Population analyses reveal heterogenous encoding in the medial prefrontal cortex during naturalistic foraging

Ji Hoon. Jeong[1] and June-Seek. Choi[1]

[1] School of Psychology, Korea University

Author Note

Ji Hoon. Jeong https://orcid.org/0000-0002-1214-0989

June-Seek. Choi https://orcid.org/0000-0002-4394-2140

Correspondence concerning this article should be addressed to June-Seek. Choi, School of Psychology, Korea University, 145 Anam-ro, Seongbuk-gu, Seoul 02841, Korea. Email: j-schoi@korea.ac.kr

Abstract

Foraging in the wild requires coordinated switching of critical functions, including goal-oriented navigation and context-appropriate action selection. Nevertheless, few studies have examined how different functions are represented in the brain during naturalistic foraging. To address this question, we recorded multiple single-unit activities from the medial prefrontal cortex (mPFC) of rats seeking a sucrose reward in the presence of an unpredictable attack posed by a robotic predator (Lobsterbot). Simultaneously recorded ensemble activities from neurons were analyzed in reference to various behavioral indices as the animal moved freely across the foraging area (F) between the nest (N) and the goal (E) area. An artificial neural network, trained with simultaneously recorded neural activity, estimated the rat's current distance from the Lobsterbot. The accuracy of distance estimation was the highest in the middle F-zone in which the dominant behavior was active navigation. The spatial encoding persisted in the N-zone when non-navigational behaviors such as grooming, rearing, and sniffing were excluded. In contrast, the accuracy decreased as the animal approached the E-zone, when the activity of the same neuronal ensembles was more correlated with events related to dynamic decision-making between food procurement and Lobsterbot evasion. A population-wide analysis confirmed highly heterogeneous encoding by the region. To further assess the decision-related activity in the E-zone, a naïve Bayesian classifier was trained to predict the success and failure of avoidance behavior. The classifier predicted the avoidance outcome as much as 6 s before the head withdrawal. In addition, two sub-populations of recorded units with distinct temporal dynamics contributed differently to the prediction. These findings suggest that an overlapping population of mPFC neurons may switch between two heterogenous modes, encoding relevant locations for goal-directed navigation or an imminent situational challenge.

Introduction

Foraging in nature entails context-appropriate action selection to resolve a mixture of challenges, including risk and value assessment, maintenance of goal-oriented navigation, and task-switching (Pyke et al., 1977; Stephens & Krebs, 1986). For example, foraging animals must update the value of the goal in a space-time domain, negotiate threats while obtaining food, and regulate various stimulus-driven defensive behaviors (Fanselow & Lester, 1988; Kim et al., 2018; Mobbs et al., 2007, 2020; Mobbs & Kim, 2015). These survival-related demands require executive functions capable of flexibly updating risk and value representations in working memory to adapt to the foraging environment. Numerous studies have identified the prefrontal cortex as a primary computational hub for hosting multiple strategies and integrating survival-related information (Fuster, 1973; Korn & Bach, 2019; Lee et al., 2014).

A critical component of successful foraging is goal-directed navigation. Although the hippocampus (HIPP) is crucial for maintaining reward location and path planning (Jarzebowski et al., 2022; Krishnan et al., 2022; Nyberg et al., 2022; Wikenheiser et al., 2013; Wikenheiser & Redish, 2015), the medial prefrontal cortex (mPFC) and its interaction with the HIPP also play a role in goal-directed navigation (Guise & Shapiro, 2017; Sapiurka et al., 2016; Spellman et al., 2015; Yu et al., 2018). Furthermore, given that the mPFC is involved in processing deliberate defensive strategies in multiple situations (see Mobbs et al., 2020 for review), naturalistic foraging where risk and reward coexist may serve as a suitable foundation for exploring adaptive information processing in the mPFC. Notably, mPFC houses spatially selective neurons that represent relative positions in a maze (Kaefer et al., 2020; Yang & Mailman, 2018), goal locations (Hok et al., 2005; but see Böhm & Lee, 2020), and precise Cartesian coordinates during food procurement (Mashhoori et al., 2018).

The mPFC also processes emotional stimuli, especially those signaling environmental

threats. Through its extensive connections with the major limbic areas including the amygdala (Gabbott et al., 2005; Hoover & Vertes, 2007; Vertes, 2004), the mPFC is causally linked to both learning and expression of fear responses. For example, the infralimbic cortex (IL) of the mPFC is associated with the extinction of conditioned fear response (Park & Choi, 2010; Sierra-Mercado et al., 2006). In contrast, the activation of the prelimbic cortex (PL) is linked to the expression of fear-conditioned responses (Burgos-Robles et al., 2009; Dejean et al., 2016; M. M. Diehl et al., 2018; Kim et al., 2013). In addition, PL neurons show increased activity in response to survival-related stimuli such as food, artificial predators, or both (Kim et al., 2018). These findings suggest that the critical information for risky decision-making might be represented and processed in the mPFC.

Currently, there is limited understanding of how the brain organizes navigation and behavior selections during naturalistic foraging. It is unclear whether spatial representation and emotional processing are separated at the single neuron level or coexist within overlapping populations of the mPFC neurons. In this study, we devised a naturalistic foraging task which allowed self-initiated shuttling between different locations and require avoiding an unpredictable attack by a robotic predator while approaching sucrose. This approach allowed continuous monitoring of the neural activity as the rat switched between navigation and strategic reward procurement. Machine learning-based populational decoding methods, alongside single-cell analyses, were employed to investigate the correlations between neuronal activity and a range of behavioral indices across different sections within the foraging arena.

Results

Mix of avoidance and escape behaviors during naturalistic foraging under threat

The foraging arena, modified from Kimm & Choi (2018), was designed to encourage naturalistic foraging that involved both approaching food and avoiding threat. The arena was divided into two distinctive zones by a wall with a pass-through opening in the middle (**Figure 1A**): the foraging (F) and the nest (N) zones. At one end of the F-zone, a motorized gate controlled the access to the robot compartment, which contained a sucrose port and the predator-like "Lobsterbot", named for its large, claw-like arms. Entry into the sucrose port was detected by a beam sensor mounted near the gate (**Figure 1B**). A semicircular area surrounding the robot compartment was designated as the encounter (E) zone where the rat's body was located during sucrose licking and Lobsterbot attack.

The experiment comprises three distinct phases (**Figure 1C**): habituation, shuttling, and Lobsterbot phase. The Lobsterbot phase was further divided into early and late phases separated by the time of surgery. During the habituation phase, a food-deprived rat was acclimated to the sucrose port that released sucrose solution. The lick port was activated by an IR-beam sensor, triggering the solenoid valve when the beam was interrupted. The rat gradually learned to obtain rewards by continuously licking the port. Once the rat maintained a stable licking frequency, typically after 3 days, the shuttling phase began.

In the shuttling phase, rats were trained to shuttle between zones to obtain sucrose. The robot compartment's gate closed six seconds after the first lick and reopened only when the rat returned to the N-zone. After three days of shuttling phase, the Lobsterbot phase was introduced. During this phase, the robot executed a fast-striking attack, 3 s (30%) or 6 s (70%) from the first lick, with the timing randomized across trials to mimic natural uncertainty and prevent simple time-counting strategies. As in the shuttling phase, the gate closed automatically after

the attack, and the rat must return to the N-zone to reopen it.

In each trial of the Lobsterbot phase, the foraging rat could respond with either **avoidance withdrawal** (AW: head retraction before the attack) or **escape withdrawal** (EW: reflexive head retraction after the attack; **Figure 1D**). **Figure 1E** illustrates a representative sequence of responses and events in two consecutive trials. AW was marked by a distinct interval between head withdrawal (end of the blue shade) and the attack (red line), whereas EW was characterized by overlap between the attack and the animal's presence in the E-zone (**Video 1 & 2**).

We examined head withdrawal timepoints to assess whether rats developed a temporal strategy to differentiate between the 3 sand 6 s attack trials. No such strategy was observed; premature withdrawals during 6 s trials were evenly distributed (see **Supplementary Figure S1**).

All five rats exhibited distinctive phasic behavioral changes. Notably, approach attempts (number of trials) dropped sharply during the first Lobsterbot session (Lob1), then stabilized in later sessions after the surgery (starting from Lob4) (**Supplementary Figure S2A**). Similar phasic patterns were observed in the number of licks (**Supplementary Figure S2B**), licks per trial (**Supplementary Figure S2C**) and lick latency after gate opening (**Supplementary Figure S2D**). Comparing the last shuttling session (Sht3) with all pre-surgery Lobsterbot sessions (Lob1-Lob3) revealed significant changes in approach attempts (Friedman test; $\chi^2(3)=9.240$, p=.017), licks ($\chi^2(3)=12.60$, $p<.001$), lick latency ($\chi^2(3)=7.800$, $p=.044$), and licks per trial ($\chi^2(3)=13.08$, $p<.001$). Specifically, both licks and licks per trial decreased during Lob1 (Dunn's multiple comparisons test; $p=.021$; $p=.043$) and returned to shuttling-phase levels by Lob3 ($p>.999$ for both).

In the early sessions of the Lobsterbot phase, the rats displayed passive defensive

behaviors such as freezing (Blanchard & Blanchard, 1969; McClelland & Colman, 1967) and stretched-attend posture (Grewal et al., 1997; Mackintosh & Grant, 1963). These behaviors diminished toward the middle of the phase. Compared to the shuttling phase—where rats typically licked until the gate closed on nearly all trials (88.96%), the Lobsterbot phase featured a mix of AW and EW trials (65.11% and 34.89% respectively). Due to the transient drop in approach attempts and the variety of defensive responses following initial threat exposure, surgery was performed after three Lobsterbot sessions (**Figure 1F**).

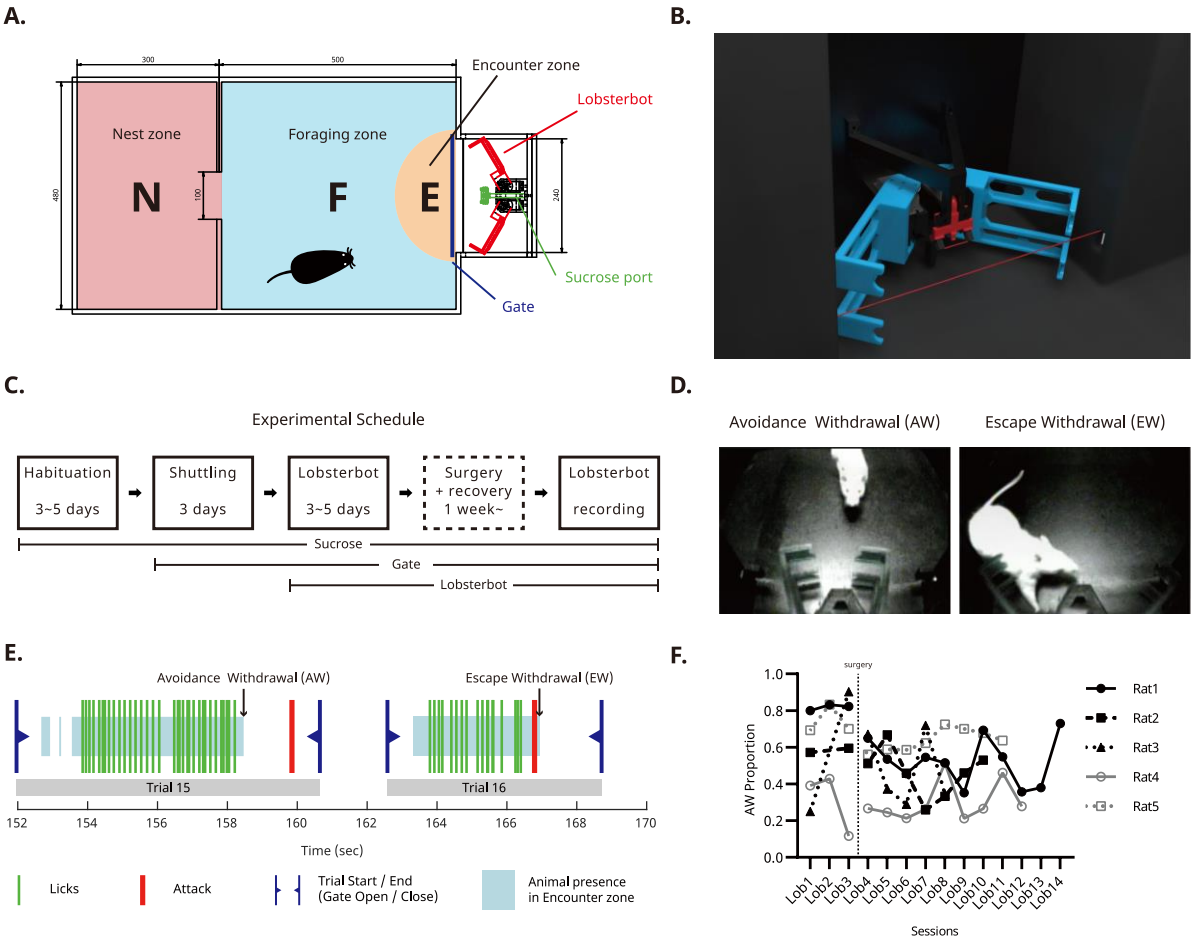


Figure 1. Behavioral procedure and apparatus.

(A) Schematic of the arena showing major sections: N, nest zone; F, foraging zone; E, encounter zone. At the end of the E-zone, the Lobsterbot (red) guards the sucrose delivery port (green). **(B)** Rendered 3-D image of the Lobsterbot. The sucrose port is positioned between the "claws". Two red lines indicate infrared detectors: a short line for lick detection (short line) and a long line for E-zone entry detection. **(C)** Experimental schedule. **(D)** Example snapshots of avoidance withdrawal (AW) and escape withdrawal (EW). In an AW trial, the rat typically retracts its head in advance and observes the Lobsterbot attack. In an EW trial, the rat reflexively flees from the attack. **(E)** Example behavior data containing two consecutive trials (Trials 15 and 16). Each trial started with a reentry to the N zone which triggers gate opening. The rat leaving the N zone typically moves toward the E-zone across the F-zone. The entry to the E zone is detected by an IR beam sensor (blue shade). Within the E zone, the rat starts licking (green lines) until being attacked by Lobsterbot (red line) 3 or 6 s after the first lick. The rat shows voluntary withdrawal behavior (AW; Trial 15) or forced escape behavior (EW; Trial 16). **(F)** Summary of the AW trial rates for each animal during Losterbot sessions. Points for Lob2 of Rat2 and Rat3 are omitted because they did not approach the robot during the entire Lob2 session.

Population activity in the mPFC predicts the distance from the Lobsterbot while navigating in the arena

Recording data were collected from a total of five rats implanted with a custom-made moveable tetrode drive. The tetrodes were initially implanted in the PL and advanced ventrally toward the IL at the end of each session (0.1~0.2 mm/session). Recording locations were reversely calculated from the marking lesion made after the last recording session. Tetrode tracks were identified by visual inspection under a microscope as indicated by black vertical lines on the matching atlas (Paxinos & Watson, 2007) as shown in **Figure 2A**. All tetrodes except for one spanned both the PL and the IL. After cluster cutting and preprocessing (see Materials and Methods), 632 units (57 ~ 198 units per rat) from the PL (n = 379) and the IL (n = 253) were selected for ensemble analysis.

Previous studies have reported spatial correlates were observed in multiple subregions of the mPFC (Hok et al., 2005; Jung et al., 1998; Ma et al., 2023; Mashhoori et al., 2018; Sauer et al., 2022; Spellman et al., 2015). To determine whether location-specific neural activity exists at the single-cell level in our dataset, a traditional place map was constructed for individual neurons. Although most neurons did not show cohesive place fields, a few neurons modulated their firing rates based on the rat's current location. However, even these neurons were not comparable to place cells in the hippocampus (O'Keefe & Dostrovsky, 1971) or grid cells in the entorhinal cortex (Hafting et al., 2005) as the place fields were patchy and irregular in some cases (**Figure 2B**; Units 66 and 125) or too broad, spanning the entire zone rather than a discrete location within it (Units 26 and 56). The latter type of neuron has been identified in other studies (e.g., Kaefer et al., 2020).

Since we did not find neural activity that can be mapped straight onto the spatial layout of the arena based on individual activity, we performed machine-learning (ML) analysis

targeting all recorded units during the given session. Among various machine learning algorithms, we selected a robust tool for decoding underlying patterns in the data, rather than to model the architecture of the mPFC. We implemented a four-layer artificial neural network regressor (ANN; see Materials and Methods for a detailed structure), as the ANN achieves significantly lower decoding errors (**Supplementary Figure S3**) and has robustness to hyperparameter changes (Glaser et al., 2020). The inputs to the regressor were the simultaneously recorded neural activity which was segmented into 50-ms bins and the predicted output was the distance between the rat's head and the center of the E-zone where reward and threat co-exist (**Figure 2C**).

The analysis of trained ANNs revealed that the distance from the rat to the center of the E-zone can be decoded from the mPFC ensemble activity. Compared to the control regressor trained on a shuffled dataset, the regressor with original data showed significantly lower errors (paired t-test, $t(39) = 10.47$, $p < .001$). However, we did not observe a difference in decoding accuracy between the PL and IL ensembles, and there were no significant interactions between regressor type (shuffled vs. original) and regions (mixed-effects model; regions: $p=.996$; interaction: $p=.782$). These results indicate that the population activity in both the PL and IL contains spatial information (**Figure 2D, Video 3**). Overall accuracy measured by mean absolute error (MAE) was 16.61 (± 3.67) cm, which was slightly higher than values from previous reports employing different spatial tasks (Ma et al., 2023; Mashhoori et al., 2018). This reduced accuracy may reflect the smaller number of neurons per ensemble in our dataset (mean = 15.8) relative to prior studies. Given the properties of algorithm, we expect that larger ensembles would improve performance.

We next asked whether the decoding accuracy varied with movement directions. Specifically, we compared the decoding accuracy of the regressor for outbound (from the N-

to E-zone) vs. inbound (from the E- to N- zone) navigation within the F-zone. There was no significant difference in decoding accuracy between outbound vs. inbound trips (paired *t*-test; $t(39) = 1.52$, $p = .136$), indicating that the stability of spatial encoding was maintained regardless of the moving direction or perceived context (**Figure 2E**).

One potential concern is that the decoder might rely on visual cues or zone-specific behaviors (e.g., Lobsterbot confrontation or licking) rather than spatial signal per se. To rule out this possibility, we analyzed a subset of data collected when the compartment door was closed, preventing visual access to the Lobsterbot and sucrose port and limiting active foraging behavior. Even under these conditions, the regressor decoded distance with a MAE of 12.14 (± 3.046) cm (paired *t*-test; $t(39) = 12.17$, $p < .001$). Notably, the regressor's performance was significantly higher with this subset than with the full dataset (paired *t*-test; $t(39) = 9.895$, $p < .001$). In summary, these findings indicate ensemble activity in the mPFC contains sufficient spatial information, detectable through ML-based decoding.

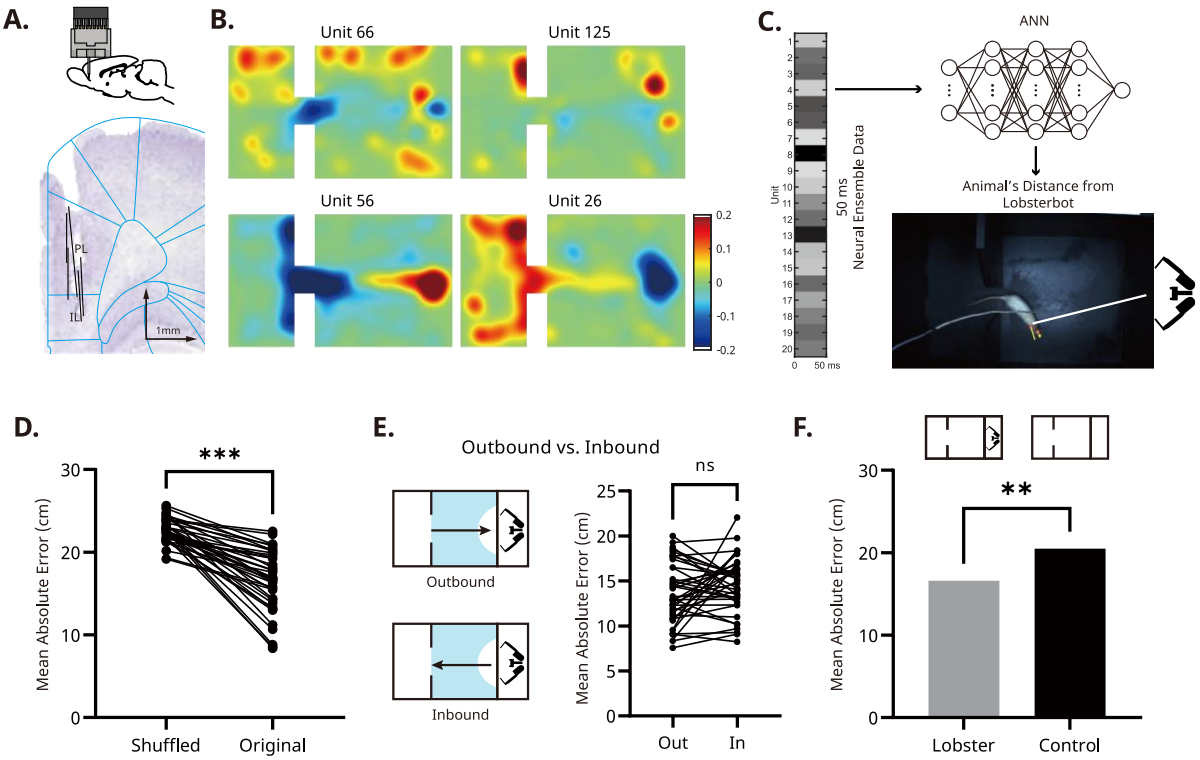


Figure 2. Ensemble activity in the mPFC predicts distance from the goal.

(A) Electrode implantation and recording sites. Top: A movable 4-tetrode microdrive was initially implanted in the PL region and lowered ventrally toward the IL after every recording session. Bottom: Representative recording tracks from all five animal are superimposed over an image of a stained coronal section of the frontal brain. Histological examination of all brain sections confirmed that the electrode tracks spanned the dorsoventral axis between the PL and IL. **(B)** Modulation of unit firing showing place-cell like activities. Units 66 and 125 exhibit fragmented place fields in all over the arena, while Units 56 and 26 display relatively large place fields surrounding particular spots such as the gates. Heat maps are calculated from z-scored spatial tuning curves. **(C)** Schematics of the ensemble decoding analysis. The 4-layer deep artificial neural network (ANN) receives populational neural data during 50ms-timewindow and is trained to predict the rat's current distance from the center of the E-zone. The example data depicted in the figure is a sample recording from 20 units when the rat is at a particular distance away from the center of the E-zone, indicated by the white bold line. **(D)** Accuracy of the distance regressor. Mean Absolute Error (MAE) was significantly smaller for the original dataset (16.61 cm, Original) compared to the shuffled dataset (Shuffled), and below the rat's body length, suggesting that the mPFC ensembles encode spatial correlates of the distance from the goal. **(E)** Prediction accuracy in the F-zone during outbound/inbound paths. Decoding accuracy in the F-zone was calculated separately for the outbound (from the N-zone to the E-zone) and inbound (from the E-zone to the N-zone) paths. The decoding accuracy remained

unchanged despite the differences in motivation and perceived visual cues due to the movement direction. **(F)** Comparison of the regressor's accuracy from the control experiment. When the Lobsterbot was removed from the robot compartment, reverting the task back to simple shuttling, the mPFC distance regressor's performance significantly decreased compared to the Lobsterbot phase.

