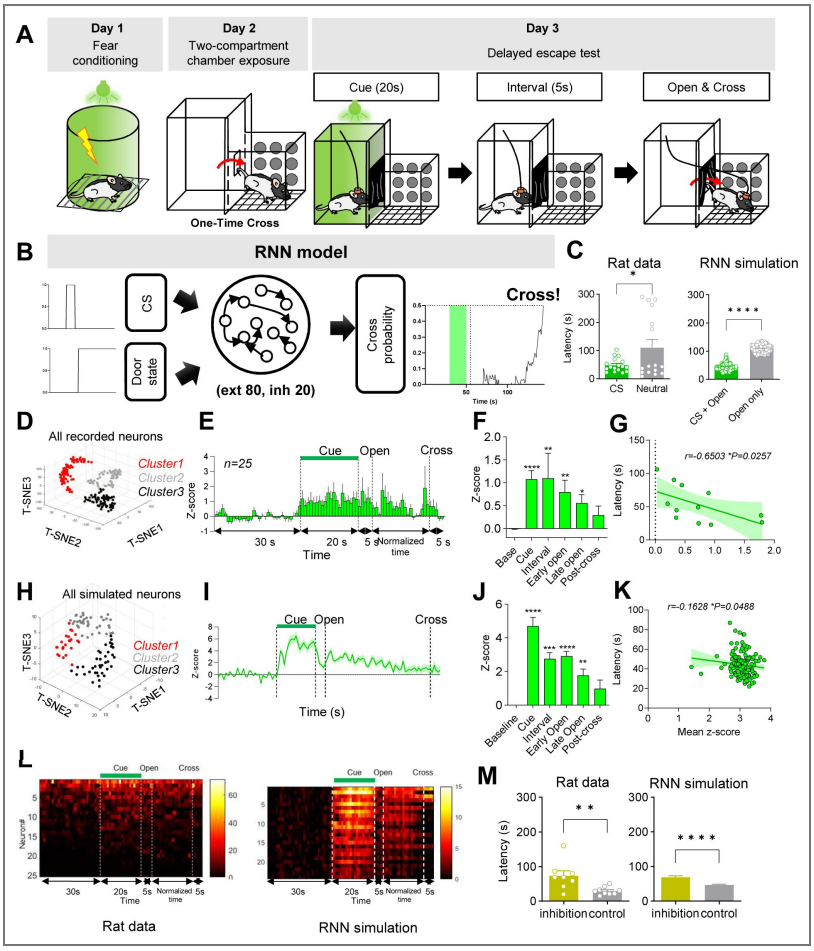

Figure 1. RNN simulation capable of performing the Delayed Escape Test.

(A) Schematic diagram of the Delayed Escape Task. On Day 1, animals were conditioned with a light cue paired with an electric footshock. On Day 2, they were adapted to the test context with only a single crossing permitted. On Day 3, animals were placed in one compartment of the test context and presented with a 20-s CS, followed by a 5-s delay, after which the outlet door was opened to allow escape. The test was conducted only once. (B) Schematic diagram of the task structure that the RNN model was designed to simulate. (C) Graphs of crossing latency, measured as the interval from door opening to crossing in rats and as derived from RNN simulations. (*p = 0.0335, unpaired *t*-test; RNN, ****p < 0.0001, Mann–Whitney). (D) Clustering of all recorded claustral neurons in rats (n = 203). (E) Mean Z-scored firing rates of non-exploratory Cluster 1 neurons in the CS group (n = 25, rat data). Non-exploratory Cluster 1 denotes the cluster showing increased persistent activity both during CS presentation and thereafter. The CS group corresponds to the CS + door-opening condition in the RNN model. (F) Average Z-score for each behavioral period shown in E. Bars: Friedman test, ****p < 0.0001; Dunn's multiple comparisons —Base vs Cue, ****p < 0.0001; Base vs Interval, **p = 0.0085; Base vs Early Open, **p = 0.0075; Base vs Late Open, *p = 0.0455; Base vs Post-cross, p > 0.9999. (G) Correlation between crossing latency and Cluster 1 activity in the CS group (n = 12; Spearman r = 0.6503, *p = 0.0257). (H) Clustering for simulated neurons (n = 100 neurons). (I) Mean Z-scored firing rates of Cluster 1 neurons in the CS + door-opening condition (n = 24 neurons, RNN) across all simulated trials. (J) Average Z-score for each behavioral period shown in I. Bars: Friedman ****p < 0.0001; Dunn's—Base vs Cue, ****p < 0.0001; Base vs Interval, ***p = 0.0001; Base vs Early Open, ****p < 0.0001; Base vs Late Open, **p = 0.0034; Base vs Post-cross, p = 0.7134. (K) Correlation between crossing latency and Cluster 1 activity in the CS + door-opening condition (147 trials; Spearman r = –0.1628, *p = 0.0488). (L) Heatmaps of single-neuron activity corresponding to panels E and I. Left: rat data (firing rate, Hz); Right: RNN data (Z-score). Neurons are ordered by overall activity. (M) Crossing latency in rats with anterior claustrum inhibition during the 5-s period between CS offset and door opening, compared with control virus–expressing animals (left, **p = 0.0079, Mann–Whitney). Corresponding results in the RNN are shown for selective inhibition of Cluster 1 neurons during the same 5-s window simulation (right, ****p < 0.0001, Mann–Whitney).

after CS presentation, and the magnitude of this increase was inversely correlated with escape latency (Fig. 1E–G). From this, we concluded that this cluster maintained CS signals until the door to the escape path opened, thereby contributing to delayed escape.

To determine whether this cluster was reproduced in the trained RNN, unit responses in the RNN representing each behavioral epoch were also subjected to dimensionality reduction and clustering. One of the clusters displayed the same persistent activation during and after CS presentation as observed in the in vivo claustrum (Fig. 1H–J). As in the biological in vivo data, stronger activation in this model cluster predicted shorter escape latency (Fig. 1K). Single-unit responses were also similar between real and simulated data: the persistence after the CS did not arise from a single unit firing continuously but rather from many units intermittently increasing their firing, which overlapped to form a persistent signal at the population level (Fig. 1L).

Even when no CS was presented and only the door opening cue was given (the door-opening condition), the population activity pattern generated by the RNN showed a pattern of gradually increasing after the door opened and then decreasing as the crossing approached (Fig. S1A). The other two model clusters exhibited different patterns: Cluster 2 was generally suppressed regardless of CS presence, and Cluster 3 gradually ramped up from the door opening cue to just before the crossing, also regardless of CS presence (Fig. S1).

Next, we quantified the contributions of the three clusters to the RNN output when the CS was presented. The claustrum-like Cluster 1 contributed very little to the output, whereas Cluster 3 was dominant in influencing cross behavior. This suggests that Cluster 1 may not directly drive the output but instead processes the CS and door opening signals and broadcasts this information to other clusters (Fig. S2). In our previous study (17), optogenetic inhibition of the anterior claustrum during the delay period between CS offset and door opening significantly increased escape latency (Fig. 1M, Fig. S3A). Introducing inhibitory input to suppress RNN Cluster 1 during the same period produced a similar increase in latency (Fig. 1M), whereas inhibiting the other clusters had no effect (Fig. S3B). Furthermore, when the interval between CS and door opening was extended from 5 s to 180 s in vivo (17), the latency difference between the CS and neutral groups disappeared. Indeed, the RNN also reproduced this result under the same extended interval (Fig. S3C).

Taken together, the claustrum-like Cluster 1 in the RNN reproduced the firing pattern and behavioral correlation of the persistent activity cluster observed in vivo and resembled several key characteristics of actual claustral neurons in almost every respect, including latency increases following inhibition. This trained RNN model therefore enables deeper analyses under experimental conditions that are otherwise difficult to implement.

Claustral neurons exhibit recurrent connectivity similar to that of the RNN

Because a recurrent neural network (RNN), by definition, contains recurrent connections among its units, finding a claustrum-like cluster whose properties mirror those of the real claustrum suggests that biological claustral neurons may likewise be interconnected through recurrent circuitry. As expected, the RNN exhibited strong recurrent weights within Cluster 1 (Fig. 2A). If a similar architecture exists in vivo, the first prediction is that excitatory connectivity should be detectable between claustral neurons.

To test this, we sparsely expressed ChrimsonR in a subset of claustral cells and prepared sagittal brain slices for patch-clamp recordings to measure synaptic currents (Fig. 2B–C). We activated ChrimsonR-expressing presynaptic neurons with 617 nm light, while patching ChrimsonR-negative neurons to isolate synaptic responses. Patching ChrimsonR-positive cells would have risked contamination by direct photocurrents in addition to synaptic responses.

Wide-field illumination induced excitatory postsynaptic currents (EPSCs) that were abolished by NBQX + D-AP5, confirming that they were mediated by glutamatergic transmission (Fig. 2D–E). One remaining concern was that these EPSCs might have originated from ChrimsonR-expressing axons projecting into the claustrum from outside regions. To rule this out, we employed high-

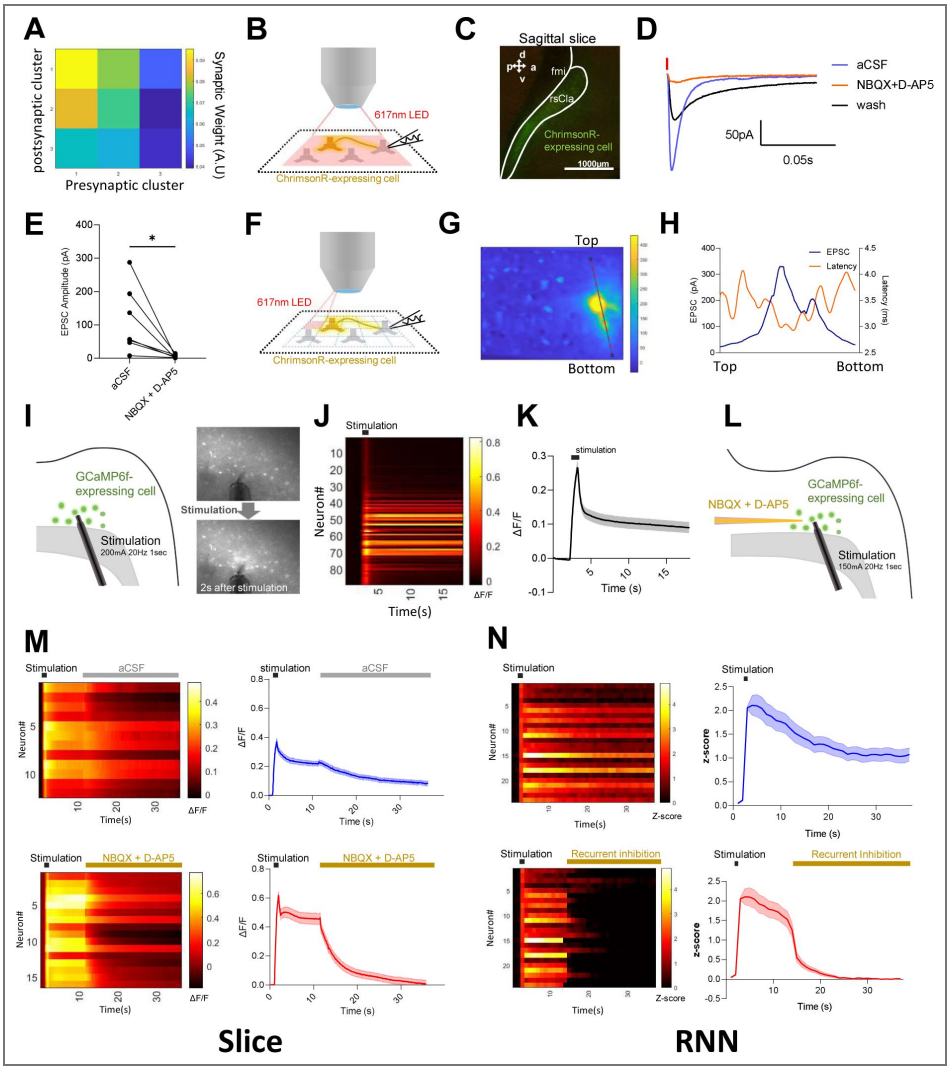

Figure 2. Comparison of recurrent connectivity in the claustrum and in the RNN simulation

(A) Mean absolute inter-cluster synaptic weight in the RNN. (B) Schematic representation of whole-cell recordings of ChrimsonR-non-expressing neurons during optogenetic stimulation using LEDs to stimulate ChrimsonR-expressing neurons across the entire optical field. (C) Confocal image showing the expression pattern of the ChrimsonR virus in sagittal claustrum slices. (D) Representative EPSCs evoked by 2ms LED pulses before and after pharmacological treatments (red bar indicates stimulation). (E) Pooled data of EPSC amplitudes (n=7) with statistical analysis using the Wilcoxon matched-pairs signed rank test (*p=0.0156). (F) Schematic depiction of local stimulation using a digital mirror device (DMD). (G) Optical stimulation of each divided part produces variable amplitudes of EPSCs in whole-cell patched neurons expressing no ChrimsonR. Heatmap showing EPSC amplitudes upon stimulation of the designated part of optical field. Please note that when stimulating closer to the recorded neuron, larger amplitude EPSCs were induced. (H) The graph displays EPSC amplitudes and latencies in the region marked with dashed lines on the heatmap shown in G. (I) Left: schematic representation showing an experimental configuration in which brief electrical stimulation produces a persistently enhanced activity in rsCla slices. Right: representative images of GCaMP6f fluorescence changes immediately before and 2 seconds after stimulation. (J) Heatmap of fluorescence changes for individual puncta in the slice shown in I. (K) Population-averaged fluorescence trace of all puncta in J (mean ± SEM). (L) Schematic representation showing an experimental configuration in which effects of the blockers for AMPA/NMDA receptors on the persistent response were examined. (M) Calcium imaging before and after an aCSF puff (top panels) and an NBQX + D-AP5 puff (bottom panels): heatmaps (left) and population-averaged traces (right). (N) RNN analogue: heatmaps (left) and averaged Z-scores (right) for Cluster 1 neurons following brief excitation (top panels) and after the addition of recurrent inhibition (bottom panels).

resolution optical mapping. The wide field was divided into a 30 × 30 grid (900 pixels), and each pixel was stimulated for 2ms (Fig. 2F). EPSCs could be elicited not only near the recorded soma but also from distant pixels, indicating the presence of local excitatory-to-excitatory connectivity within the claustrum itself (Fig. 2G–H). These findings are consistent with the previous study by Orman et al. (2015) (32) and Shelton et al. (2025) (2) but contrast with reports that the posterior claustrum exhibits little excitatory-to-excitatory coupling (33). In particular, Orman et al. (2015) reported that persistent increases in claustral activity were observed only when brain slices were prepared at a specific slicing angle.

A more stringent prediction of recurrent wiring is that a brief, strong stimulus should induce sustained neuronal activity. To test this, we expressed the calcium sensor GCaMP6f in the anterior claustrum and delivered a 1 s train of electrical stimulation (20 Hz, 200 μA) in horizontal slices (Fig. 2I–K, Fig. S4C). This stimulation evoked calcium signals that remained elevated for over 10 s. Consistently, patch-clamp recordings from GCaMP6f-positive cells confirmed a prolonged increase in firing rate over a similar period (Fig. S4B). Notably, coronal slices failed to show such persistence, suggesting that horizontal slicing better preserves recurrent excitatory circuits. To determine whether this persistent activity depends on synaptic reverberation, we locally pressure-injected NBQX + D-AP5 at the stimulation site during the sustained phase of the calcium signal. As predicted, the persistent response collapsed rapidly (Fig. 2L–M, Fig. S4D–F), indicating that excitatory synaptic transmission within the claustrum is essential for generating the sustained activity.

The RNN reproduced a similar phenomenon (Fig. 2N, Fig. S4F). Isolating Cluster 1 from the trained network, we reduced all inhibitory weights by 60%. When a 1 s CS-strength input was delivered, the cluster exhibited >10 s of persistent activity, matching the slice data. Reducing inhibitory weights facilitates the persistent activity, which may reflect the reduced inhibitory connectivity often present in horizontal slice preparations. Next, reducing 60% of excitatory weights for only 10% of Cluster 1 units—mimicking the spatially restricted application of NBQX and D-AP5—largely abolished the persistence (Fig. 2N, Fig. S4D–F). Together, these slice and RNN experiments suggest that the anterior claustrum contains recurrent excitatory connections capable of supporting prolonged activity.

PCA-based trajectory analysis characterizes RNN dynamics

Having established that claustral neurons exhibit recurrent connectivity capable of sustaining activity, we next examined how such dynamics manifest at the population level in the RNN. Based on the results so far, the claustrum-like cluster in the RNN resembles the cluster showing persistent activity after CS presentation in the anterior claustrum. To visualize the population activity dynamics of the RNN clusters—particularly the claustrum-like cluster— we applied principal component analysis (PCA) to z-scored firing rates from three conditions: CS + door-opening, door-opening only, and CS only (Fig. 3A–E, Fig. S5–6). At each time point, the ensemble firing rates across all neurons in each cluster were treated as a population activity vector and projected into a three-dimensional space defined by the top 3 principal components. Because cross-latency (the time from door opening to actual crossing) varied across trials, the interval from door opening to crossing was time-normalized within each trial. For the CS only condition, in which cross-latency was undefined, we used the mean cross-latency from the CS + door-opening condition to define a notional crossing time.

In the CS + door-opening condition, the mean trajectory moved away from the start region after CS presentation and progressed along a curved path. Notably, a short-loop pattern appeared immediately after the door opened—5 seconds after CS offset (Fig. 3A). In contrast, in the door-opening only group, the trajectory remained near the start region until the door opened and then followed a long-curved path toward the crossing point (Fig. 3B). As expected, the CS only group exhibited a trajectory similar to that of the CS + door-opening condition until 5 seconds after CS offset, after which it returned toward the start region (Fig. 3C). Interestingly, in two

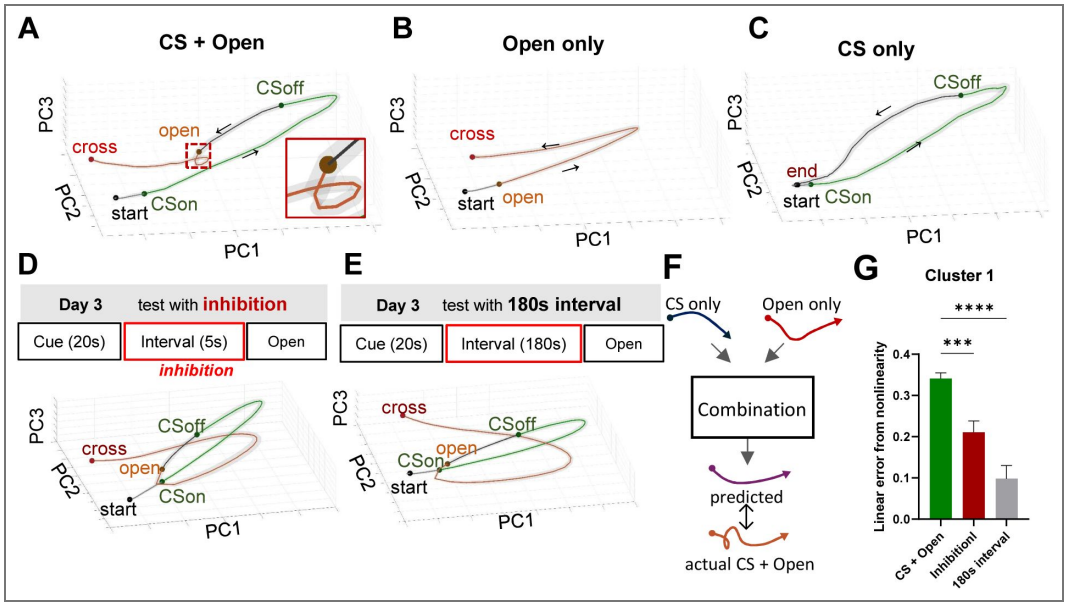

Figure 3. RNN PCA Trajectories.

