Open access • Posted Content • DOI:10.1101/2021.05.06.442972

Subpopulations of neurons in lOFC encode previous and current rewards at time of choice — Source link

David L Hocker, Carlos D. Brody, Carlos D. Brody, Cristina Savin ...+2 more authors

Institutions: Center for Neural Science, Princeton University, Howard Hughes Medical Institute, New York University

Published on: 07 May 2021 - bioRxiv (Cold Spring Harbor Laboratory)

Subpopulations of neurons in lOFC encode previous and current rewards at time of choice

David Hocker[1*], Carlos D. Brody[2,3,4], Cristina Savin[1,5†], and Christine M Constantinople[1†]

[1]Center for Neural Science, New York University, New York, NY 10003, USA
[2]Princeton Neuroscience Institute, Princeton University, Princeton, NJ, 08544, USA
[3]Department of Molecular Biology, Princeton University, Princeton, NJ, 08544, USA
[4]Howard Hughes Medical Institute, Princeton University, Princeton, NJ, 08544, USA
[5]Center for Data Science, New York University, New York, NY 10011, USA
[*]Corresponding author
[†]Co-senior authors

1 Abstract

Studies of neural dynamics in lateral orbitofrontal cortex (lOFC) have shown that subsets of neurons that encode distinct aspects of behavior, such as value, may project to common downstream targets. However, it is unclear whether reward history, which may subserve lOFC's well-documented role in learning, is represented by functional subpopulations in lOFC. We analyzed neural recordings from rats performing a value-based decision-making task, in which we previously documented trial-by-trial learning that required lOFC. We found five distinct clusters of lOFC neurons, either based on clustering of their trial-averaged peristimulus time histograms (PSTHs), or a feature space defined by their average conditional firing rates aligned to different task variables. We observed weak encoding of reward attributes, but stronger encoding of reward history, the animal's left or right choice, and reward receipt across all clusters. Only one cluster, however, encoded the animal's reward history at the time shortly preceding the choice, suggesting a possible role in integrating previous and current trial outcomes at the time of choice.

2 Introduction

Previous experience can profoundly influence subsequent behavior and choices. In trial-based tasks, the effects of previous choices and outcomes on subsequent ones are referred to as "sequential effects," and while they are advantageous in dynamic environments, they produce suboptimal biases when outcomes on each trial are independent. The orbitofrontal cortex (OFC) has been implicated in updating behavior based on previous experience, particularly when task contingencies are partially observable [1–4]. However, it is unclear whether behavioral flexibility in OFC is mediated by dedicated subpopulations of neurons exhibiting distinct encoding and/or connectivity. We previously trained rats on a value-based decision-making task, in which they chose between explicitly cued, guaranteed and probabilistic rewards on each trial [5, 6]. Despite the fact that outcomes were independent on each trial, we observed several distinct sequential effects that contributed to behavioral variability. Optogenetic perturbations of the lateral orbitofrontal cortex (lOFC) eliminated one particular sequential effect, an increased willingness to take risks following risky wins, but spared other types of trial-by-trial learning, such as spatial "win-stay/lose-switch" biases. We interpreted this data as evidence that (1) different sequential effects may be mediated by distinct neural circuits, and (2) lOFC promotes learning of abstract biases that reflect the task structure (here, biases for the risky option), but not spatial ones [6].

Electrophysiological recordings from lOFC during this task revealed encoding of reward history and reward outcomes on each trial at the population level, which could in principle support sequential effects [6]. Recent studies of rodent OFC have suggested that despite the apparent heterogeneity of neural responses in prefrontal cortex, neurons can be grouped into distinct clusters that exhibit similar task-related responses and, in some cases, project to a common downstream target [7, 8]. In light of these results, we hypothesized that reward history might be encoded by a distinct cluster of neurons in lOFC. This would suggest that the lOFC-dependent sequential effect we observed (an increased willingness to take risks following risky wins) may derive from a subpopulation of neurons encoding reward history, and potentially projecting to a common downstream target.

Here, we analyzed an electrophysiological dataset from lOFC during a task in which independent and variable sensory cues conveyed dissociable reward attributes on each trial (reward probability and amount; [5, 6]). We found that clustering of lOFC neurons based on either their trial-averaged peristimulus time histograms (PSTHs), or a feature space defined by their average conditional firing rates aligned to different task variables [7], both revealed five clusters of neurons with different response profiles. We exploited the temporal variability of task events to fit a generalized linear model (GLM) to lOFC firing rates, generating a rich description of the encoding of various task variables over time in individual neurons. All of the clusters exhibited weak encoding of reward attributes; and stronger encoding of reward history, the animal's left or right choice, and reward receipt. Only one cluster, however, encoded the animal's reward history at the time shortly preceding the choice. This distinct encoding was observable by three independent metrics (coefficient of partial determination, mutual information, and discriminability or d') and two separate clustering methods. Moreover, the subpopulation of neurons that represented reward history before the choice also exhibited the strongest encoding of reward outcomes, suggesting that these neurons are well-situated to integrate previous and current reward experience. We hypothesize that this subpopulation of neurons, which were identifiable based on their temporal response profiles alone, may mediate sequential learning effects by integrating previous and current trial outcomes at the time of choice.