Removing the Lobsterbot decreases spatial decoding accuracy

Earlier studies have suggested that PL neurons do not exhibit any spatially correlated firing patterns in an empty round apparatus (Poucet, 1997). This absence of spatial correlates may result from a lack of complex goal-oriented navigation behavior, which requires deliberate planning to acquire more rewards and avoid potential threats. To evaluate how the Lobsterbot's presence impacted spatial decoding accuracy, additional single-unit recording experiments (Ctrl-Exp) were conducted and compared with the previous results (Lob-Exp). In this experiment, three rats followed the same protocol until the microdrive implant surgery. After the surgery, unlike the Lob-Exp group, the Ctrl-Exp group returned to the shuttling phase, during which the Lobsterbot was removed. With this protocol, both groups experienced sessions with the Lobsterbot, but the Ctrl-Exp group's task became less complex, as it was reduced to mere reward collection. During a total of nine recording sessions, 112 units were identified and ANN regressors were built and trained using these data. Statistical tests indicated that the original dataset had significantly lower decoding errors (20.48 ±2.31 cm) compared to the shuffled dataset (22.61 ±1.95 cm; paired *t*-test; $t(8) = 6.20$, $p < .001$). Nevertheless, when the error was compared with that of Lob-Exp, the Ctrl-Exp group showed a higher decoding errors (**Figure 2F**; unpaired t-test; $t(47) = 3.02$, $p = .004$). This did not result from differences in activity levels, as there was no significant difference in total travel distance between the Lob-Exp and Ctrl-Exp groups (unpaired *t*-test; $t(47) = 0.07$, $p = .941$).

Another factor potentially contributing to the increase in decoding errors could be differences in recording quality across experiments. To discriminate the effects, a generalized linear model (GLM) was constructed with predictors of 1) the Lobsterbot's presence, 2) the number of units recorded, and 3) the recording site (PL vs. IL) on decoding error. The results showed that the GLM predicted the outcome variable with significant accuracy ($F(3,45) =$

14.06, p <.001), explaining 48.38% of the variance in decoding error. The regression coefficients were as follows: presence of the Lobsterbot (2.76, standard error [SE] = 1.11, t = 2.49, p =.016), number of recorded cells (−0.43, SE =.08, t = 5.29, p <.001), and recording location (0.91, SE =.92, p =.329). The results indicated that the number of recorded cells significantly influenced decoding accuracy. Most importantly, the presence of the Lobsterbot significantly decreased the overall decoding error, even after accounting for the number of recorded cells. In summary, the removal of the Lobsterbot decreased spatial decoding accuracy.

Non-navigational behavior reduces the accuracy of decoded location

Although initial analysis showed that the distance of the rat to the center of the E-zone could be decoded with reasonable accuracy, the spatial precision was unevenly distributed throughout the arena (**Figure 3A**). The magnitude of error was highest in the N-zone, followed by the E-zone, and lowest in the F-zone. A one-way repeated measures ANOVA revealed a significant main effect of the zone (F(1.68, 65.52) = 40.36, p <.001). Post hoc analysis showed that the MAE in the F-zone (13.86 ± 2.56 cm) was significantly lower than that in the N-zone (22.72 ± 6.38 cm; t(39) = 10.69; p <.001) and the E-zone (19.98 ± 7.30 cm; t(39) = 6.34; p <.001). However, there was no significant difference between the N- and E-zones (t(39) = 2.29; p =.081; **Figure 3B**). To correct for the unequal distribution of location visits (more visits to the F- than to other zones), the regressor was trained using a subset of the original data, which was equalized for the data size per distance range (see Materials and Methods). Despite the correction, there was a significant main effect of the zone (F(1.16, 45.43) = 119.2, p <.001) and the post hoc results showed that the MAEs in the N-zone (19.52 ± 4.46 cm; t(39) = 10.45; p <.001) and the E-zone (26.13 ± 7.57 cm; t(39) = 11.40; p <.001) had a significantly higher errors when compared to the F-zone (14.10 ± 1.64 cm).

In the N-zone, the rats engaged in stereotypic behaviors such as grooming, climbing, and rearing, during which they stopped locomotive behavior (**Figure 3C**). To determine whether these non-navigational behaviors resulted in higher decoding errors, these behaviors inside the N-zone is manually labelled and the MAEs during navigational vs. non-navigational behaviors were compared. Due to the imbalanced nature of the dataset resulting from varying preference for non-navigational behaviors, data from 21 sessions were included, each containing at least 20% of one type of behavior. The paired t-test showed a significant difference between the MAEs during the non-navigational and navigational behaviors in the N-zone (paired t-test; $t(20) = 9.381$, $p < .001$; **Figure 3D**). Moreover, excluding the non-navigational data of the N-zone from the original dataset improved the overall accuracy of the regressor (paired t-test; $t(20) = 8.562$, $p < .001$; **Figure 3E**). These data indicate that erroneous decoding result in the N-zone might have resulted from the disruption induced by non-navigational behaviors.

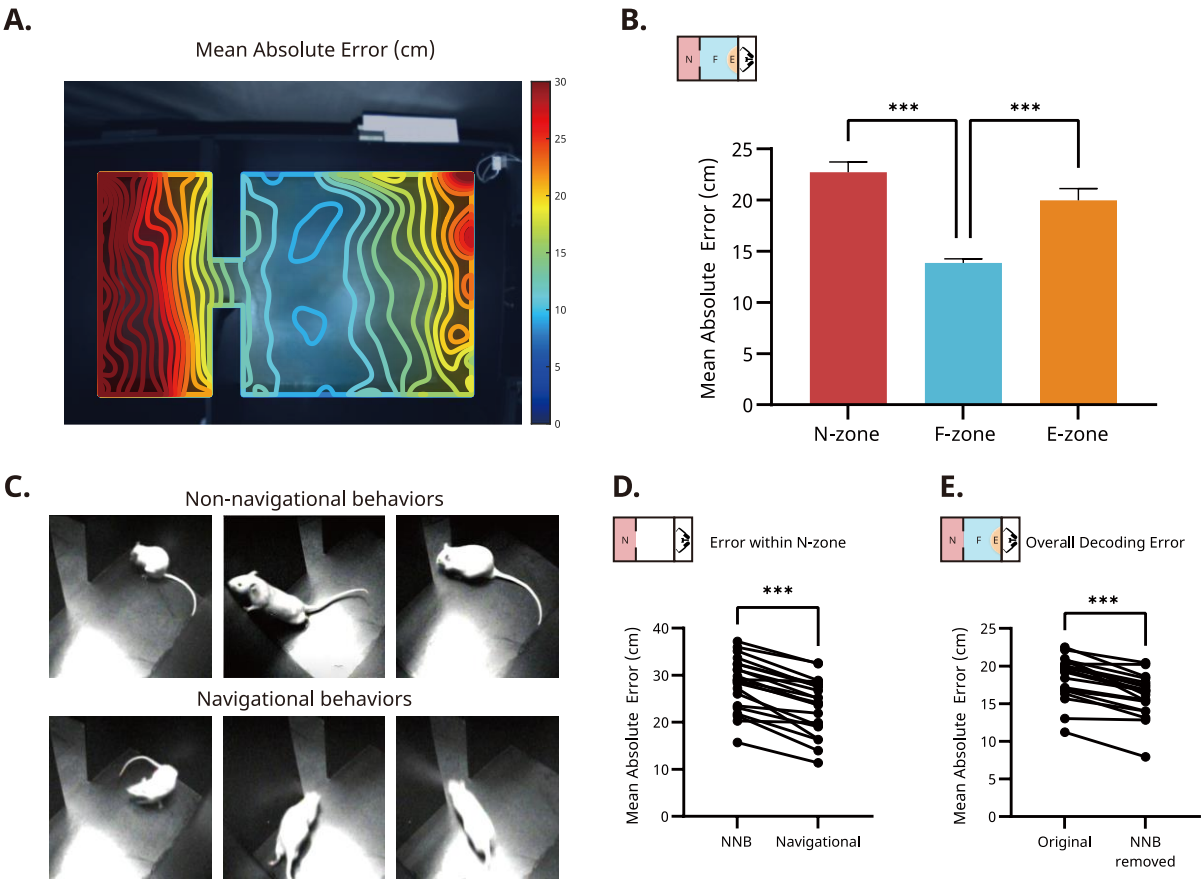


Figure 3. Spatial encoding is disrupted by non-navigational behaviors

(A) Spatial distribution of the prediction accuracy. The heatmap indicates MAE of a fully trained ANN, superimposed over the entire foraging arena. Prediction accuracy was highest in the F-zone and lower in the N- and E-zone. **(B)** Mean prediction accuracy by the zones. The MAE in F-zone was significantly lower than the other zones. Error bars represent the SEM. **(C)** Examples of the non-navigational behaviors in the N-zone. The top three snapshots depict grooming, rearing and sniffing. The bottom three snapshots show typical goal-directed navigational movements. **(D)** Comparison of decoding errors (N-zone) during navigational vs. non-navigational behaviors. Errors were significantly larger when the rat was engaged in non-navigational behaviors within the N-zone. **(E)** Comparison between regressors trained with vs. without data from non-navigational behaviors. The overall decoding error was significantly smaller when the regressor was trained without the data from non-navigational behaviors.

Population analysis reveals a different functional mode in the E-zone

To assess the functional coherence of population activity (Durstewitz et al., 2010) within and across different zones, principal component analysis (PCA) was performed on the ensemble data collected throughout the session. **Figure 4A** illustrates the first two principal components of the transformed neural data from a representative recording session (number of units = 19). First, within- and between-zone distances in population neural activity were compared to determine whether activity within the same zone exhibited similar properties, distinct from those of other zones A paired *t*-test revealed a significant difference between distances ($t(119) = 10.46, p < .001$), confirming that the neural activity within a given zone was distinct from those of other zones.

To quantitatively measure functional resemblance among ensemble activity within different zones, we measured the distance between pairs of centroids projected onto the latent space (N-F, N-E, and F-E). A Friedman test showed a significant difference between centroid pairs ($\chi^2(2)=51.05, p < .001$). Further post hoc analysis revealed a significant difference between N-F and N-E pairs ($p < .001$) and N-E and F-E pairs ($p < .001$) but not in N-F and F-E pairs ($p = .353$), suggesting an ordinal similarity with N-F (D =.72) being most similar, followed by F-E (D = .94) and N-E (D = 1.44) (**Figure 4**Figure 4**B**).

Not surprisingly, the firing pattern in the E-zone was the most dissimilar, reflecting a functional state different from those in other zones. In contrast, the distance between centroids of the N- and F-zones was relatively small, indicating that the overlap between the neural subspaces in the two zones was not negligible.

Although PCA can capture population differences between neural ensembles in the zones, a small subset of neurons highly tuned to salient sensory cues could solely drive this separation. Particularly in the E-zone, which encompasses two significant events—head-entry

and head-withdrawal—neurons responsive to the delivery of rewards or attacks from the robot could dominate the principal components. To minimize this influence, we defined "critical event times" as ±1 second from head-entry and head-withdrawal, excluded neural data during these periods, and then recalculated the principal components. The results still demonstrated the same sequential ensemble differences ($F(1.57, 61.34) = 73.56$, $p < .001$), indicating that the population differences are not driven by events specific to any zone (**Supplementary Figure S4**). In summary, populational activity inside the E-zone showed distinctive activity patterns unlike those of active navigation.

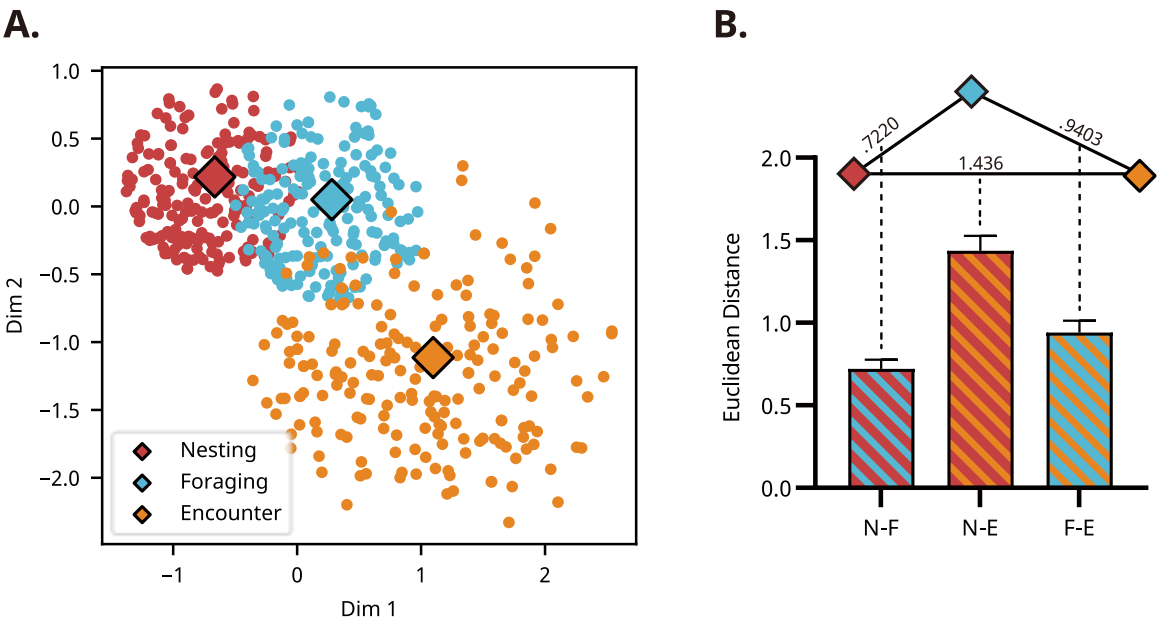


Figure 4. PCA results reveal distinctive population activity in the E-zone

(A) Representative recording session depicting the first two principal components of population neural activity. Each dot represents a 50-ms segment of multi-unit activity, color-coded by the rat's location. Diamonds mark the centroids of neural representations for each zone. To clustering, only data points near each centroid have been plotted. **(B)** Distances between all centroid pairs across recording sessions. The centroid of the E-zone is distinctly separated from those of the other two zones, indicating a unique neural state within the E-zone. The triangle above the bar plot illustrates the relative distances between the centroids of each zone; Longer edges indicate greater dissimilarity between neural ensemble activities. Error bars represent SEM.

Hierarchical clustering identifies functionally distinctive sub-populations in the E-zone

In the E-zone, navigational behavior was largely replaced by approach attempts and avoidance behaviors. To characterize the temporal dynamics of single-unit activity in the E-zone, we examined the modulation of firing rate to the two key behaviors; head-entry and head-withdrawal. The temporal dynamic of neuronal firing relative to these two behavioral events was visualized using color-coded raster plots and peri-event time histograms (**Figure 5A & 5C**), featuring spiking activities from all behavior-responsive units (see Materials and Methods). Overall, the units showed modulated firing with varying degrees of temporal relations to the two events. Most strikingly, overall modulation patterns were clearly different between head-entry and head-withdrawal.

During head-entry, a large portion of the units increased the firing before the event (**Figure 5A, top)**, a pattern confirmed by pooled analysis across neurons (**Figure 5A, bottom)**. The mean activity showed a peak approximately 150 ms before the head-entry, with the peak exhibiting substantial width (FWHM of ~200 ms), suggesting the presence of varied neuron populations differing in peak position. In contrast, this anticipatory modulation was not clearly identified when the activities were aligned to the head-withdrawal. Instead, activity sharply increased around the event itself (**Figure 5C, top)**, producing a narrow peak at 0 ms (FWHM=~75 ms, **Figure 5C, bottom)**.

To categorize recorded units into multiple sub-populations based on their functional connectivity in reference to the key events, the unsupervised hierarchical clustering method was applied to all recorded units (Boehlen et al., 2016; Di Miceli et al., 2020). We identified two sub-populations when we classified neurons according to their activity around the head-entry (**Figure 5B**). The first sub-population (HE1: n = 366, 57.91%) showed an anticipatory increase with a peak located at 350 ms before and a negative modulation after entry and

remained low for an extended period. The other sub-population (HE2: n = 189, 29.91%) showed a more concentrated positive modulation at around the time of entry. When we used the head withdrawal as criterion, three sub-populations were identified (**Figure 5D**). The first (HW1: n = 380, 60.13%) showed a complex modulation pattern: lower than the baseline in the beginning, followed by a peak firing timed at the head-withdrawal. The second (HW2: n = 147, 23.26%) maintained increased activity until the onset of head-withdrawal and returned to zero afterward. The last sub-population, HW3, which consisted of a relatively small number of units (n = 62, 9.81%), showed an inhibitory response at the time of head-withdrawal.

Well-trained rats typically run across the F-zone at high velocity and arrive at the sucrose port in the E-zone, stopping abruptly. Moreover, the head-entry is always aligned with the rats' movements, including both subtle (AW) and reflexive (EW) movement. Consequently, it is plausible that the sudden movement might result sudden change in neural activity around the time of head-entry and/or head-withdrawal. To exclude this possibility, a “run-and-stop” event was correlated with the neural data. The event was defined as the transitional moment between a directional locomotion of 2 s or longer and a subsequent pause of 1 s or longer (see Materials and Methods). Unit activities were aligned around this run-and-stop event, and the mean activity of each sub-population was plotted for visual inspection. As shown in the **Supplementary Figure S5**, there was no detectable neural modulation around the run-and-stop events. The lack of significant modulation at the time of action suggests that the neural modulation occurring around the head-entry and head-withdrawal is not a simple motoric response.

An interesting overlap was found between sub-populations characterized by head-entry and head-withdrawal. In the HW1 group, 78.68% (n = 299) of units were classified as HE1. On the other hand, 63.95% (n = 94) of the HW2 units were from HE2 (**Supplementary**

Figure S6A). Based on this, we further characterized the temporal dynamics of spike activity in relation to the two behavioral events using the combined categories: Type 1 (HE1-HW1) and Type 2 (HE2-HW2). These two subpopulations demonstrated distinctive firing patterns within the E-zone (**Supplementary Figure S6B & S7**). Specifically, Type 1 neurons exhibited a rapid decline in activity following a short peak 200 ms before the head-entry, whereas Type 2 neurons displayed a sustained high peak activity ($z \approx 2$) from the onset of the head-entry until the head-withdrawal.

We also examined whether there is a significant difference between the PL and IL in the proportion of Type 1 and Type 2 neurons. In the PL, among 379 recorded units, 143 units (37.73%) were labeled as Type 1, and 75 units (19.79%) were labeled as Type 2. In contrast, in the IL, 156 units (61.66%) and 19 units (7.51%) of 253 recorded units were labeled as Type 1 and Type 2, respectively. A Chi-square analysis revealed that the PL contains a significantly higher proportion of Type 2 neurons ($\chi^2(1, 632) = 34.85$, $p < .001$), while the IL contains a significantly higher proportion of Type 1 neurons compared to the other region ($\chi^2(1, 632) = 18.07$, $p < .001$). Taken together, these findings provide support for the existence of at least two distinct subpopulations in the PL and IL that may be involved in different cognitive functions within the E-zone.

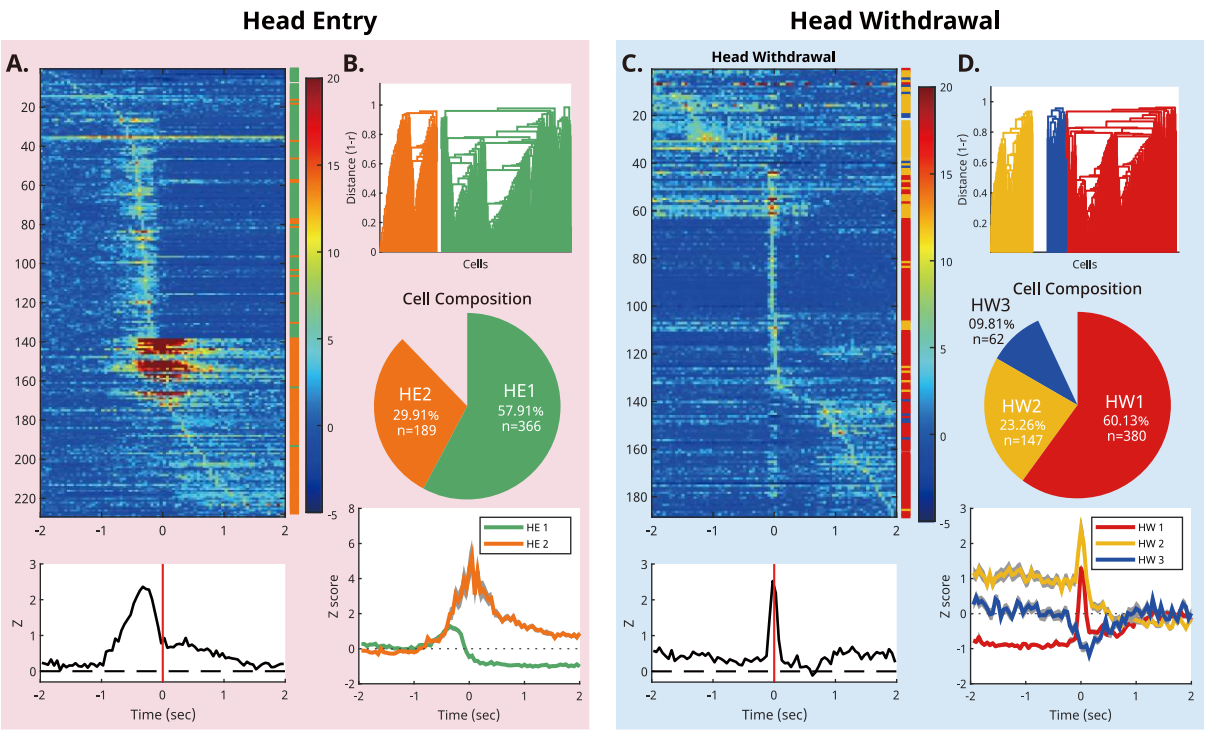


Figure 5. Multiple subpopulations in the mPFC react differently to head entry and head withdrawal.

(A) Top**:** The PETH of head entry-responsive units is color-coded based on the Z-score of activity. Bottom**:** The red vertical lines mark the timing of the head-entry. The peak latency of each unit varies from as early as 2 s before and to 1~2 s after the head-entry. **(B)** Functional segregation of all recorded units. Top and middle: Two sub-populations of units based on hierarchical cell clustering analysis. Bottom**:** The averaged activity for each sub-population. **(C)** The PETH of head withdrawal-responsive units is color-coded based on the Z-score of activity. **(D)** Functional segregation of all recorded units. Top and middle: Three sub-populations of units based on hierarchical cell clustering analysis. Bottom: The averaged activity for each sub-population.

Population activities in the E-zone encode success and failure of avoidance with high accuracy

The previous analysis confirmed that a significant number of neurons modulated their activity before or around the time of approach toward and withdrawal from the E-zone. We then further analyzed the correlation between neural activity and success and failure of avoidance behavior. If the mPFC neurons participate in the avoidance decisions, avoidance withdrawal (AW; withdrawal before the attack) and escape withdrawal (EW; withdrawal after the attack) may be distinguishable from decoded population activity even prior to motor execution. A naïve Bayesian classifier (see Materials and Methods) was trained to distinguish AW from EW using 2-s neural ensemble data around the onset of the event (**Figure 6A).** We corrected the accuracy against the sample imbalance owing to varying degrees of the AW/EW response ratio across different sessions. Our analysis showed that the classifier based on the original dataset achieved a significantly higher prediction accuracy (72.71%) than a theoretical chance-level prediction (50%) or a shuffled dataset (51.62%) (paired *t*-test; $t(39) = 10.06$, $p < .001$) in distinguishing AW from EW (**Figure 6B)**. Furthermore, we analyzed whether there is a difference in prediction accuracy between sessions with different recorded regions, the PL and the IL. A repeated two-way ANOVA revealed no significant difference between recorded regions, nor any interaction (regions: $F(1, 38) = 0.1828$, $p = 0.671$; interaction: $F(1, 38) = 0.1614$, $p = 0.690$).