(A-C) Time-normalized, trial-averaged PCA trajectories for each simulation condition. All trajectories begin at the Start event (black circle), proceed through CS onset and offset (CSon & CSoff; green circle) in the CS-containing and inhibition simulations, then through the Open event (orange circle) in the Open-containing and inhibition simulations, and terminate either at the Cross event (red circle) for crossing conditions or at the End event (black circle) for the CS-only condition. (A) PCA trajectory for the CS+door-opening condition. The mean latency from door opening to crossing was 44.16 ± 0.9583 s. (B) PCA trajectory for the door-opening only condition. (C) PCA trajectory for the CS only condition. (D) Simulated inhibition applied during a 5-s interval between CSoff and door-opening. Top: schematic of the simulated task. Bottom: time-normalized, trial-averaged PCA trajectory. (E) Simulation with a 180-s interval between CSoff and door-opening (no inhibition applied). Top: schematic of the simulated task. Bottom: time-normalized, trial-averaged PCA trajectory. (F) Schematic of the trajectory-combination model: the predicted CS + door-opening trajectory (purple line)—obtained by model of which input are the CS only and door-opening only trajectories—is plotted against the actual CS + door-opening trajectory (orange line). (G) Model-fit comparison for cluster 1: difference in residual sum of squares (ΔRSS) between the linear regression and MLP models, normalized to the mean RSS of the linear model. Bar colors denote condition (CS + door-opening = green, inhibition = red, 180-s interval = gray). One-way ANOVA with Holm–Sidak's multiple comparisons test: CS + door-opening vs. inhibition, **p = 0.0063; CS + Open vs. 180-s interval, ****p < 0.0001. Bars show mean ± SEM. Solid lines represent mean PCA trajectories; shaded areas denote SEM.

manipulations known to delay crossing latency—(1) inhibition of neural activity during the 5-second interval between CS offset and door opening, and (2) extension of this interval to 180 seconds—the short loop observed in the CS + door-opening group was not observed (Fig. 3D, E).

Sudden changes in trajectory shape, such as the short loop observed in the CS + door-opening condition, have been reported in previous studies of other brain regions to be associated with specific brain functions (34). Therefore, based on the assumption that the persisting CS signal and the door-opening signal combine to drive integration, we examined their interaction during this period. To test this, we assumed that the post-CS segment of the CS only trajectory (starting 5 seconds after CS offset) and the post-door-opening segment of the door-opening only trajectory each represent the respective inputs to the integration process. We then tested whether their combination could explain the post-door-opening trajectory in the CS + door-opening condition (Fig. 3F-G, Fig S6). A linear combination model failed to reconstruct the early time bins immediately after door opening, likely due to the distinctive short-loop shape. In contrast, a nonlinear model (multilayer perceptron, MLP) achieved a lower residual sum of squares (RSS), outperforming the linear model (Fig. 3G, Fig S6A). The difference in model fit between the nonlinear and linear models was significantly larger in the CS + door-opening condition than in either the inhibition or 180-second interval conditions (Fig. 3G). Notably, the difference in the CS + door-opening condition was specific to the claustrum-like cluster, as the other two clusters showed no comparable difference (Fig. S6B–D). Taken together, these results suggest that nonlinear integration occurs in the claustrum-like cluster specifically during the short-looping period immediately following door opening in the CS + door-opening condition.

Trajectory features predicted by the RNN are present in biological recordings

We next asked whether these trajectory features predicted by the RNN are also present in biological claustral recordings. As mentioned earlier, our previous study (17) was limited by the use of single-trial testing, a small number of recorded units per rat, and a limited total number of animals, which restricted both the overall population size and the feasibility of in-depth analyses. To address these limitations, we constructed a trained RNN model that produced reproducible and testable trajectories, particularly those exhibiting short-loop dynamics suggestive of information integration. To compare the model-derived trajectories with empirical neural data, we applied Gaussian Process Factor Analysis (GPFA) (35, 36), a dimensionality reduction method well-suited for single-trial time-series data. Although GPFA is typically optimized with multiple trials, it can still provide a useful low-dimensional representation of neural activity. By applying GPFA to single-trial unit data collected from multiple animals and embedding it into a three-dimensional latent space, we were able to visualize population dynamics in a manner comparable to the trajectories generated by the RNN, facilitating comparison between model dynamics and experimental observations (Fig. 4A).

Interestingly, the trajectory derived from units recorded in CS rats—the experimental group corresponding to the CS + door-opening condition in the RNN model—resembled the short-loop structure observed in the RNN (Fig. 4A, left). The onset of the short loop was markedly delayed compared to that in the claustrum-like cluster of the RNN. In contrast, trajectories from neutral rats, which received a neutral CS (a light cue not associated with an electrical shock) during testing, exhibited a gentler curvature (Fig. 4A, right). To confirm the reliability of this structure, we partitioned the neurons into two subgroups and independently repeated the GPFA analysis. Neurons with comparable basal firing rates were paired and subsequently classified into two groups. To compare trajectories before and after the split, the behavioral reference points on the post-split trajectories were aligned to those on the pre-split trajectories. The resulting latent trajectories remained qualitatively similar in shape and evolution (Fig. 4B), suggesting that the short-loop structure reflects coordinated activity across the neuronal population rather than being driven by a small subset of neurons.

To further examine this similarity quantitatively, we performed PCA on the trajectories and quantified their local geometry using an eigenvalue-based metric ($\lambda_3/\Sigma\lambda$) (Fig. 4C–E). This metric reflects the proportion of variance not explained by the plane defined by PC1 and PC2, thereby capturing deviations from planarity in the trajectory. Importantly, this geometric metric does not rely on recovering neuron-to-neuron noise correlations. Nevertheless, the estimation of the latent space still depends on the covariance structure among claustral neurons, suggesting that the inferred trajectories remain tied to biologically meaningful population dynamics. When applied to the PCA trajectories of the RNN, the metric was significantly higher in the CS + Open condition than in the Open-only condition (Fig. 4C). Similarly, when the GPFA trajectories from real claustral neurons were reanalyzed using PCA and the same metric was applied, the CS group showed significantly higher values than the Neutral group (Fig. 4D). Furthermore, a leave-one-neuron-out analysis revealed that claustral neurons in the CS group contributed more uniformly to this metric compared to those in the Neutral group (Fig. 4E). These findings support the interpretation that the observed short-loop structure did not arise from a small subset of neurons, outliers, or chance fluctuations.

Partial information decomposition uncovers neurons carrying synergy between CS and door-opening signals

Having established that nonlinear integration emerges at the population level, we next asked how this integration is implemented at the level of individual neurons. Because this analysis requires full population observability and repeated trials, which are not feasible in the biological dataset, we performed the following analyses using the trained RNN model. To examine the contribution of individual neurons to this nonlinear integration, we measured the synergistic information for each neuron—i.e., information encoded only when both stimuli, CS and door opening, were present. To this end, we performed Partial Information Decomposition (PID) analysis (37, 38). PID quantifies how two input variables (CS and door opening) contribute to the information contained in the output neural activity by decomposing the total mutual information into three components: (1) unique information provided independently by each input, (2) redundant information shared by both inputs, and (3) synergistic information generated only when both inputs are present simultaneously (Fig. 5C).

In the trained RNN model (as well as in real claustral neurons), the effect of an input stimulus persists in neural activity even after the physical stimulus has ended. Therefore, to define stimulus strength for PID analysis, we used the output of trained decoders—which infer the presence or absence of each stimulus from neural activity—as the effective input variable in the PID computation. We trained two decoders using the z-scored activity of neurons in the RNN claustrum-like cluster: one to distinguish between CS-present and CS-absent trials, and another to classify door-open versus no-door-open trials (Fig. 5A–B, Fig. S7; see Methods). The CS decoder performed at chance level during the baseline period but sharply increased in accuracy immediately after CS onset, with performance remaining high well beyond the end of CS presentation (Fig. 5A). The door-opening decoder showed high accuracy during the period when the door signal was present (Fig. 5B).

We used the CS decoding accuracy score and door-opening decoding accuracy score as inputs, and the z-scored activity of claustrum-like cluster neurons in each trial as the output. PID was computed for each neuron over time (Fig. 3D–H, see Methods). As expected, CS-unique information rose following CS onset and gradually decayed, whereas door-opening-unique information increased after door opening (Fig. S7C–D). Importantly, synergistic information—representing new information generated only when CS and door-opening representations were jointly present—increased immediately after door opening and then gradually decreased (Fig. 5E). This suggests that neurons in the claustrum-like cluster integrate CS and door-opening signals to produce new, combined information. Moreover, synergistic signals were reduced in two conditions where nonlinear integration was disrupted: when inhibition was applied to the network, and when the CS–door interval was extended to 180 seconds (Fig. 3F–G).

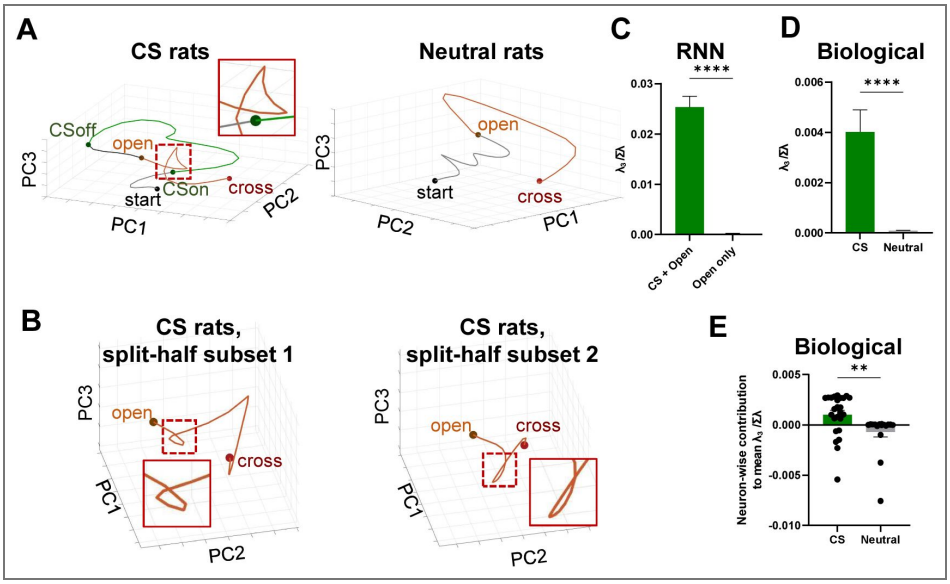

Figure 4. Biological Neural GPFA Trajectories and Quantification of Local Geometry.

(A) GPFA trajectories of in vivo single-unit recording data: (left) CS rats, (right) neutral rats (n = 25 neurons, 16 rats). The mean latency from door opening to crossing was 56.363 ± 8.22 s. (B) Validation of GPFA trajectories in CS group recordings. Neurons were ranked in descending order of mean z-scored firing rates (baseline to 5-s post-crossing) and split into odd-and even-indexed subsets. The trajectory from subset 1 (left; n = 13 neurons, 6 rats) served as the reference, while the trajectory from subset 2 (right; n = 12 neurons, 10 rats) was aligned accordingly. For clarity, only the segment from door opening to crossing is displayed. The mean latency from door opening to crossing for subset 1 was 53.306 ± 14.857 s. The mean latency from door opening to crossing for subset 2 was 50.969 ± 7.417 s. (C) Third eigenvalue divided by the sum of all eigenvalues (i.e., the proportion of total three-dimensional variance explained by PC3), computed using sliding windows from the post-open RNN trajectory, comparing the CS+Open (green; n = 20 windows) and Open-only (gray; n = 20 windows) conditions. **** Unpaired t test, $p < 0.0001$. (D) $\lambda_3/\Sigma\lambda$ of the biological neural trajectory within the sliding window, comparing the CS (green) and Neutral (gray) group pseudopopulations. $\lambda_3/\Sigma\lambda$ (the third eigenvalue divided by the sum of all eigenvalues) was computed within a sliding window of 10 time bins across the 20 post-open time bins of the averaged PCA trajectory (RNN) or the GPFA trajectory (biological pseudopopulation). **** Mann–Whitney test, $p < 0.0001$. (E) Neuron-wise contribution to the mean $\lambda_3/\Sigma\lambda$, defined using a leave-one-neuron-out analysis of the biological neural trajectory, comparing CS (green; n = 25 neurons) and Neutral (gray; n = 17 neurons). For the leave-one-neuron-out analysis, the GPFA trajectory was recomputed separately for each biological group after excluding one neuron at a time. For each neuron, its contribution was defined as the difference between the full-population mean $\lambda_3/\Sigma\lambda$ and the leave-one-out mean $\lambda_3/\Sigma\lambda$. The mean $\lambda_3/\Sigma\lambda$ was calculated by averaging $\lambda_3/\Sigma\lambda$ across sliding windows. ** Mann–Whitney test, $p = 0.0012$. Raw data from Fig. 4 of Han et al. (12), *Cell Reports*, were reanalyzed using GPFA.

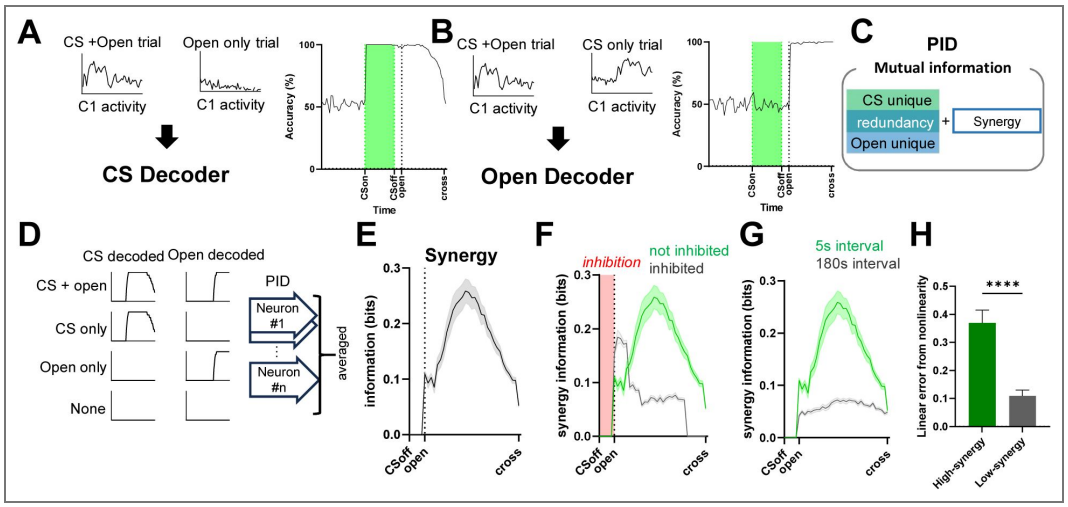

Figure 5. Decoder Accuracy, and Partial Information Decomposition of Cluster 1 RNN Neurons under different simulation conditions.

(A) Left: Schematic of the CS decoder. A classifier was trained to distinguish CS + door-opening trials from door-opening only trials using cluster 1 neuron activity. Right: CS decoding accuracy (%) across time bins. (B) Left: Schematic of the door-opening decoder. A classifier was trained to distinguish CS + door-opening trials from CS only trials using cluster 1 neuron activity. Right: Door-opening decoding accuracy (%) across time bins. (C) Conceptual diagram of Partial Information Decomposition. Mutual information about CS and door-opening is decomposed into CS-unique information, door-opening-unique information, redundancy, and synergy. (D) Schematic of the PID analysis. CS decoding accuracy and door-opening decoding accuracy were computed for each trial type (CS + door-opening, CS only, door-opening only, and None), and used as input variables for neuron-wise PID. PID terms were then averaged across neurons. (E) Synergy information of cluster 1 neurons. (F) Comparison of synergy between inhibition and no-inhibition simulations. (G) Comparison of synergy between the 180-s interval and the 5-s interval simulations. Green line: 5-s interval; green shaded box indicates CS presentation window. Gray line: 180-s interval. Solid lines represent means; shaded areas indicate SEM. (H) Model fit comparison between cluster 1 RNN neurons with high synergy and those with low synergy. The difference in residual sum of squares (ΔRSS) between linear regression and MLP models is shown, normalized to the mean RSS of the linear model. Bar colors indicate condition (high synergy = green, low synergy = gray). (High synergy: n = 147 trials; Low synergy: n = 147.) Mann-Whitney test: high vs. low synergy, ****$p < 0.0001$. Bars show mean ± SEM. In this figure, Open denotes door-opening.

We further divided claustrum-like cluster neurons into the top 25% with the highest synergy ("high-synergy" neurons) and the bottom 25% ("low-synergy" neurons) and analyzed their trajectory features (i.e., the short loop). In high-synergy neurons, the curvature of the neural trajectory immediately after door opening exhibited a more pronounced turning-angle change compared to low-synergy neurons (Fig. S8A–B). Furthermore, when CS + door-opening trials were predicted using trajectories from CS-only and door-opening-only conditions, nonlinear prediction errors were greater for high-synergy neurons than for low-synergy neurons (Fig. 5H). Together, these findings suggest that synergistic information at the single-neuron level is related to the short loop and that, within the claustrum-like cluster, certain neurons play a more dominant role in the nonlinear integration process.

Trajectory-based dynamic coding underlies integration

We next used the RNN model to examine how this integrated representation is dynamically encoded over time. To assess the temporal evolution and stability of the integrated representation, we applied cross-temporal decoding. This approach allowed us to examine whether the combined information from CS and door-opening is encoded in a static manner or in a time-varying fashion.