3 Results

Rats' behavior on this task has been previously described [5, 6]. Briefly, rats initiated a trial by poking their nose in a central nose port (Fig. 1A-C). They were then presented with a series of pseudo-randomly timed light flashes, the number of which conveyed information about reward probability on each trial. Simultaneously, they were presented with randomly timed auditory clicks, the rate of which conveyed the volume of water reward (6-48 μl) baited on each side. After a cue period of $\sim 2.5 - 3$s, rats reported their choice by poking in one of the two side ports. Reward volume and probability of reward (p) were randomly chosen on each trial. On each trial, the left or right port was randomly designated as risky ($p < 1$) or safe ($p = 1$). Well-trained rats tended to choose the side offering the greater subjective value, and trial history effects contributed to behavioral variability [5, 6]. We analyzed 659 well-isolated single-units in the lOFC with minimum firing rates of 1Hz, obtained from tetrode recordings [6].

3.1 Single units in lOFC belong to clusters with distinct temporal response profiles

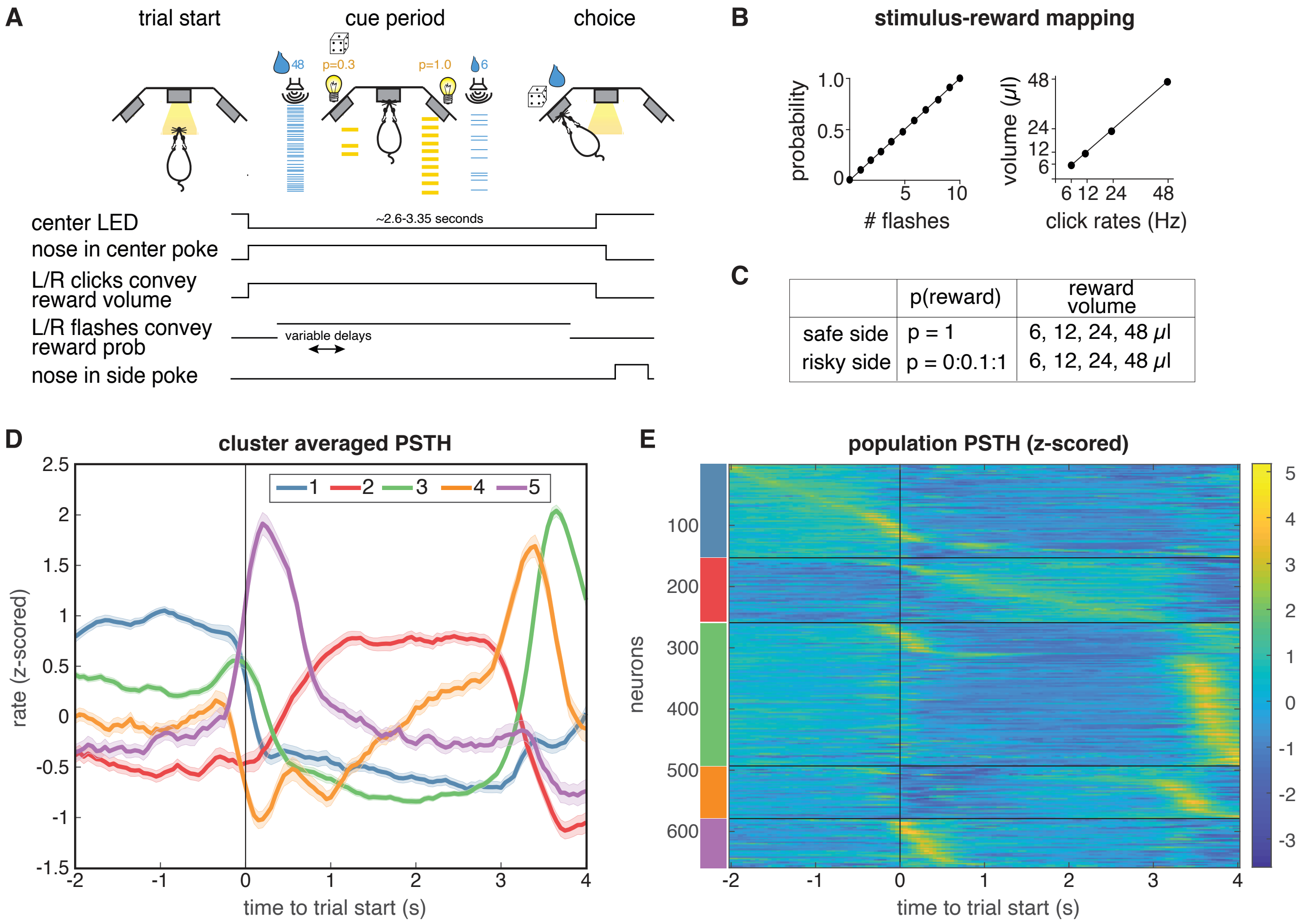
To characterize the temporal dynamics of the neural responses during this task, we performed k-means clustering on trial-averaged PSTHs ("PSTH clustering"), aligned to trial initiation. We used the gap statistic [9], which quantifies the improvement in cluster quality with additional components relative to a null distribution (Methods), to choose a principled number of clusters, and identified five distinct clusters of responses (Fig. 1D). Each cluster has a stereotyped period of elevated activity during the task: at trial start, during the cue period, and at or near reward delivery, with the largest cluster (cluster 3) having activity just before the rat entered the reward port (Fig. S1). Cluster 2 is the only cluster that exhibited persistent activity during the cue period. These data indicate that despite the well-documented heterogeneity of neural responses in prefrontal cortex, responses in lOFC belong to one of a relatively small, distinct set of temporal response profiles aligned to different task events.