Because the behavioral reactions following the two events were quite different (bigger reactions during the EW as if the animal failed to predict the attack), it is not clear whether the naïve Bayesian classifier discriminated the two types of responses based on the post-event population activity, which would have been a reflection of the different reactions in the two responses rather than an anticipatory or value-driven encoding. To clarify this issue, we trained

naïve Bayesian classifiers using 2-s ensemble activity at varying time points before the execution of the head-withdrawal response. Multiple t-tests between original and shuffled datasets revealed that AW/EW discrimination accuracy began to increase above the chance level 6 s before the actual response and remain significant at all time points (multiple paired *t*-tests with Šídák-Boferroni correction; *ps* < .01; **Figure 6C, Supplementary Data 1**).

Based on the previous clustering analysis that identified functionally cohesive sub-populations of neurons (Types 1 and 2), we further investigated the predictive encoding by the ensemble activity of the two subpopulations. The naïve Bayesian classifier uses class likelihood, P(unit|AW) and P(unit|EW), to classify a particular type of withdrawal response. By comparing the class likelihoods, we can assess the ability of these subgroups to distinguish between AW and EW. This ability is quantified by a 'Class Discrimination Index,' calculated as a log odds ratio. This index reflects the likelihood that a given feature set is associated with a particular type of withdrawal response (AW or EW). To test whether the unit activity right before the head-withdrawal (from 2 s before the head-withdrawal to the onset of the head-withdrawal) was correlated with subsequent avoidance behavior, we calculated the log odds ratio from the trained classifier. The results showed that Type 2 neurons had a significantly higher log odds ratio, indicating a high correlation between unit activity and success in avoidance, whereas no discriminatory modulation was observed for Type 1 neurons or others (one sample t-test, $t(298) = .564$; $p = .573$, $t(94) = 5.63$; $p < .001$, $t(238) = .531$; $p = .596$, Type 1, Type 2, and others, respectively). These results indicate that Type 2 units carries information that is relevant for classifying trial as AW. To summarize, the populational neural activity of the mPFC in rats within the E-zone distinguished between two types of head-withdrawal behavior, and increased activity of unit subgroups was strongly associated with the execution of AW.

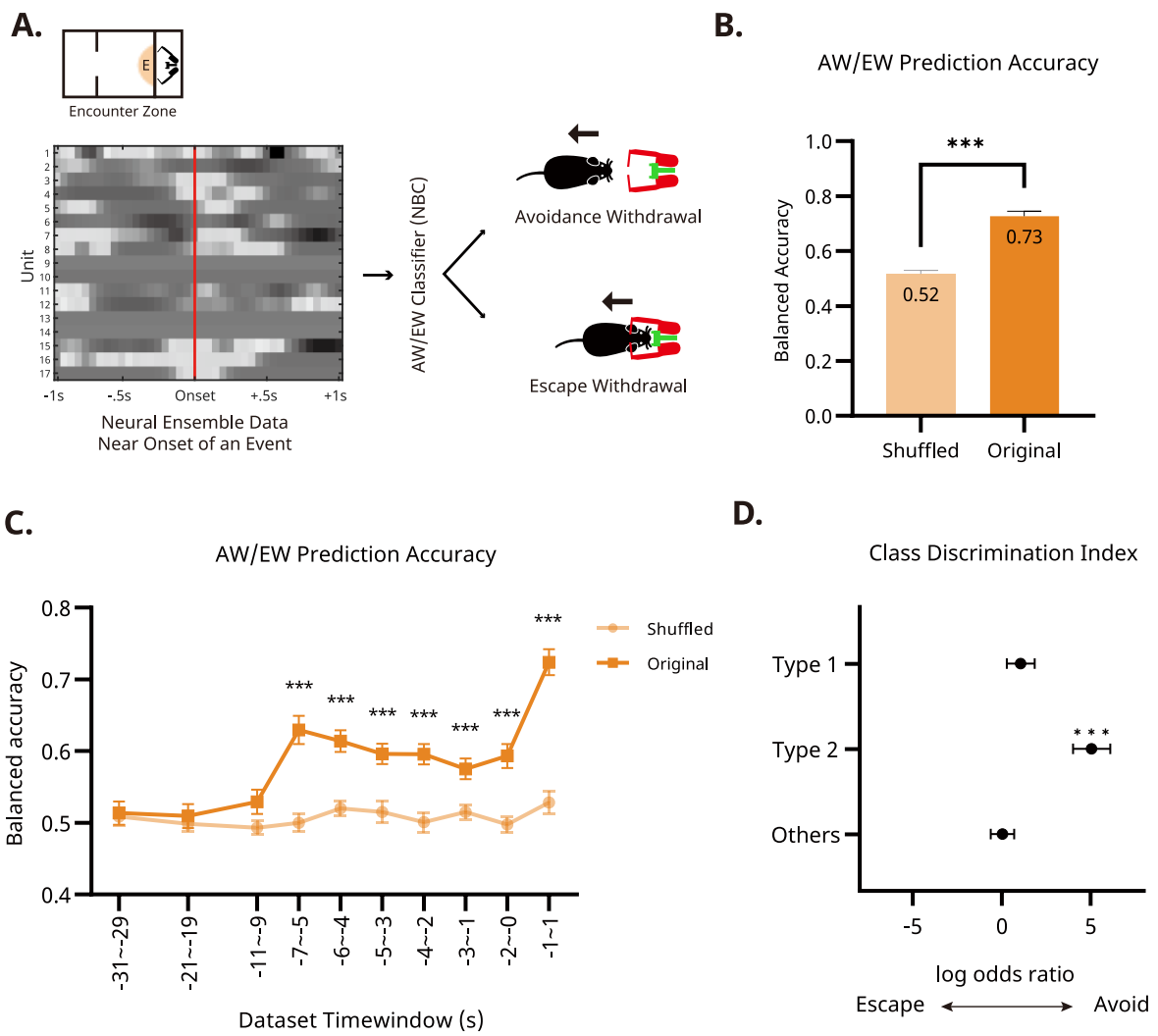


Figure 6. Neural ensemble activity predicts failure and success of avoidance response. **(A)** Schematics of the event decoding analysis. The Naïve Bayesian decoders are trained with 2 s window of neural activity to discriminate avoidance or escape on every trial (AW/EW classifier). The grayscale image depicts an example firing pattern of 17 units on a given trial, arranged to the onset of the withdrawal response. The decoder classifies whether this trial is AW or EW based on this data. **(B)** Accuracy of the naïve Bayesian classifier. The decoding accuracy of the classifier was significantly higher than that from the shuffled data. **(C)** Temporal characteristics of prediction accuracy of the naïve Bayesian classifier. Prediction accuracy was significantly higher at the time points as early as 5-7 s before the head-withdrawal. **(D)** Class discrimination index by the two sub populations of neurons. The class discrimination index indicates that the Type 2 neurons showed a significant discriminatory power towards AW. Neurons in the Type 2 and the Others group did not exhibit significant discriminatory power.

Overlapping populations of mPFC neurons adaptively encode spatial information and defensive decision

Through a series of analyses, we found that both distance and foraging decision can be decoded from the same mPFC neural ensemble. PCA results support the that the mPFC ensemble activity forms a distinctive functional subspace when the rat is engaged in foraging decision facing the robotic predator in the E-zone. Previous studies have also observed different firing patterns from the same mPFC neural ensembles when the rat switches between tasks (Durstewitz et al., 2010; Nigro et al., 2023). Therefore, it would be interesting to see whether the distance and foraging decision are encoded by two exclusively subsets of neurons or by the whole ensemble.

To quantitatively assess each neuron's contribution to the distance regressor and event classifier, we calculated the feature importance scores for each decoder (**Figure 7A**). This metric measures the increase in error when a neuron's data is shuffled. For example, if a neuron holds substantial data and is heavily relied upon by the decoder, shuffling its signal will significantly increase the error. We analyzed the permutation feature importance scores for each neuron for both distance regression and event classification (see Materials and Methods).

First, for the distance regressor, the median permutation feature importance score was 0.73 cm (**Figure 7B**). This suggests that shuffling one unit’s neural activity increases the decoding error by less than 1 cm. Moreover, 95% of units exhibited a permutation feature importance score below 1.998 cm, indicating only a negligible increase. To determine whether a subset of neurons with high feature importance scores solely drives distance decoding, we tested whether removing the top 20% of neurons with the highest scores in a given session affects distance decoding. The results showed that although removal significantly increased the decoding error, the regressor trained with the remaining data still achieved a significantly lower

error (17.35cm; Wilcoxon signed rank test, $p<.001$). Next, we applied a similar method to the event classifier. This analysis revealed a median permutation feature importance score of 0.01 and a 95th percentile of 0.06. We also tested the effect of removing the top 20% of neurons with the highest scores and found that the decoder still exhibited a significantly lower error (paired *t*-test, $t(39)=6.899$, $p<.001$). This result indicates that there is no sub-population which dominates in encoding distance information.

Furthermore, we tested whether there is a correlation between the feature importance scores of the distance regressor and the event classifier. If there were two subgroups of neurons, each dedicated to encoding distance and events respectively, we might expect to see a negative correlation. To verify this, we calculated Pearson's correlation for each session. We found that there was no significant correlation between the feature importance scores of the two encoders ($r=0.077$, $p=.062$), indicating the absence of two distinct subpopulations that exclusively encode either distance or event.

Additionally, we found no evidence that Type I or Type II neurons have statistically higher or lower feature importance in the distance regressor. These results support the notion that distance and avoidance outcomes are not encoded separately by subsets of neurons; rather, the entire ensemble encodes two different concepts according to the given context.

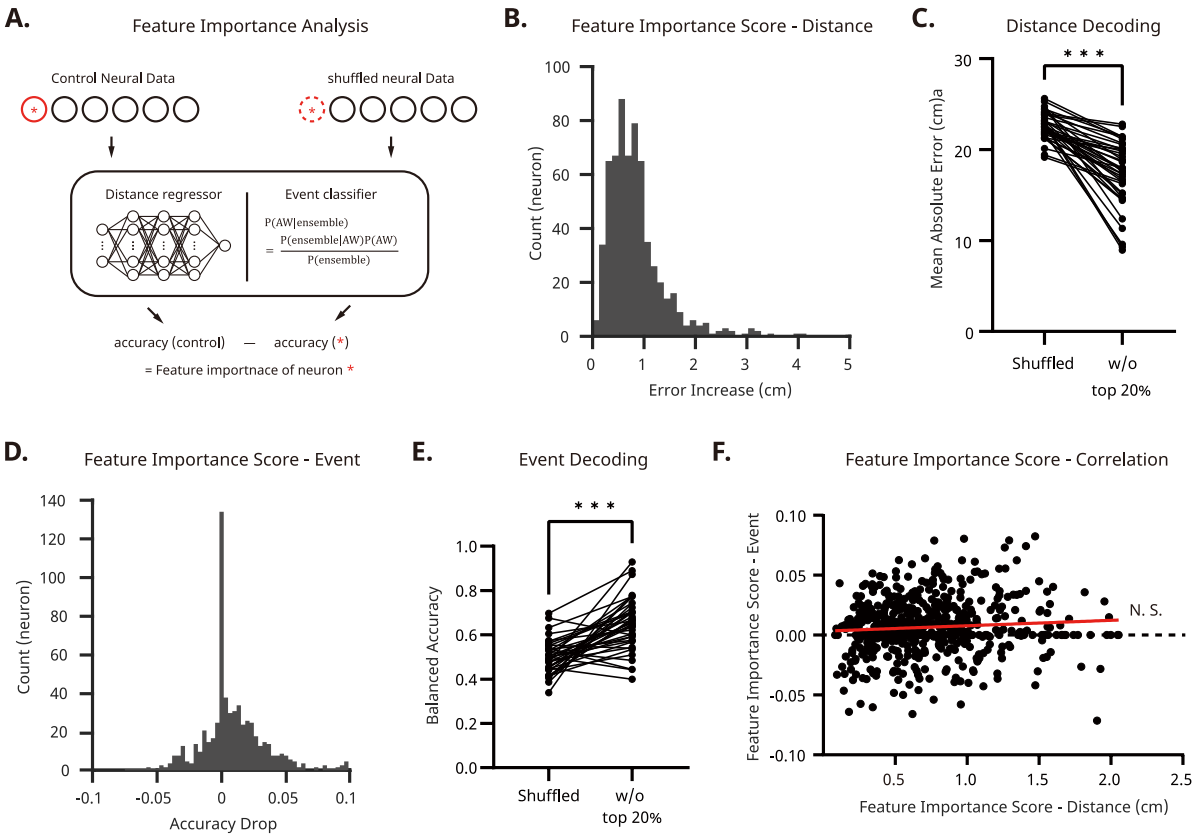


Figure 7. Feature importance analysis shows no evidence of dedicated neural subsets for distance or event encoding.

(A) Schematic of the computational protocol for the feature importance analysis.

(B) Frequency distribution of all recorded units for their feature importance (measured as error increase) in distance encoding. Only a few units produced non-negligible increase in error when removed, indicating no dedicated distance-encoding neurons. The red line marks the 95th percentile.

(C) Accuracy of distance regressor after removing the 20% of high-importance units. Even without these units, the regressor accurately decoded the rat's location.

(D) Frequency distribution of all recorded units by their feature importance (measured in accuracy drop) in event classification. Only a few units caused a non-negligible accuracy decrease when removed, indicating no dedicated event-encoding neurons. Shuffling unit data resulted in a small decrease in accuracy, and even some cases, an increase. The red line marks the 95th percentile.

(E) Accuracy of the event classifier after removing the top 20% of high-performance units. Even without these units, the classifier decoded the type of the event (AW vs. EW).

(F) Correlation between feature importance scores for the distance regressor and the event classifier. No correlation was found between the two measures.

Discussion

We found distinct ensemble activity states in the mPFC (PL and IL) of a foraging rat, each linked to specific spatial zones or task demands within the foraging environment. The rat's distance from the conflict zone could be decoded with stable accuracy during navigation, but ensemble activity switched to encode defensive behavior when the rat entered the immediate reward-threat zone (E-zone). In this high-conflict context, the rat's threat-evading behavior (AW), coupled with forgoing the reward, could be predicted with a naïve Bayesian classifier. These findings show that mPFC neurons engage in two distinct functional modes of processing allocentric spatial information and egocentric reward-threat information.

In the present study, spatial information was extracted from the mPFC neuron in terms of distance. Previous studies on mPFC spatial coding ranged from reporting no spatial information (Poucet, 1997) to precise coordinates encoding with errors as small as a few centimeters (Ma et al., 2023; Mashhoori et al., 2018). In some studies, the spatial coding mechanism was directly compared between the hippocampus and the mPFC (Kaefer et al., 2020; Powell & Redish, 2014), and the mPFC cells were found to encode a generalized form of space rather than specific locations. Therefore, we focused on distance as a more generalized spatial index that also reflects an internal representation of value and risk, and built an ANN regressor to target it.

Brief navigation was sufficient to recruit most recorded mPFC neurons into spatial encoding, with decoding accuracy consistent across the F-zone regardless of navigational direction (inbound vs. outbound) or visual cues during foraging. The spatial resolution encoded by the ensemble of neurons was precise enough to provide navigational utility (13.86 cm in the F-zone) considering an adult rat’s body length (~20 cm without tail), supporting the spatial encoding hypothesis.

However, spatial encoding was disrupted when the rat engaged with the Lobsterbot in the E-zone, paralleling findings that a disfavored reward can distort location decoding (Mashhoori et al., 2018). Non-navigational behaviors such as grooming and sniffing also degraded accuracy, suggesting that spatial coding in the mPFC is tightly bound with task-demands such as goal-directed navigation or threat avoidance. Consistent with this, removing the Lobsterbot—reducing the task to simple reward collection—lowered decoding accuracy, echoing previous reports that complex, goal-driven navigation sharpens mPFC spatial selectivity (Hok et al., 2005; Ma et al., 2023).

Consistent with previous findings that mPFC neurons are sensitive to changes in valence—whether reward (Horst & Laubach, 2013), threat (Burgos-Robles et al., 2009), or both (Kim et al., 2018)—the recorded neurons modulated their firing rate around entry into and withdrawal from the E-zone, where sucrose reward and robotic threat coexisted. In line with the mPFC's established role in executive functions (Dalley et al., 2004; G. W. Diehl & Redish, 2023), an ensemble classifier trained on the recorded unit activity predicted AW success with high accuracy (0.73). This prediction remained significantly above the chance level (0.5), even several seconds before the action.

Interestingly, decoding accuracy for distance and avoidance decision did not differ between the PL an IL, despite their well-established functional dissociation in fear conditioning and extinction studies: the PL has been linked to fear expression, and the IL to fear extinction learning (Burgos-Robles et al., 2009; Dejean et al., 2016; Kim et al., 2013; Quirk et al., 2006; Sierra-Mercado et al., 2011; Vidal-Gonzalez et al., 2006). On the other hand, more Type 2 neurons were found in the PL and more Type 1 neurons were found in the IL. To recap, typical Type 1 neurons increased the activity briefly after the head entry and then remained inhibited, while Type 2 neurons showed a burst of activity during head entry and sustained increased

activity. One study employing context-dependent fear discrimination task (Kim et al., 2013) also identified two distinct types of PL units: short-latency CS-responsive units, which increased firing during the initial 150 ms of tone presentation, and persistently firing units, which maintained firing for up to 30 seconds. Given the temporal dynamics of Type 2 neurons, it is possible that our unsupervised clustering method may have merged the two types of neurons found in Kim et al.'s study.

We did not observe decreased IL activity during dynamic foraging, whereas prior studies have shown that IL excitability decreases after fear conditioning (Santini et al., 2008) and that increased IL activity is required for fear extinction learning. In our paradigm, extinction learning was unlikely, as the threat persisted throughout the experiment. Future studies with direct manipulation of these subpopulations, particularly examining head withdrawal timing after such interventions, could provide insight into how these subpopulations guide behavior.

We also found that the activity of Type 2 neurons was highly correlated with the success of AW. What is the role of the Type 2 sub-population in the execution of active defensive behavior? One straightforward hypothesis is that the increased activity simply represents internal motivation of fear or anxiety. This premise is supported by multiple experiments where the PL is vital for the expression of freezing responses (Burgos-Robles et al., 2009; Kim et al., 2018). Consequently, Type 2 units may represent the perceived threat level, which would strongly correlate with the rat's voluntary head-withdrawal. Alternatively, the Type 2 sub-population may contribute to the computation or decision-making processes underlying defensive behaviors. Supporting this, Halladay and Blair (2015) identified two distinct sub-populations in the PL selectively responsive to specific defensive behaviors—conditioned movement inhibition and conditioned movement excitation. Given that Type 2

neurons increased firing during AW trials, they may overlap with the excitation-related neuronal subgroups in Halladay and Blair's study. Despite this interesting conjecture, any analysis based on recording data is only correlational, mandating further studies with direct manipulation of the subpopulation to confirm its functional specificity.

Encoded representations in the mPFC range from prediction of upcoming decisions (Passecker et al., 2019) to rule switching (Laskowski et al., 2016). This diversity of cognitive functions in the mPFC has also been observed in primates, giving rise to the *adaptive coding model* (Duncan, 2001). This model posits that the mPFC neurons are fine-tuned to encode task-relevant information depending on the contexts. Our data support this framework: over 80% of units that encoded distance in the F-zone modulated their firing during defensive decision-making in the E-zone, reflecting a context-dependent functional switching. PCA further revealed that the manifold of populational activity in the E-zone was distinct from that of the F- and N-zones, underscoring the heterogeneity of mPFC activity during threat-reward conflict (for review, see Ebitz & Hayden, 2021).

Importantly, upon exiting the E-zone, neurons resumed distance encoding in the same manner as before. Because the regressor was trained on data from the entire session, such stable decoding would not be possible if encoding schemes changed each time the rats re-entered the F-zone. In our dynamic foraging task, rats rapidly alternated between navigational behavior in the F-zone and confrontational behavior in the E-zone, yet spatial decoding in the F-zone remained stable. This suggests selective, reversible switching between functional modes rather than global reorganization of coding with each context change.

These results point to a dual, context-sensitive encoding scheme in the mPFC, accommodating both spatial navigation and defensive decision-making within the same behavioral session (**Figure 8A**). This multifunctionality helps reconcile conflicting views of

the mPFC's role in spatial orientation by showing that spatial and non-spatial coding can coexist in a task-dependent manner. The "department store" analogy of the mPFC—where diverse functions emerge from learned associations between contexts, locations, events, and adaptive responses—is consistent with our findings.

In our multi-task design, decoder accuracy varied systematically with context, raising the question of how the mPFC determines which information to encode. We consider three possibilities: (1) a bottom-up hypothesis, in which encoding priorities are dictated by sensory inputs or upstream signals; (2) a top-down hypothesis, in which an internal "arbitrator" selects the encoding mode and coordinates relevant information; and (3) a competitive hypothesis, in which bottom-up and top-down influences operate simultaneously, with the dominant input determining the active encoding mode (**Figure 8B**).

Although our study cannot resolve the mechanism, a control experiment in which rats performed a simple shuttling task provides some insight. Removing the Lobsterbot degraded spatial encoding rather than enhancing it, suggesting that reducing task demands alone does not bias the mPFC toward one encoding mode. This aligns with the top-down or arbitrator hypothesis, in which mPFC neurons encode heterogeneous information when task demands are high and require coordinated, goal-directed behavior.

Which brain region might act as this arbitrator? Evidence from human neuroimaging studies implicates the anterior cingulate cortex (ACC) as a central hub for switching cognitive modes. During task switching, the ACC shows increased activation (Hyafil et al., 2009), enhances connectivity with task-specific regions (Aben et al., 2020), correlates with multitask performance (Kondo et al., 2004), and monitors the reliability of competing decision systems (Lee et al., 2014). Collectively, these findings point to a pivotal role for the ACC in coordinating task assignment. Rodent studies also link the ACC to strategic mode switching (Tervo et al.,

2014), suggesting that the rodent ACC could similarly arbitrate between strategies, determining which task-relevant variables are represented in the ventral mPFC, including the PL and IL. Future studies combining multi-context tasks with causal manipulations will be essential to determine whether these functional shifts are driven primarily by top-down arbitration or by bottom-up sensory inputs.