Using cross-temporal decoding, we examined how the neuronal population encodes information unique to the CS + door-opening condition over time. For each corresponding time bin across the three conditions (CS+Door, CS-only, and Door-only), we trained a decoder to distinguish the CS+Door condition from the other two, and evaluated its classification accuracy across all time bins. In the entire population of claustrum-like cluster, cross-temporal decoding revealed a mixed coding regime: shortly after door opening, decoders trained at specific time bins maintained discrimination power across relatively broad temporal windows (Fig. 6A). However, as time progressed, the discriminative power of decoders became temporally confined, performing well only near the training bin.

When neurons in the claustrum-like cluster were divided into high- and low-synergy groups and decoders were trained and tested separately on each, high-synergy neurons exhibited a dynamic coding pattern. Decoders trained on specific time bins in this group showed strong performance at that time point, with limited generalization to neighboring bins. In contrast, decoders trained on low-synergy neurons maintained more temporally stable performance. These results suggest that high synergy neurons, which contribute critically to nonlinear integration, encode CS and door-opening information using a dynamic coding scheme, whereas low-synergy neurons may support a more stable representation (Fig. 6B–C).

This distinction was further supported by raw firing-rate heatmaps: both high and low synergy neurons showed increased activity following CS onset, but after the door opened, only high-synergy neurons displayed variable firing patterns (Fig. 6D, left). Notably, a similar pattern of transient, variable spiking was observed in our in vivo recordings of claustral single units after door opening in the CS + door-opening group (Fig. 6D, right).

Temporal evolution of the population state can be framed by two hypotheses (Fig. 6E). The attractor hypothesis posits that neural dynamics consist of discrete or continuous attractors, with the integrated state represented by transitions among them in response to varying inputs. The trajectory hypothesis proposes that integration is embodied in the evolving trajectory itself, independent of fixed attractors. To distinguish these views, we performed fixed-point analysis on the RNN, locating states where the velocity vector is nearly zero (39). Fixed points were found to lie far from the empirical trajectories, indicating that the network's dynamics are trajectory-based rather than attractor-based when integrating CS and door-opening cues (Fig. 6F).

Discussion

In this study, we trained a vanilla recurrent neural network (RNN) solely on behavioral metrics from the delayed escape task, which had previously been shown to require rsCla (17). The trained network developed a claustrum-like cluster whose dynamics showed similarities to the experimentally observed population activity patterns and reflected the relationship between

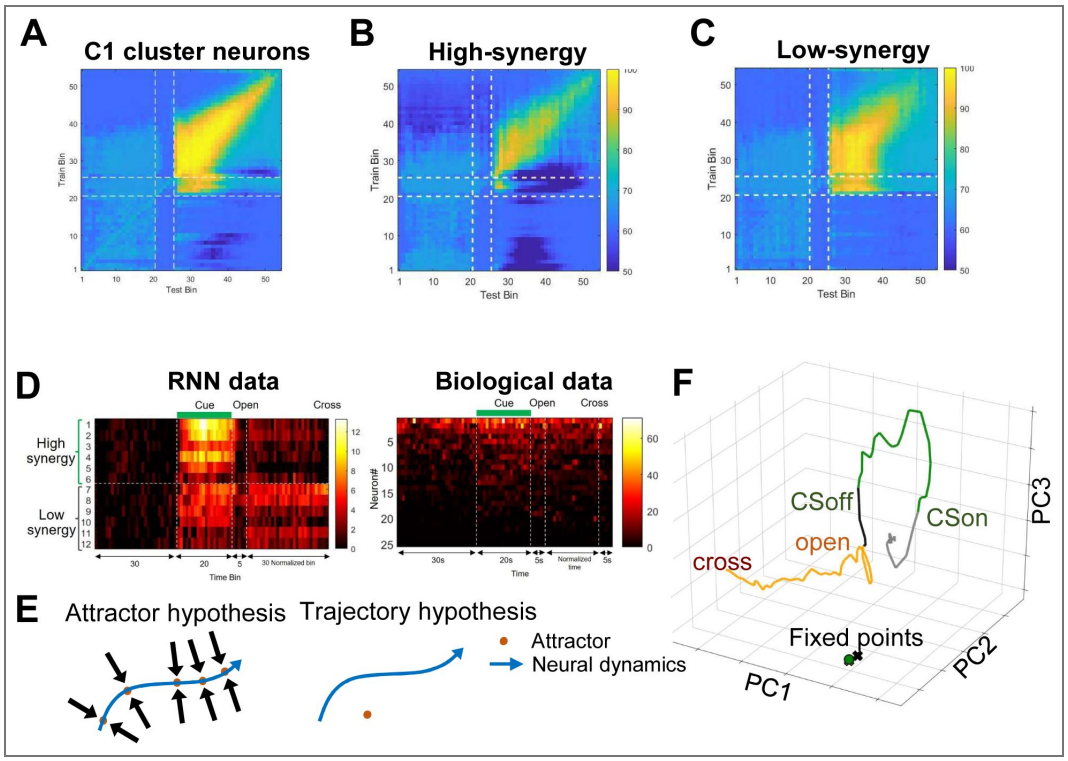

Figure 6. Trajectory Coding Hypothesis and Biological Neural Trajectories.

(A) Cross-temporal decoding of integration-specific information from cluster 1 RNN neurons. (B) Cross-temporal decoding of integration-specific information from cluster 1 RNN neurons with high synergy. (C) Cross-temporal decoding of integration-specific information from cluster 1 RNN neurons with low synergy. (D) Left: Heatmaps of raw firing rates for RNN neurons from a representative CS + door-opening trial. Right: Heatmaps for biological cluster 1 non-exploratory neurons in the CS group during the delayed escape task (right). Neurons in the right heatmap are ordered by overall activity. (E) Hypotheses on dynamic coding. Left: Attractor hypothesis: Multiple or continuous stable converging points exist, and changes in input cause neural dynamics to move from one point to another, thereby encoding varying states. Right: Trajectory hypothesis: Neural dynamics are not, or are only minimally, related to attractors (which may not exist); instead, changes in input cause changes in trajectory, which encode varying states. (F) Fixed-point analysis of a representative trial in an RNN. Green dots represent stable fixed points (CS period), while black and yellow X-marks denote unstable fixed points (interval and open periods).

neural activity and behavior. Low-dimensional trajectory analysis revealed a nonlinear integration of the conditioned stimulus (CS) and the temporally delayed door-opening signal, accompanied by an increase in synergistic information. Rather than converging to stable attractor states, the network encoded these signals through continuously evolving trajectories. In the post-integration phase, the trajectories formed a sharply curved "short loop" immediately after door opening—a trajectory pattern also observed in real claustrum recordings, thereby supporting the biological plausibility of the model. Such temporal integration through dynamic coding enables flexible responses to inputs that arrive in the claustrum with temporal gaps and offers the advantage of generating richer representations compared with stable coding (27–29, 31, 34, 40). Moreover, contrary to the intuitive expectation that dynamically coded information might be difficult to decode, it can, in fact, be easily read out using simple linear decoders (41), which provides an additional advantage.

It is important to emphasize that behavior-trained recurrent neural networks (RNNs) can admit multiple internal solutions capable of producing the same behavioral output. Accordingly, the network analyzed in this study represents only one possible computational realization. Nevertheless, the dynamical regime observed in the present model converged with several independent lines of evidence from claustral recordings. These include persistent neural activity during the delay period, the correlation between neural activity and escape latency, recurrent connectivity, and the consistent geometric changes in population trajectories revealed through population-level analyses. Taken together, these findings suggest that the present computational model should not be interpreted as a direct implementation of the claustrum's computational mechanism, but rather as a candidate model that captures a plausible computational principle that may operate in the claustrum. In particular, conclusions derived from the RNN analysis, such as increased synergy and dynamic coding, remain hypotheses that could not be directly validated in the biological claustrum due to experimental constraints, and therefore require future experimental testing. At the same time, the present model was specifically designed to capture the computations required for the delayed escape task and does not aim to account for the full range of claustral functions. In particular, it is not intended to replace or encompass the diverse roles of the claustrum proposed in previous studies, including those from the Mathur and Citri research groups (3–14). Future studies will be needed to determine whether dynamical principles similar to those identified in the present model also contribute to other proposed functions of the claustrum, including attention, salience detection, cognitive control, sleep, and premotor control.

Previous studies have shown that the anterior claustrum is largely unresponsive to simple sensory stimuli, becoming active only when such stimuli acquire behavioral relevance possibly through convergent inputs from regions encoding memory, value, context, and motivational states (7, 17, 23, 42–45). It also exhibits stronger responses to cross-modal stimuli when they are semantically congruent (42, 46). These findings suggest that the claustrum preferentially integrates behaviorally meaningful inputs via nonlinear operations to generate unified representations. Within this framework, both the conditioned stimulus (CS) and the door-opening cue in the delayed escape task may carry escape-relevant information. However, since the behavioral meaning of the door-opening cue may require additional processing in the biological circuit, the integration process may require additional time, which could explain the delayed emergence of the short-loop trajectory in real neural data (see Fig. 4G, left). This interpretation may be viewed as broadly compatible with prior theoretical proposals suggesting that the claustrum contributes to the integration of semantically related information (1). Furthermore, such integrated information is likely broadcast to higher-order cognitive regions such as the PFC, ACC, and OFC, where it can be read out and interpreted (47), a possibility that resonates with broader theoretical accounts of large-scale integrative brain function (1, 48).

Although our conclusions are mainly derived from modeling, the correspondence between model predictions and empirical observations provides a rationale for future experimental validation. Multi-trial behavioral paradigms and large-scale single-neuron recordings will be essential to rigorously test these predictions. Furthermore, elucidating the nature of the preprocessing that

shapes inputs to the claustrum, as well as the downstream pathways through which its integrated representations are accessed and utilized, will be critical for understanding its role in brain-wide computation.

Materials and Methods

Animals

The following behavioral procedures were previously reported in Han et al. (2024) and are summarized here for clarity (17). Male Long-Evans rats (Japan SLC Inc.) were dual-housed for 5–11 days before the experimental procedures began, under 12-hours inverted light/dark cycle (light off at 9:00 a.m.) with ad libitum access to food and water. All behavioral experiments were conducted during the dark portion of the cycle. The Institutional Animal Care and Use Committee (IACUC) of the Seoul National University approved all experimental procedures, and all experiments were performed following the guidelines for care and use of laboratory animals of the Seoul National University. The in vivo single-unit recording experiments for each of the CS and Neutral groups were performed separately.

Delayed escape task with in vivo single-unit recording

The followingrecording procedures were previously reported in Han et al. (2024) and are summarized here for clarity (17). Rats (7 weeks old) were anesthetized with an intraperitoneal (i.p.) injection of sodium pentobarbital (50 mg/kg) and maintained with isoflurane (1–1.5%) in O2. Rats were mounted on a stereotaxic apparatus (Stoelting Co.). Fixed-wire electrodes were bilaterally implanted into the rsCla (AP +3.35/ML ± 2.10/DV-4.50). A ground wire was implanted in the cerebellum. The electrodes consisted of eight individually insulated nichrome microwires (50 μm outer diameter, impedance 1-3 MΩ; California Fine Wire) contained in a stainless steel guide cannula. The electrodes were affixed to the skull with screws using Poly-F zinc polycarboxylate cement (Konstanz), vertex self-curing (vertex-dental, Zeist), and bond (Loctite 411, Henkel). Rats were allowed to recover for a week before they underwent experimental procedures.

The rats in both the CS and Neutral groups were presented with the green light cue (10 s × 5 times) in the fear conditioning context with modification (a foamex floor instead of a metal grid floor), which allowed for adaptation to the recording cable attachment. After 24 h, the rats were presented with the green light cue (10 s × 5 times, average inter-trial interval: 90 s) in the same context, to optimize recording parameters for each channel. Approximately 1 h later, the rats in the CS group were fear-conditioned, and the rats in the Neutral group underwent pseudo-conditioning in which the green light cues were presented but electrical foot shocks were omitted.

On day 2, rats underwent the two-compartment chamber (54 × 23 cm area in total) exposure session. The two compartments had identical floor sizes, but they were differently constructed (compartment A vs. B: foamex floor vs. mesh grill floor; no visual cued walls vs. visual cued walls with circles; 50 cm vs. 30 cm-height; a green LED bulb in the middle of the side wall and a speaker right above the LED vs. no accessories installed on the wall, respectively). The two compartments were connected by opposing square-shaped (7 × 7 cm) outlets, located 12 cm above the floor. The outlet could be opened or closed by sliding. To ensure that the subject localized a source of light within compartment A, the walls of the compartment were made of non-reflective material (black foamex). Consequently, compartments A and B were illuminated to a lesser intensity than in the fear conditioning performed on day 1, in which the walls were made of reflective materials.

The rats were placed in compartment A to acclimate for 2 min on average on day 2. The outlet in the wall (without any curtains) was then opened simultaneously with a tone pip sound (2.8 kHz, 200 ms, 85 dB). The rats were allowed to cross over to compartment B for a duration of 5 min. Once they crossed, the outlet was closed and the rats were allowed to stay in compartment B for 2 min. The rats were then returned to their home cages. If a rat did not cross in 5 min, it was removed from compartment A and temporarily moved to a small box near the two-compartment chamber. Around 30 s later, the rat was re-introduced to compartment A and the same procedure was performed. The rats that did not cross after 10 trials were excluded from further experiments.

We measured crossing latency and the number of trials in which the rat succeeded in crossing. Crossing latency was defined as the duration from outlet opening to the moment when all four paws of the rat touched the floor of compartment B. The chamber was cleaned with 70% ethanol and then with distilled water prior to the beginning of each trial.

For the delayed escape test session (day 3), the two-compartment chamber was modified to further prohibit any potential association between the outlet opening and crossing behavior. The opened outlet was visually blocked by a black curtain installed at the outlet side of compartment B. The rat head could go through the curtain to see compartment B since the curtain consisted of two pieces of fabric adjoined at the centreline of the outlet. The rats repetitively put their heads into compartment B through the curtain during the test session, but this behavior did not appear to be associated with outlet crossing. In the modified chamber, a 20 s CS was presented 220 s after the placement of the rat in compartment A. The CS presentation was delayed when the rat showed immobility at the time of CS onset. The outlet was opened with a tone pip sound (2.8 kHz, 200 ms, 85 dB) 5 s after the CS offset, and the rat was allowed to cross the outlet for a duration of 5 min. The test session was performed only once. The rats that did not cross during the test session were considered to omit the task, and they were excluded from data analysis.

RNN modeling

The RNN model was implemented in PsychRNN (24, 25) as a continuous-time network with a 1s integration step. Each network consisted of a single hidden layer of 100 ReLU units, of which 80% were excitatory and 20% inhibitory.

Hidden-state $h(t)$ dynamics followed

$$\tau \boldsymbol{h}(t) = -\boldsymbol{h}(t) + f\left[\boldsymbol{W}_{\text{rec}}\,\boldsymbol{h}(t) + \boldsymbol{W}_{\text{in}}\,x(t) + \boldsymbol{b}_{\text{rec}}\right] + \boldsymbol{\eta}(t)$$
$$f(x) = \max(x, 0)$$
$$\boldsymbol{y}(t) = W_{\text{out}}\,\boldsymbol{h}(t) + b_{\text{out}}$$

An internal time constant of τ = 1000ms was adopted, in accordance with electrophysiological measurements (Fig. S4), and ReLU activation f(x)=max(0,x) was used throughout the network. Here, W_{rec}, W_{in}, and W_{out} denote the recurrent, input, and output weight matrices, respectively, while b_{rec} and b_{out} are constant biases applied to the recurrent and output units. Gaussian noise η(*t*) with variance $\eta^2 = 0.05^2$ was injected into the recurrent layer. Simulations were conducted using a continuous-time recurrent neural network with a discretized time step of 1000ms. To ensure stable signal propagation and gradient flow during early training, the recurrent weight matrix (W_{rec}) and output weight matrix (W_{out}) were initialized using a Glorot normal distribution. Specifically, initial weights were drawn from a normal distribution centered on 0 with a standard deviation of $\sigma = \sqrt{2}/(N_{in} + N_{out})$ where N_{in} and N_{out} denote the number of input and output units for the given connectivity matrix. Input weights (W_{in}) were initially drawn uniformly from the interval [0, 0.15], after which all rows corresponding to inhibitory units were zeroed out to enforce Dale's principle. To strictly enforce Dale's principle throughout training, rows in (W_{out}) corresponding to inhibitory units were zeroed out, and recurrent weights were constrained such that excitatory and inhibitory neurons maintained their respective positive and negative synaptic projections.

We trained the RNN in a supervised manner using the behavioral data from rat experiments (49). We provided the RNN with sequential inputs that matched the experimental timeline: a 30-second baseline period with no input, followed by a 20-second CS input period, then a 5-second interval with no input, and finally a sustained door-open signal until the end of the trial. In the control group (neutral group), no input was provided during the period corresponding to CS presentation, and trials from the CS and neutral groups were presented in a randomized order. The network's output was treated as the probability of crossing; when this probability exceeded 0.5, the model was considered to have initiated a cross, allowing us to compute latency for each trial.

The network received four input channels, denoted as $x(t) = [x_{CS}, x_{door}, x_{inh}, x_{ctx}]^\top$. To reflect the intense and salient nature of the fear-conditioning experience, the magnitude of the CS channel ($x_{CS}(t)$) was set to 2 during the presentation of the CS, specifically from t = 70 s to t = 90 s (20 s

duration), serving as a relatively stronger sensory drive compared to the simple mechanical cue of the door opening. The door signal ($x_{door}(t)$) was set to 1 starting at t = 90 + d s, where d was selected from [0, 1, 2.5, 5] seconds depending on curriculum learning, and remained 0 beforehand. The inhibition signal ($x_{inh}(t)$) was set to −1 for the 5 seconds immediately following the CS offset (t = 90-95 s) on inhibition trials. Finally, the context input ($x_{ctx}(t)$) was maintained at 1 throughout the entire trial.

Input weights were first drawn uniformly from the interval [0, 0.15], after which all rows corresponding to inhibitory units were zeroed out to enforce Dale's principle.

A cross was defined as the first time point t^* at which the network output $y(t^*) \geq 0.5$. The desired output trajectory $y\neg(t)$ was specified as a step function transitioning at the target escape

$$\tilde{y}(t) = \begin{cases} 0, & t < t_{tar}, \\ 1, & t \geq t_{tar}. \end{cases}$$

time t_{tar}, which combines the door-opening time and the experimentally measured behavioral latency measured in rats (latency for CS present: 48.7s, latency for CS absent: 111.3s):

$$t_{\text{tar}} = \begin{cases} t_{\text{door}} + \ell_{\text{CS}} = 90 + d + 48.7\ \text{s}, & \text{CS present}, \\ t_{\text{door}} + \ell_{\text{noCS}} = 90 + d + 111.3\ \text{s}, & \text{CS absent}, \end{cases}$$

Using fixed empirical latencies anchored the network to the average rat behavior, such that remaining variability in output timing arose solely from the network's internal dynamics.