A recent study in rat lOFC similarly found discrete clusters of neural responses by clustering the conditional firing rates of neurons for different task-related variables ("conditional clustering") [7]. We sought to compare results from clustering PSTHs and conditional firing rates. Therefore, we generated conditional response profiles for each neuron (Fig. 2B; Methods), performed clustering on these conditional responses via the same procedure. This also revealed five distinct clusters that exhibited qualitatively similar temporal response profiles as the clusters that were based on the average PSTHs (Fig. 2C-E, Suppl. Fig. S2). Moreover, clusters 2 and 3 were comprised of highly similar groups of neurons across both procedures (66% overlap of cluster 2 neurons and 70% of cluster 3 neurons), which indicates that neurons in these two clusters exhibit similar temporal response profiles and also similar encoding of task variables (Fig. 2E). Additionally, assessing the distance among clusters verified that clusters obtained by both clustering procedures are generally well separated from one another (Suppl. Fig. S3).

3.2 Neural responses in lOFC are well-captured by a generalized linear model that includes attributes of offered rewards, choice, and reward history

We next sought to characterize the encoding properties of neurons in each cluster. To that end, we modeled each neuron's trial-by-trial spiking activity with a generalized linear model with Poisson noise (Fig. 3). This approach has been previously used in sensory areas to model the effects of external stimuli upon neural firing [10, 11], and more recently has been extended to higher-order areas of cortex during decision-making tasks [12]. In this model, the time-dependent firing rate (λ_t) of a neuron is the exponentiated sum of a set of linear terms

$$\lambda_t = \exp\left[\sum_{s=1}^{S}\left(X_{t-\tau:t}^{(s)} * k_s(\tau)\right)\right], \qquad (1)$$



	p(reward)	reward volume
safe side	p = 1	6, 12, 24, 48 μl
risky side	p = 0:0.1:1	6, 12, 24, 48 μl

Figure 1: A. Behavioral task: rats chose between a guaranteed and probabilistic water reward on each trial, that could each be one of four water volumes. Rats were well-trained and tended to choose the option with the greater subjective value. Reward probability and volume were cued by visual flashes and auditory click rates, respectively. B. Mapping between the visual flashes/auditory clicks and reward attributes. C. Range of reward attributes. D. Cluster-specific, mean responses of the trial-averaged, z-scored PSTHs. Error bars denote s.e.m. E. Z-scored PSTHs for neurons in each cluster, sorted by time to peak within each cluster. Colored bars on the left indicate the cluster identities from panel D. Panels A-C reproduced and modified from [6].