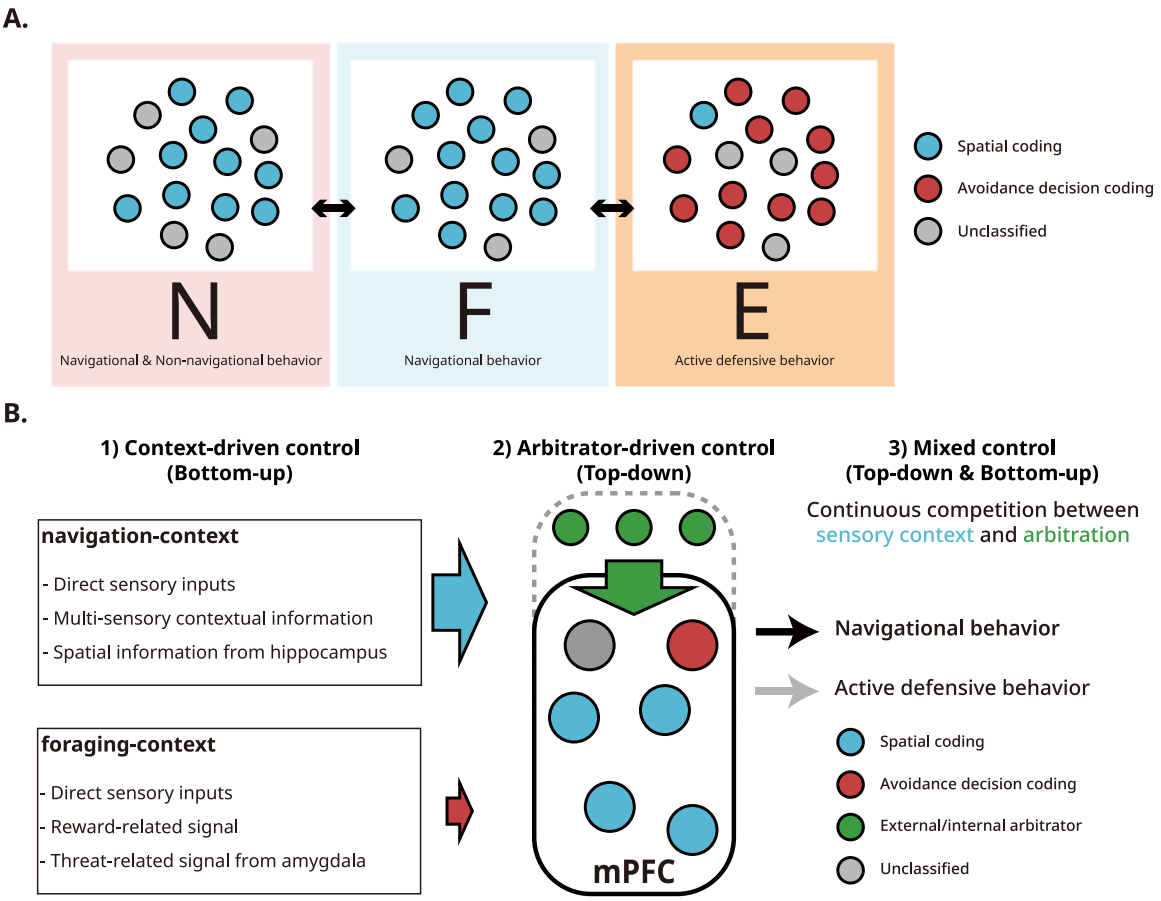


Figure 8. Hypothetical control models by which mPFC neurons assume different functional states.

(A) In the F-zone, navigational behaviors enhance the mPFC's encoding of spatial information compared to other zones. In the N-zone, spatial coding diminishes when the rat engages in non-navigational behaviors. However, in the E-zone, these neurons shift their encoding strategy and become involved in coding for active foraging. We did not find a subset of neurons dedicated exclusively to either spatial coding or active foraging throughout the session. Instead, neurons changed their encoding scheme in a population-wide manner.

(B) Three hypotheses about how the switch is manifested. In this example, most mPFC neurons encode spatial information (blue circles). Information encoded in the mPFC can be regulated by internal/external arbitration signal (top-bottom blue arrow from green circles), or influenced by direct sensory inputs and navigation-related signals (left-right blue arrow) that prompt mPFC neurons to encode spatial information. A third possibility is that both signals compete to gain control.

Materials and Methods

Subjects

Eight male Sprague Dawley rats (250-300 g, 5-6 weeks of age, Orient Bio.) were used in this study. The rats were housed in pairs on a 12-h light/dark cycle with food and water available ad libitum for a one-week acclimation period. In the following week, rats were handled for 10 min each day until they displayed reduced fear responses towards the experimenter.

Apparatus and procedures

The apparatus followed the same dimensions as in Kimm & Choi (2018), measuring 800 mm × 480 mm and a height of 500 mm excluding the robot compartment. It was covered in gray EMF-shield fabric to reduce high-frequency noise during experiments. Battery-powered LED strip lighting illuminated the apparatus and −60 dB white noise masked background sounds. The system employed four Arduino Uno boards and a Raspberry Pi board for experiment control and behavioral tracking. All sensor and control signals were relayed to an RV5 multiprocessor (Tucker-Davis Technologies, FL, USA) for further analysis. Bird's-eye view videos were captured during sessions, and the rat's location was tracked using the butter package (https://github.com/knowblesse/butter). All control schematics and codes are publicly available (https://github.com/knowblesse/Lobster/Apparatus).

Before the training, food and water were gradually restricted until the rats' weight reached 80% of their initial value. Over three days, they acclimated to the apparatus for 20 min. Training began with the habituation phase, during which rats had unrestricted sucrose port access for 20 min. The sucrose port has an IR sensor which was activated by a single lick. The rat usually stays in front of the lick port and continuously lick up to a rate of 6.3 times per

second to obtain sucrose. Any sucrose droplets dropped in the bottom sink were immediately removed by negative pressure so that the rat's behavior was focused on the licking. Upon reaching 1,500 licks per trial, rats proceeded to the three-day shuttling phase. The shuttling session was 30 min long, and the gate was closed 6 s after the first lick, forcing rats into the N-zone to reopen it. The subsequent Lobsterbot phase introduced the Lobsterbot, which executed two grabbing motions after 3 s (30%) or 6 s (70%) from the initial lick. After three days of Lobsterbot sessions, the rats underwent microdrive implant surgery, and recording data were collected from subsequent sessions, either Lobsterbot or shuttling sessions, depending on the experiment. For all post-surgery sessions, those with fewer than 20 approaches in 30 minutes were excluded from further analysis.

Surgery

Before surgery, rats were anesthetized using isoflurane (3%-5%) and oxygen (500 mL/min). Enrofloxacin (5 mg/kg) and meloxicam (1 mg/kg) were administered to prevent infection and mitigate postsurgical pain. A craniotomy was performed at AP +3.2 mm and ML 0.5 mm from the bregma using a carbide burr (Komet, Kempten, Germany) mounted on a high-speed drill (Strong 207A; Saeshin, Daegu, Korea). The tip of each microdrive was lowered to the following coordinates: AP +3.2 mm, ML right 0.3 ~ 0.6 mm, DV −4 mm from the bregma. The precise ML coordinate was determined after the craniotomy to avoid damage to the superior sagittal sinus. Six to seven screws were securely affixed to the skull, and two additional screws above the cerebellum were used as the ground and reference points. Finally, dental cement was applied around the microdrive to secure the implant. Rats were allowed at least one week of recovery before the recordings. All procedures were approved by the Korea University Institutional Animal Care & Use Committee.

Data collection

Two types of microdrives were used in the recordings: a custom-made 4-tetrode microdrive and a pre-assembled 16-channel silicon probe (A2x2-tet-10 mm-150-150-121-HZ16_21 mm, NeuroNexus, MI, USA) mounted on a 3D printed microdrive. The microdrive featured an M1.6-18 mm screw with a 0.35 mm pitch. The recording site was retroactively determined by calculating the number of rotations from the final recording site, identified by a marking lesion.

During each recording session, the rat's microdrive was connected to a ZIF-Clip digital headstage (Tucker-Davis Technologies, FL, USA). The neural signals were collected at a 25 kHz sampling rate, amplified, digitized from the head stage, and transferred to the RZ5 (Tucker-Davis Technologies, FL, USA) for data acquisition. The acquired data stream was further bandpass filtered between 300 and 5000 kHz, and all spikes above 3 to 4 sigma compared to the stream were collected for further clustering analysis. The Offline Sorter (Plexon, TX, USA) was used for the manual clustering process. All units with firing rates below 0.5 Hz or exactly 60 Hz (power noise) were considered artifacts and removed from further analysis. Sessions with fewer than 10 recorded units were also excluded.

Spatial decoding data preparation

For the neural data used in the spatial encoding analysis of the mPFC, the following method was employed. First, a Gaussian kernel with a standard deviation of 100 ms and a length of 1000 ms was prepared. Each unit's spike data were down-sampled and serialized to a 1000 Hz vector, and the kernel was applied to smooth the signal. The mean and standard deviation values of this vector were computed and used to normalize the vector. After normalization, the vector was binned into 50 ms windows, resulting in 20 data points per second. Finally, each datapoint was paired with the corresponding location data.

The spatial firing map was constructed using the same normalized activity data. Neural activity was averaged for each pixel and then smoothed using a Gaussian filter with a sigma of 15 px and a filter size of 1001 px.

ANN regressor

A neural network regressor was built using the PyTorch package. It comprised seven layers in total, including the output layer; four of these were fully connected (FC) layers and two were drop-out layers. The initial weights for the FC layers were normalized with a mean of zero and a standard deviation of 0.2. Mean squared error loss was used in conjunction with the stochastic gradient descent method. For each regressor, a control regressor was built, which used the same data-label set, but the relationships were randomly shuffled. The evaluation of the regression results used a stratified 5-fold cross-validation while ensuring that each subset maintained the same proportion of data from the three zones. The five regression results per dataset were combined, and these results were compared with the true dataset using MAE. For the training/testing dataset, preprocessed 50 ms-long neural data were used. The binned neural data from all units (n) were concatenated to form a $1 \times n$ size vector. These vectors were paired with the corresponding distance during the 50 ms time window. In the dataset adjusted for uneven location visits, we divided distance values into five equally sized bins. Then, a sub-dataset was created that contains an equal number of data points for each of these bins. Two different GPU servers were employed for training: a Tesla V100 from the National IT Industry Promotion Agency (NIPA, Seoul, Korea) and an in-house A4000.

Decoding error heatmap construction

To construct the decoding error heatmap, the following method was used. First, error data from each pixel were collected and mean values were computed. To accommodate pixels not visited by a rat during an entire session, an interpolation method was employed. A custom

Matlab script, built using the *scatteredInterpolant* function, was used to interpolate errors for unknown locations. The 2D interpolated decoding error was subsequently smoothed using a Gaussian filter with a sigma value of 15 px and a filter size of 1001 px. These smoothed error values were then converted to cm, and a contour plot was overlaid using 25 levels.

PCA

A custom Python script for PCA was created using the *sci-kit-learn* package. Neural data vectors generated for the spatial decoding analysis were used for this analysis. The behavior and head tracking data were matched with neural data vectors, and the corresponding zone was labeled to the vectors. All neural data vectors from a single session were projected onto a new space using the PCA. For each zone, the centroid of the projected neural data vectors was calculated using Euclidean distance. The within-zone distance ($d^z_{within}(x)$) and the between-zone distance ($d^z_{between}(x)$) were calculated using the equations below.

$$d^z_{within}(x) = \| C^z - x^z \|_2,$$

$$d^z_{between}(x) = \frac{\sum_{z'=\neg z} \| C^{z'} - x^z \|_2}{2},$$

where C^z denotes the centroid of the zone z and x^zrepresents a neural data vector from the zone z.

Representative figures were drawn in a manner to highlight the cluster differences. Among all projected neural vectors designated to a zone, the 2000 vectors with the shortest distance from the centroid were selected. From these vectors, 200 were randomly chosen and the first two dimensions of the projected value were plotted.

Behavior-responsive unit analysis

Behavior-responsive units were identified using the following method. For each trial,

spikes within a 4-s window (2 s before and after the onset of three events: gate opening upon returning to the N-zone, head-entry, and head-withdrawal) were counted and registered in 50 ms-long time bins, resulting in 80 data points per trial. These neural activities were then averaged across all trials, producing a 1 × 80 vector for each event. The binned neural activity ± 2 s from the entrance to the N-zone was chosen as a baseline, considering that the rat was consistently oriented towards areas other than the robot. The mean and standard deviation of the baseline were used to normalize activity during the head-entry and the head-withdrawal. The overall method created two event vectors, head-entry-vector and head-withdrawal-vector, per unit. A unit was considered behavior-responsive to an event if any value within the 80 data points of the event vector was greater or less than 3 sigma.

Hierarchical cell clustering analysis

To cluster cells that exhibited similar responses during the head-entry and/or head-withdrawal, an unsupervised hierarchical cell clustering method was employed. For every pair of units, the similarity score using Pearson's r was calculated, and the pair with the highest similarity score was grouped to form a new pseudo-unit that had the average response of the two units. This procedure was repeated until all units were grouped into eight pseudo-units. Among each pseudo-unit, if the number of units belonging to the group was equal to or greater than 50, the pseudo-unit was considered a valid cluster. Otherwise, the pseudo-unit was classified as an artifact.

To compensate for the limited number of neurons recorded per session, the hyperparameter set was chosen to generalize their activity and categorize them into major types, allowing us to focus on neurons that appeared across multiple recording sessions. Although changes in the hyperparameter sets resulted in different numbers of clusters, the major activity types remained consistent (**Supplementary Figure S8**). However, there is a chance that this

method may not differentiate smaller subsets of neurons, particularly those with fewer than 50 recorded neurons.

Run-and-stop event analysis

To examine whether a change in movement velocity was related to neural activity, a "run-and-stop" event was defined by the following criteria: 1) to qualify as "run" at a given moment, the rat must be moving at least 7.5 cm/s for at least 2 s or longer; 2) to qualify as the "stop" at a given moment, the rat must be moving at most 5 cm/s for at least 1 s or longer; and 3) the "run-and-stop" event was the time point between "run" and "stop" outside the E-zone. The temporal relationship between the run-and-stop event and neural activity was visualized using the same method as in the behavior-responsive unit analysis.

Naïve Bayes AW/EW classifier and data preparation

For the neural data used in the head-withdrawal-type (AW/EW) classification analysis, a data preparation method similar to that of the spatial decoding analysis was employed. First, the same Gaussian kernel was generated and applied to the down-sampled and serialized neural data. Mean and standard deviation values of this vector were computed for later normalization. Unlike the 50 ms-length neural data used in spatial regression, 2-s-length neural data was used. For each trial, all spikes within a 2-s window around the onset of the head-withdrawal were collected and serialized into a 1000 Hz vector. The Gaussian kernel was applied and normalized using precalculated values. This vector was then binned into 50 ms intervals, resulting in 40 data points per event. Binned vectors from all recorded units (n) were concatenated to form a $1 \times (40n)$ size vector per trial. These vectors were then paired with their corresponding head-withdrawal-type, AW or EW. In the analysis using varying time windows, the previously employed data preparation method was the same, but the time range of the neural data varied relative to the head-withdrawal.

A naïve Bayesian classifier based on the multivariate Bernoulli model was built using the *sci-kit-learn* package. A custom Python script was written to train and evaluate the classifiers. The classifier employed a uniform prior and smoothing parameter, alpha, which was set to 1 in all classifiers used in the study. Similar to the analysis using the ANN regressor, a control classifier was built that was trained with shuffled data. The evaluation of the classifier results employed the stratified 5-fold cross-validation method with each subset maintaining the same proportion of head-entry/head-withdrawal. Five classification results per dataset were combined, and these results were compared with the true dataset using balanced accuracy, calculated using the equation below.

$$\text{balanced accuracy} = \frac{1}{2}\left(\frac{n_{correctly\ labeled\ as\ HE}}{n_{HE}} + \frac{n_{correctly\ labeled\ as\ HW}}{n_{HW}}\right)$$

To interpret how the classifier anticipated upcoming head-withdrawal behavior, class discrimination index was employed, focusing on a dataset with a time window of −2 to 0 s. To calculate the index, the class likelihood for every dimension was extracted during each test dataset evaluation, and a natural logarithm was applied to balance the small likelihood values. With each unit's neural data divided into 40 bins, the log likelihood values were added to create a joint probability. This process yielded five joint likelihood values per unit, and their average was computed. Following Mladenić & Grobelnik's (1999) method, the log odds ratio between P(unit|AW) and P(unit|EW) was calculated for all units using the equation below.

$$\text{LogOddsRatio}(unit) = \log\left(\frac{P(unit|AW)\big(1 - P(unit|EW)\big)}{\big(1 - P(unit|AW)\big)P(unit|EW)}\right),$$

where P(unit|AW) denotes joint likelihood (probability of observing unit activity given the AW) and P(unit|EW) denotes the same for the EW. These values were then compared according to the unit's assigned group: Type 1, Type 2, or others. Taking the natural log of the odds ratio

brings about a symmetricity in the results. A positive value represented a unit's relatively increased activity during the AW compared to the EW, and a negative value represented the opposite.

Statistical analysis

Statistical analyses were performed using Prism (GraphPad, MA, USA). To confirm normality of the data, Shapiro-Wilk test was applied to all data we obtained from the experiment. For all statistical tests, *p*-value of <.05 were used to assess the significance of differences and all *t*-tests were two tailed. The correlation analysis comparing feature importance scores from two decoders excluded the 1% of outliers before computing Pearson's r. Both ANOVA and mixed-effect model analysis applied the Geisser-Greenhouse correction to account for violations of the sphericity assumption. The corrected degree of freedom is presented alongside F values. For multiple comparison tests, Sidak's correction was used to control the family-wise error rate. For multiple linear regression analysis of the distance decoding error, the following model was incorporated.

$$\begin{aligned}\mathrm{MSE} = \beta_0 &+ \beta_1(\text{number of units used in the decoder}) \\ &+ \beta_2(\text{presence of Lobsterbot}) \\ &+ \beta_3(\text{recording site; PL, IL})\end{aligned}$$

The model fitting method incorporated the least squares method assuming a Gaussian distribution of residuals. Boolean variables, *presence of Lobsterbot*, and *recording site* used dummy codes (0: Lobster, 1: Control; 0: PL, 1: IL) to represent the state of the experiment. All error bars in graphs represent one standard error of the mean. Statistically significant differences are denoted as * for *p* <.05, ** for *p* <.01, and *** for *p* <.001.

Code availability

The Matlab and Python scripts used for classification and behavioral analysis are openly accessible at the following repository: https://github.com/knowblesse/Lobster.

Data Availability

All raw data including behavior, neural, and video are available upon request.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF-2021M3E5D2A01023887) funded by the Korean Government, MSIT. GPU resource was supported by High-performance computing support project by NIPA, Seoul, Korea.

Author Contributions

J. H. J and J-S. C designed the study. J. H. J conducted the experiment and analyzed the data. J. H. J and J-S. C wrote the manuscript.

Competing Interest Statement

The authors declare no competing interests.

References

Aben, B., Buc Calderon, C., Van Den Bussche, E., & Verguts, T. (2020). Cognitive Effort Modulates Connectivity between Dorsal Anterior Cingulate Cortex and Task-Relevant Cortical Areas. *The Journal of Neuroscience*, *40*(19), 3838–3848. https://doi.org/10.1523/JNEUROSCI.2948-19.2020

Blanchard, R. J., & Blanchard, D. C. (1969). Crouching as an index of fear. *Journal of Comparative and Physiological Psychology*, *67*(3), 370–375. https://doi.org/10.1037/h0026779

Boehlen, A., Heinemann, U., & Henneberger, C. (2016). Hierarchical spike clustering analysis for investigation of interneuron heterogeneity. *Neuroscience Letters*, *619*, 86–91. https://doi.org/10.1016/j.neulet.2016.03.024

Böhm, C., & Lee, A. K. (2020). Canonical goal-selective representations are absent from prefrontal cortex in a spatial working memory task requiring behavioral flexibility. *eLife*, *9*, e63035. https://doi.org/10.7554/eLife.63035

Burgos-Robles, A., Vidal-Gonzalez, I., & Quirk, G. J. (2009). Sustained Conditioned Responses in Prelimbic Prefrontal Neurons Are Correlated with Fear Expression and Extinction Failure. *Journal of Neuroscience*, *29*(26), 8474–8482. https://doi.org/10.1523/JNEUROSCI.0378-09.2009

Dalley, J. W., Cardinal, R. N., & Robbins, T. W. (2004). Prefrontal executive and cognitive functions in rodents: Neural and neurochemical substrates. *Neuroscience & Biobehavioral Reviews*, *28*(7), 771–784. https://doi.org/10.1016/j.neubiorev.2004.09.006

Dejean, C., Courtin, J., Karalis, N., Chaudun, F., Wurtz, H., Bienvenu, T. C. M., & Herry, C. (2016). Prefrontal neuronal assemblies temporally control fear behaviour. *Nature*,

535(7612), 420–424. https://doi.org/10.1038/nature18630

Di Miceli, M., Husson, Z., Ruel, P., Layé, S., Cota, D., Fioramonti, X., Bosch-Bouju, C., & Gronier, B. (2020). In silico Hierarchical Clustering of Neuronal Populations in the Rat Ventral Tegmental Area Based on Extracellular Electrophysiological Properties. *Frontiers in Neural Circuits*, *14*, 51. https://doi.org/10.3389/fncir.2020.00051

Diehl, G. W., & Redish, A. D. (2023). Differential processing of decision information in subregions of rodent medial prefrontal cortex. *eLife*, *12*, e82833. https://doi.org/10.7554/eLife.82833

Diehl, M. M., Bravo-Rivera, C., Rodriguez-Romaguera, J., Pagan-Rivera, P. A., Burgos-Robles, A., Roman-Ortiz, C., & Quirk, G. J. (2018). Active avoidance requires inhibitory signaling in the rodent prelimbic prefrontal cortex. *eLife*, *7*, e34657. https://doi.org/10.7554/eLife.34657

Duncan, J. (2001). An adaptive coding model of neural function in prefrontal cortex. *Nature Reviews Neuroscience*, *2*(11), 820–829. https://doi.org/10.1038/35097575

Durstewitz, D., Vittoz, N. M., Floresco, S. B., & Seamans, J. K. (2010). Abrupt Transitions between Prefrontal Neural Ensemble States Accompany Behavioral Transitions during Rule Learning. *Neuron*, *66*(3), 438–448. https://doi.org/10.1016/j.neuron.2010.03.029

Ebitz, R. B., & Hayden, B. Y. (2021). The population doctrine in cognitive neuroscience. *Neuron*, *109*(19), 3055–3068. https://doi.org/10.1016/j.neuron.2021.07.011

Fanselow, M. S., & Lester, L. S. (1988). A functional behavioristic approach to aversively motivated behavior: Predatory imminence as a determinant of the topography of defensive behavior. In *Evolution and learning.* (pp. 185–212). Lawrence Erlbaum Associates, Inc.