The per-trial loss function minimized during training was:

$$\mathscr{L} = -\sum_t [\tilde{y} \ln y + (1 - \tilde{y}) \ln(1 - y)] + \lambda \left(\|\mathbf{W}_{\text{in}}\| + \|\mathbf{W}_{\text{rec}}\| \right) + \lambda_{FR} \langle \| \text{h(t)} \|_2^2 \rangle_t$$

The L2 weight regularization coefficient λ = 0.01 and the firing rate regularization coefficient λ_{FR} = 0.95. were systematically optimized via a grid search procedure to maintain biologically plausible firing rates while preventing overfitting. The network was trained using TensorFlow 2.1.0 and the Adam optimizer (learning rate $\alpha = 5 \times 10^{-4}$) for 1.2 million iterations with a batch size of 256. Each training batch comprised trials assigned randomly to CS-present or CS-absent conditions with equal probability (50% each). The delay d between CS offset and door-opening was incrementally increased (0,1, 2.5, and 5 s) once the batch-wise accuracy exceeded thresholds of 0.5, 0.5, 0.5 and 1, respectively. All simulations were executed on a single RTX-2080 Ti GPU using XLA just-in-time compilation.

After training was completed, the network was saved for downstream testing under various experiment conditions, including inhibition, extended (180-s) interval, CS-only, no-CS/no-Door-opening, and stimulation of the slice experiments. The resulting cross latencies t^*, neuronal activity $h(t)$, and weight matrices were exported to MATLAB for further analyses such as clustering, principal component analysis (PCA), and Partial Information Decomposition (PID).

RNN Population Analysis

The recurrent neural network (RNN) simulation was processed with the same population-level pipeline used for the experimental data (49).The first 10 s of each simulation were discarded to eliminate variability caused by initial states. Activity from the next 30 s served as the baseline for Z-score normalization.

For every CS trial (n = 147), Z-scores were averaged neuron-wise and then summarized across five task epochs—cue, interval, early-open, late-open, and the 5 s after the cross. These epoch-wise vectors were embedded into a three-dimensional t-SNE space (perplexity = 24, exaggeration = 48, RNG = mt19937ar). The optimal number of clusters was determined with the gap statistic and applied in a k-means clustering step. Through this procedure, we obtained three clusters consisting of 24, 38, and 38 units (with 24 units in the claustrum-like cluster).

Here, the term claustrum-like refers to the fact that the mean response of the units remained persistent not only during the CS presentation but also throughout the subsequent period until crossing. The main difference between the empirical experiment and the RNN results is that, in the actual experiment, each rat was tested with only one latency value, so correlations had to be

calculated across animals, whereas in the simulation, multiple trials were generated from a single RNN, allowing correlations to be calculated across trials. Specifically, for each trial, the mean z-scored activity across the 24 units during the cue-to-cross window was calculated, and this value was paired with the crossing latency from that trial., and these z-scores were paired with the crossing latency from that trial. With 100 trials, this yielded 100 data points. The variability of model latency reflects both the injected noise and differences in initial states.

Notably, only a subset of the supervised RNNs we trained produced a claustrum-like cluster; among 100 independent runs, only 5 met this criterion, and all results reported here are based on one representative RNN network from those 5 cases.

As shown in Fig S2, the output contribution of each cluster for each time points was computed as the product of the hidden-state vector $h(t)$ and the output-weight matrix W_{out}.

Slice electrophysiology

Male Long-Evans rats (3–4 weeks old) were anesthetized (sodium pentobarbital, 50 mg/kg, i.p.) and mounted in a stereotaxic frame. For optogenetic circuit mapping, AAV encoding ChrimsonR (pAAV-CamKIIa-ChrimsonR-mScarlet-KV2.1; Addgene; 1×10^{13}–1×10^{14} vg/mL) was diluted 1:4 in sterile saline for sparse labeling and injected unilaterally (200 nL) into the rsCla (AP +2.95, ML ±1.95, DV −3.85 mm). For calcium imaging, AAV encoding GCaMP6f (pENN.AAV.CamKII.GCaMP6f.WPRE.SV40; Addgene; ≥1×10^{13} molecules/mL) was injected bilaterally (400 nL) into the same coordinates. Viruses were delivered at 20 nL/min via a pulled glass capillary using a pressure injection system (Nanoject II, Drummond; or Nanoliter 2010, WPI). The pipette was left in place for 10 min post-infusion to ensure diffusion. Animals were dual-housed for 3–4 weeks for viral expression. Subsequently, rats were anesthetized with isoflurane and decapitated for acute slice preparation. The isolated whole brains were placed in the slicing chamber filled with a warm (34∼36 °C) artificial cerebrospinal fluid (aCSF) solution containing 120 mM NaCl, 3.5 mM KCl, 1.25 mM NaH_2PO_4, 26 mM $NaHCO_3$, 1.3 mM $MgCl_2$, 2 mM $CaCl_2$, and 11 mM D-(+)-glucose, and continuously bubbled at room temperature with 95% O_2/5% CO_2. Acute sagittal or horizontal slices containing the rsCla (300 μm thickness) were prepared using a vibratome (VT1200, Leica). Slices were then stored under submerged conditions at 34 °C for 1.5 hour before recordings. For recording, slices were transferred to the recording chamber in which aCSF was continuously perfused. The perfused aCSF was continuously aerated with 95% O_2/5% CO, and maintained at 32 °C.

Whole-cell patch clamp recordings were performed using micropipettes (4∼6 MΩ) which were pulled from borosilicate glass capillaries (1.2 mm OD, 0.94 mm ID, Warner Instruments; Massachusette, USA) with a micropipette puller (Pc-10, Narishige). The pipettes were filled with the following solution: 120 mM potassium D-gluconate, 0.2 mM EGTA, 10 mM HEPES, 5 mM NaCl, 2 mM Mg-ATP, 0.3 mM Na-GTP, and 1 mM $MgCl_2$, with the pH adjusted to 7.2 with KOH and osmolality adjusted to approximately 297 mmol/kg with sucrose. Recordings were made under infrared differential interference contrast (IR-DIC)–enhanced visual guidance from neurons located in the rsCla slices. Neurons were first current-clamped neurons and then used with membrane potentials lower than −50 mV. The neurons were voltage-clamped at −70 mV, and solutions were delivered to slices via superfusion driven by gravity at a flow rate of 1.3 ml/min. The pipette series resistance was monitored throughout the experiments, and if it changed by >20%, the data were discarded. To avoid direct light stimulation, whole cell recordings were conducted using neurons in which the expression of ChrimsonR was absent.

Wide-field photostimulation was executed via a microscope's objective lens (40×/0.80NA) using a power of 6.4 mW (measured at the lens level) and a wavelength of 617 nm. This stimulation was achieved through a digital mirror device (Polygon 400, mightex) for a duration of 2 ms. To stimulate each of the divided parts of the entire optical field, the same digital mirror device (DMD) was used. The entire optical field was partitioned into a 30 by 30 grid, yielding a total of 900 equivalent sections. Each section received 2 ms of stimulation, and the heatmap was generated using the Matlab interpolation algorithm with the EPSC data. The resultant EPSC amplitudes were used to generate an EPSC heatmap encompassing the entire field. Whole-cell currents were filtered at 2 kHz, digitized at up to 10 kHz, and stored on a microcomputer (Clampex 10.7 software,

Molecular Devices). We used 10 μM of NBQX (Tocris), and 50 μM of D-AP5 (Tocris). One or two neurons were recorded per animal (a single neuron per slice). All recordings were completed within 5 hours after slice preparation.

In brain slices where GCaMP6f was expressed, electrical stimulation was applied using a Concentric Bipolar Electrode (FHC) via an isolator (WPI, A360). Stimulation was delivered at 150-200 μA, 20 Hz, 1sec duration. The puffer pipette was positioned 75 μm below the slice surface and 50 μm away from the electrode. At the 10-second mark post electrical stimulation, pressure injection was carried out under a pressure of 20 psi. We used 100 μM of NBQX (Sigma) and 500 μM of D-AP5 (Tocris) for this experiment. The analysis of fluorescence changes in the recorded images follows these steps: Initially, cell boundaries were manually marked for the 35 recorded GCaMP6f slice samples. Subsequently, a U-Net deep learning model was trained on these images to develop a cell segmentation model. Using this model, the attached cells were segmented and separated utilizing the watershed algorithm. Finally, all cells were manually reviewed once more through video inspection.

RNN simulation adjusted for slice-stimulation conditions

To mimic the effect of electrical stimulation and applied NBQX+D-AP5 in the acute slice, we constructed a reduced circuit composed exclusively of Cluster 1 (C1) units. All recurrent weights outside C1 were zeroed, leaving only C1 → C1 connections in W_{rec}. Because inhibitory neurons may lose efficacy in acute slices, every inhibitory weight in W_{rec} was scaled to 40 % of its original value.

Stimulation was modeled by injecting an external drive $x_{\text{stim}}(t) = 2$ for 1 s. The corresponding input weights were drawn uniformly from 0 to 0.15 for excitatory neurons only. Input weights to the inhibitory neurons were set to zero.

In our model, we simulated the effects of applying NBQX + D-AP5 by introducing recurrent inhibition, which began 10 seconds after the initial stimulation. The inhibition was targeted, affecting only 10% of the neurons in the network. For these selected neurons, the strength of the excitatory inputs they received from other neurons (W_{rec}) was reduced by 60%. This simulation was designed to replicate key findings from our slice experiments (Fig. S4): First, the effect of NBQX + D-AP5 is localized and does not spread widely; and second, drug application resulted in an approximately 60% reduction in the amplitude of EPSCs.

Visualization of PCA trajectories

To visualize neural dynamics, trial-wise Z-scored firing rates from a specific neuronal cluster were extracted for each simulation group (CS + Open, Open-only, and CS-only). Trials were concatenated across groups and reshaped into a 2D matrix of size (trial x time) × neuron for principal component analysis (PCA). PCA was performed on the combined matrix, and the first three PCs were used for visualization.

Since each trial had a different latency to the cross event, trajectories were time-normalized. Each trial was divided into two segments: a fixed-length pre-open segment (Segment A, from the start to the open event; 55s) and a variable-length post-open segment (Segment B, from the open event to the cross event). For trials without a defined Cross event (e.g., CS-only group), the average cross latency from the CS + Open group was used. Segment B was interpolated to a fixed length of 30 bins using shape-preserving piecewise cubic interpolation. The two segments were concatenated and smoothed using a moving average. The smoothing window size was determined heuristically to attenuate approximately 25% of the energy of the PCA scores for each trial.

Time-normalized trajectories were averaged within each group to produce group-level mean trajectories and corresponding standard errors of the mean (SEM). These trajectories, uniform in the number of time bins, were visualized either separately for each group or overlaid for direct comparison.

For both the inhibition simulation and the 180s interval simulation, corresponding trial data were combined with CS + Open trials. These combined datasets were reshaped and processed using the same PCA and time-normalization procedures described above.

Prediction of CS+Open Trajectories from Component Conditions

To evaluate how well CS+Open trajectories could be reconstructed from their component conditions, we extracted time-normalized PCA trajectories from the CS+Open, Open-only, and CS-only groups. For the CS+Open condition, trajectories were aligned to the door-opening event, which occurred 5 seconds after CS offset; for the CS-only condition, trajectories were taken beginning at the corresponding time point—5 s after CS offset. All trajectories were preprocessed using the same procedures described in the trajectory visualization section. To standardize across conditions, trajectories were trimmed to include only the period from the open event to the cross event (30 bins), which served as the input for modeling.

We trained two models to predict individual CS+Open trial trajectories from the group-averaged Open-only and CS-only trajectories:

Linear regression model

At each time bin, the trial trajectory was regressed onto the corresponding bins of the mean Open-only and CS-only trajectories, yielding a bin-wise linear estimate.

Multilayer perceptron (MLP)

To test for nonlinear integration, we implemented a feedforward neural network with two hidden layers ([6, 3] units) and L2 regularization ($\lambda \in [0, 0.25, 0.5, 0.75, 1]$), with the optimal λ chosen to minimize the RSS for each condition. For each condition (CS+Open, inhibition, 180 s interval), the λ yielding the lowest residual sum of squares (RSS) was selected. The MLP received concatenated mean trajectories from the Open-only and CS-only conditions as inputs and was trained to predict the trial trajectory of the CS+Open condition. Training was performed independently for each trial using a 70/15/15% split of bins for training, validation, and test sets, with min–max normalization applied to the inputs.

Prediction performance was quantified as the RSS at each time bin. For the MLP, training was repeated 10 times per trial, and RSS values were averaged across repetitions to yield a trial-wise mean RSS time series. To assess the benefit of nonlinear integration, improvement by the MLP over the linear model was calculated as the bin-wise difference between their RSS values, normalized by the linear model RSS. These normalized scores were averaged across bins to obtain a single summary value for each trial.

Decoding analysis (for Fig. 5A-B)

CS Decoding

We first gathered the time-normalized Z-scores of all cluster neurons from two trial types: CS + Open (cue present) and Open-only (cue absent). Time normalization from open to crossing was identical to the procedure described above.

For each time bin we formed a trial-by-neuron matrix and trained a linear discriminant model with fitcdiscr('linear') in MATLAB R2022b. Trials in which a cue was delivered were labelled 1 from cue onset onward, whereas cue-absent trials were labelled 0 throughout. Model performance was assessed by five-fold cross-validation, and accuracy was reported as 1 – kfoldLoss. Repeating this procedure over all bins produced a temporal profile of CS-decoding accuracy.

Door Open Decoding

A Door-open decoder was generated in the same way, but the training data comprised CS + Open versus CS-only trials. Here, bins occurring after the outlet opened were labelled 1, and all earlier bins (as well as every bin in CS-only trials) were labelled 0. Five-fold cross-validated linear discriminant analysis again yielded a time-resolved accuracy curve that reflected when RNN activity best predicted door opening.

Partial Information Decomposition (PID) Analysis

To quantify how CS and door-open information were distributed across RNN neurons, we applied Partial Information Decomposition (PID) analysis to decoding accuracies as X variables and neural activity as a Y in cluster 1–3. For each cluster, z-scored activity was extracted and time-normalized using a fixed window of 85 bins, which included a 55s pre-open segment and a 30-bin post-open segment interpolated to a fixed length. Trials from four conditions were included: CS + Open, Open-only, CS-only, and None.

For each neuron and each time bin, we computed PID terms (redundancy, CS-unique information, Open-unique information, and synergy) using MATLAB code(the Neuroscience Information Theory Toolbox (37)).

At each time bin, we used the neuron's z-scored activity rounded to one decimal place across all trials as the Y (time × trial), and the CS and door-open decoding scores as X_1 and X_2, respectively (time × trial) (see Fig. 5D).

Specifically, the construction of X_1 (cue-related events) was as follows.

For the CS+Open and CS-only conditions, decoding accuracy were computed as described above. Trials in which a cue was delivered were labeled 1 from cue onset onward, whereas cue-absent trials were labeled 0 throughout. A decoder was trained to classify the z-scored firing rates in each time bin of CS+Open and Open-only trials according to these labels. To express decoding accuracy as a score ranging from 0 to 1, we subtracted 0.5 from the decoding accuracy, multiplied the result by 2, and set any negative values to 0. The resulting scores were smoothed with a Gaussian kernel (window size = 15). To compute the probability of coincident events, smoothed scores were rounded to the first decimal place. For Open-only and None conditions, all time bins were assigned a value of 0.

The construction of X_2 (door-open-related events) was as follows.

For CS+Open trials, as described above, time bins after the outlet opening were labeled 1, and all earlier bins as well as bins in CS-only trials were labeled 0. A decoder was trained to classify each time bin's z-scored activity in CS+Open and CS-only trials based on these labels.

Decoding scores were then computed using the same method as described for X_1.

For Open-only trials, time bins following outlet opening were labeled 1, and all earlier bins as well as bins in None trials were labeled 0. A decoder was trained to classify z-score in Open-only and None trials, and decoding scores were computed identically.

For CS-only and None conditions, all bins were assigned a value of 0.

For each discrete state j of Y and each combination of states k from a source A (either X_1 alone, X_2 alone, or their joint distribution [X_1, X_2]), we computed the corresponding conditional probabilities required for PID analysis.

$$P(A = k \mid Y = j) = P(Y = j, A = k)/P(Y = j), P(Y = j \mid A = k) = P(Y = j, A = k)/P(A = k)$$

The information that source A conveys about Y when Y is in state j is then

$$I(Y = j; A) = \sum_k P(Y = j, A = k) \log_2 \frac{P(Y=j|A=k)}{P(Y=j)}$$

We defined its redundancy by, for each j, taking the minimum of $I(Y = j; A)$ over the X1 and X2, and then computing the dot product of that vector of minima with the state-probabilities $P(Y = j)$. X_1-unique information is the mutual information between Y and X1 minus the redundancy. X2-unique information is the mutual information between Y and X2 minus the redundancy.

Synergy is obtained by taking the joint mutual information, subtracting both unique-information terms, and then adding back the redundancy.

PID terms were averaged across neurons within each cluster. For visualization, time-resolved PID terms were shown as mean ± SEM across neurons.

Cross-temporal Decoding Analysis

Cross-temporal decoding of CS+Open information was performed as follows: For each time bin, a decoder was trained to classify z-scored neural activity from CS+Open trials (all time bins labelled as 1) versus CS-only and Open-only trials (all time bins labelled as 0). The trained decoder was then tested across all time bins to predict trial type. Decoding accuracy was computed as (correct trials / total trials × 100) and visualized as a two-dimensional heatmap indexed by training and testing time bins.

Fixed Point Analysis

We implemented the FixedPointFinder toolbox to identify fixed points in the trained RNN model(39).

Fixed point analysis was performed separately for the following input epochs, defined by specific timestep intervals and corresponding input vectors:

Baseline (40–69 s): [0, 0, 0, 1, 0, 0, 0, 0]

CS (70–89 s): [2, 0, 0, 1, 0, 0, 0, 0]

Interval (90–94 s): [0, 0, 0, 1, 0, 0, 0, 0]

Open (95–239 s): [0, 1, 0, 1, 0, 0, 0, 0]

The recurrent and input weights, bias vector, and the alpha parameter ($\alpha = \Delta t/\tau$) were used to represent the trained dynamics throughout the epochs.

For each epoch, the neural state trajectories x and corresponding input vector u were constructed. Initial conditions for fixed point searches were formed by combining all neural states within the epoch with a randomly sampled subset of global network states (n_trial= (number of epoch states)+ 512 randomly sampled states; x: n_trial × n_neurons matrix,u:n_trial x 8 matrix). The input vector u was tiled across trials to match the dimensionality.