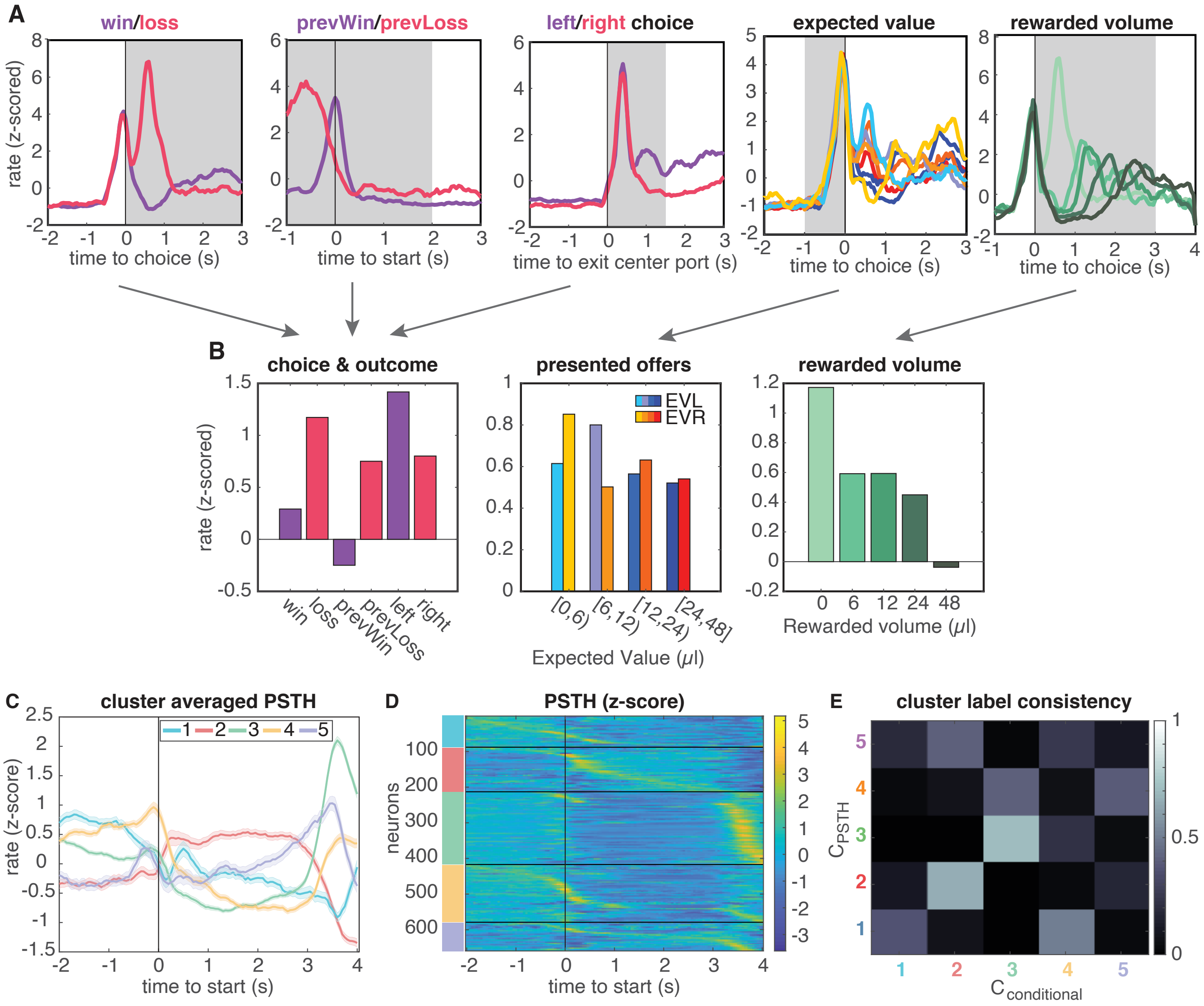


Figure 2: A. Feature-conditioned PSTHs of an example neuron. The average z-scored firing rate in each gray time window comprises an element in the feature space. B. Feature space for conditional clustering for the sample neuron from A. The average firing rate for each condition is concatenated to yield a 19-dimensional feature space. Here, features are grouped into three qualitative types (for presentation purposes only): responses for choice and trial outcome (left); responses to expected value of left and right offers (middle); and responses to received reward volume (right). C. Cluster-averaged PSTHs, aligned to trial initiation, using this conditional feature space. Error bars denote s.e.m. D. Z-scored PSTHs for all neurons, sorted by time to peak within each cluster. E. Consistency of cluster labeling, calculated as the conditional probability $P(C_{\text{conditional}}|C_{\text{PSTH}})$ of belonging in any 'conditional' cluster ($C_{\text{conditional}}$), given that a unit belongs to a certain PSTH-defined cluster (C_{PSTH}).

where the variables $X^{(s)}_{t-\tau:t}$ are the stimuli and behavioral events against which neural activity is regressed; $k_s(\tau)$ denote the response kernels that are convolved with the task variables to model time-dependent responses to each variable (Fig. 3C). The probability that a given number of spikes would occur in a discrete time bin of $\Delta t = 50$ms is given by a homogeneous Poisson process with mean $\lambda_t \Delta t$.

The task variables and example kernels for our model are shown for a sample neuron in Fig. 3C. Variables were binarized such that 1 (0) denoted the occurrence (absence) of an event. The task variables included reward (win or loss on the current or previous trial), choice (left or right), and the timing of cues indicating reward attributes for the left and right offers (reward volume conveyed by auditory clicks and reward probability conveyed by visual flashes). We also included the average reward rate prior to starting each trial and the location of the trial within the session (not shown). Given the large number of model parameters, we used a smooth, log-raised cosine temporal basis and L2 regularization to prevent overfitting (Methods). Based on model comparison, we found that including a spike history term as in other GLM approaches did not improve our model, presumably due to the fact that we are modeling longer timescale responses.

The chosen model parameters were selected by model comparison against several alternative models using cross-validation. Model comparison favored a simpler binary win/loss representation of rewarded outcomes over richer representations of reward volume on the current or previous trial (Suppl. Fig. S4). The model reproduced the trial-averaged PSTHs of individual neurons (Fig. 3E). To quantify model performance, we calculated the proportion of variance explained (R^2) for held-out testing data (Fig. 3F). The model captured a high percentage of variance for most of the neurons in our dataset. A small fraction of neurons exhibited negative R^2 values (69 units, Suppl. Fig. S5), indicating that the model produced a worse fit of the test data than the data average. Our liberal inclusion criteria did not require neurons to exhibit task-modulation of their firing rates, so these neurons were likely not task-modulated, and were excluded from subsequent analyses.

3.3 Clustering reveals distributed encoding of most task variables across subpopulations of lOFC neurons

We next sought to characterize the extent to which neurons that belonged to different clusters might be "functionally distinct," and encode different task-related variables [7, 8]. To address this question, we computed two complementary but independent metrics, both based on the GLM fit: the coefficient of partial determination and mutual information. The coefficient of partial determination (CPD) quantifies the contribution of a single covariate (e.g., left/right choice) to a neuron's response by comparing the goodness-of-fit of the encoding model with and without that covariate. In other words, the CPD quantifies the amount of variance that is explained by kernels representing different aspects of the model. We computed the CPD in a sliding time window throughout the trial, and compared CPD values to a shuffle test, in which trial indices were randomly shuffled relative to the design matrix before generating model predictions. CPD values that were within the 95% confidence intervals of the shuffled distribution were set to 0 before averaging CPD values over neurons in a cluster. The average CPD values for different task variables is shown for clusters based on each clustering procedure in Fig. 4A and Fig. 4C. Note that CPD plots are aligned to different events, depending on the covariate.