Fuster, J. M. (1973). Unit activity in prefrontal cortex during delayed-response performance: Neuronal correlates of transient memory. *Journal of Neurophysiology*, *36*(1), 61–78. https://doi.org/10.1152/jn.1973.36.1.61

Gabbott, P. L. A., Warner, T. A., Jays, P. R. L., Salway, P., & Busby, S. J. (2005). Prefrontal cortex in the rat: Projections to subcortical autonomic, motor, and limbic centers. *The Journal of Comparative Neurology*, *492*(2), 145–177. https://doi.org/10.1002/cne.20738

Glaser, J. I., Benjamin, A. S., Chowdhury, R. H., Perich, M. G., Miller, L. E., & Kording, K. P. (2020). Machine Learning for Neural Decoding. *Eneuro*, *7*(4), ENEURO.0506-19.2020. https://doi.org/10.1523/ENEURO.0506-19.2020

Grewal, S. S., Shepherd, J. K., Bill, D. J., Fletcher, A., & Dourish, C. T. (1997). Behavioural and pharmacological characterisation of the canopy stretched attend posture test as a model of anxiety in mice and rats. *Psychopharmacology*, *133*(1), 29–38. https://doi.org/10.1007/s002130050367

Guise, K. G., & Shapiro, M. L. (2017). Medial Prefrontal Cortex Reduces Memory Interference by Modifying Hippocampal Encoding. *Neuron*, *94*(1), 183-192.e8. https://doi.org/10.1016/j.neuron.2017.03.011

Hafting, T., Fyhn, M., Molden, S., Moser, M.-B., & Moser, E. I. (2005). Microstructure of a spatial map in the entorhinal cortex. *Nature*, *436*(7052), 801–806. https://doi.org/10.1038/nature03721

Halladay, L. R., & Blair, H. T. (2015). Distinct ensembles of medial prefrontal cortex neurons are activated by threatening stimuli that elicit excitation vs. Inhibition of movement. *Journal of Neurophysiology*, *114*(2), 793–807. https://doi.org/10.1152/jn.00656.2014

Hok, V., Save, E., Lenck-Santini, P. P., & Poucet, B. (2005). Coding for spatial goals

in the prelimbic/infralimbic area of the rat frontal cortex. *Proceedings of the National Academy of Sciences*, *102*(12), 4602–4607. https://doi.org/10.1073/pnas.0407332102

Hoover, W. B., & Vertes, R. P. (2007). Anatomical analysis of afferent projections to the medial prefrontal cortex in the rat. *Brain Structure and Function*, *212*(2), 149–179. https://doi.org/10.1007/s00429-007-0150-4

Horst, N. K., & Laubach, M. (2013). Reward-related activity in the medial prefrontal cortex is driven by consumption. *Frontiers in Neuroscience*, *7*. https://doi.org/10.3389/fnins.2013.00056

Hyafil, A., Summerfield, C., & Koechlin, E. (2009). Two Mechanisms for Task Switching in the Prefrontal Cortex. *The Journal of Neuroscience*, *29*(16), 5135–5142. https://doi.org/10.1523/JNEUROSCI.2828-08.2009

Jarzebowski, P., Hay, Y. A., Grewe, B. F., & Paulsen, O. (2022). Different encoding of reward location in dorsal and intermediate hippocampus. *Current Biology*, *32*(4), 834-841.e5. https://doi.org/10.1016/j.cub.2021.12.024

Jung, M. W., Qin, Y., McNaughton, B. L., & Barnes, C. A. (1998). Firing characteristics of deep layer neurons in prefrontal cortex in rats performing spatial working memory tasks. *Cerebral Cortex*, *8*(5), 437–450. https://doi.org/10.1093/cercor/8.5.437

Kaefer, K., Nardin, M., Blahna, K., & Csicsvari, J. (2020). Replay of Behavioral Sequences in the Medial Prefrontal Cortex during Rule Switching. *Neuron*, *106*(1), 154-165.e6. https://doi.org/10.1016/j.neuron.2020.01.015

Kim, E. J., Kim, N., Kim, H. T., & Choi, J.-S. (2013). The prelimbic cortex is critical for context-dependent fear expression. *Frontiers in Behavioral Neuroscience*, *7*. https://doi.org/10.3389/fnbeh.2013.00073

Kim, E. J., Kong, M.-S., Park, S. G., Mizumori, S. J. Y., Cho, J., & Kim, J. J. (2018). Dynamic coding of predatory information between the prelimbic cortex and lateral amygdala in foraging rats. *Science Advances*, *4*(4), eaar7328. https://doi.org/10.1126/sciadv.aar7328

Kimm, S., & Choi, J.-S. (2018). Sensory and motivational modulation of immediate and delayed defensive responses under dynamic threat. *Journal of Neuroscience Methods*, *307*, 84–94. https://doi.org/10.1016/j.jneumeth.2018.06.023

Kondo, H., Osaka, N., & Osaka, M. (2004). Cooperation of the anterior cingulate cortex and dorsolateral prefrontal cortex for attention shifting. *NeuroImage*, *23*(2), 670–679. https://doi.org/10.1016/j.neuroimage.2004.06.014

Korn, C. W., & Bach, D. R. (2019). Minimizing threat via heuristic and optimal policies recruits hippocampus and medial prefrontal cortex. *Nature Human Behaviour*, *3*(7), 733–745. https://doi.org/10.1038/s41562-019-0603-9

Krishnan, S., Heer, C., Cherian, C., & Sheffield, M. E. J. (2022). Reward expectation extinction restructures and degrades CA1 spatial maps through loss of a dopaminergic reward proximity signal. *Nature Communications*, *13*(1), 6662. https://doi.org/10.1038/s41467-022-34465-5

Laskowski, C. S., Williams, R. J., Martens, K. M., Gruber, A. J., Fisher, K. G., & Euston, D. R. (2016). The role of the medial prefrontal cortex in updating reward value and avoiding perseveration. *Behavioural Brain Research*, *306*, 52–63. https://doi.org/10.1016/j.bbr.2016.03.007

Lee, S. W., Shimojo, S., & O'Doherty, J. P. (2014). Neural Computations Underlying Arbitration between Model-Based and Model-free Learning. *Neuron*, *81*(3), 687–699. https://doi.org/10.1016/j.neuron.2013.11.028

Ma, X., Zheng, C., Chen, Y., Pereira, F., & Li, Z. (2023). Working memory and reward increase the accuracy of animal location encoding in the medial prefrontal cortex. *Cerebral Cortex*, *33*(5), 2245–2259. https://doi.org/10.1093/cercor/bhac205

Mackintosh, J. H., & Grant, E. C. (1963). A Comparison of the Social Postures of Some Common Laboratory Rodents. *Behaviour*, *21*(3–4), 246–259. https://doi.org/10.1163/156853963X00185

Mashhoori, A., Hashemnia, S., McNaughton, B. L., Euston, D. R., & Gruber, A. J. (2018). Rat anterior cingulate cortex recalls features of remote reward locations after disfavoured reinforcements. *eLife*, *7*, e29793. https://doi.org/10.7554/eLife.29793

McClelland, W. J., & Colman, F. D. (1967). Activity and different types of electric shock stimuli. *Psychonomic Science*, *7*(11), 391–392. https://doi.org/10.3758/BF03331139

Mladenić, D., & Grobelnik, M. (1999). *Feature Selection for Unbalanced Class Distribution and Naive Bayes.* 258–267.

Mobbs, D., Headley, D. B., Ding, W., & Dayan, P. (2020). Space, Time, and Fear: Survival Computations along Defensive Circuits. *Trends in Cognitive Sciences*, *24*(3), 228–241. https://doi.org/10.1016/j.tics.2019.12.016

Mobbs, D., & Kim, J. J. (2015). Neuroethological studies of fear, anxiety, and risky decision-making in rodents and humans. *Current Opinion in Behavioral Sciences*, *5*, 8–15. https://doi.org/10.1016/j.cobeha.2015.06.005

Mobbs, D., Petrovic, P., Marchant, J. L., Hassabis, D., Weiskopf, N., Seymour, B., Dolan, R. J., & Frith, C. D. (2007). When Fear Is Near: Threat Imminence Elicits Prefrontal-Periaqueductal Gray Shifts in Humans. *Science*, *317*(5841), 1079–1083. https://doi.org/10.1126/science.1144298

Nigro, M., Tortorelli, L. S., Dinh, K., Garad, M., Zlebnik, N. E., & Yang, H. (2023). *Prefrontal dynamics and encoding of flexible rule switching* [Preprint]. Neuroscience. https://doi.org/10.1101/2023.12.13.571356

Nyberg, N., Duvelle, É., Barry, C., & Spiers, H. J. (2022). Spatial goal coding in the hippocampal formation. *Neuron*, *110*(3), 394–422. https://doi.org/10.1016/j.neuron.2021.12.012

O'Keefe, J., & Dostrovsky, J. (1971). The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Research*, *34*(1), 171–175. https://doi.org/10.1016/0006-8993(71)90358-1

Park, J., & Choi, J.-S. (2010). Long-term synaptic changes in two input pathways into the lateral nucleus of the amygdala underlie fear extinction. *Learning & Memory*, *17*(1), 23–34. https://doi.org/10.1101/lm.1482910

Passecker, J., Mikus, N., Malagon-Vina, H., Anner, P., Dimidschstein, J., Fishell, G., Dorffner, G., & Klausberger, T. (2019). Activity of Prefrontal Neurons Predict Future Choices during Gambling. *Neuron*, *101*(1), 152-164.e7. https://doi.org/10.1016/j.neuron.2018.10.050

Paxinos, G., & Watson, C. (2007). *The rat brain in stereotaxic coordinates* (6th ed). Elsevier.

Poucet, B. (1997). Searching for spatial unit firing in the prelimbic area of the rat medial prefrontal cortex. *Behavioural Brain Research*, *84*(1–2), 151–159. https://doi.org/10.1016/S0166-4328(96)00144-1

Powell, N. J., & Redish, A. D. (2014). Complex neural codes in rat prelimbic cortex are stable across days on a spatial decision task. *Frontiers in Behavioral Neuroscience*, *8*. https://doi.org/10.3389/fnbeh.2014.00120

Pyke, G. H., Pulliam, H. R., & Charnov, E. L. (1977). Optimal Foraging: A Selective Review of Theory and Tests. *The Quarterly Review of Biology*, *52*(2), 137–154.

Quirk, G. J., Garcia, R., & González-Lima, F. (2006). Prefrontal Mechanisms in Extinction of Conditioned Fear. *Biological Psychiatry*, *60*(4), 337–343. https://doi.org/10.1016/j.biopsych.2006.03.010

Santini, E., Quirk, G. J., & Porter, J. T. (2008). Fear Conditioning and Extinction Differentially Modify the Intrinsic Excitability of Infralimbic Neurons. *The Journal of Neuroscience*, *28*(15), 4028–4036. https://doi.org/10.1523/JNEUROSCI.2623-07.2008

Sapiurka, M., Squire, L. R., & Clark, R. E. (2016). Distinct roles of hippocampus and medial prefrontal cortex in spatial and nonspatial memory. *Hippocampus*, *26*(12), 1515–1524. https://doi.org/10.1002/hipo.22652

Sauer, J.-F., Folschweiller, S., & Bartos, M. (2022). Topographically organized representation of space and context in the medial prefrontal cortex. *Proceedings of the National Academy of Sciences*, *119*(6), e2117300119. https://doi.org/10.1073/pnas.2117300119

Sierra-Mercado, D., Corcoran, K. A., Lebrón-Milad, K., & Quirk, G. J. (2006). Inactivation of the ventromedial prefrontal cortex reduces expression of conditioned fear and impairs subsequent recall of extinction. *European Journal of Neuroscience*, *24*(6), 1751–1758. https://doi.org/10.1111/j.1460-9568.2006.05014.x

Sierra-Mercado, D., Padilla-Coreano, N., & Quirk, G. J. (2011). Dissociable Roles of Prelimbic and Infralimbic Cortices, Ventral Hippocampus, and Basolateral Amygdala in the Expression and Extinction of Conditioned Fear. *Neuropsychopharmacology*, *36*(2), 529–538. https://doi.org/10.1038/npp.2010.184

Spellman, T., Rigotti, M., Ahmari, S. E., Fusi, S., Gogos, J. A., & Gordon, J. A. (2015).

Hippocampal–prefrontal input supports spatial encoding in working memory. *Nature*, *522*(7556), 309–314. https://doi.org/10.1038/nature14445

Stephens, D. W., & Krebs, J. R. (1986). *Foraging theory*. Princeton University Press.

Tervo, D. G. R., Proskurin, M., Manakov, M., Kabra, M., Vollmer, A., Branson, K., & Karpova, A. Y. (2014). Behavioral Variability through Stochastic Choice and Its Gating by Anterior Cingulate Cortex. *Cell*, *159*(1), 21–32. https://doi.org/10.1016/j.cell.2014.08.037

Vertes, R. P. (2004). Differential projections of the infralimbic and prelimbic cortex in the rat. *Synapse*, *51*(1), 32–58. https://doi.org/10.1002/syn.10279

Vidal-Gonzalez, I., Vidal-Gonzalez, B., Rauch, S. L., & Quirk, G. J. (2006). Microstimulation reveals opposing influences of prelimbic and infralimbic cortex on the expression of conditioned fear. *Learning & Memory*, *13*(6), 728–733. https://doi.org/10.1101/lm.306106

Wikenheiser, A. M., & Redish, A. D. (2015). Hippocampal theta sequences reflect current goals. *Nature Neuroscience*, *18*(2), 289–294. https://doi.org/10.1038/nn.3909

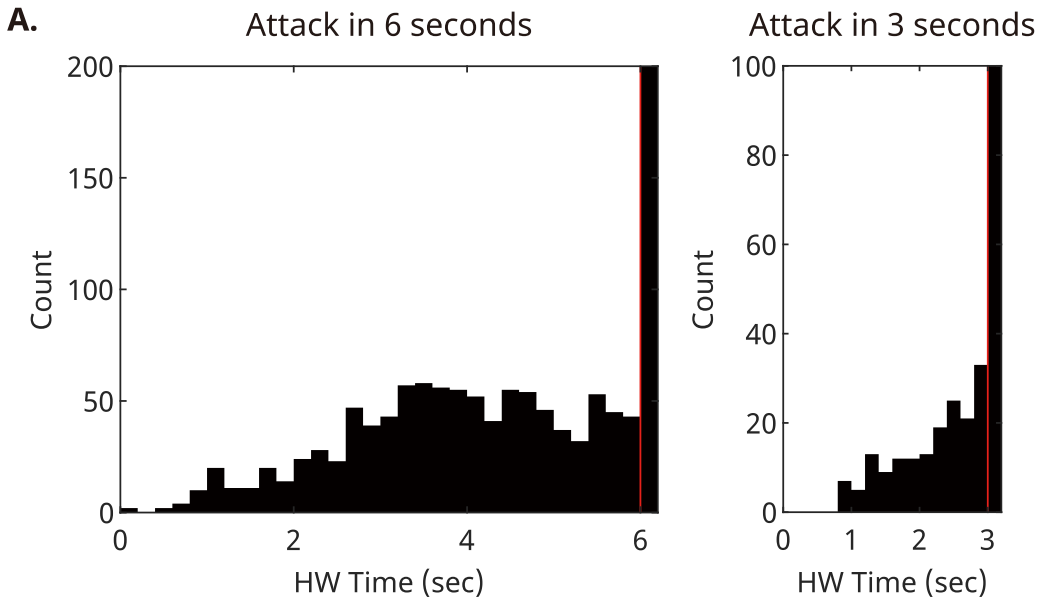
Wikenheiser, A. M., Stephens, D. W., & Redish, A. D. (2013). Subjective costs drive overly patient foraging strategies in rats on an intertemporal foraging task. *Proceedings of the National Academy of Sciences*, *110*(20), 8308–8313. https://doi.org/10.1073/pnas.1220738110

Yang, Y., & Mailman, R. B. (2018). Strategic neuronal encoding in medial prefrontal cortex of spatial working memory in the T-maze. *Behavioural Brain Research*, *343*, 50–60. https://doi.org/10.1016/j.bbr.2018.01.020

Yu, J. Y., Liu, D. F., Loback, A., Grossrubatscher, I., & Frank, L. M. (2018). Specific hippocampal representations are linked to generalized cortical representations in memory.

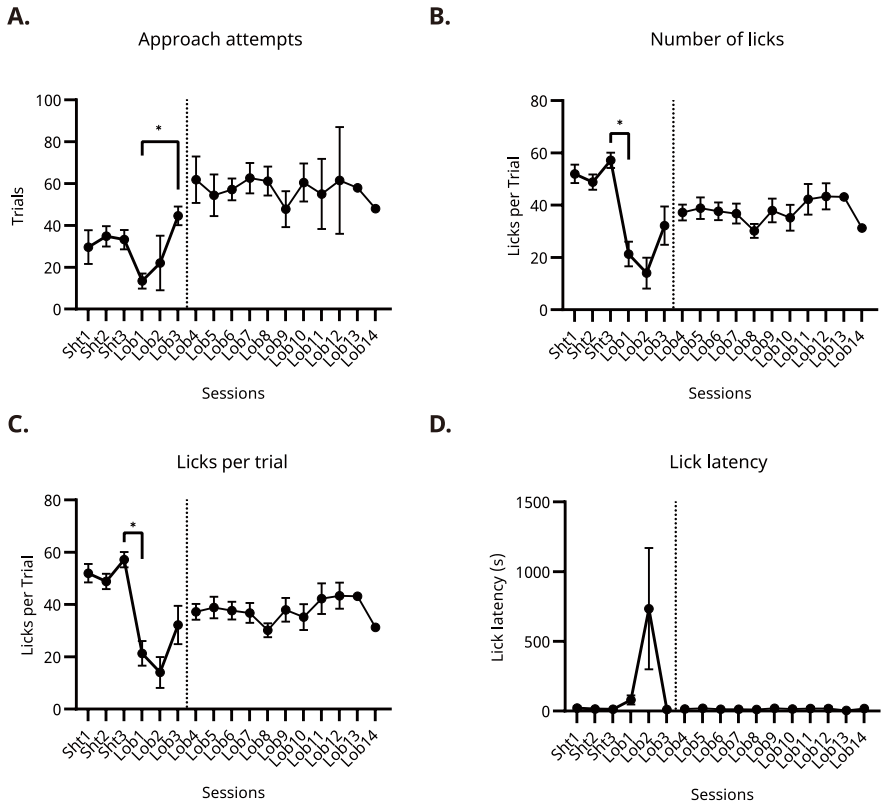
Nature Communications, *9*(1), 2209. https://doi.org/10.1038/s41467-018-04498-w

Supplementary Figures



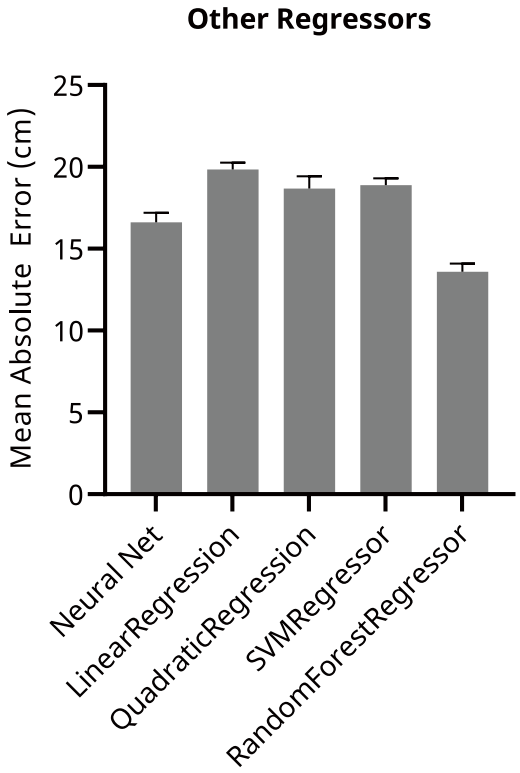
Supplementary Figure S1 Head withdrawal time distribution across all subjects, categorized by trial type

Despite the use of two distinct attack times (3s and 6s), there was no noticeable increase in head with drawals around the 3-second mark, indicating that the rats did not rely on a 3-second countdown. Inst ead, they exhibited a relatively stable distribution leading up to the 6-second attack.



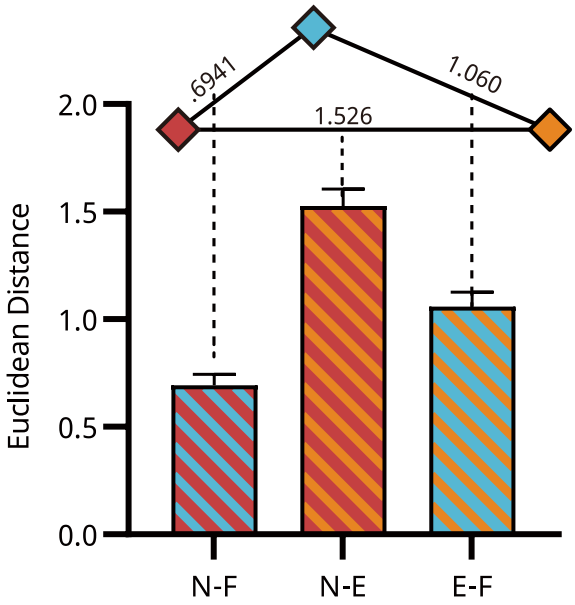
Supplementary Figure S2 Foraging-related behavioral indices fluctuate upon the initial encounter w ith the Lobsterbot but stabilize after 3 sessions.

(A) Number of approaches. The number of approaches, measured in total trials, decreased after the in itial encounter with the Lobster (Lob1), but later increased after 3 Lob sessions. **(B)** Number of lickin g behaviors. The number of licking behaviors significantly decreased during the first encounter but re turned after 3 sessions. **(C)** Number of licks per trial. The number of licking behaviors per trial was d ecreased after the encounter. **(D)** Lick latency. The lick latency increased after encounter but returne d to pre-encounter level after 3 sessions. The black dotted line indicates the timepoint of surgery. Error bars represent SEM.

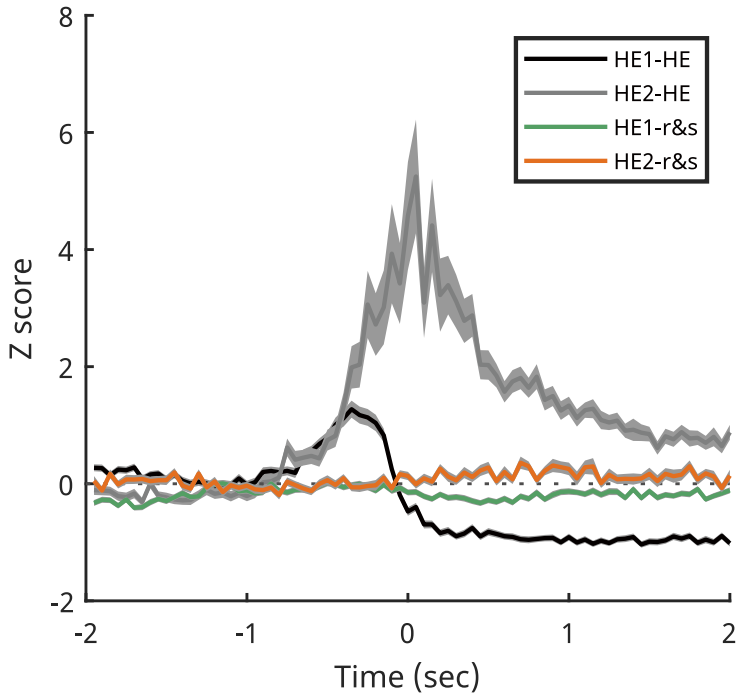


Supplementary Figure S3 Comparison between distance regressor algorithms

A comparison of distance regressors using the same dataset showed that the artificial neural network (ANN) and the random forest regressor outperformed the others, but ANN was chosen for its strong generalization to noisy neural data and robustness to hyperparameters. Error bars represent SEM.