Fixed points were then identified jointly via optimization To find fixed point candidates, x and their one-step forward states F were computed, and the difference between x and F was minimized in an iterative gradient descent loop, where x was updated using the Adam optimizer. If change in loss became smaller than a convergence tolerance(1×10^{-4}) times the learning rate or if the loss is smaller than tolerance, then the iteration was stopped. The maximum number of iterations was set to 10,000. We discarded the candidate as an outlier if the distance between a fixed point candidate x* and the centroid of the initial states was larger than 10.0 times the average distance between the centroid and each initial state. Redundant candidates were further merged if the Euclidean distance between any pair of x* was less than 1×10^{-2}.The remaining fixed points x* and their Jacobian eigenvalues were collected. Fixed points were classified based on their dynamical stability: points were labeled as stable (attractors) if all eigenvalues had absolute magnitudes less than 1.0, and as saddle points otherwise.

Trajectory Analysis of Biological Data Using GPFA

For the biological data, z-scored activity pooled from in vivo recordings of nonexploratory cluster 1 neurons was reduced to a three-dimensional space using Gaussian-Process Factor Analysis (GPFA), with Gaussian smoothing kernel size of 20 bins (implemented using MATLAB code from Lakshmanan et al. 2015 (50)). Each pseudopopulation trajectory from CS or Neutral rats was subsequently smoothed with a Gaussian kernel (10-bin window) and visualized. We employed GPFA because it estimates latent states shared across neurons while explicitly modeling the noise characteristics of individual neurons, making it well-suited for single-trial trajectory visualization. Unlike PCA, which identifies orthogonal axes that explain maximal variance without incorporating temporal structure, GPFA accounts for temporal dynamics, providing a key advantage in analyzing population activity over time.

Validation was performed by independently applying GPFA to two neuronal subsets derived from in vivo CS group recordings. Neurons were ranked in descending order based on their mean z-scored firing rates during the baseline to 5-second post-crossing window. The ranked list was then divided into two subsets comprising odd- and even-indexed neurons, respectively. GPFA was applied to each subset using identical parameter settings as described above. For visualization, neural trajectories were smoothed with a 10-bin Gaussian window. The trajectory of one subset (subset 1) was chosen as the reference, and the others were aligned to it using the Procrustes method (MATLAB function). Specifically, anchor points from each trajectory were mapped to the corresponding anchors in the reference trajectory by minimizing the sum of squared errors under a linear transformation. The procedure included centering and normalization, followed by singular value decomposition (SVD) to estimate the optimal rotation, and finally applying translation and scaling. Anchor points consisted of the trial start, CS onset, CS offset, door opening, and crossing time.

Figure supplements

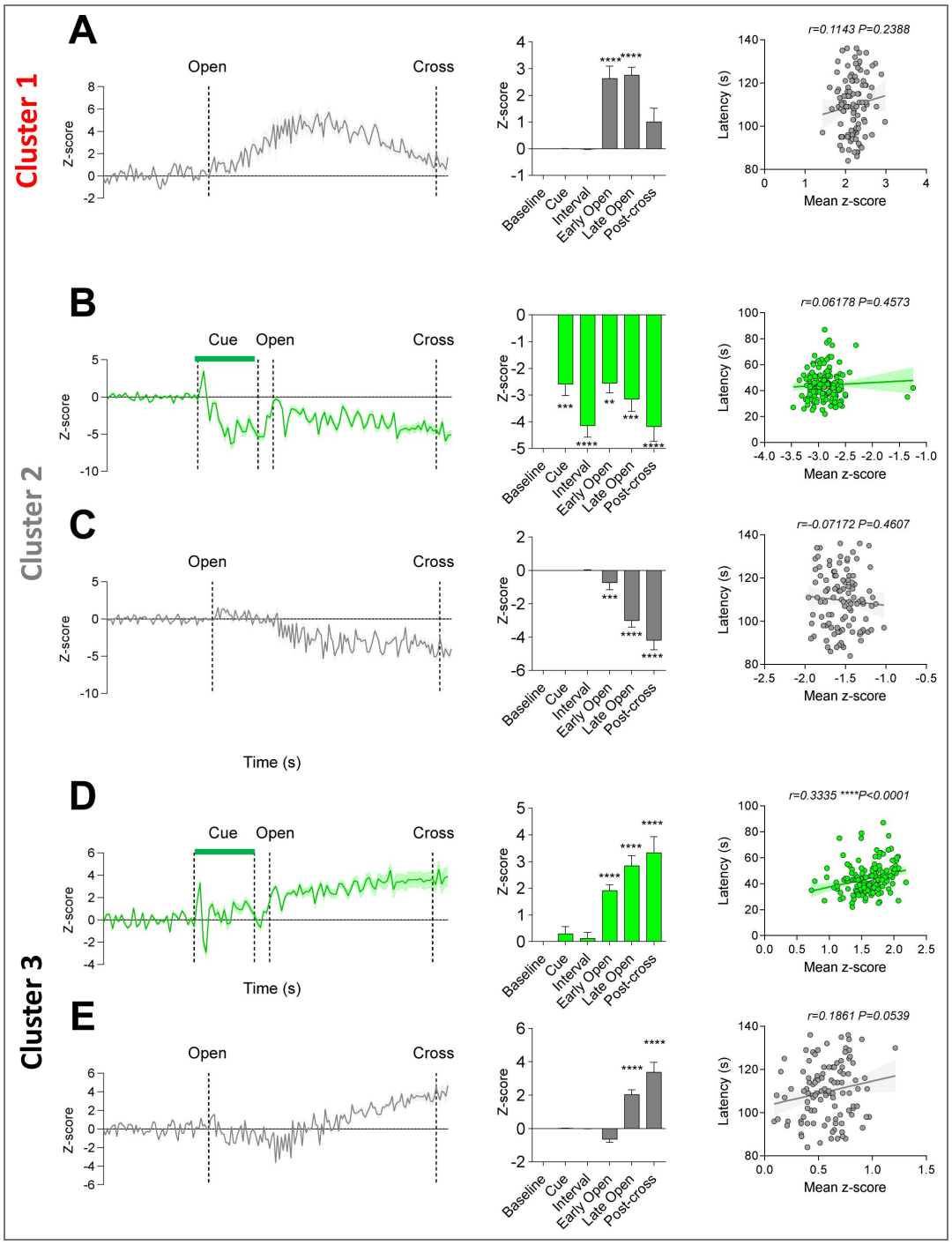

Figure S1. RNN activity of Clusters 1–3 under CS and door-opening only conditions (A) **Cluster 1, door-opening only.** Left, Z-scored firing rate of Cluster 1 neurons (n = 24) in a representative trial. Middle, average Z-score for each behavioral period across all trials; Friedman test, ****p < 0.0001; Dunn's post hoc—Base vs Cue, ****p < 0.0001; Base vs Interval, ***p = 0.0001; Base vs Early Open, ****p < 0.0001; Base vs Late Open, **p = 0.0034; Base vs Post-escape, p = 0.7134. Right, Spearman correlation between crossing latency and Cluster 1 activity (n = 147; r = 0.1143, p = 0.2388). (B) **Cluster 2, CS.** Left, Z-scored firing rate of Cluster 2 neurons (n = 38) in a representative trial. Middle, average Z-score for each behavioral period across all trials; Friedman ****p < 0.0001; Dunn's—Base vs Cue, ***p < 0.003; Base vs Interval, ****p < 0.0001; Base vs Early Open, **p = 0.0072; Base vs Late Open, ***p = 0.0002; Base vs Post-escape, ****p < 0.0001. Right, correlation between crossing latency and Cluster 2 activity (n = 147; r = 0.0618, p = 0.4573). (C) **Cluster 2, door-opening only.** Left, Z-scored firing rate of

Cluster 2 neurons (n = 38) in a representative trial. Middle, average Z-score for each behavioral period across all trials; Friedman ****p < 0.0001; Dunn's—Base vs Cue, p > 0.9999; Base vs Interval, p > 0.9999; Base vs Early Open, ***p = 0.0009; Base vs Late Open, ****p < 0.0001; Base vs Post-escape, ****p < 0.0001. Right, correlation between crossing latency and Cluster 2 activity (n = 147; r = –0.0717, p = 0.4607). (D) **Cluster 3, CS.** Left, Z-scored firing rate of Cluster 3 neurons (n = 38) in a representative trial. Middle, average Z-score for each behavioral period across all trials; Friedman ****p < 0.0001; Dunn's—Base vs Cue, p > 0.9999; Base vs Interval, p > 0.9999; Base vs Early Open, ****p < 0.0001; Base vs Late Open, ****p < 0.0001; Base vs Post-escape, ****p < 0.0001. Right, correlation between crossing latency and Cluster 3 activity (n = 147; r = 0.3335, ****p < 0.0001). (E) **Cluster 3, door-opening only.** Left, Z-scored firing rate of Cluster 3 neurons (n = 38) in a representative trial. Middle, average Z-score for each behavioral period across all trials; Friedman ****p < 0.0001; Dunn's—Base vs Cue, p = 0.2867; Base vs Interval, p > 0.9999; Base vs Early Open, p > 0.9999; Base vs Late Open, ****p < 0.0001; Base vs Post-escape, ****p < 0.0001. Right,, correlation between crossing latency and Cluster 3 activity (n = 147; r = 0.1861, p = 0.0539).

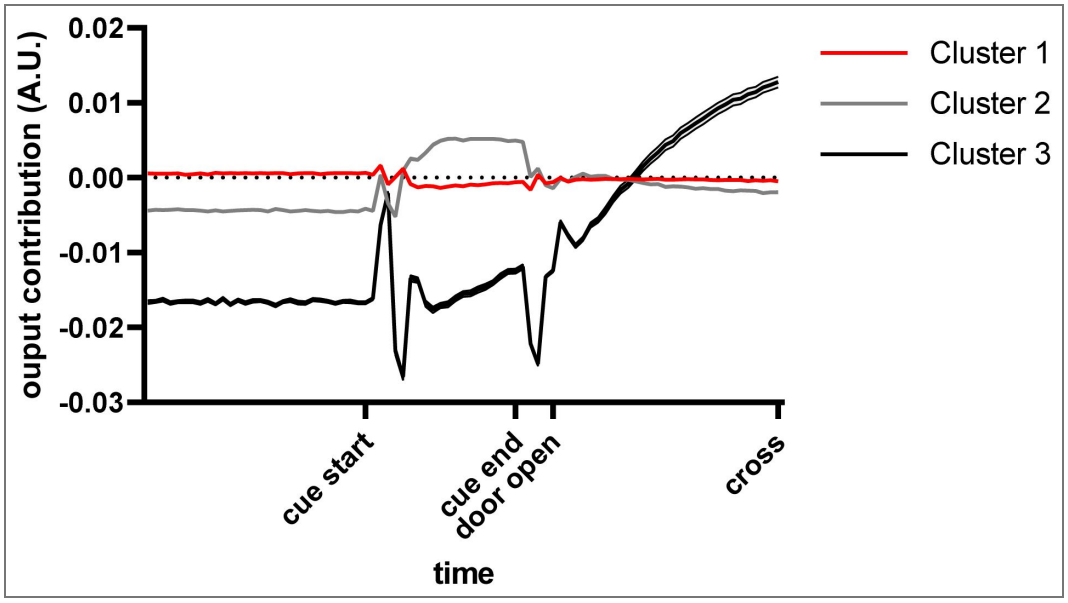

Figure S2. Contribution of each cluster to the output during the delayed escape test

Relative contributions of Clusters 1–3 to the network output across behavioral periods in the delayed escape test.

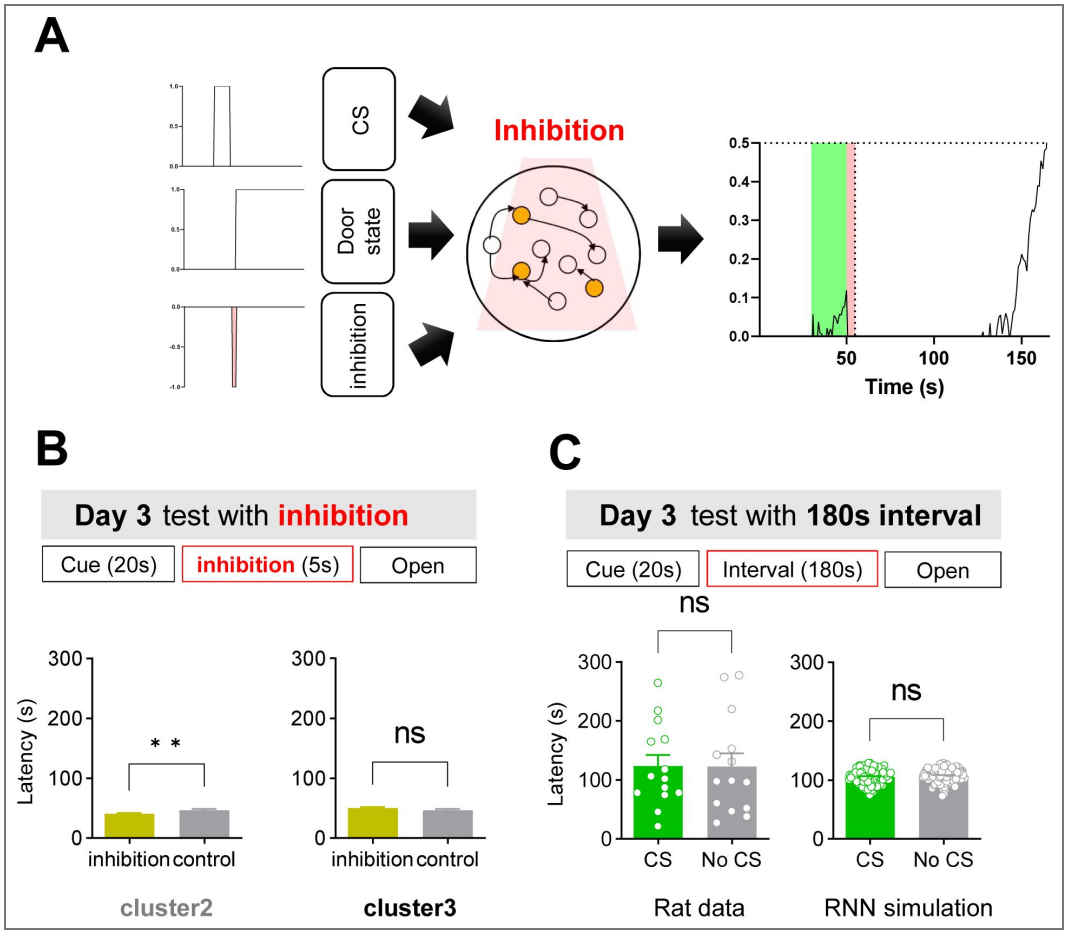

Figure S3. RNN activity of Clusters 2 and 3 under inhibition and escape latency under 180-s delay condition

(A) Schematic of the inhibition-simulation protocol. In the RNN model, inhibition was applied for 5 s during the delay interval after CS offset and before the door-opening signal. (B) Bar graphs showing crossing latency under inhibition of Cluster 2 or Cluster 3 neurons.: Cluster 2, **p = 0.0023; Cluster 3, p = 0.1511 (both Mann–Whitney). (C) Cross-latency under a prolonged 180-s CS–door-opening interval. Bar graphs show crossing latencies in rats (left) and in the RNN model (right): rats, p = 0.9720 (unpaired *t*-test); RNN, p = 0.3574 (unpaired *t*-test).

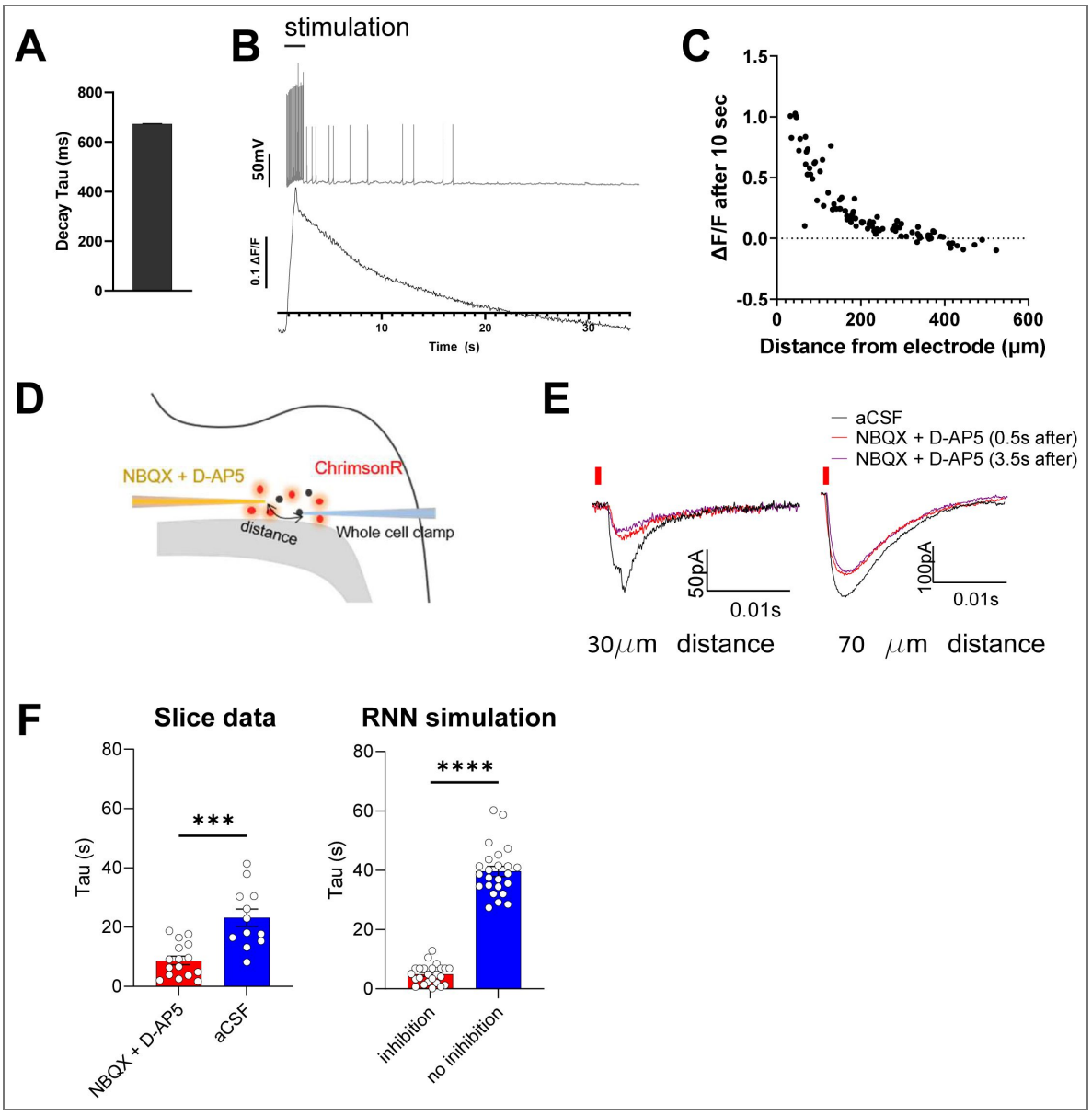

Figure S4. Slice physiology and pharmacological manipulation of claustral persistent activity

(A) Membrane-potential decay τ during whole-cell recordings. (B) Top: voltage trace showing persistent spiking after 20 Hz, 200 μA, 1-s stimulation; Bottom: corresponding GCaMP6f fluorescence changes recorded from the same cell. (C) Relationship between the distance from the stimulating electrode to the recorded cell and the persistence ratio (peak fluorescence / fluorescence at 10 s), calculated from the dataset shown in Fig. 2J. (D) Schematic of pressure injection of drugs onto whole-cell patch-clamped neurons in horizontal claustral slices. (E) Representative EPSCs recorded under control (aCSF, no pressure injection) and after NBQX + D-AP5 pressure injection with the drug pipette positioned 30 μm or 70 μm from the recorded cell. EPSCs were evoked by electrical stimulation at 0.5 s and 3.5 s after injection. (F) Decay kinetics measured from the graphs in Fig. 2M and N under NBQX + D-AP5 or aCSF treatment. Left, decay τ after application of NBQX + D-AP5 (n = 16) or aCSF (n = 12) 10 s after electrical stimulation; **p = 0.0012, Mann–Whitney. Right, decay τ with versus without recurrent inhibition in the RNN (n = 24 each); ****p < 0.0001, Mann–Whitney.