We also computed the mutual information (MI) between neural spike trains and different model covariates (Methods). Our approach relates the statistics of trial-level events to firing rates, which allows us to assess the information content for each stimulus throughout the entire time course of a trial. Such an approach does not easily generalize to the statistics of within-trial events such as the information represented in stimulus clicks and flashes, so we restricted the MI analysis to the other covariates: reward history, choice, and reward outcome. The average MI for these variables is shown for clusters based on the trial-averaged PSTHs in Fig. 4B.

In general, regardless of the metric (CPD or MI) or clustering procedure (PSTH clustering or conditional clustering), task variables appear to be broadly encoded across neural subpopulations, with similar temporal dynamics and average CPD/MI values across clusters. All clusters encoded cues conveying reward volume and probability during the cue period (Fig. 4C), although it is worth noting that the magnitude of CPD for clicks and flashes was an order of magnitude lower than for the other task variables. Therefore, encoding of cues representing reward attributes was substantially weaker than encoding of reward history, choice, and outcome. It may be surprising that neurons with strikingly different PSTHs appear to encode task variables with similar time courses. However, PSTHs marginalize out all conditional information, so a PSTH carries no information about encoding of task or behavioral variables, per se.

Reward history was most strongly encoded at the time of trial initiation and decayed over the course of the trial, consistent with previous analyses [6]. Neurons belonging to cluster 2, which strongly overlap in both clustering procedures (Fig. 2E), seem to exhibit slightly more pronounced encoding of reward history during the trial, compared to the other clusters, although these neurons were also persistently active during this time period. Encoding of choice and reward outcomes were phasic, peaking as the animal made his choice and received outcome feedback, respectively, and was broadly distributed across clusters (Fig. 4A-B).