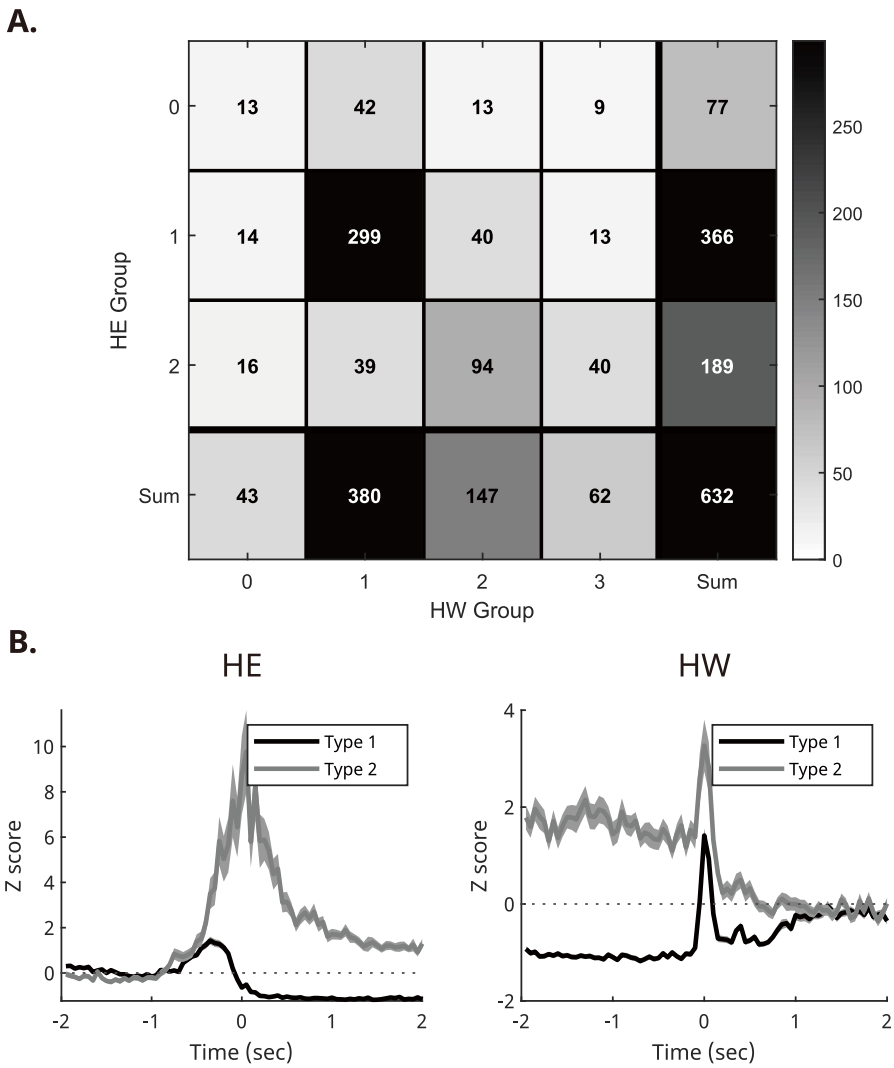
Supplementary Figure S4 Distances between each centroid pairs from all recording sessions. Figure 4B result was reanalyzed using partial dataset which excluded "critical event times" defined as ±1 second from head-entry and head-withdrawal. Even removing behaviorally significant data, E-zone's populational activity was distinctly positioned compare to other zones. Error bars represent SE M.



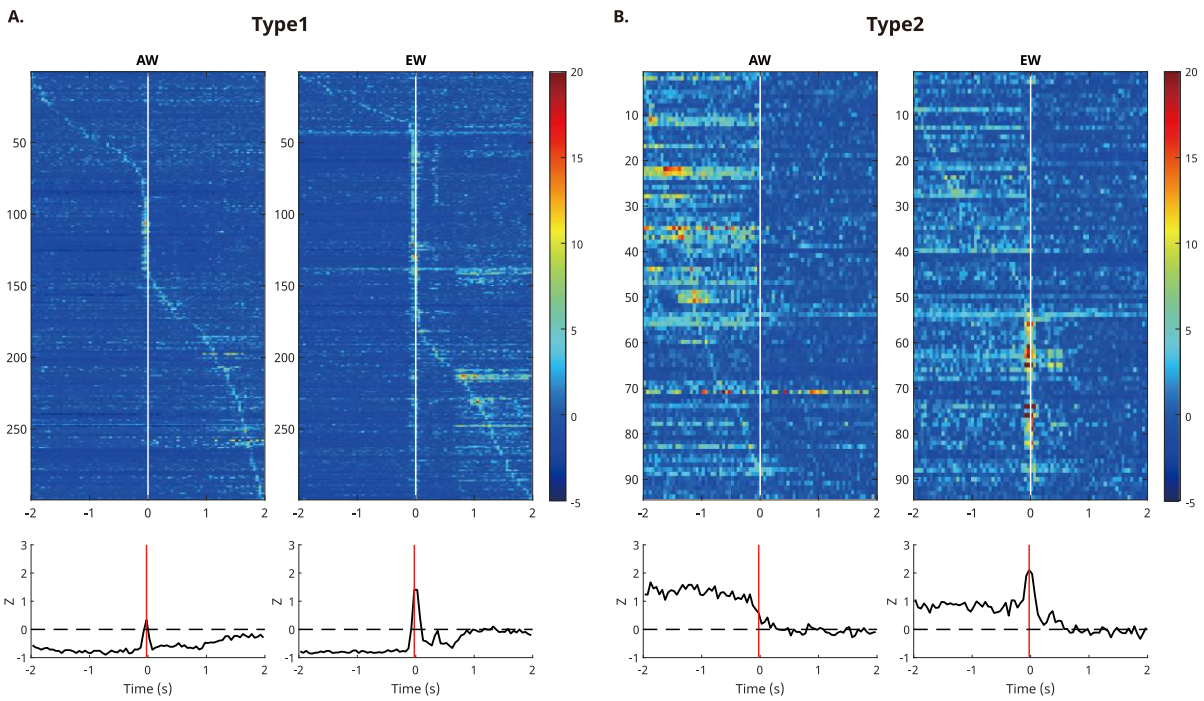
Supplementary Figure S5 A run-and-stop event (sudden velocity drop outside the E-zone) does not evoke neural modulation.

Normalized activity of HE1 and HE2 units during run-and-stop events (colored; HE1-r&s and HE2-r&s) show no modulation of neural activity compared to highly modulated activity around the head-entry (black and gray; HE1-HE and HE2-HE). Gray shadings indicate SEM.

1031



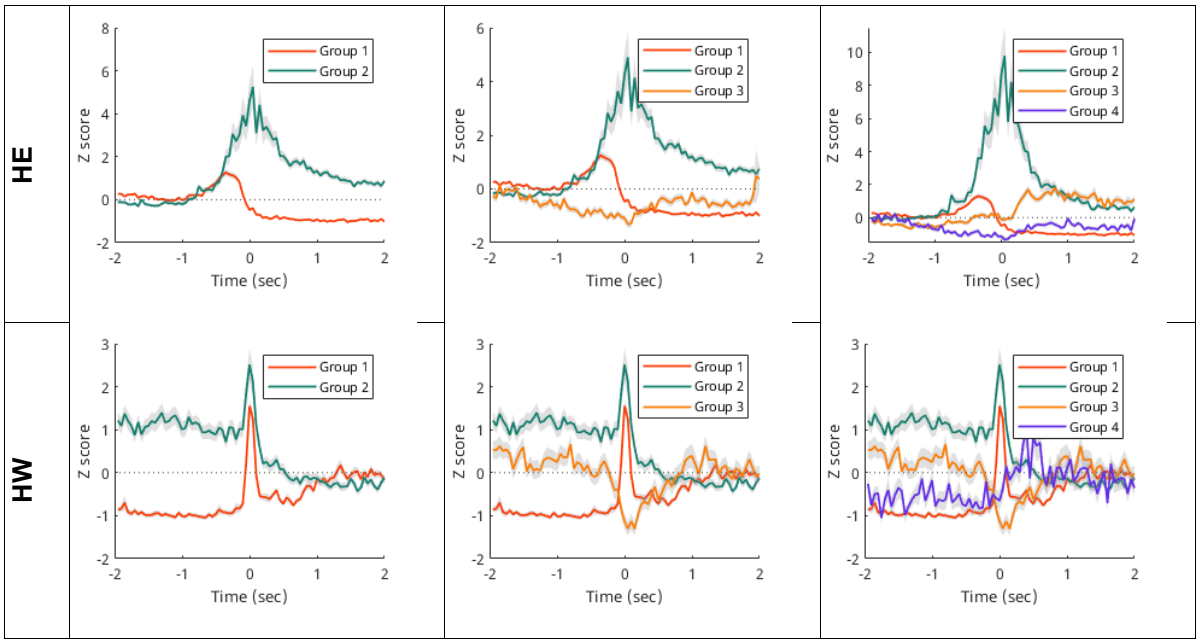
Supplementary Figure S6 Most units are classified into either the HE1-HW1 or HE2-HW2 groups **(A)** Confusion matrix comparing the Head Entry and Head Withdrawal groups. A large proportion of units fall into either the HE1-HW1 category (n=299) or the HE2-HW2 category (n=94). **(B)** Normali zed neural activity of Type 1 (HE1-HW1) and Type 2 (HE2-HW2) neurons during the head-entry and the head-withdrawal. Gray shadings represent SEM.



Supplementary Figure S7 Type1 and Type 2 neurons' PETHs around head withdrawal separated by AW and EW

(A) (Top) Type1 neurons' PETH around head withdrawal separated by AW and EW. Each PETH is sorted by neuron's peak timepoint. (Bottom) Average of PETH. **(B)** Same as (A) with Type 2 neurons.

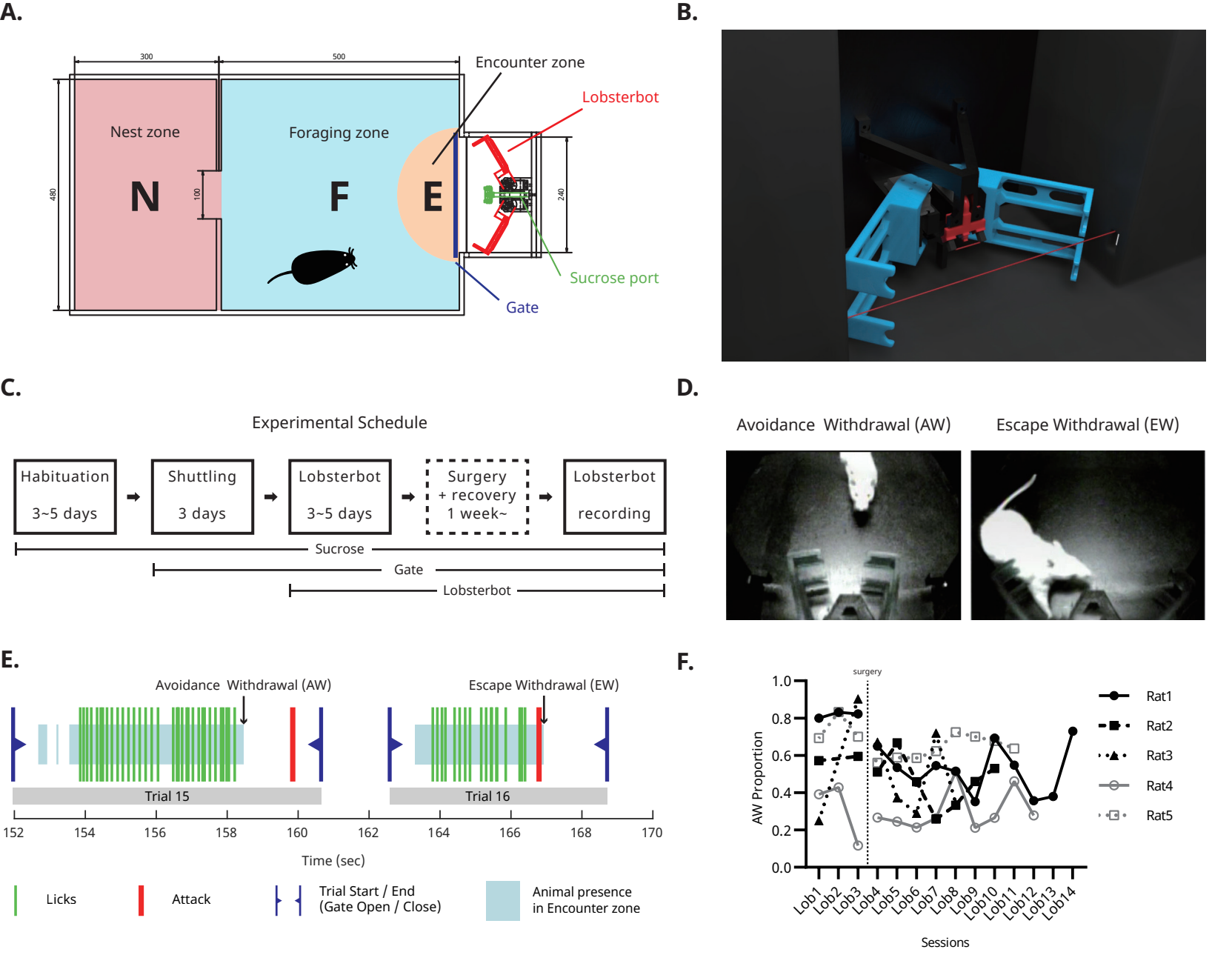
1033

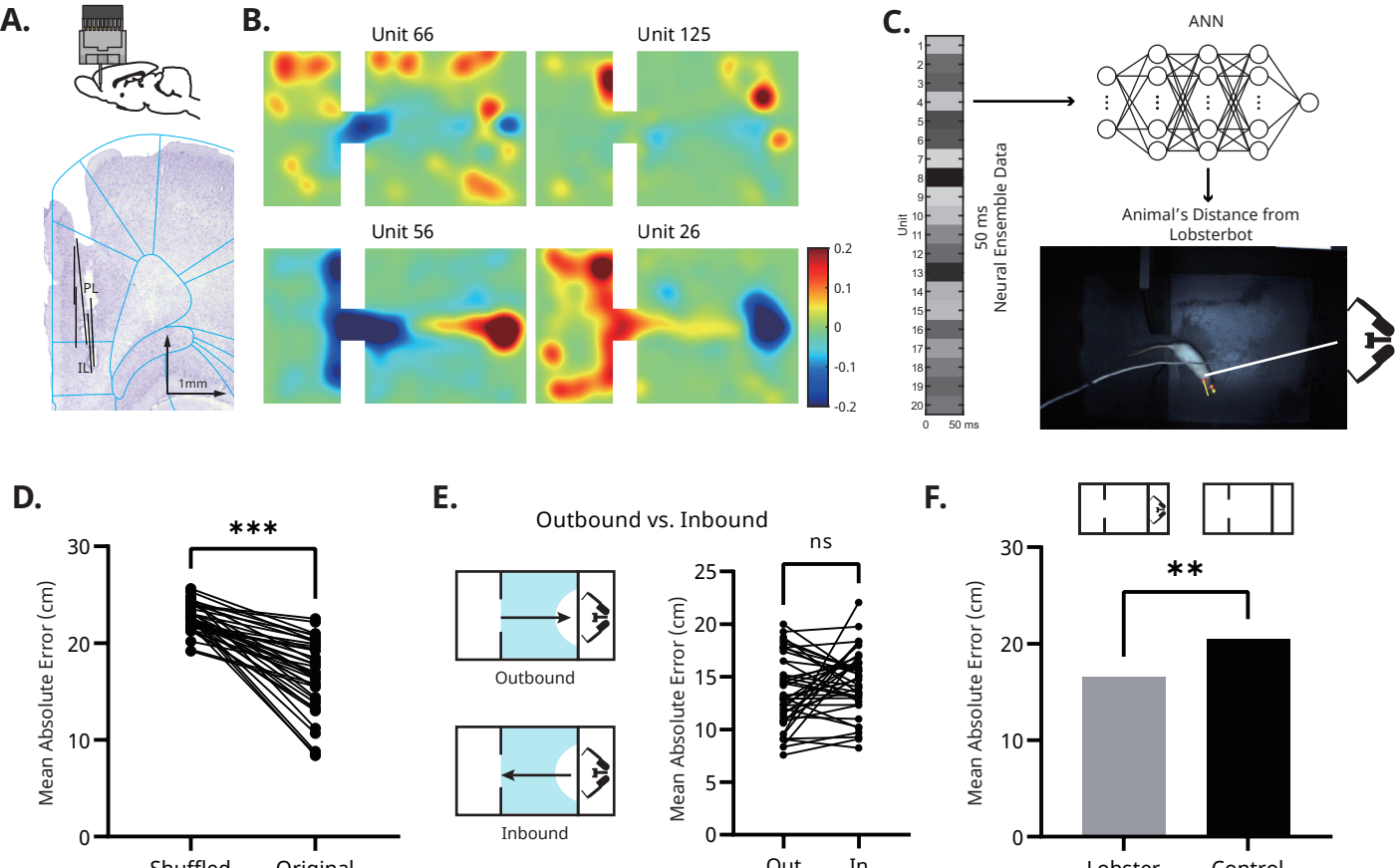


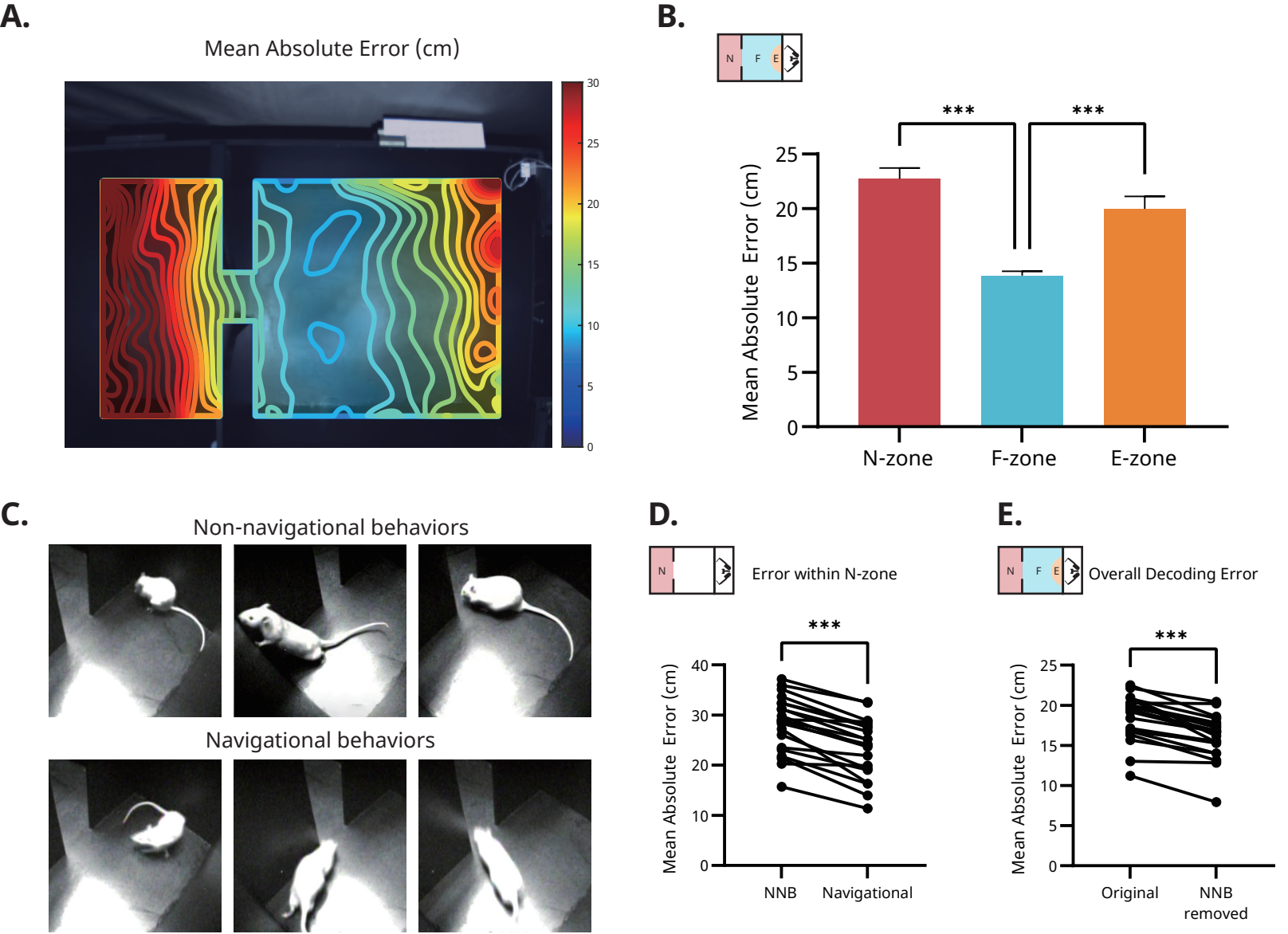
Supplementary Figure S8 Hierarchical clustering results with different hyperparameter sets Hierarchical clustering uses two hyperparameters: cutoff limit and the number of initial clusters. These variables were varied to assess the clustering results' dependence on them. While changes in these variables affected the number of groups, the response characteristics of the top two groups (which were used to further classify Type 1 and Type2) remained consistent. Gray shadings represent SEM.

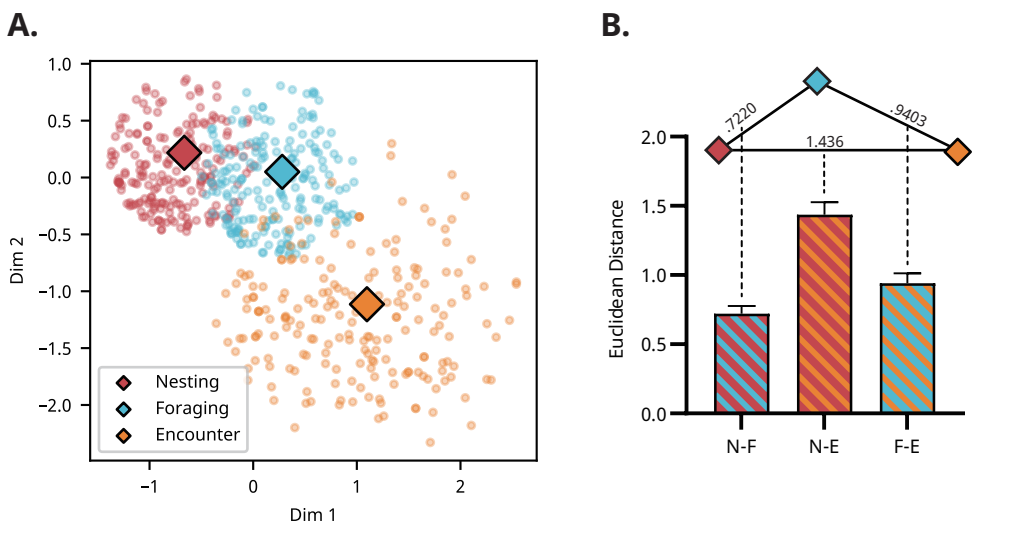
Supplementary Data 1. AW/EW prediction accuracy as a function of dataset time window.