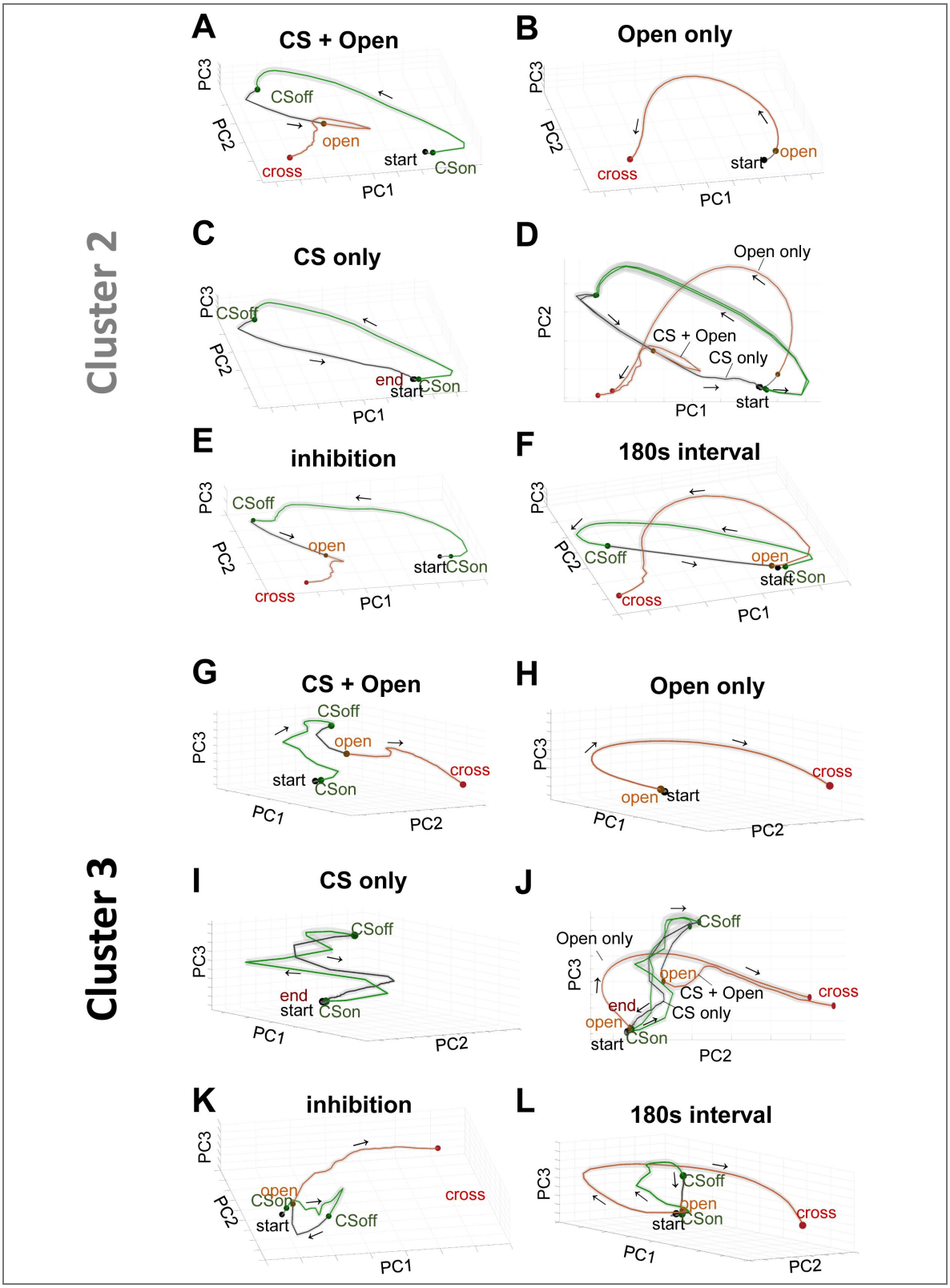

Figure S5. PCA trajectories of Clusters 2 and 3 under various simulation conditions

(A–F) Cluster 2: time-normalized, trial-averaged PCA trajectories for (A) CS + door-opening, (B) door opening only, (C) CS only, (D) overlay of A–C, (E) inhibition simulation, and (F) 180 s delay between CS offset and door-opening. (G–L) Cluster 3 trajectories arranged as in panels A–F.

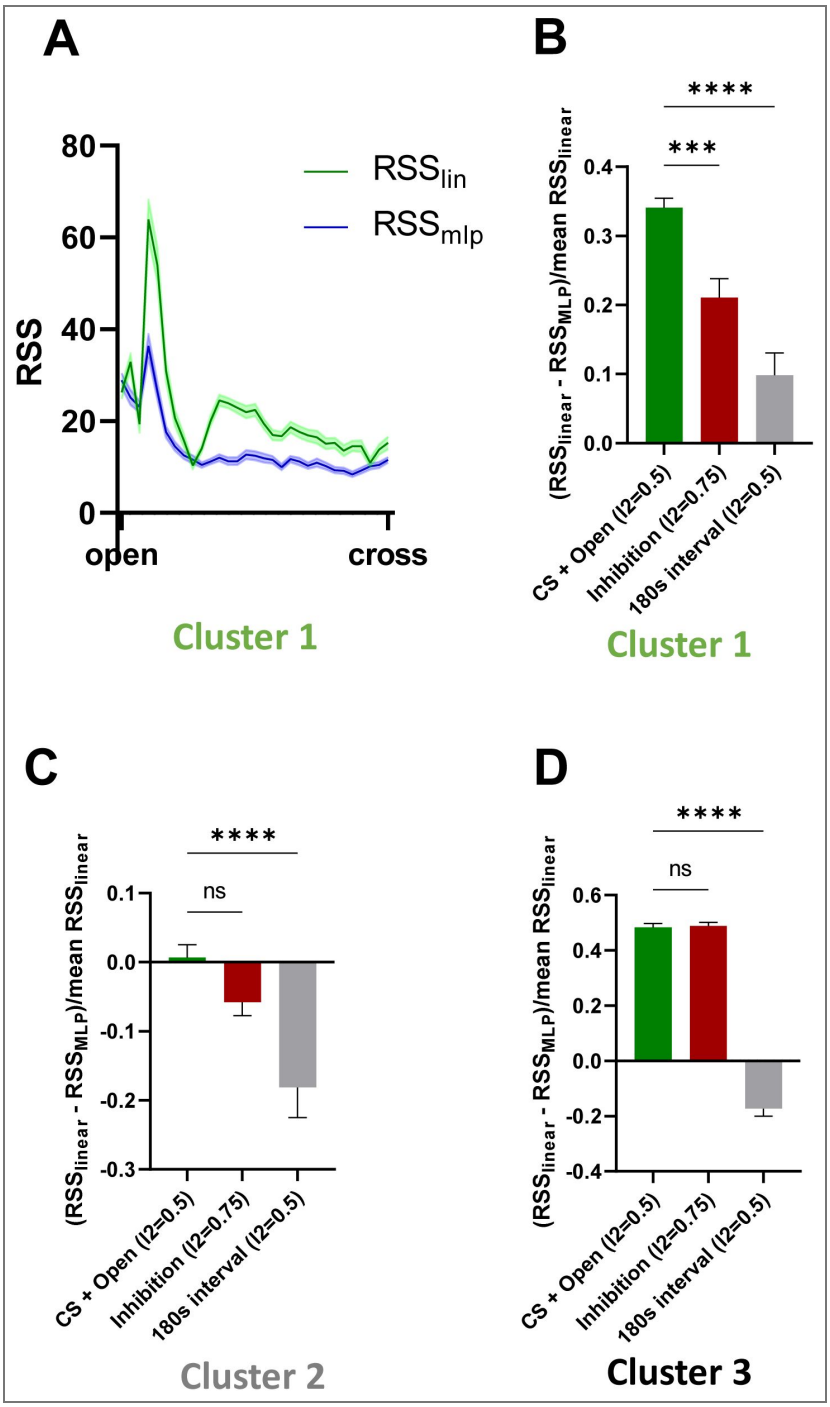

Figure S6. Model-fit comparison across clusters

(A) Residual sum of squares (RSS) from a linear regression model and a multilayer perceptron (MLP) across normalized time bins. Data are taken from the Open event to the Cross event (for the CS only case, from the open event to the mean cross latency of the CS + door-opening condition). Green line: RSS of the linear regression model; blue line: RSS of the MLP. (B–D) Difference in residual sum of squares (ΔRSS) between linear regression and MLP fits, normalized by the mean RSS of the linear model, across conditions.; bars, mean ± SEM. (B) Cluster 1: CS + door-opening vs inhibition, **p = 0.0063; CS + door-opening vs 180 s delay, ****p < 0.0001. (C) Cluster 2: CS + door-opening vs inhibition, p = 0.6735; CS + door-opening vs 180 s delay, **p = 0.0021. (D) Cluster 3: CS + door-opening vs inhibition, p = 0.4067; CS + door-opening vs 180 s delay, ****p < 0.0001 (one-way ANOVA with Holm–Šidák).

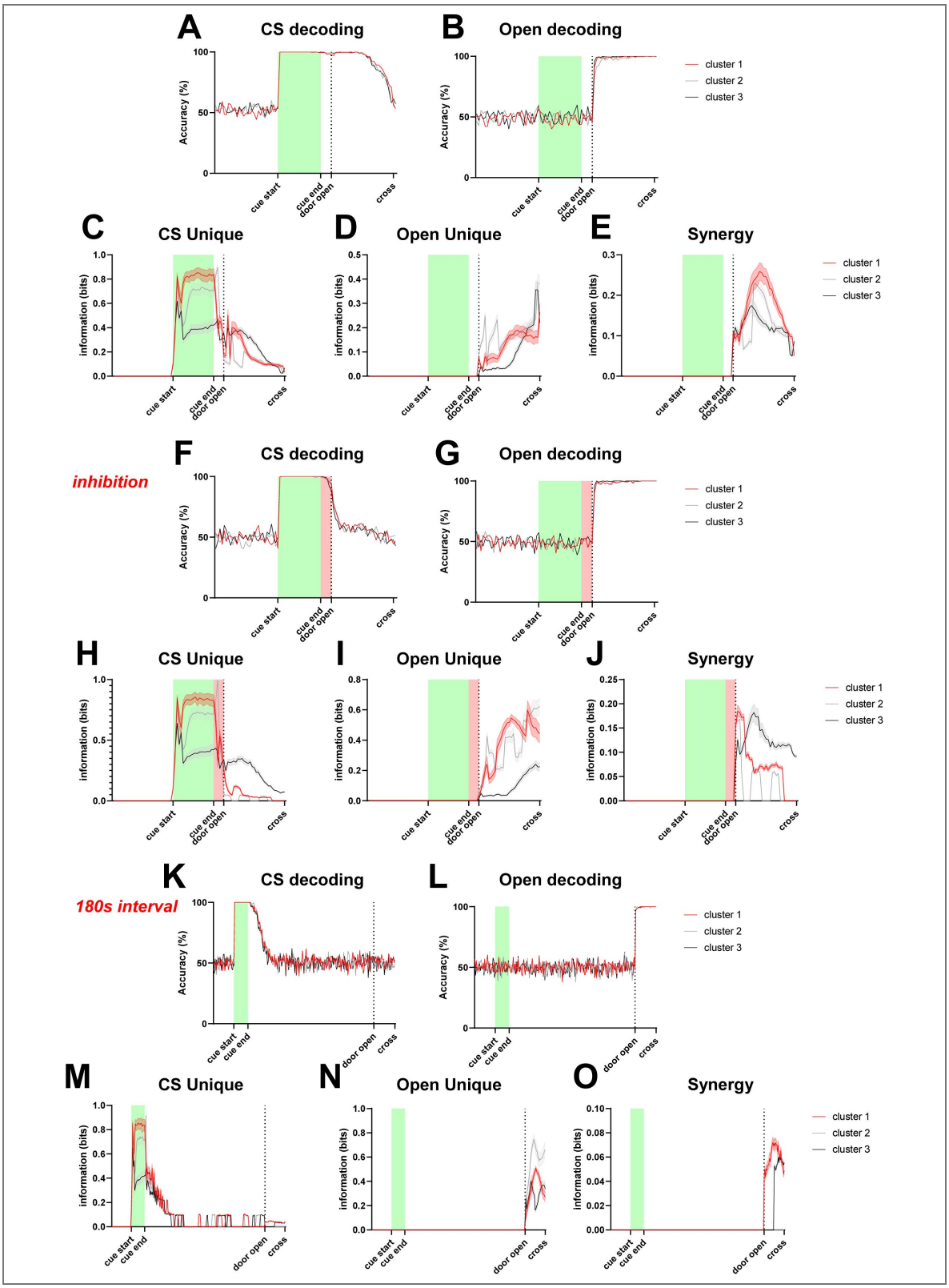

Figure S7. Decoding performance and PID analysis across clusters

(A) Decoding accuracy for classifying CS-present versus CS-absent trials using cluster-specific decoders. (B) Decoding accuracy for classifying Open-present versus Open-absent trials using cluster-specific decoders. (C–E) PID metrics (CS-unique, Open-unique, synergy) for neurons in each cluster. (F–J) Same analyses as in panels A–E under the inhibition condition. (K–O) Same analyses as in panels A–E under a 180 s CS–Open interval.

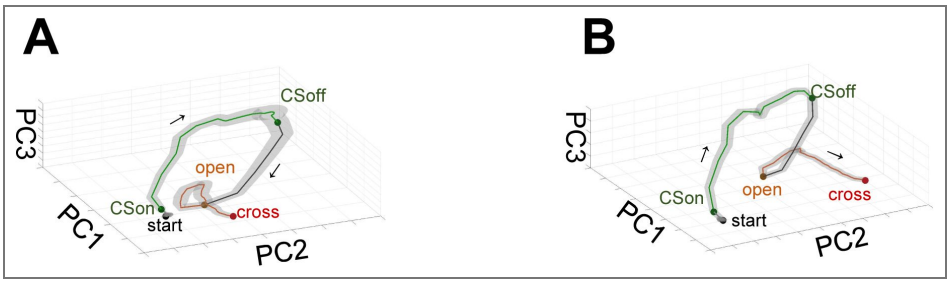

Figure S8. PCA trajectories of high- and low-synergy neurons

(A–B) Time-normalized, trial-averaged PCA trajectories for (A) high-synergy and (B) low-synergy Cluster 1 RNN neurons.

Data availability

All data and code used in this study are available upon request.

Acknowledgements

This work was supported by the NRF (*https://www.nrf.re.kr*) of Korea grant funded by the Korean Government Ministry of Education (*https://english.moe.go.kr*), Science and Technology (NRF-2021R1A2B5B03002345), awarded to SC.