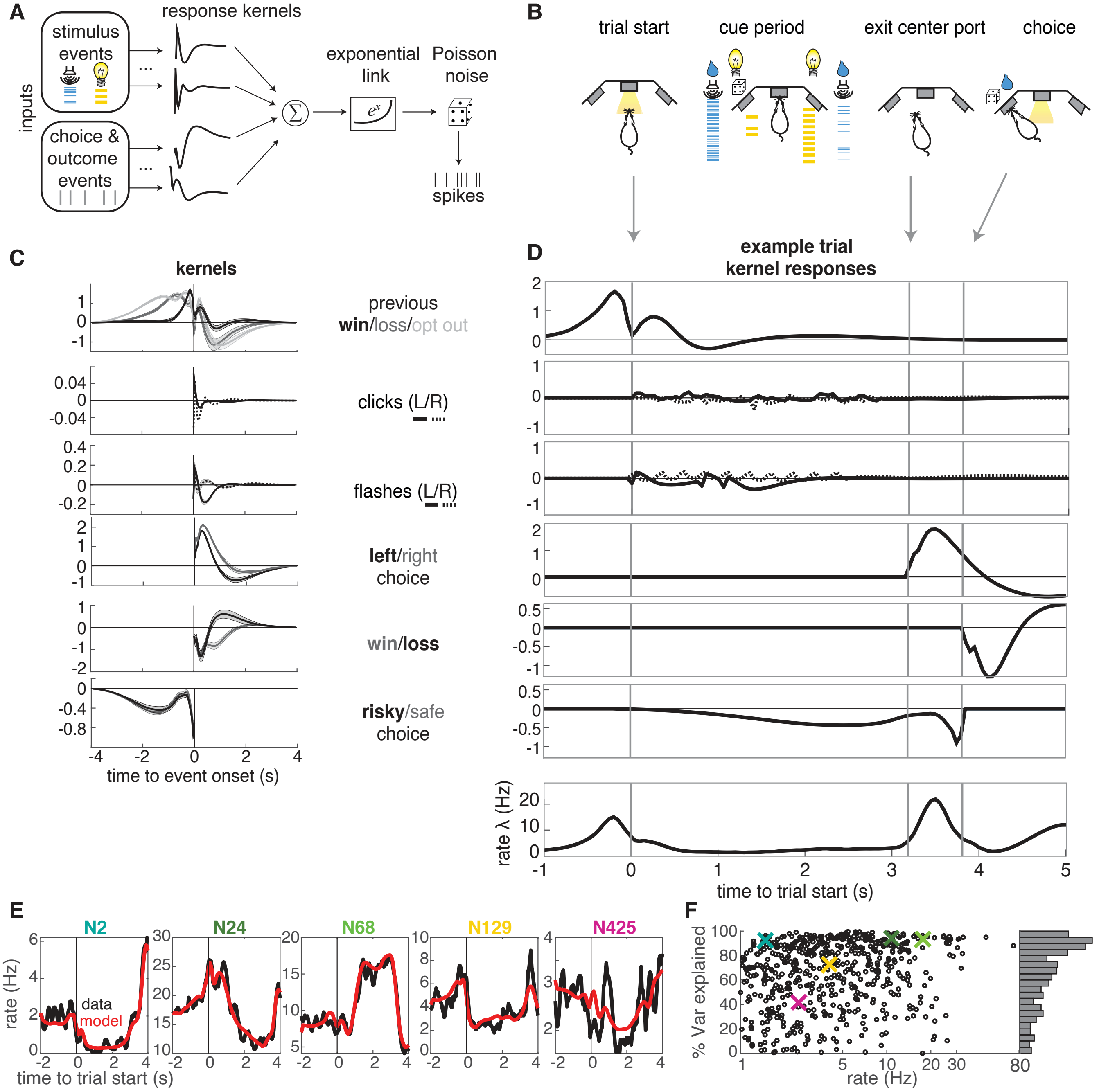


Figure 3: GLM analysis. A. Model schematic: the timings of external stimuli conveying reward attributes, as well as timings of choices and outcomes on each trial serve as model inputs. Nonlinear response kernels convolved with each input generates time dependent responses that are summed and exponentiated to give a mean firing rate, λ_t, in each time bin. Spikes are generated from a Poisson process with mean firing rate λ_t. B. Task schematic illustrating the key choice and outcome inputs to the model. C. Kernel fits for a sample neuron. Kernels are grouped by the aspects of the task that they model. Error bars denote estimated kernel standard deviation (Methods). D. Timing of each kernel's contribution in an example trial. The kernels in bold from panel C are the kernels that are active in this trial. The resulting model-predicted firing rate is shown in the bottom row. E. Representative PSTHs to held-out testing data from 5 different neurons (black) and model prediction (red). F. Variance explained for each neuron, with sample neurons from E denoted by correspondingly colored crosses. The distribution of R^2 values is presented along edge of the panel.

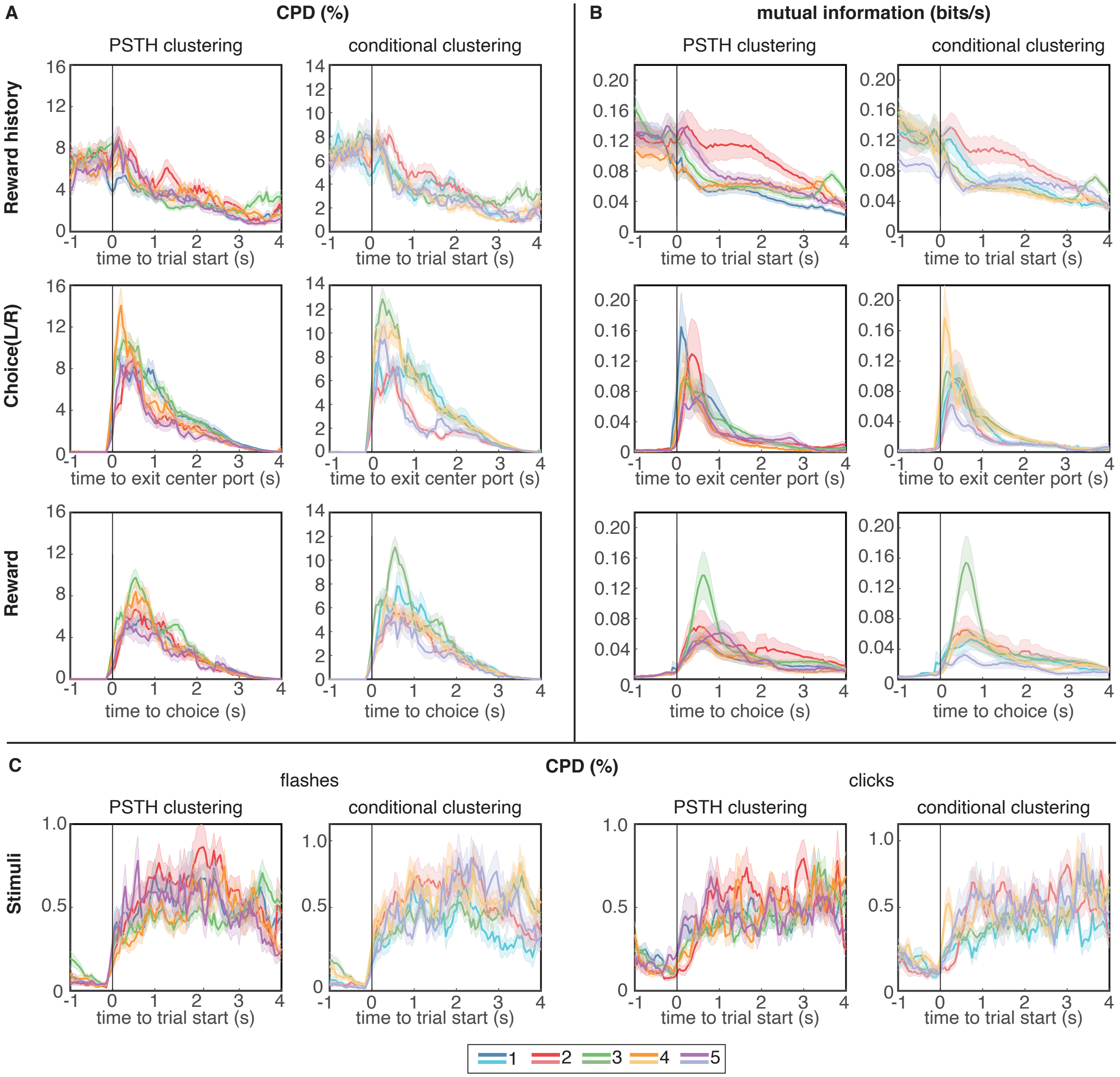


Figure 4: Coefficient of partial determination (CPD) and mutual information (MI) reveal broad encoding of all task parameters in each cluster. A. CPD for choice and reward outcome encoding: Reward history panels quantify encoding of outcome on the previous trial (previous win/loss/opt out), choice panels convey encoding of choices on the current trial (left/right), and reward panels quantify encoding of outcomes on the current trial (win/loss). B. Mutual information for same task parameters. C. CPD of flashes (left panels) and clicks (right panels) that encode reward probability and reward volume, respectively. CPD and MI values are averaged over neurons in each cluster; error bars show s.e.m. Results for PSTH clustering are in the left columns, and results for conditional clustering are in the right columns.