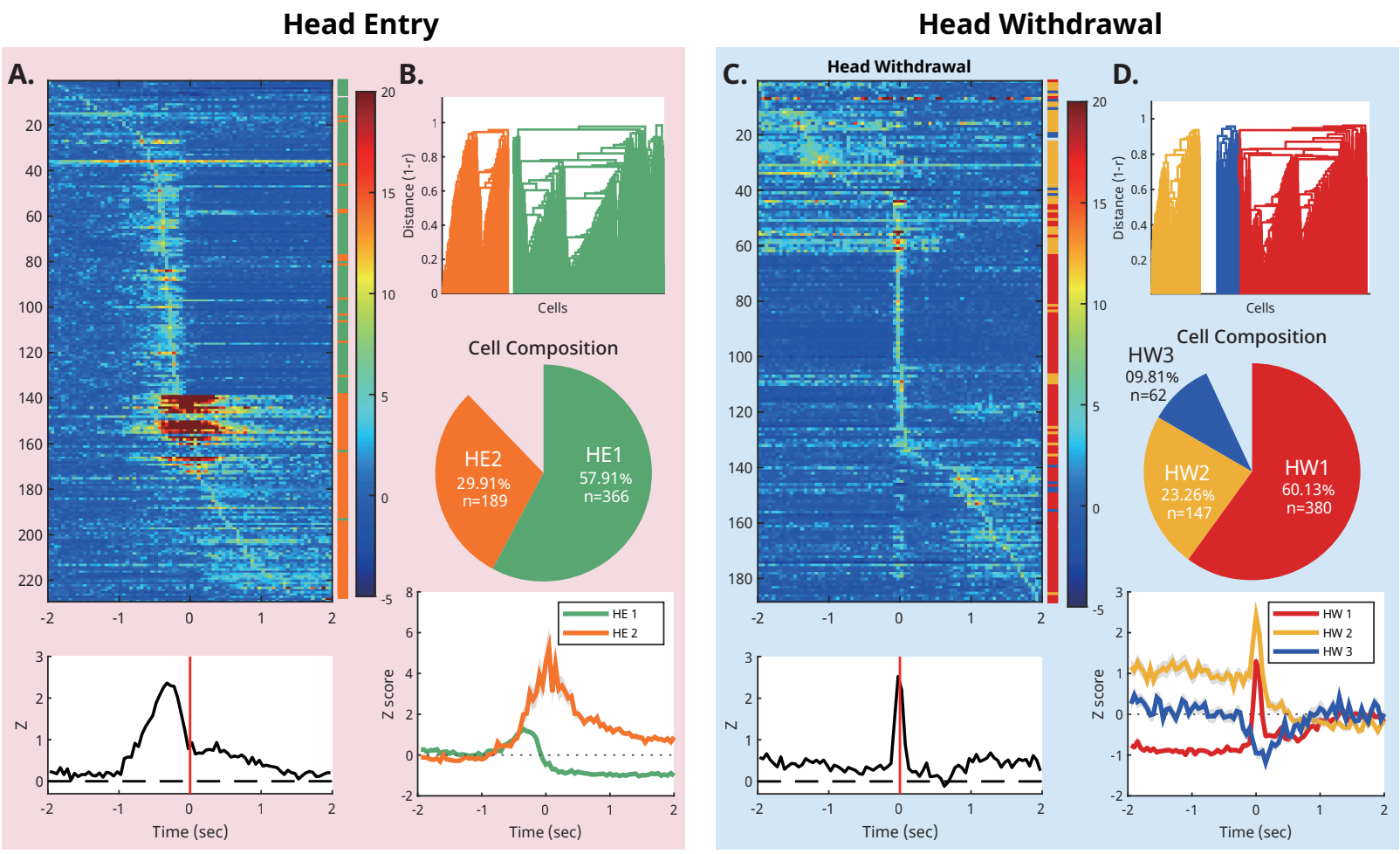
Dataset time window	Statistical result	Significance
−31 – −29	$t(39) = .24$, $p > .999$	NS
−11 – −9	$t(39) = 1.62$, $p = .701$	NS
−7 – −5	$t(39) = 5.381$, $p < .001$	***
−6 – −4	$t(39) = 5.328$, $p < .001$	***
−5 – −3	$t(39) = 4.610$, $p < .001$	***
−4 – −2	$t(39) = 4.631$, $p < .001$	***
−3 – −1	$t(39) = 4.544$, $p = .008$	**
−2 – 0	$t(39) = 4.544$, $p < .001$	***
−1 – +1	$t(39) = 8.505$, $p < .001$	***

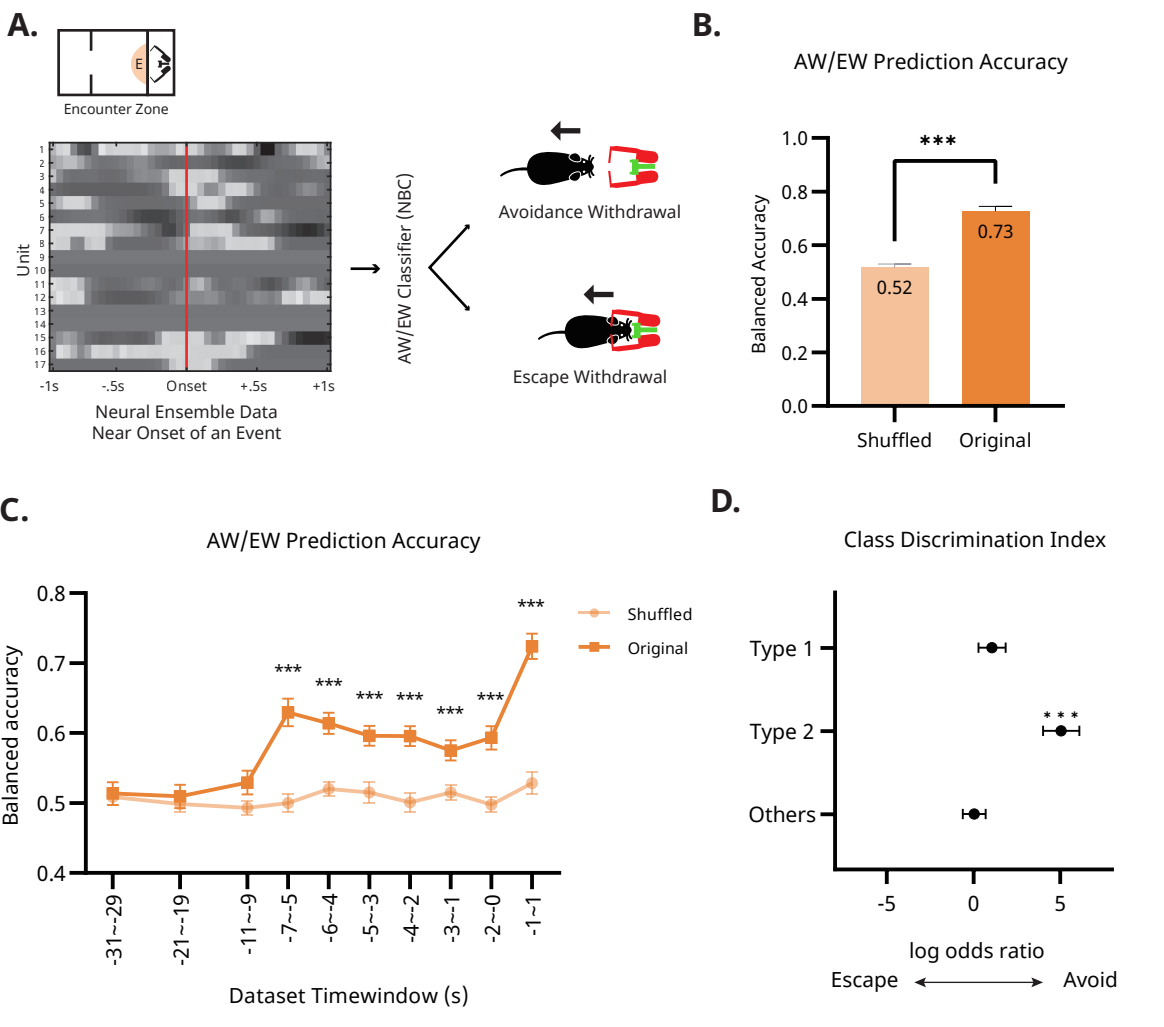

A.
300
500
Encounter zone
Lobsterbot
Nest zone
Foraging zone
N
F
E
480
100
240
Sucrose port
Gate
B.
C.
Experimental Schedule
Habituation
3~5 days
Shuttling
3 days
Lobsterbot
3~5 days
Surgery
+ recovery
1 week~
Lobsterbot
recording
Sucrose
Gate
Lobsterbot
D.
Avoidance Withdrawal (AW)
Escape Withdrawal (EW)
E.
Avoidance Withdrawal (AW)
Escape Withdrawal (EW)
Trial 15
Trial 16
152
154
156
158
160
162
164
166
168
170
Time (sec)
Licks
Attack
Trial Start / End
(Gate Open / Close)
Animal presence
in Encounter zone
F.
surgery
AW Proportion
1.0
0.8
0.6
0.4
0.2
0.0
Lob1
Lob2
Lob3
Lob4
Lob5
Lob6
Lob7
Lob8
Lob9
Lob10
Lob11
Lob12
Lob13
Lob14
Sessions
Rat1
Rat2
Rat3
Rat4
Rat5

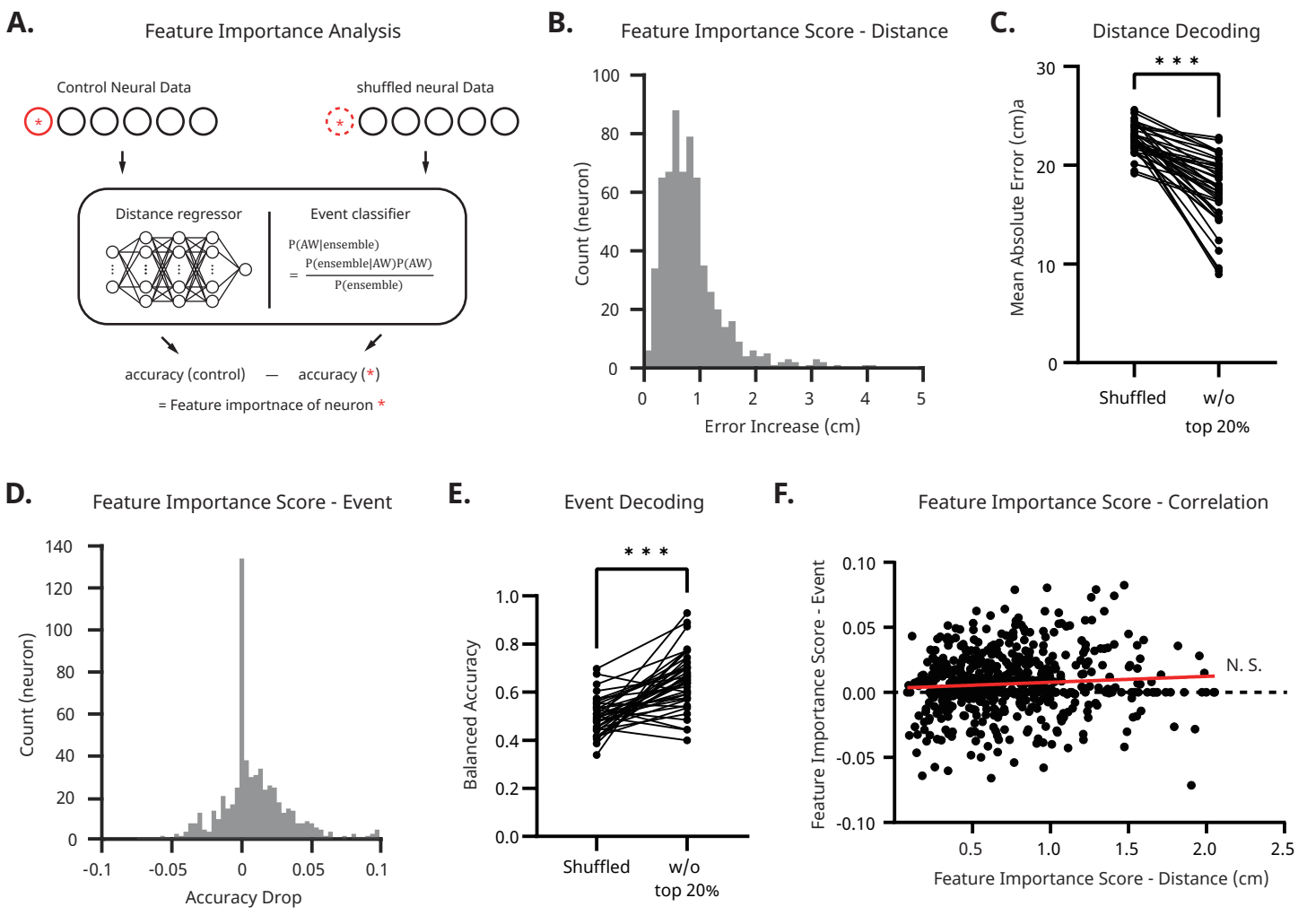

A.
PL
IL
1mm
B.
Unit 66
Unit 125
Unit 56
Unit 26
0.2
0.1
0
-0.1
-0.2
C.
Unit
50 ms
Neural Ensemble Data
ANN
Animal's Distance from Lobsterbot
D.
***
Mean Absolute Error (cm)
Shuffled
Original
E.
Outbound vs. Inbound
Outbound
Inbound
ns
Mean Absolute Error (cm)
Out
In
F.
**
Mean Absolute Error (cm)
Lobster
Control

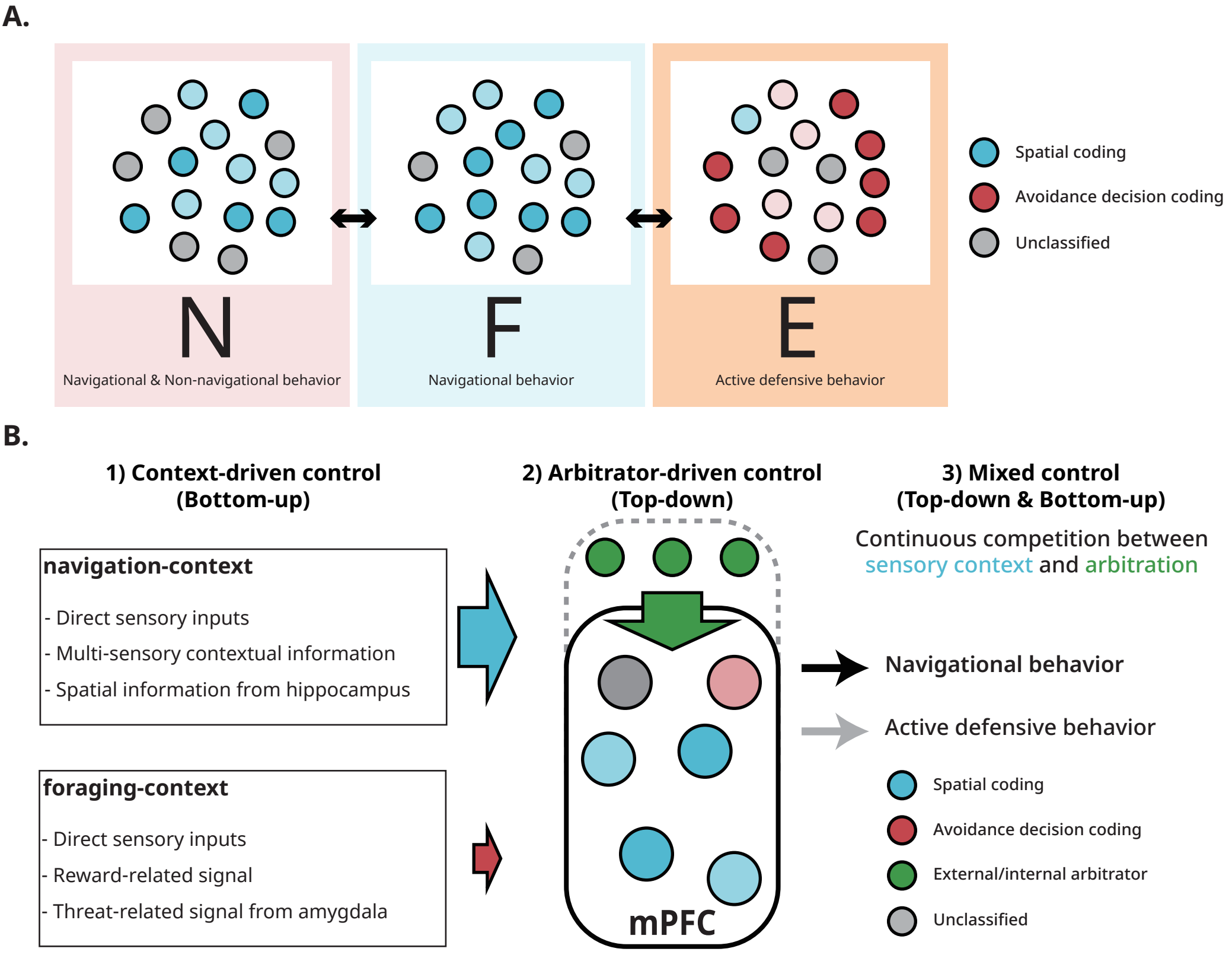

A.
Mean Absolute Error (cm)
30
25
20
15
10
5
0
B.
N
F
E
***
***
Mean Absolute Error (cm)
25
20
15
10
5
0
N-zone
F-zone
E-zone
C.
Non-navigational behaviors
Navigational behaviors
D.
N
Error within N-zone
***
Mean Absolute Error (cm)
40
30
20
10
0
NNB
Navigational
E.
N
F
E
Overall Decoding Error
***
Mean Absolute Error (cm)
25
20
15
10
5
0
Original
NNB removed

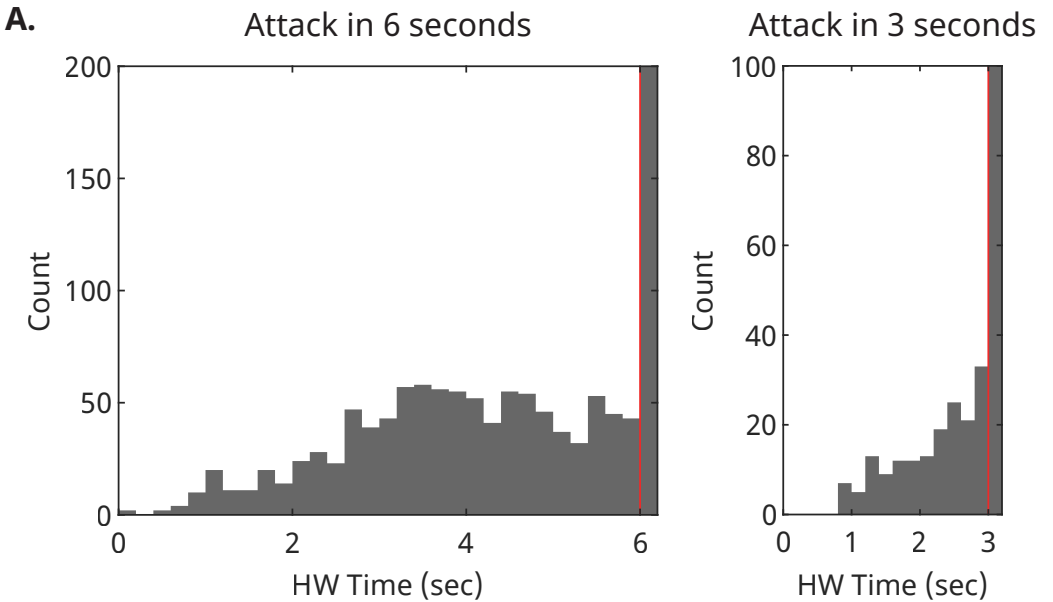

A.
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
Dim 2
−1
0
1
2
Dim 1
Nesting
Foraging
Encounter
B.
.7220
.9403
1.436
2.0
1.5
1.0
0.5
0.0
Euclidean Distance
N-F
N-E
F-E

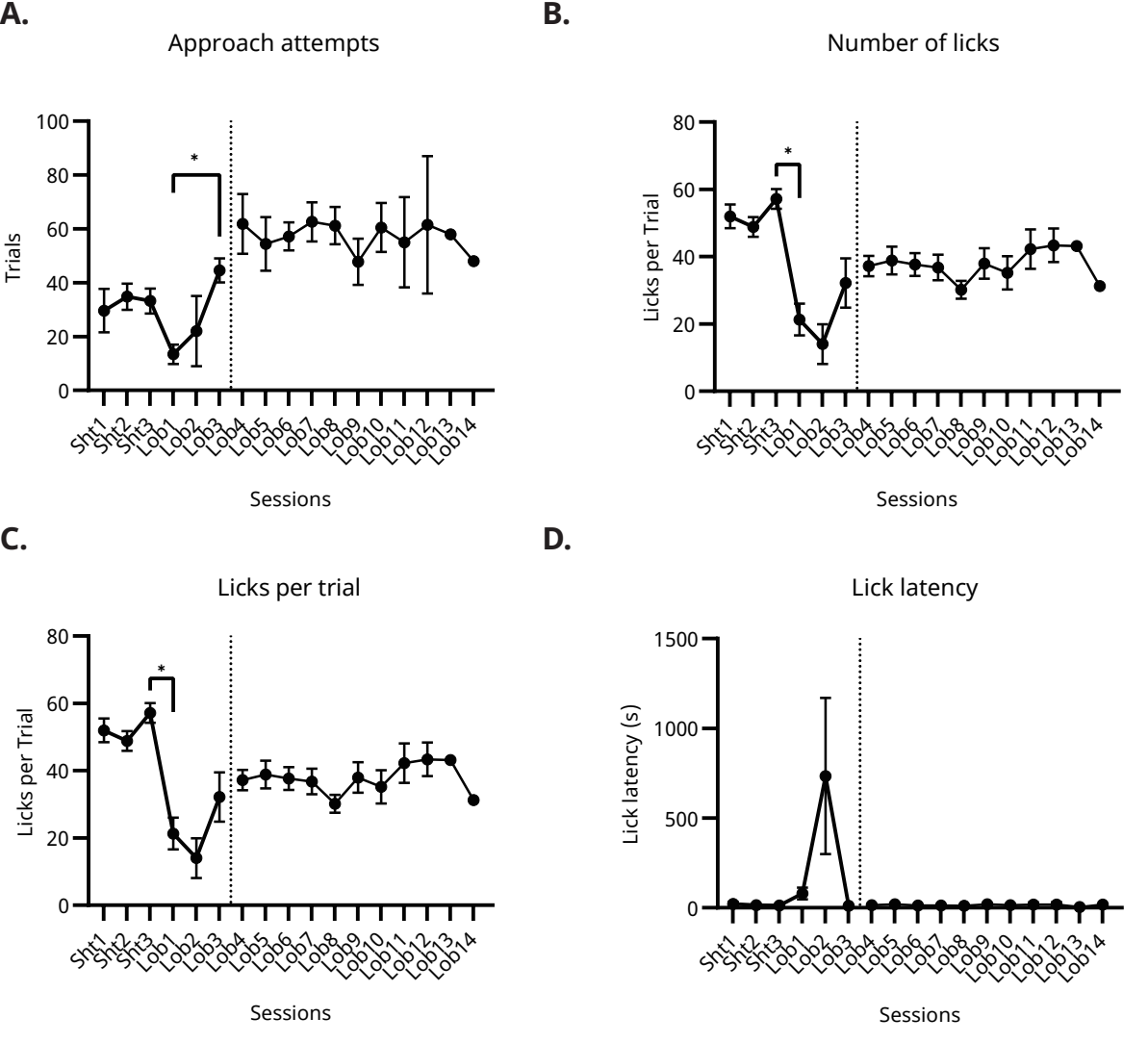

Head Entry
Head Withdrawal
A.
B.
C.
D.
Head Withdrawal
Distance (1-r)
Cells
Cell Composition
HE2
29.91%
n=189
HE1
57.91%
n=366
HW3
09.81%
n=62
HW2
23.26%
n=147
HW1
60.13%
n=380
HE 1
HE 2
HW 1
HW 2
HW 3
Z score
Z
Time (sec)