Additional information

Funding

Funder	Grant reference number	Author
National Research Foundation of Korea (NRF)	NRF-2021R1A2B5B03002345	Junghwa Lee

Author ORCID iDs

Kuenbae Sohn: https://orcid.org/0000-0001-6301-6403
Sukwoo Choi: https://orcid.org/0000-0002-6445-4912

References

1. **Crick F. C.,** Koch C. (2005) What is the function of the claustrum?. *Philos Trans R Soc Lond B Biol Sci* **360**:1271-1279 https://doi.org/10.1098/rstb.2005.1661 | PubMed
2. **Shelton A. M.,** et al. (2025) Single neurons and networks in the mouse claustrum integrate input from widespread cortical sources. *eLife* **13** https://doi.org/10.7554/elife.98002 | PubMed
3. **White M. G.,** et al. (2018) Anterior Cingulate Cortex Input to the Claustrum Is Required for Top-Down Action Control. *Cell Rep* **22**:84-95 https://doi.org/10.1016/j.celrep.2017.12.023 | PubMed
4. **Atlan G.,** et al. (2024) Claustrum neurons projecting to the anterior cingulate restrict engagement during sleep and behavior. *Nat Commun* **15**:5415 https://doi.org/10.1038/s41467-024-48829-6 | PubMed
5. **Faig C. A.,** et al. (2024) Claustrum projections to the anterior cingulate modulate nociceptive and pain-associated behavior. *Curr Biol* **34**:1987-1995.e1984 https://doi.org/10.1016/j.cub.2024.03.044 | PubMed
6. **Niu M.,** et al. (2022) Claustrum mediates bidirectional and reversible control of stress-induced anxiety responses. *Sci Adv* **8**:eabi6375 https://doi.org/10.1126/sciadv.abi6375 | PubMed
7. **Chevee M.,** Finkel E. A., Kim S. J., O'Connor D. H., Brown S. P. (2022) Neural activity in the mouse claustrum in a cross-modal sensory selection task. *Neuron* **110**:486-501.e487 https://doi.org/10.1016/j.neuron.2021.11.013 | PubMed
8. **Narikiyo K.,** et al. (2020) The claustrum coordinates cortical slow-wave activity. *Nat Neurosci* **23**:741-753 https://doi.org/10.1038/s41593-020-0625-7 | PubMed
9. **White M. G.,** et al. (2020) The Mouse Claustrum Is Required for Optimal Behavioral Performance Under High Cognitive Demand. *Biol Psychiatry* **88**:719-726 https://doi.org/10.1016/j.biopsych.2020.03.020 | PubMed
10. **Qadir H.,** et al. (2022) The mouse claustrum synaptically connects cortical network motifs. *Cell Rep* **41** https://doi.org/10.1016/j.celrep.2022.111860 | PubMed
11. **Goll Y.,** Atlan G., Citri A. (2015) Attention: the claustrum. *Trends Neurosci* **38**:486-495 https://doi.org/10.1016/j.tins.2015.05.006 | PubMed
12. **Atlan G.,** et al. (2018) The Claustrum Supports Resilience to Distraction. *Curr Biol* **28**:2752-2762.e2757 https://doi.org/10.1016/j.cub.2018.06.068 | PubMed

13. **Terem A.,** et al. (2020) Claustral Neurons Projecting to Frontal Cortex Mediate Contextual Association of Reward. *Curr Biol* **30**:3522-3532.e3526 https://doi.org/10.1016/j.cub.2020.06.064 | PubMed
14. **Madden M. B.,** et al. (2025) Serotonin and psilocybin activate 5-HT(1B) receptors to suppress cortical signaling through the claustrum. *Nat Commun* **16**:7733 https://doi.org/10.1038/s41467-025-62980-8 | PubMed
15. **Chong M. H.,** Gămănuţ R. (2024) Anatomical and physiological characteristics of claustrum neurons in primates and rodents. *Frontiers in Mammal Science* **3** https://doi.org/10.3389/fmamm.2024.1309665
16. **Wang Q.,** et al. (2017) Organization of the connections between claustrum and cortex in the mouse. *J Comp Neurol* **525**:1317-1346 https://doi.org/10.1002/cne.24047 | PubMed
17. **Han Y.,** et al. (2024) Delayed escape behavior requires claustral activity. *Cell Rep* **43** https://doi.org/10.1016/j.celrep.2024.113748 | PubMed
18. **Dillingham C. M.,** et al. (2019) The Anatomical Boundary of the Rat Claustrum. *Front Neuroanat* **13** https://doi.org/10.3389/fnana.2019.00053 | PubMed
19. **Zhang X.,** et al. (2001) Susceptibility to kindling and neuronal connections of the anterior claustrum. *J Neurosci* **21**:3674-3687 https://doi.org/10.1523/jneurosci.21-10-03674.2001 | PubMed
20. **Jankowski M. M.,** O'Mara S. M. (2015) Dynamics of place, boundary and object encoding in rat anterior claustrum. *Front Behav Neurosci* **9** https://doi.org/10.3389/fnbeh.2015.00250 | PubMed
21. **Jankowski M. M.,** Islam M. N., O'Mara S. M. (2017) Dynamics of spontaneous local field potentials in the anterior claustrum of freely moving rats. *Brain Res* **1677**:101-117 https://doi.org/10.1016/j.brainres.2017.09.021 | PubMed
22. **Grasby K.,** Talk A. (2013) The anterior claustrum and spatial reversal learning in rats. *Brain Res* **1499**:43-52 https://doi.org/10.1016/j.brainres.2013.01.014 | PubMed
23. **Park S.,** Sohn K., Yoon D., Lee J., Choi S. (2025) Single-unit activity in the anterior claustrum during memory retrieval after trace fear conditioning. *PLoS One* **20**:e0318307 https://doi.org/10.1371/journal.pone.0318307 | PubMed
24. **Ehrlich D. B.,** Stone J. T., Brandfonbrener D., Atanasov A., Murray J. D. (2021) PsychRNN: An Accessible and Flexible Python Package for Training Recurrent Neural Network Models on Cognitive Tasks. *eNeuro* **8** https://doi.org/10.1523/eneuro.0427-20.2020 | PubMed
25. **Ehrlich D. B.,** Murray J. D. (2022) Geometry of neural computation unifies working memory and planning. *Proc Natl Acad Sci U S A* **119**:e2115610119 https://doi.org/10.1073/pnas.2115610119 | PubMed
26. **Kim R.,** Sejnowski T. J. (2021) Strong inhibitory signaling underlies stable temporal dynamics and working memory in spiking neural networks. *Nat Neurosci* **24**:129-139 https://doi.org/10.1038/s41593-020-00753-w | PubMed
27. **Stroud J. P.,** Watanabe K., Suzuki T., Stokes M. G., Lengyel M. (2023) Optimal information loading into working memory explains dynamic coding in the prefrontal cortex. *Proceedings of the National Academy of Sciences* **120**:e2307991120 https://doi.org/10.1073/pnas.2307991120 | PubMed
28. **Brody C. D.,** Romo R., Kepecs A. (2003) Basic mechanisms for graded persistent activity: discrete attractors, continuous attractors, and dynamic representations. *Current opinion in neurobiology* **13**:204-211 https://doi.org/10.1016/s0959-4388(03)00050-3 | PubMed
29. **Chaisangmongkon W.,** Swaminathan S. K., Freedman D. J., Wang X.-J. (2017) Computing by robust transience: how the fronto-parietal network performs sequential, category-based decisions. *Neuron* **93**:1504-1517.e1504 https://doi.org/10.1016/j.neuron.2017.03.002 | PubMed
30. **Mohan K.,** Zhu O., Freedman D. J. (2021) Interaction between neuronal encoding and population dynamics during categorization task switching in parietal cortex. *Neuron* **109**:700-712.e704 https://doi.org/10.1016/j.neuron.2020.11.022 | PubMed
31. **Stroud J. P.,** Duncan J., Lengyel M. (2024) The computational foundations of dynamic coding in working memory. *Trends Cogn Sci* **28**:614-627 https://doi.org/10.1016/j.tics.2024.02.011 | PubMed

32. **Orman R.** (2015) Claustrum: a case for directional, excitatory, intrinsic connectivity in the rat. *J Physiol Sci* **65**:533-544 https://doi.org/10.1007/s12576-015-0391-6 | PubMed

33. **Kim J.,** Matney C. J., Roth R. H., Brown S. P. (2016) Synaptic Organization of the Neuronal Circuits of the Claustrum. *J Neurosci* **36**:773-784 https://doi.org/10.1523/jneurosci.3643-15.2016 | PubMed

34. **Voigts J.,** et al. (2025) Spatial reasoning via recurrent neural dynamics in mouse retrosplenial cortex. *Nat Neurosci* **28**:1293-1299 https://doi.org/10.1038/s41593-025-01944-z | PubMed

35. **Yu B. M.,** et al. (2009) Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. *J Neurophysiol* **102**:614-635 https://doi.org/10.1152/jn.90941.2008 | PubMed

36. **Lakshmanan K. C.,** Sadtler P. T., Tyler-Kabara E. C., Batista A. P., Yu B. M. (2015) Extracting Low-Dimensional Latent Structure from Time Series in the Presence of Delays. *Neural Comput* **27**:1825-1856 https://doi.org/10.1162/neco_a_00759 | PubMed

37. **Timme N. M.,** Lapish C. (2018) A Tutorial for Information Theory in Neuroscience. *eNeuro* **5** https://doi.org/10.1523/eneuro.0052-18.2018 | PubMed

38. **Timme N.,** Alford W., Flecker B., Beggs J. M. (2014) Synergy, redundancy, and multivariate information measures: an experimentalist's perspective. *J Comput Neurosci* **36**:119-140 https://doi.org/10.1007/s10827-013-0458-4 | PubMed

39. **Golub M. D.,** Sussillo D. (2018) FixedPointFinder: A Tensorflow toolbox for identifying and characterizing fixed points in recurrent neural networks. *Journal of open source software* **3**:1003 https://doi.org/10.21105/joss.01003

40. **Brody C. D.,** Hernández A., Zainos A., Romo R. (2003) Timing and neural encoding of somatosensory parametric working memory in macaque prefrontal cortex. *Cerebral cortex* **13**:1196-1207 https://doi.org/10.1093/cercor/bhg100 | PubMed

41. **Meyers E. M.** (2018) Dynamic population coding and its relationship to working memory. *J Neurophysiol* **120**:2260-2268 https://doi.org/10.1152/jn.00225.2018 | PubMed

42. **Naghavi H. R.,** Eriksson J., Larsson A., Nyberg L. (2007) The claustrum/insula region integrates conceptually related sounds and pictures. *Neurosci Lett* **422**:77-80 https://doi.org/10.1016/j.neulet.2007.06.009 | PubMed

43. **Reus-Garcia M. M.,** et al. (2021) The Claustrum is Involved in Cognitive Processes Related to the Classical Conditioning of Eyelid Responses in Behaving Rabbits. *Cereb Cortex* **31**:281-300 https://doi.org/10.1093/cercor/bhaa225 | PubMed

44. **Ollerenshaw D. R.,** et al. (2021) Anterior claustrum cells are responsive during behavior but not passive sensory stimulation. *BioRxiv* https://doi.org/10.1101/2021.03.23.436687

45. **Stewart B. W.,** et al. (2024) Pathological claustrum activity drives aberrant cognitive network processing in human chronic pain. *Curr Biol* **34**:1953-1966.e1956 https://doi.org/10.1016/j.cub.2024.03.021 | PubMed

46. **Froesel M.,** et al. (2024) Macaque claustrum, pulvinar and putative dorsolateral amygdala support the cross-modal association of social audio-visual stimuli based on meaning. *Eur J Neurosci* **59**:3203-3223 https://doi.org/10.1111/ejn.16328 | PubMed

47. **Seguin C.,** Sporns O., Zalesky A. (2023) Brain network communication: concepts, models and applications. *Nat Rev Neurosci* **24**:557-574 https://doi.org/10.1038/s41583-023-00718-5 | PubMed

48. **Dehaene S.,** Kerszberg M., Changeux J. P. (1998) A neuronal model of a global workspace in effortful cognitive tasks. *Proc Natl Acad Sci U S A* **95**:14529-14534 https://doi.org/10.1073/pnas.95.24.14529 | PubMed

49. **Han Y.,** et al. (2024) Delayed escape behavior requires claustral activity. *Cell Reports* **43** https://doi.org/10.1016/j.celrep.2024.113748 | PubMed

50. **Lakshmanan K. C.,** Sadtler P. T., Tyler-Kabara E. C., Batista A. P., Yu B. M. (2015) Extracting low-dimensional latent structure from time series in the presence of delays. *Neural computation* **27**:1825-1856 https://doi.org/10.1162/neco_a_00759 | PubMed

Peer reviews

Reviewer #1 (Public review):

Summary:

In this manuscript, the authors investigate how the anterior claustrum may integrate temporally separated task-relevant signals to guide behavior in a delayed escape paradigm. Because in vivo neural recordings from claustrum during this task are extremely limited-comprising single-trial data with small neuronal samples-the authors adopt a modeling-driven approach. They train recurrent neural networks (RNNs) using only behavioral data (escape latency) to reproduce task performance and then analyze the internal dynamics of the trained networks. Within these networks, they identify a subset of units whose activity exhibits persistent responses and strong correlations with behavior, which the authors label as "claustrum-like." Using dimensionality reduction, decoding, and information-theoretic analyses, they argue that these units dynamically integrate conditioned stimulus (CS) and door-opening signals via nonlinear, trajectory-based population dynamics rather than fixed-point attractor states.

To bridge model predictions and biology, the authors complement the modeling with in vitro slice experiments demonstrating recurrent excitatory connectivity and prolonged activity in the anterior claustrum that depends on glutamatergic transmission. They further compare latent neural trajectories derived from previously published in vivo claustrum recordings to those observed in the RNN, reporting qualitative similarities. Based on these results, the authors propose that the claustrum implements temporal signal integration through recurrent excitatory circuitry and dynamic population trajectories, potentially supporting broader theories of integrative brain function.

Strengths:

This study addresses an important and challenging problem: how to infer population-level computation in a brain structure for which in vivo data are sparse and experimentally constrained. The authors are commendably transparent about these limitations and seek to overcome them through a principled modeling framework. The integration of behavioral modeling, RNN analysis, and slice electrophysiology is ambitious and technically sophisticated.

Several aspects stand out as strengths. First, the behavioral RNN is carefully trained and interrogated using a rich set of modern analytical tools, including cross-temporal decoding, trajectory analysis, and partial information decomposition, providing multiple complementary views of network dynamics. Second, the slice experiments convincingly demonstrate recurrent excitatory connectivity in anterior claustrum, lending biological plausibility to the model's reliance on recurrent dynamics. Third, the manuscript is clearly written, logically organized, and conceptually engaging, and it offers a coherent mechanistic hypothesis that could guide future large-scale recording experiments.

Importantly, the work has significant heuristic value: rather than merely fitting data, it attempts to generate testable computational ideas about claustral function in a regime where direct empirical access is currently limited.

Weaknesses:

Despite these strengths, the manuscript suffers from a recurring and substantial conceptual issue: systematic over-interpretation of model-data correspondence. While the modeling results are potentially insightful, the extent to which they are presented as recapitulating real claustral neural mechanisms goes beyond what the available data can support.

A fundamental limitation is that the RNN is trained solely on behavioral output, without being constrained by neural data at either single-unit or population levels. As a result, the internal network dynamics are underdetermined and non-unique. Many distinct internal solutions could plausibly generate identical behavior. However, the manuscript frequently treats the specific internal solution discovered in the RNN as if it were a close approximation of the actual claustrum circuit.

This issue is compounded by the sparse nature of the in vivo data used for comparison. The GPFA-based trajectory analyses rely on pseudo-populations and single-trial recordings, yet are interpreted as evidence for robust population-level dynamics. Because neurons were not recorded simultaneously, the inferred trajectories necessarily lack true population covariance and shared trial-to-trial variability, limiting their interpretability as genuine population dynamics. Similarly, conclusions about trajectory-based versus attractor-based computation are drawn almost exclusively from model analyses and then generalized to the biological system.

Overall, while the modeling framework is appropriate as a hypothesis-generating tool, the manuscript repeatedly crosses the line from proposing plausible mechanisms to asserting explanatory or even causal equivalence between the model and the brain. This undermines the otherwise strong contributions of the work.

Below are several specific points that warrant further clarification or revision:

(1) Tone of model-data correspondence

Numerous statements describe the RNN as "closely mimicking," "recapitulating," or being "nearly identical" to claustral neural dynamics, sometimes extending to claims about causal relationships between neural activity and behavior. Given that neural data were not used to train the model, and that only a small subset of trained networks showed the reported dynamics, these statements should be substantially softened throughout the manuscript. The RNN should be framed as providing one possible computational realization consistent with existing data, not as a close instantiation of the biological circuit.

(2) Non-uniqueness of RNN solutions

The fact that only a small fraction of trained networks exhibited "claustrum-like" clusters deserves deeper discussion. This observation raises the possibility that the identified solution is fragile or highly specific rather than canonical. The authors should explicitly discuss the non-uniqueness of internal solutions in behavior-trained RNNs, including the range of alternative network dynamics that can reproduce the same behavior. In particular, it should be clarified why the specific network exhibiting "claustrum-like" clusters is informative about claustral computation, rather than representing one arbitrary solution among many.

(3) GPFA trajectory comparisons

The qualitative similarity between RNN trajectories and GPFA-derived trajectories from sparse in vivo data is interesting but insufficient to support claims of robustness or population-level structure. Statements suggesting that these patterns are unlikely to arise from noise or random fluctuations are not justified given the single-trial, pseudo-population nature of the data. Either additional quantitative controls should be added, or the interpretation should be substantially tempered.

(4) Scope of functional claims

The discussion connecting the findings to broad theories of claustral function, global workspace, or consciousness extends well beyond the data presented. These speculative links should be clearly labeled as such and significantly reduced in strength and prominence.

The manuscript repeatedly describes the delayed escape task as an "inference-based behavioral paradigm" and states that animals "infer that a value-neutral alternative space is likely to be safer" when the CS is presented in a novel environment. While I appreciate that the US-CS association was established in a different context and that the CS is then presented in a new environment, I am not convinced that the current behavioral evidence uniquely supports an inference interpretation.

First, it is not clear that this task is widely recognized in the literature as a canonical inference task, in the sense of, for example, sensory preconditioning, transitive inference, or model-based inference paradigms. Rather, the observed effect-that CS animals escape faster to a neutral compartment than neutral-CS controls-can be parsimoniously interpreted in terms of generalized threat value, heightened fear/anxiety, or a bias toward avoidance/escape under elevated threat, without requiring an explicit inferential step about the specific safety of the alternative compartment. The fact that no prior training is needed is compatible with flexible generalization, but does not by itself demonstrate inference in a more formal computational sense.

Second, the inference claim becomes central to the manuscript's conceptual framing (e.g., the idea that rsCla supports "inference-based escape"), yet the behavioral analyses presented here and in the cited prior work do not clearly rule out simpler accounts. Clarifying this distinction would help avoid overstating both the inferential nature of the behavior and the specific role of rsCla and the RNN's "claustrum-like" cluster in supporting inference per se, as opposed to more general integration of threat-related signals with an opportunity for escape.

This manuscript presents an interesting and potentially valuable modeling-based framework for thinking about temporal integration in the claustrum, supported by solid slice physiology. However, in its current form, it overstates the degree to which the proposed RNN dynamics reflect actual claustral neural mechanisms. With substantial revision-especially a more cautious interpretation of model-data similarity and a clearer articulation of modeling limitations-the study could make a meaningful contribution as a hypothesis-generating work rather than a definitive mechanistic account.

Comments on revisions:

The authors have carefully addressed the concerns raised in the initial review. In particular, the manuscript has been substantially improved in terms of tone, conceptual clarity, and the interpretation of the modeling results. The revised version now presents a well-balanced and appropriately framed account of the work.

The study offers a compelling and useful hypothesis-generating framework for understanding temporal integration in the claustrum, and I support its publication. As a minor point, given the acknowledged limitations of pseudo-population and single-trial data, it would be preferable to slightly soften a few remaining statements that describe trajectory structure as directly "reflecting" population-level dynamics (e.g., using "consistent with" instead).

https://doi.org/10.7554/eLife.109539.2.sa2

Reviewer #2 (Public review):

This manuscript reports the behavior of a computational model of rat claustral neurons during the performance of a behavioral task known as the delayed escape task (in this reviewer's understanding, this behavioral task was created and implemented by this group only). These authors have argued in a prior manuscript (Han et al.) that a group of neurons located "rostral to striatum" are part of the claustrum. The group names the region the

"rostral to striatum claustrum." Additionally, in the Han et al. paper, the authors argue that these cells are responsible for maintaining a signal that lasts through the delay period.

The main findings of the current paper are:

(1) The authors have built a model network that was trained to show firing similar to what was reported for rats in their prior paper.

(2) The authors' analysis of model behavior is used to suggest that the model network recapitulates biological activity, including the existence of a cluster of cells mainly responsible for the delay period firing.

(3) The authors offer evidence from patch clamp recordings for excitatory interconnections among claustral neurons that are an essential feature of the model network.

A major value of the computational network is that "trials" of the network can be performed. In experiments on animals, only single trials can be used.