3.4 Reward history re-emerges before choice in a distinct subset of lOFC neurons

The CPD and MI analysis revealed one aspect of encoding that is unique to cluster 3. Both clustering methods identified largely overlapping populations of neurons in this cluster (Fig. 2E), indicating that these neurons exhibited similar temporal response profiles as well as encoding. Like neurons in the other clusters, neurons in cluster 3 encoded reward history at trial initiation, and that encoding decreased through the cue period. However, for neurons in this cluster, encoding of reward history re-emerged at the time preceding the animal's choice on the current trial (Fig. 5A-B, green lines). This "bump" of reward history encoding late in the trial was unique to cluster 3 regardless of the clustering method, and was observable in both CPD and MI (Fig. 4A-B). This result was further corroborated by computing the average discriminability index or (d') for reward history, which is a model-agnostic metric that quantifies the difference in mean firing rate in each condition, accounting for response variance over trials. Cluster 3 was unique from the other clusters for having a subset of neurons with high d' values for reward history at the time preceding the animal's choice (Fig. 5C). We additionally found that encoding of previous wins and losses during this time period only extended to the previous trial, and did not encode the outcome of additional past trials (Suppl. Fig. S7).

Notably, cluster 3 also exhibited the most prominent encoding of reward outcome compared to the other clusters (Fig. 4A-B, green, bottom row). This suggests that this subset of neurons may be specialized for representing or even integrating information about reward outcomes on previous and current trials. We wondered if this might reflect adaptive value coding, in which value representations in OFC have been observed to dynamically reflect the statistics of offered rewards [13–18]. Adaptive value coding, which is thought to reflect a divisive normalization or range adaptation mechanism, allows the circuit to efficiently represent values in diverse contexts in which their magnitudes may differ substantially. As such, it provides key computational advantages, such as efficient coding, or the maximization of mutual information between neural signaling and the statistics of the environment [14–17, 19–22].

According to divisive normalization models of subjective value, the value of an option or outcome is divided by a recency-weighted average of previous rewards [14, 21, 22]. Therefore, if neurons in OFC exhibit adaptive value coding, we would predict that they would exhibit stronger reward responses following unrewarded trials, and weaker responses following rewarded trials (Fig. 6C). Put another way, regressing the firing rate against current and previous rewards should reveal coefficients with opposite signs [23]. Neurons with regression coefficients for current and previous rewards with the same sign would be modulated by current and previous rewards, but not in a way that is consistent with the adaptive value coding hypothesis (Fig. 6D).

To test this hypothesis, we regressed the firing rates of neurons in a 1 second window after reward receipt against reward win/loss outcomes on the current trial, as well as the win/loss outcome from the previous trial (Fig. 6, Methods). We found that 180 neurons (27% of population) exhibited significant coefficients for both current reward and reward history regressors, and 45% of these units demonstrated adaptive value coding conveyed by the opposite signs of regression coefficients for current and previous reward outcomes (81 units, Fig. 6A, blue). To determine if these adaptive neurons preferentially resided in a particular cluster, we calculated the cluster-specific probability of a neuron demonstrating either significant coefficients for rewarded volume and reward history, or having coefficients with opposite signs, consistent with adaptive value coding (Fig. 6B). We found that no cluster preferentially contained adaptive units with higher probability (comparison of 95% binomial confidence intervals). We also investigated whether adaptive value coding was present for rewarded volume representations, and similarly found that neurons did not preferentially reside in any one cluster (30 adaptive units out of 79 significant units, Suppl. Fig. S9). Notably, activity during the reward epoch did not appear to encode a reward prediction error (Suppl. Fig. S8), defined as the difference between the expected value and the outcome of the chosen option, as has been previously reported in rat and primate OFC [23–27].

4 Discussion

We have analyzed neural responses in lOFC from rats performing a value-based decision-making task, in which they chose between left and right offers with explicitly cued reward probabilities and volumes. Despite the apparent response heterogeneity in our dataset, two independent clustering methods revealed that neurons belonged to one of a small number of distinct clusters. We clustered based on the PSTHs over all trials, and found that subpopulations of neurons exhibited characteristic temporal response profiles. To our knowledge, this is a novel approach for identifying distinct neural response profiles in prefrontal cortex. We also clustered based on a conditional feature space for each neuron, consistent with previous work [7, 8]. The feature space for each neuron corresponded to its average firing rate on different trial types, in select time windows that most closely corresponded to the differentiating covariate on each trial (e.g., wins vs. losses at the time of reward feedback). Notably, these independent clustering methods identified highly similar groups of neurons for two of the clusters, indicating that neurons in these clusters exhibited similar temporal response profiles and task encoding, as defined by the feature space.