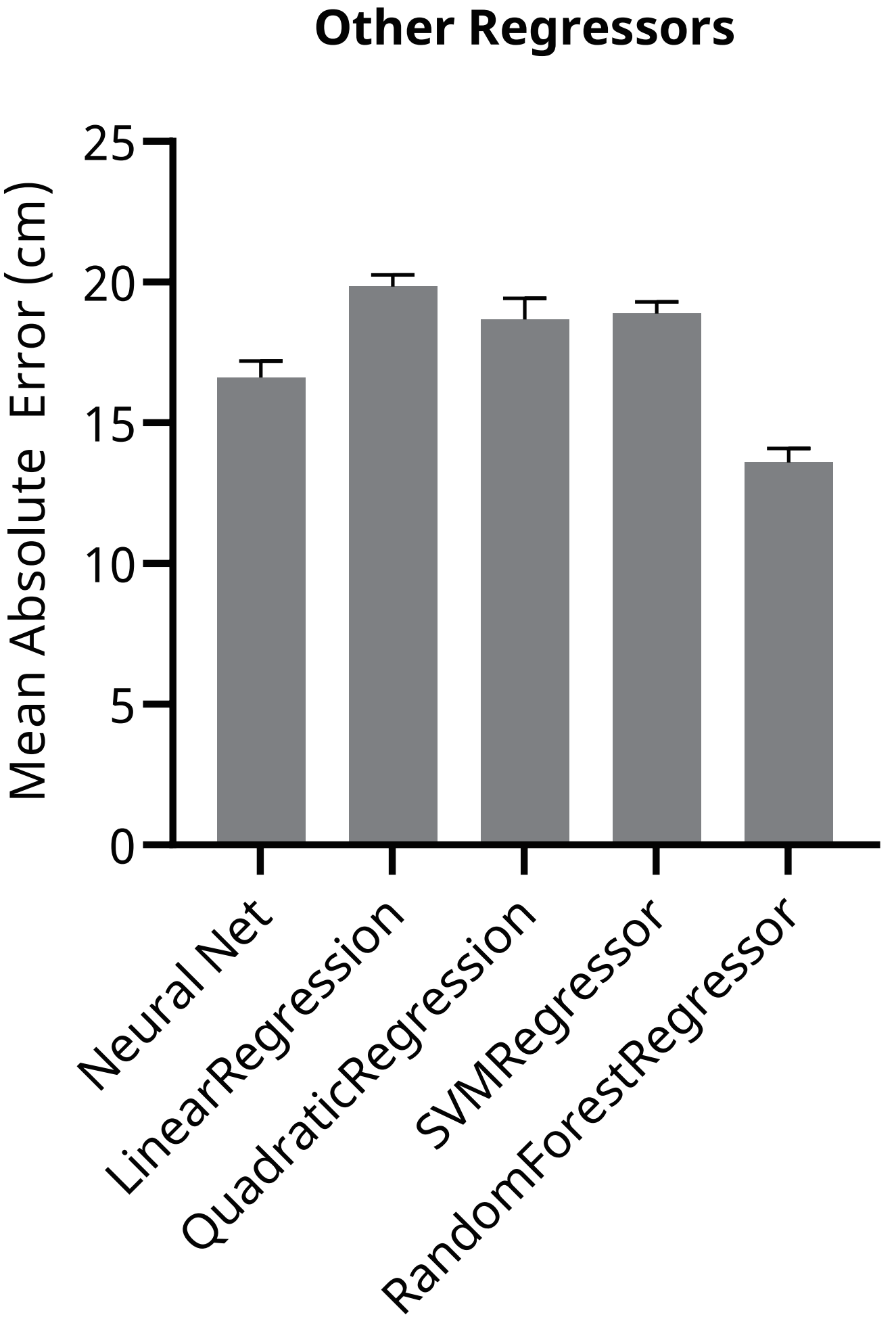

A.
E
Encounter Zone
Unit
-1s
-.5s
Onset
+.5s
+1s
Neural Ensemble Data
Near Onset of an Event
AW/EW Classifier (NBC)
Avoidance Withdrawal
Escape Withdrawal
B.
AW/EW Prediction Accuracy
Balanced Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
***
0.52
0.73
Shuffled
Original
C.
AW/EW Prediction Accuracy
Balanced accuracy
0.8
0.7
0.6
0.5
0.4
Shuffled
Original
-31~-29
-21~-19
-11~-9
-7~-5
-6~-4
-5~-3
-4~-2
-3~-1
-2~-0
-1~1
Dataset Timewindow (s)
D.
Class Discrimination Index
Type 1
Type 2
Others
-5
0
5
log odds ratio
Escape
Avoid

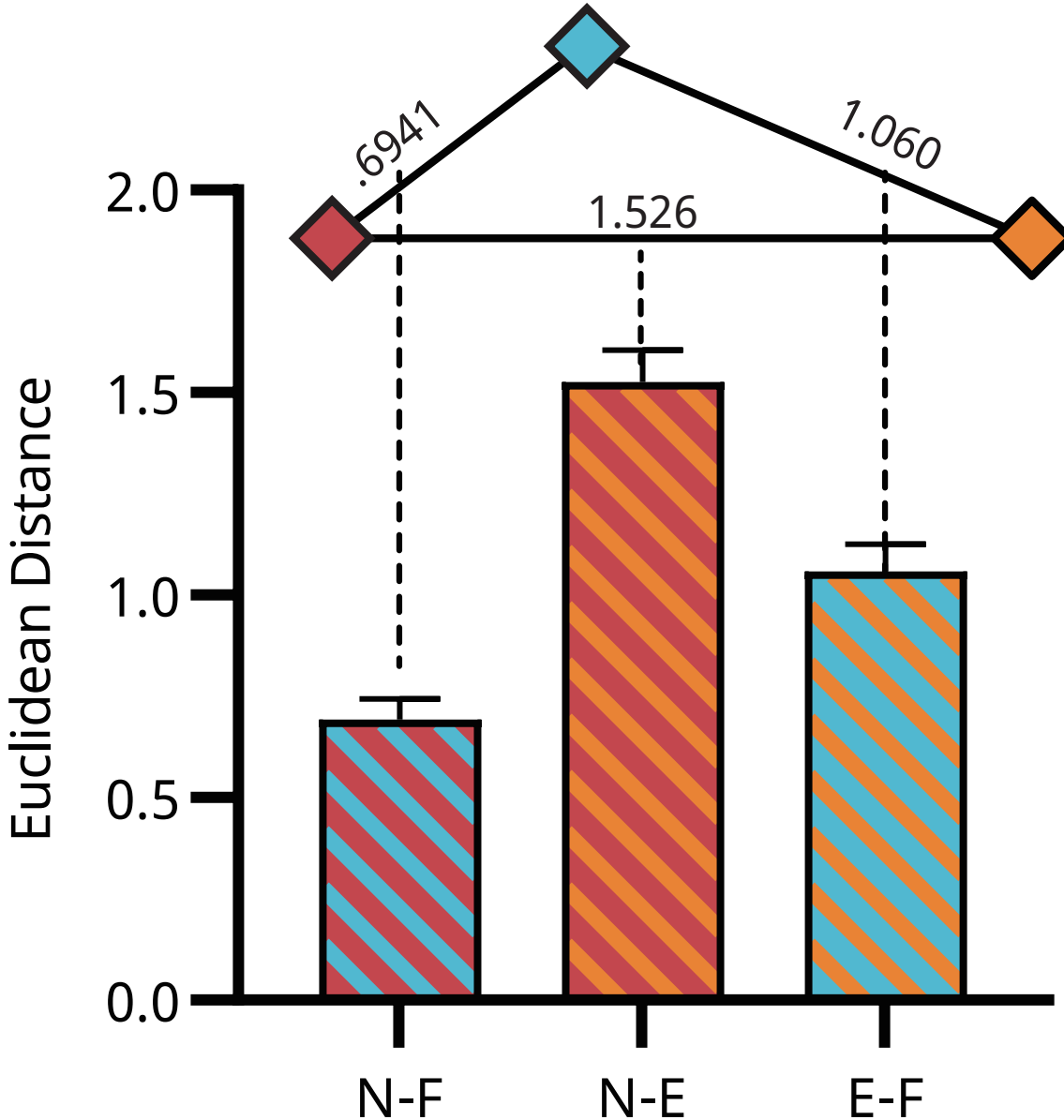

A.
Feature Importance Analysis
Control Neural Data
shuffled neural Data
Distance regressor
Event classifier
P(AW|ensemble)
= P(ensemble|AW)P(AW) / P(ensemble)
accuracy (control) — accuracy (*)
= Feature importnace of neuron *
B.
Feature Importance Score - Distance
Count (neuron)
Error Increase (cm)
C.
Distance Decoding
Mean Absolute Error (cm)a
Shuffled
w/o top 20%
D.
Feature Importance Score - Event
Count (neuron)
Accuracy Drop
E.
Event Decoding
Balanced Accuracy
Shuffled
w/o top 20%
F.
Feature Importance Score - Correlation
Feature Importance Score - Event
N. S.
Feature Importance Score - Distance (cm)

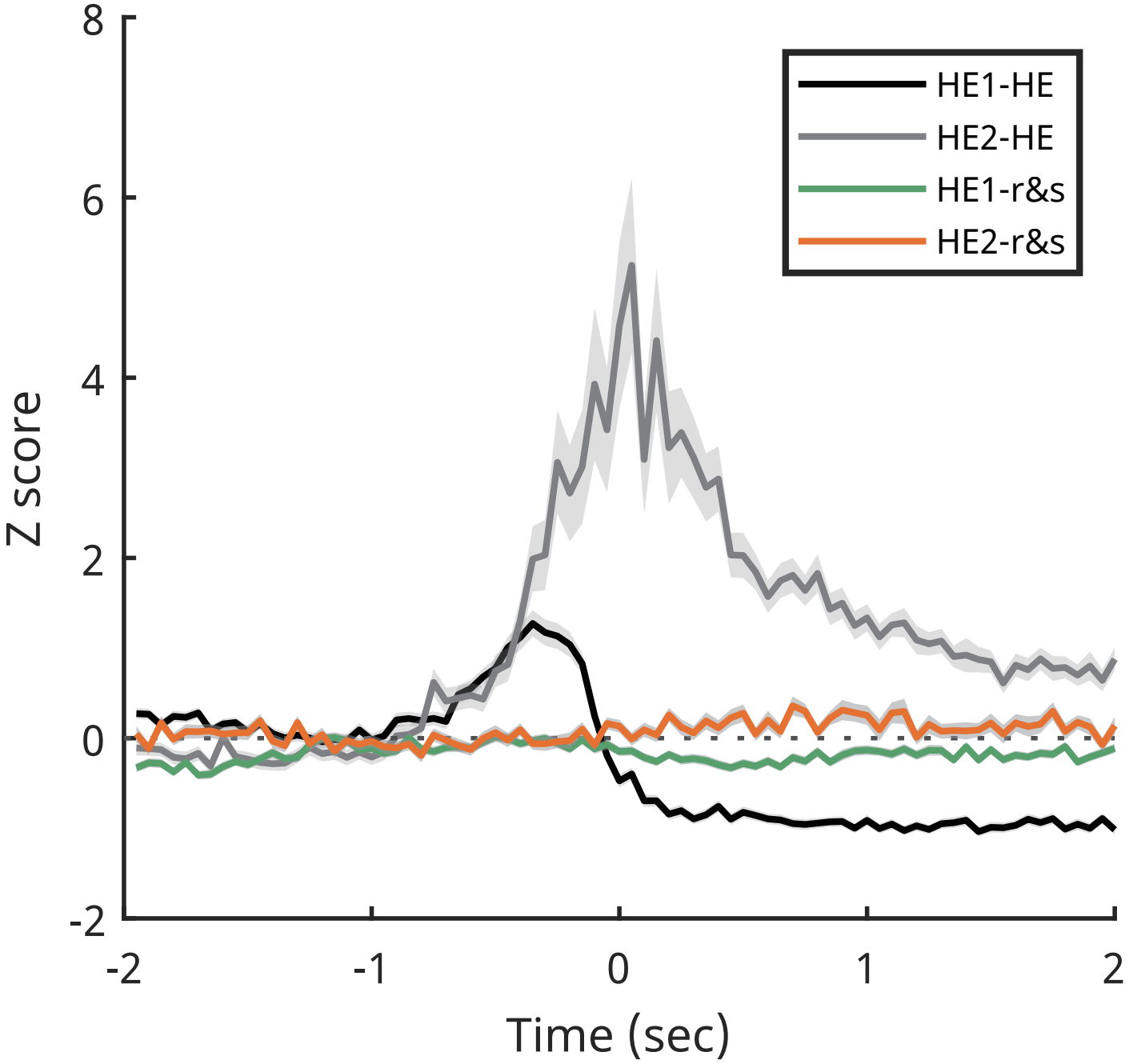

A.
N
Navigational & Non-navigational behavior
F
Navigational behavior
E
Active defensive behavior
Spatial coding
Avoidance decision coding
Unclassified
B.
1) Context-driven control
(Bottom-up)
navigation-context
- Direct sensory inputs
- Multi-sensory contextual information
- Spatial information from hippocampus
foraging-context
- Direct sensory inputs
- Reward-related signal
- Threat-related signal from amygdala
2) Arbitrator-driven control
(Top-down)
mPFC
3) Mixed control
(Top-down & Bottom-up)
Continuous competition between
sensory context and arbitration
Navigational behavior
Active defensive behavior
Spatial coding
Avoidance decision coding
External/internal arbitrator
Unclassified


A.
Attack in 6 seconds
Attack in 3 seconds
Count
HW Time (sec)
0
50
100
150
200
2
4
6
20
40
60
80
1
3


A.
Approach attempts
Trials
100
80
60
40
20
0
*
Sht1
Sht2
Sht3
Lob1
Lob2
Lob3
Lob4
Lob5
Lob6
Lob7
Lob8
Lob9
Lob10
Lob11
Lob12
Lob13
Lob14
Sessions
B.
Number of licks
Licks per Trial
80
60
40
20
0
*
Sessions
C.
Licks per trial
Licks per Trial
80
60
40
20
0
*
Sessions
D.
Lick latency
Lick latency (s)
1500
1000
500
0
Sessions


Other Regressors
Mean Absolute Error (cm)
25
20
15
10
5
0
Neural Net
LinearRegression
QuadraticRegression
SVMRegressor
RandomForestRegressor

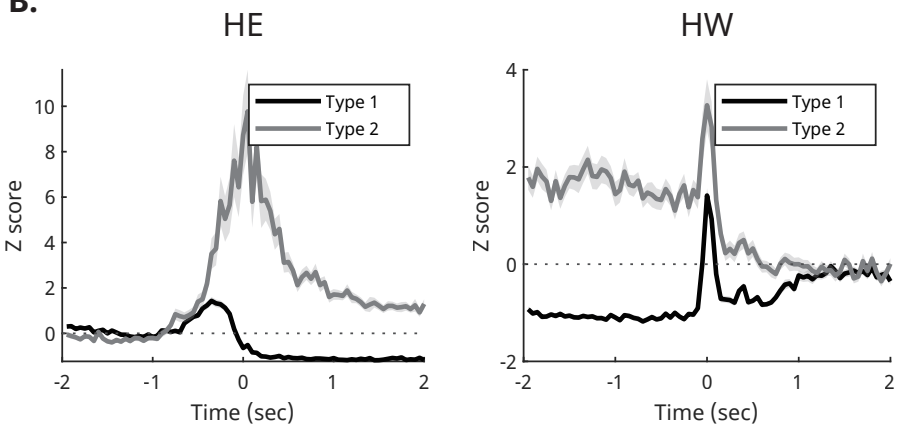

.6941
1.060
1.526
2.0
1.5
1.0
0.5
0.0
Euclidean Distance
N-F
N-E
E-F

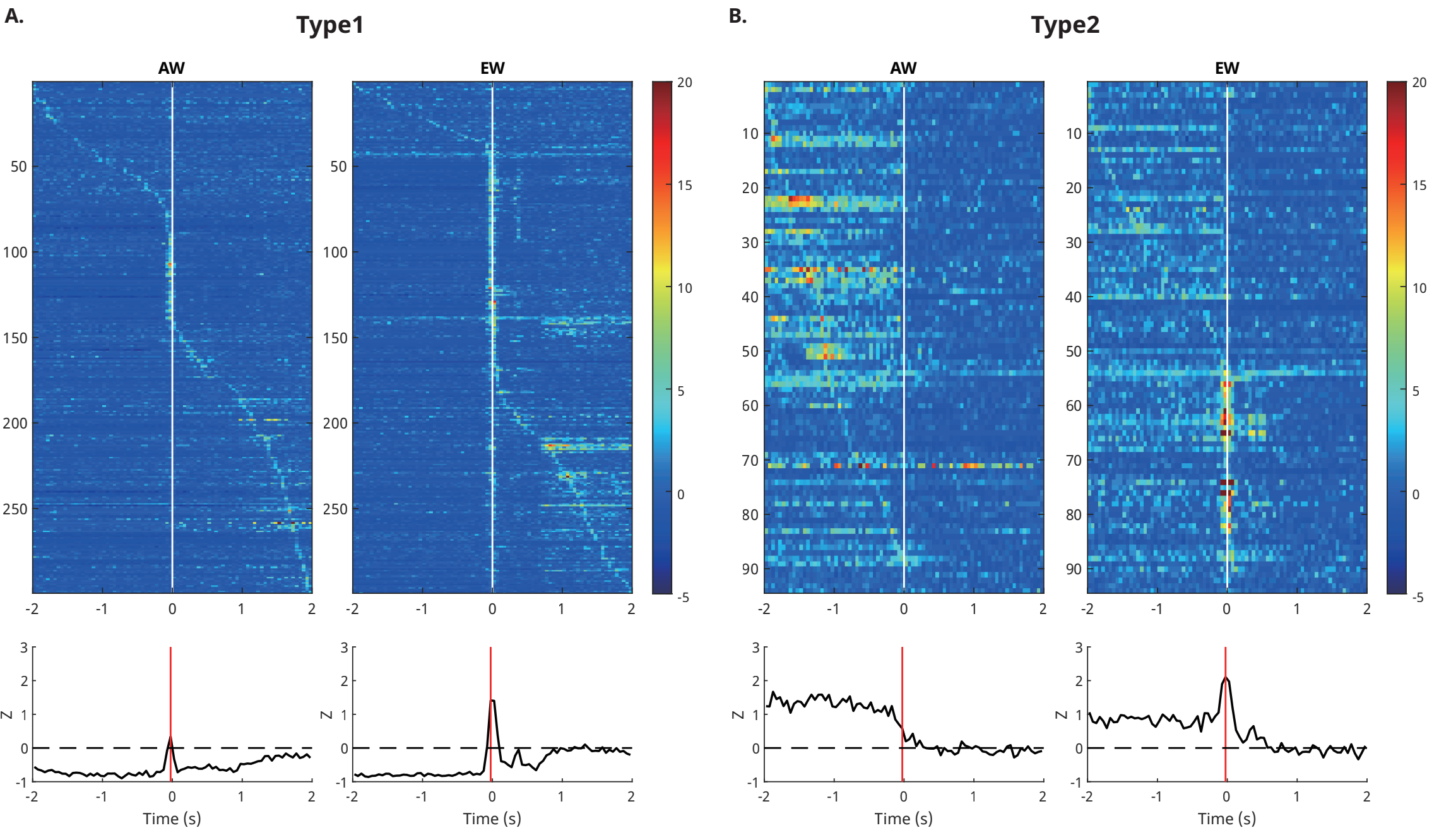

HE1-HE
HE2-HE
HE1-r&s
HE2-r&s
Z score
Time (sec)
8
6
4
2
0
-2
-2
-1
0
1
2

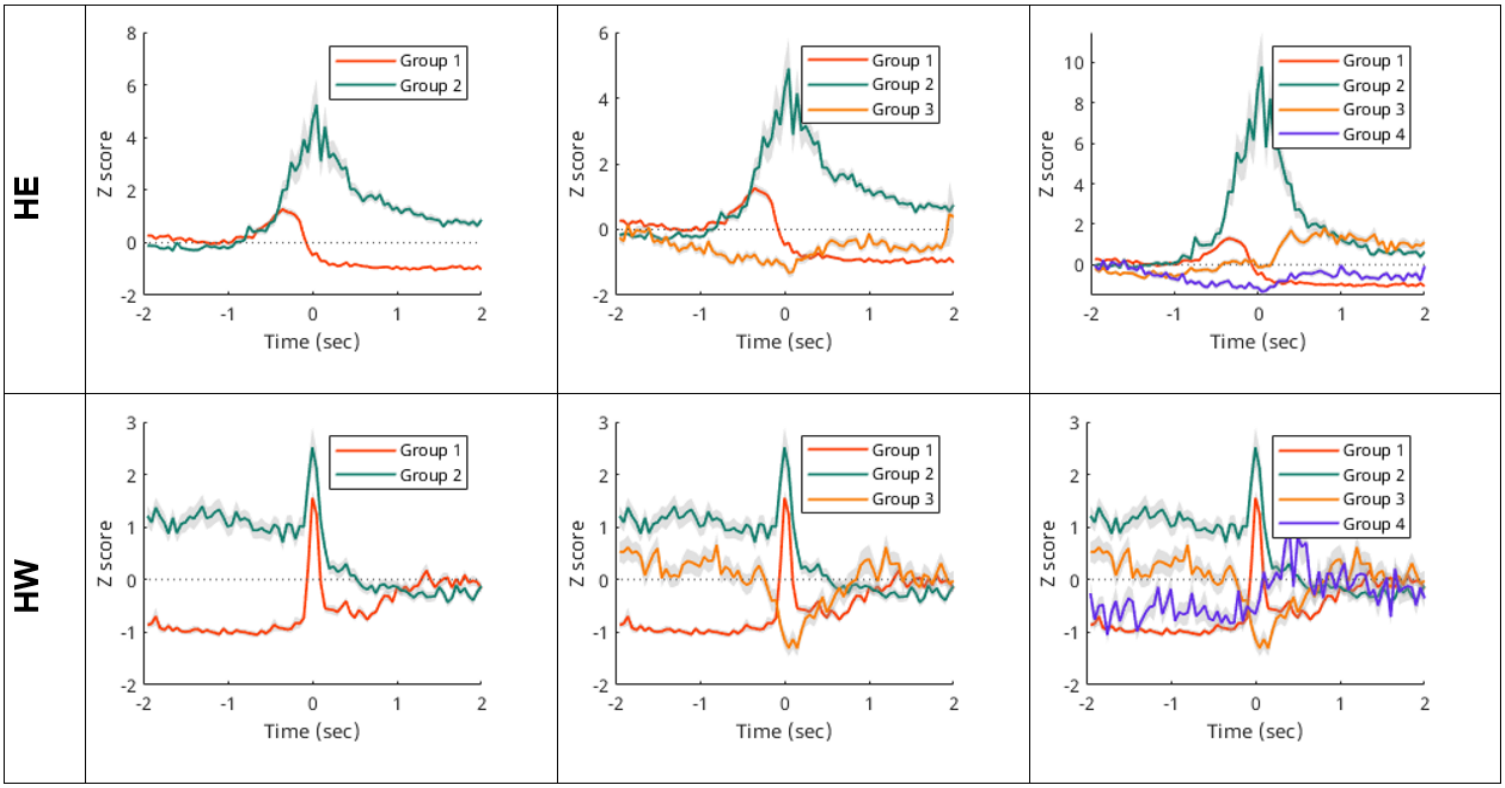
A.

HE Group \ HW Group	0	1	2	3	Sum
0	13	42	13	9	77
1	14	299	40	13	366
2	16	39	94	40	189
Sum	43	380	147	62	632

B.




A.
Type1
AW
EW
B.
Type2
AW
EW
Z
Time (s)


HE
HW
Group 1
Group 2
Group 3
Group 4
Z score
Time (sec)