Concerns:

(1) This paper is based on behavioral results and neural recordings from their prior paper (Han et al.), but data, e.g. in figure 1, are not clearly identified as new or as coming from that source. Figure 1A, for example, appears to be taken directly from Han et al. No methods are given in this manuscript for the behavioral testing or the in vivo electrophysiology.

(2) Many other details are unclear. Examples include model training, the weight matrices and how these changed with training (p. 13), the equations 2 and 3 (p. 13), the sources for the constants in the equations (p. 14), the methods (anesthesia, stereotaxic coordinates, injection specifics and details for "sparse expression") for the ChrimsonR injections.

(3) The explorations of model behavior are a catalog of everything tried rather than an organized demonstration of what the model can and cannot do. The figures could be reduced in number to emphasize the key comparisons of the different clusters and the model's behavior under different conditions intended to "test" the model.

(4) On page 6, the E-E connectivity is argued from Shelton et al. (2025) and against Kim et al. (2016), but ignores Orman (2015), which to this reviewer's knowledge was the first to demonstrate such connectivity, including the long duration events and impact of planes of section.

(5) Whereas the authors are entitled to their own opinion of prior work (references 3-8), it is inappropriate to misrepresent prior work as only demonstrating a "limited function" of claustum. Additional papers by Mathur's group and Citri's group are ignored.

In summary, the authors have made a computational model that recapitulates the firing of a subset of potentially claustral neurons during a particular behavioral task (delayed escape is certainly not the only behavior that involves claustrum - see e.g., attention, salience, sleep). If the conclusion is that excitatory claustral cells must be connected to other excitatory claustral cells, such a conclusion is not new and the electrophysiological E-E metrics are not well quantified (e.g., connectivity frequency, strength of connection). If the model is intended to predict how claustrum might accomplish any other task, there is insufficient detail to evaluate the model beyond the evidence that the model creates a subset of cells that can sustain firing during the delay period in the delayed escape task.

All relevant work must be appropriately cited throughout the manuscript.

Comments on revisions:

The authors have adequately addressed the concerns that were raised in response to the first version of the manuscript.

https://doi.org/10.7554/eLife.109539.2.sa1

Author response:

The following is the authors' response to the original reviews

> ***Public Reviews:***
>
> ***Reviewer #1 (Public review):***

We thank the reviewer for their constructive and insightful comments and agree with the importance of the points raised. We recognize that aspects of our original presentation may have been unclear or overly strong in their interpretation. We have therefore revised the manuscript to clarify our intended scope, moderate our claims, and strengthen the analysis. In the second paragraph of the Discussion, we have explicitly acknowledged the concerns raised by the reviewer and outlined how they have been addressed in the revised manuscript. Our detailed responses are provided below.

> *(1) Tone of model-data correspondence*
>
> *Numerous statements describe the RNN as "closely mimicking," "recapitulating," or being "nearly identical" to claustral neural dynamics, sometimes extending to claims about causal relationships between neural activity and behavior. Given that neural data were not used to train the model, and that only a small subset of trained networks showed the reported dynamics, these statements should be substantially softened throughout the manuscript. The RNN should be framed as providing one possible computational realization consistent with existing data, not as a close instantiation of the biological circuit*

We agree with the reviewer's comment. The expressions noted by the reviewer (e.g., closely mimicked, nearly identical, recapitulate) will be replaced with alternative wording that conveys a more moderate meaning (Line16-17, 65-66, 83, 96, 120, 212).

> *(2) Non-uniqueness of RNN solutions*
>
> *The fact that only a small fraction of trained networks exhibited "claustrum-like" clusters deserves deeper discussion. This observation raises the possibility that the identified solution is fragile or highly specific rather than canonical. The authors should explicitly discuss the non-uniqueness of internal solutions in behavior-trained RNNs, including the range of alternative network dynamics that can reproduce the same behavior. In particular, it should be clarified why the specific network exhibiting "claustrum-like" clusters is informative about claustral computation, rather than representing one arbitrary solution among many.*

As the reviewer pointed out, behaviorally trained RNNs can admit multiple internal solutions that produce the same behavioral output, and we acknowledge the non-uniqueness of such internal solutions. However, we do not interpret the fact that only a subset of trained RNNs exhibit dynamics similar to those observed in the claustrum as evidence that this solution is fragile. Notably, the claustrum-like dynamics emerged spontaneously during training and were not explicitly enforced. Furthermore, our finding suggests that the emergence of this particular dynamical regime depends on relatively specific structural constraints.

Our criterion for selecting RNNs that could inform the computational principles of the claustrum was their ability to reproduce the behavioral and physiological observations

obtained in the delayed escape experiments. RNNs that were excluded may reflect information-processing strategies used by other brain regions or may rely on artificial logical structures. The computational demand of the task, which integrates temporally separated signals, naturally drives convergence toward networks with recurrent excitatory connectivity capable of maintaining persistent activity. Indeed, all networks that exhibited a claustrum-like cluster shared a common structural feature: strong recurrent excitatory connectivity within Cluster 1. This property is consistent with biological characteristics observed in the slice experiments shown in Fig 2.

Importantly, the computational principles derived from this RNN were found to be quantitatively consistent with in vivo single-neuron activity patterns. Specifically, analysis using an eigenvalue-based metric ($\lambda_3/\Sigma\lambda$) revealed the same directional effect in both the RNN and the claustrum neuron data. In addition, a leave-one-neuron-out analysis showed that this pattern was broadly distributed across in vivo claustral neurons rather than being driven by a small subset (see Fig. 4).

Taken together, these convergent lines of evidence suggest that the computational model is not simply one arbitrary solution among many possible alternatives, but rather implements a computational principle that may underlie claustral functions.

> *(3) GPFA trajectory comparisons*
>
> *The qualitative similarity between RNN trajectories and GPFA-derived trajectories from sparse in vivo data is interesting but insufficient to support claims of robustness or population-level structure. Statements suggesting that these patterns are unlikely to arise from noise or random fluctuations are not justified, given the single-trial, pseudo-population nature of the data. Either additional quantitative controls should be added, or the interpretation should be substantially tempered.*

As the reviewer pointed out, the GPFA trajectory comparison presented in the original manuscript remained largely qualitative, and we agree that this alone was insufficient to establish robustness or provide convincing evidence for population-level structure. In the revised manuscript, we have therefore added the requested quantitative analysis (see Fig. 4).

Before describing the analysis, we would like to clarify several methodological limitations associated with pseudopopulation and single-trial data. GPFA estimates latent trajectories based on assumptions about covariance structure among neurons and temporal smoothness. In pseudopopulation datasets, the true simultaneously recorded covariance structure cannot be fully reconstructed, which is an inherent limitation. Because our dataset is based on single trials, the analysis does not directly exploit trial-to-trial variability. Nevertheless, the estimation of the latent space still depends on the covariance structure among real claustral neurons, suggesting that the inferred trajectories remain tied to biologically meaningful population dynamics.

Accordingly, the quantitative metric we introduce is not entirely independent of the GPFA estimation step. Rather, it is intended to evaluate the geometric structure of the single-trial latent trajectories estimated by GPFA. We acknowledged this limitation in the revised manuscript.

Specifically, for the biological data, we reanalyzed the GPFA-derived latent trajectories in PCA space and computed an eigenvalue-based metric ($\lambda_3/\Sigma\lambda$). For each of the 20 time bins, we applied a sliding window of 10 bins and calculated the covariance matrix within that window. The eigenvalues of PC1, PC2, and PC3 were then obtained, and the third eigenvalue (λ_3) was normalized by the total variance ($\Sigma\lambda = \lambda_1 + \lambda_2 + \lambda_3$). This metric quantifies the degree to which the trajectory locally deviates from a planar structure that can be explained by two dominant axes. An increase in $\lambda_3/\Sigma\lambda$ indicates that the population-state trajectory forms a higher-dimensional geometric structure beyond a simple two-dimensional combination.

For the RNN data, in contrast, the activity of all units can be observed simultaneously and sufficient trial repetitions are available. Therefore, GPFA was not applied; instead, PCA was performed directly on the population activity for each trial. We then computed an average trajectory across trials and applied the same $\lambda_3/\Sigma\lambda$ metric. Thus, although the initial dimensionality reduction steps differ between the two systems, the definition and calculation of the final quantitative metric are identical. The focus of the comparison is therefore not the dimensionality reduction technique itself, but the geometric dimensional structure of the population trajectories evolving over time.

Importantly, within the biological dataset, the GPFA estimation procedure, preprocessing steps, pseudopopulation construction, subsampling strategy, temporal alignment criteria, and smoothing parameters were applied identically across conditions. Likewise, the same analysis pipeline was used for all conditions in the RNN. If structural biases had been introduced during covariance estimation or dimensionality reduction, they would be expected to affect all conditions within each system similarly. Nevertheless, the $\lambda_3/\Sigma\lambda$ value was consistently and significantly higher in the CS condition than in the Neutral condition, and this directional pattern was observed in both the RNN and the claustral neuron data. This suggests that the effect reflects condition-specific differences in population dynamical structure rather than artifacts arising from a particular dimensionality reduction method.

To further test whether the observed effect might be driven by a small subset of neurons or specific neuron combinations, we performed a leave-one-neuron-out analysis on the claustrum dataset. Recomputing $\lambda_3/\Sigma\lambda$ while removing one neuron at a time showed that, in the CS group, most neurons contributed relatively evenly to this metric, whereas the Neutral group did not show such a distributed contribution pattern. This indicates that the observed three-dimensional structure is not driven by a few outlier neurons or incidental covariance patterns, but rather reflects an organized population-level phenomenon.

If the result were primarily due to structural artifacts introduced by the pseudopopulation construction or dimensionality reduction procedures, it would be unlikely for consistent selective differences to repeatedly emerge between conditions under identical analysis pipelines. The consistently higher $\lambda_3/\Sigma\lambda$ values observed in the CS condition therefore provide indirect support that this pattern reflects condition-specific population dynamics rather than estimation bias.

Taken together, these results suggest that the observed three-dimensional structure reflects condition-specific population dynamics rather than analysis artifacts. The fact that the same quantitative metric yields consistent effects in both the RNN and claustral data further strengthens the correspondence between the two systems.

> *(4) Scope of functional claims*
>
> *The discussion connecting the findings to broad theories of claustral function, global workspace, or consciousness extends well beyond the data presented. These speculative links should be clearly labeled as such and significantly reduced in strength and prominence.*

We agree with the reviewer and stated that references to these theories are speculative, while substantially reducing both their emphasis and prominence in the manuscript (Line444-446, 451).

> *(5) Comment on Conceptual Interpretation of the Behavioral Paradigm:*
>
> *The manuscript repeatedly describes the delayed escape task as an "inference-based behavioral paradigm" and states that animals "infer that a value-neutral alternative space is likely to be safer" when the CS is presented in a novel environment. While I appreciate that the US-CS association was established in a different context and that the CS is then presented in a new environment, I am not convinced that the current behavioral evidence uniquely supports an inference interpretation.*
>
> *First, it is not clear that this task is widely recognized in the literature as a canonical inference task, in the sense of, for example, sensory preconditioning, transitive inference, or model-based inference paradigms. Rather, the observed effect-that CS animals escape faster to a neutral compartment than neutral-CS controls-can be parsimoniously interpreted in terms of generalized threat value, heightened fear/anxiety, or a bias toward avoidance/escape under elevated threat, without requiring an explicit inferential step about the specific safety of the alternative compartment. The fact that no prior training is needed is compatible with flexible generalization, but does not by itself demonstrate inference in a more formal computational sense.*
>
> *Second, the inference claim becomes central to the manuscript's conceptual framing (e.g., the idea that rsCla supports "inference-based escape"), yet the behavioral analyses presented here and in the cited prior work do not clearly rule out simpler accounts. Clarifying this distinction would help avoid overstating both the inferential nature of the behavior and the specific role of rsCla and the RNN's "claustrum-like" cluster in supporting inference per se, as opposed to more general integration of threat-related signals with an opportunity for escape.*

We agree with the reviewer's concern. First, we referred to the delayed escape behavioral task as "a behavioral paradigm that requires integration of temporally separated task-relevant signals." (Line 7-8). We also removed references to the term inference throughout the manuscript (Line 46, 51, 67, 397).

> ***Reviewer #2 (Public review):***

We sincerely thank the reviewer for their constructive and insightful comments. Through the revision process, the manuscript has been substantially improved, with increased reproducibility, more appropriate acknowledgment of prior work, and a clearer and more logical presentation of the study.

> *(1) This paper is based on behavioral results and neural recordings from their prior paper (Han et al.), but data, e.g., in Figure 1, are not clearly identified as new or as coming from that source. Figure 1A, for example, appears to be taken directly from Han et al. No methods are given in this manuscript for the behavioral testing or the in vivo electrophysiology.*

We agree with the reviewer that this distinction should be made clearer. In the original manuscript, we indicated in the Figure 1 legend that panels A, D, E, F, and L (left) were reproduced from Han et al. (2024). To further clarify this point, we explicitly noted this distinction again in the main text (Line 74, 85). In addition, we described the behavioral experiments and in vivo electrophysiological recordings performed in Han et al. (2024) in the Methods section and include the appropriate citation (Line 463-530).

> *(2) Many other details are unclear. Examples include model training, the weight matrices and how these changed with training (p. 13), equations 2 and 3 (p. 13), the sources for the constants in the equations (p. 14), the methods (anesthesia, stereotaxic coordinates, injection specifics and details for "sparse expression") for the ChrimsonR injections.*

We agree with the reviewer's comment and have revised the manuscript to provide a more detailed description of the model training procedure, weight initialization, and parameter selection.

We expanded the explanation of the model training procedure and weight initialization. Specifically, the recurrent (W_{rec}) and output (W_{out}) weight matrices were initialized using a Glorot normal distribution with a standard deviation of $\sigma = \sqrt{2/(N_{in} + N_{out})}$ to ensure stable signal propagation during early training. In addition, we now explicitly describe the training algorithm and optimization procedure. The network was trained using the Adam optimizer implemented in TensorFlow (v2.1.0) with a batch size of 256 for 1.2 million training iterations, minimizing the per-trial loss function defined in the manuscript. We also explicitly stated how Dale's principle was maintained throughout training: rows in W_out corresponding to inhibitory units were zeroed out, and recurrent weights were continuously constrained so that excitatory and inhibitory neurons preserved their respective positive and negative synaptic projections. To illustrate how the weight structure evolved during training, we explicitly reference Figure 2A, which visualizes the final mean inter-cluster synaptic weights and highlights the strong recurrent connectivity that emerged within Cluster 1. Regarding Equations 2 and 3 and their constants, we clarified that the target escape times used to anchor the network were based on experimentally measured behavioral latencies (48.7 s for the CS-present condition and 111.3 s for the CS-absent condition). Furthermore, the regularization coefficients (λ = 0.01 and λ_{FR} = 0.95) were selected through a grid search procedure to maintain biologically plausible firing rates while preventing overfitting.

We detailed the surgical procedures that were previously omitted. This includes the specific anesthesia protocol (sodium pentobarbital, 50 mg/kg, i.p.), stereotaxic mounting, and the exact coordinates for the rsCla (AP +2.95, ML ±1.95, DV -3.85 mm). To define "sparse expression," we specified that the AAV was diluted 1:4 in sterile saline. Finally, we included the precise injection parameters: delivery at 20 nL/min via a pressure injection system, with the pipette left in place for 10 minutes post-infusion to ensure adequate diffusion. (Line 635, 636-639, 641-643). We have added these contents in the Methods section.

> *(3) The explorations of model behavior are a catalog of everything tried rather than an organized demonstration of what the model can and cannot do. The figures could be reduced in number to emphasize the key comparisons of the different clusters and the model's behavior under different conditions, intended to "test" the model.*

We agree with the reviewer's comment and have reorganized the figures to focus on the key results. Specifically, we separated the original figures so that they correspond to (1) Presentation of an RNN model consistent with the results of actual claustral recordings, (2) identification of dimensionality-reduced population activity patterns in the model, (3) comparison of these patterns with population activity patterns derived from recorded claustral neurons, (4) proposal of a nonlinear integration mechanism, and (5) the suggestion that such integration may be implemented through dynamic coding. Using this figure organization, we first identify RNN models trained on behavioral metrics whose dynamics are consistent with experimental claustral recordings. We then compare the dimensionality-reduced population activity patterns of these models with those derived from recorded claustral neurons to evaluate their biological plausibility. After selecting the models that satisfy this criterion, we perform further analyses that would be difficult to achieve using real neural recordings alone. These analyses ultimately allow us to propose dynamic coding exhibiting nonlinear integration as a plausible computational mechanism.

> *(4) On page 6, the E-E connectivity is argued from Shelton et al. (2025) and against Kim et al. (2016), but ignores Orman (2015), which, to this reviewer's knowledge, was the first to demonstrate such connectivity, including the long-duration events and impact of planes of section.*

We agree with the reviewer's suggestion and will include a reference to Orman (2015). We have clarified that neuronal activity can persist for extended periods and that such persistent activity has been observed in claustral slices prepared at a specific slicing angle (Line 144).

> *(5) Whereas the authors are entitled to their own opinion of prior work (references 3-8), it is inappropriate to misrepresent prior work as only demonstrating a "limited function" of claustrum. Additional papers by Mathur's group and Citri's group are ignored.*

We agree with the reviewer's comment and have revised the relevant sentences in the Introduction section. We also included and acknowledged the contributions of previous studies by the Mathur group and the Citri group by adding additional references to their works (Line 36, 429).

> *In summary, the authors have made a computational model that recapitulates the firing of a subset of potentially claustral neurons during a particular behavioral task (delayed escape is certainly not the only behavior that involves claustrum - see e.g., attention, salience, sleep). If the conclusion is that excitatory claustral cells must be connected to other excitatory claustral cells, such a conclusion is not new, and the electrophysiological E-E metrics are not well quantified (e.g., connectivity frequency, strength of connection). If the model is intended to predict how the claustrum might accomplish any other task, there is insufficient detail to evaluate the model beyond the evidence that the model creates a subset of cells that can sustain firing during the delay period in the delayed escape task.*
>
> *All relevant work must be appropriately cited throughout the manuscript.*

Regarding the E–E metric, we obtained the following result. When including recordings in which the whole-cell recording could not be completed, optogenetically evoked responses were observed in 38 out of 43 patched cells. This suggests that approximately 90% of the cells receive intra-claustral excitatory input. However, the current dataset does not allow us to quantify the connection probability or the strength of these connections.

As the reviewer pointed out, the RNN developed in this study is specifically designed for the delayed escape task, and we do not intend to claim direct generalization to other proposed functions of the claustrum, such as attention, salience, or sleep. The goal of this study is to computationally characterize the temporal integration mechanism of the claustrum observed in this specific task. We have included this in the Discussion section. In the second paragraph of the Discussion, we have explicitly acknowledged the concerns raised by the reviewer and outlined how they have been addressed in the revised manuscript.

https://doi.org/10.7554/eLife.109539.2.sa0