Previous studies that clustered using a conditional feature space identified a larger number of clusters than we did here [7, 8]. This could reflect the different metrics used to select the number of clusters (the gap statistic in

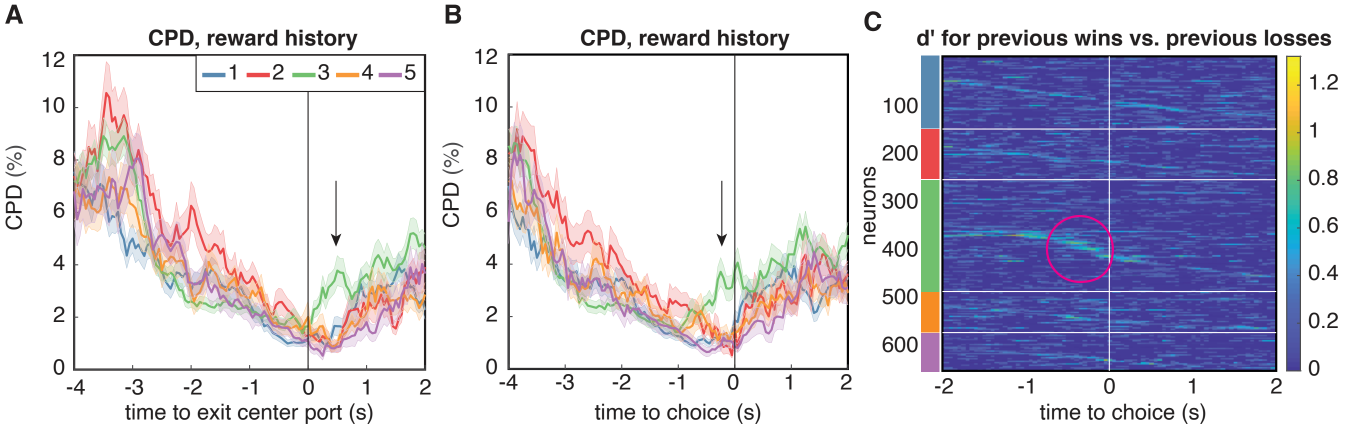


Figure 5: Encoding of reward history emerges late in the trial for cluster 3. A. CPD for reward history, aligned to leaving the center poke after the cue period. Note the peak in encoding that is isolated to cluster 3 (black arrow). B. CPD similarly aligned to choice, demonstrates that this encoding occurs before reward delivery on the current trial. C. Sensitivity, d', across time and neurons for reward history has activity isolated to cluster 3 (magenta circle). d' is sorted within the PSTH-based clusters by time-to-peak.

this study; compared to PCA and silhouette scores [8], or adjusted rand index [7]). Another difference is the time window that was used to generate the conditional feature space. In [7], for instance, the authors used the same time window for each feature, which was after the rat made a choice but before it received feedback about the outcome of that choice. There was no such epoch in our behavioral paradigm, precluding a more direct comparison.

4.1 Rat lOFC weakly encodes reward attributes

Our task design allowed us to isolate neural responses to sensory cues that conveyed information about distinct reward attributes, because these cues were presented independently and variably in time. However, our GLM revealed weak encoding of these reward attributes – reward probability and volume – across all clusters, regardless of the clustering method. This was observable by the CPD metric, and also by examination of the relative magnitude of the flash and click kernels in individual neurons. Indeed, the average flash and click CPD values were an order of magnitude smaller than for the other covariates, indicating that flashes and clicks did not contribute substantially to neural firing. Therefore, while behavioral analyses have shown that rats used the flashes and clicks to guide their choices [5], these cues were not strongly represented in lOFC. This weak encoding of reward attributes, whose combination would specify the subjective value of each offer [5], is consistent with a recent study of rat lOFC during a multi-step decision-making task that enabled dissociation of choice and outcome values [27, 28]. Recordings in that study revealed weak encoding of choice values, but strong encoding of outcome values, and optogenetic perturbations suggested a critical role for lOFC in guiding learning but not choice [27]. Other studies in rat lOFC have reported strong encoding of reward and outcome values specifically following action selection, but not preceding choice [7, 27, 29–32]. Notably, this is in contrast to studies in non-human primates, which have reported strong encoding of offer values in OFC after presentation of a stimulus, and before choice [23, 33–35]. A recent study in mouse OFC used olfactory cues to convey reward attributes (juice identity and volume) and reported encoding of offer values before choice [36]. This could either reflect a species difference between mice and rats, or perhaps encoding of olfactory stimuli in OFC. The lOFC receives prominent input from the mediodorsal nucleus of the thalamus (MD), which is strongly innervated by olfactory structures including the piriform cortex [37, 38], and some MD projection neurons are tuned to odor identity [39]. It is unclear what other factors may account for the pronounced encoding of offer value before choice in [36], but minimal encoding of offer value before choice in other rodent studies, including the present one.

4.2 Adaptive coding in rat lOFC

Previous studies in primate medial [15] and central-lateral [16, 17, 23, 40] OFC have reported subsets of neurons that adjust the gain of their firing rates to reflect the range of offered rewards, a phenomenon referred to as adaptive value coding. This type of coding is efficient because it would allow OFC to accurately encode values in diverse contexts that may vary substantially in reward statistics [14], analogous to divisive normalization in sensory systems [19, 20].

According to divisive normalization models of subjective value, the value of an option is divided by a recency-weighted average of previous rewards [14, 21, 22]. Therefore, we would predict that neurons implementing adaptive value coding would exhibit stronger responses following unrewarded trials, and weaker responses following rewarded

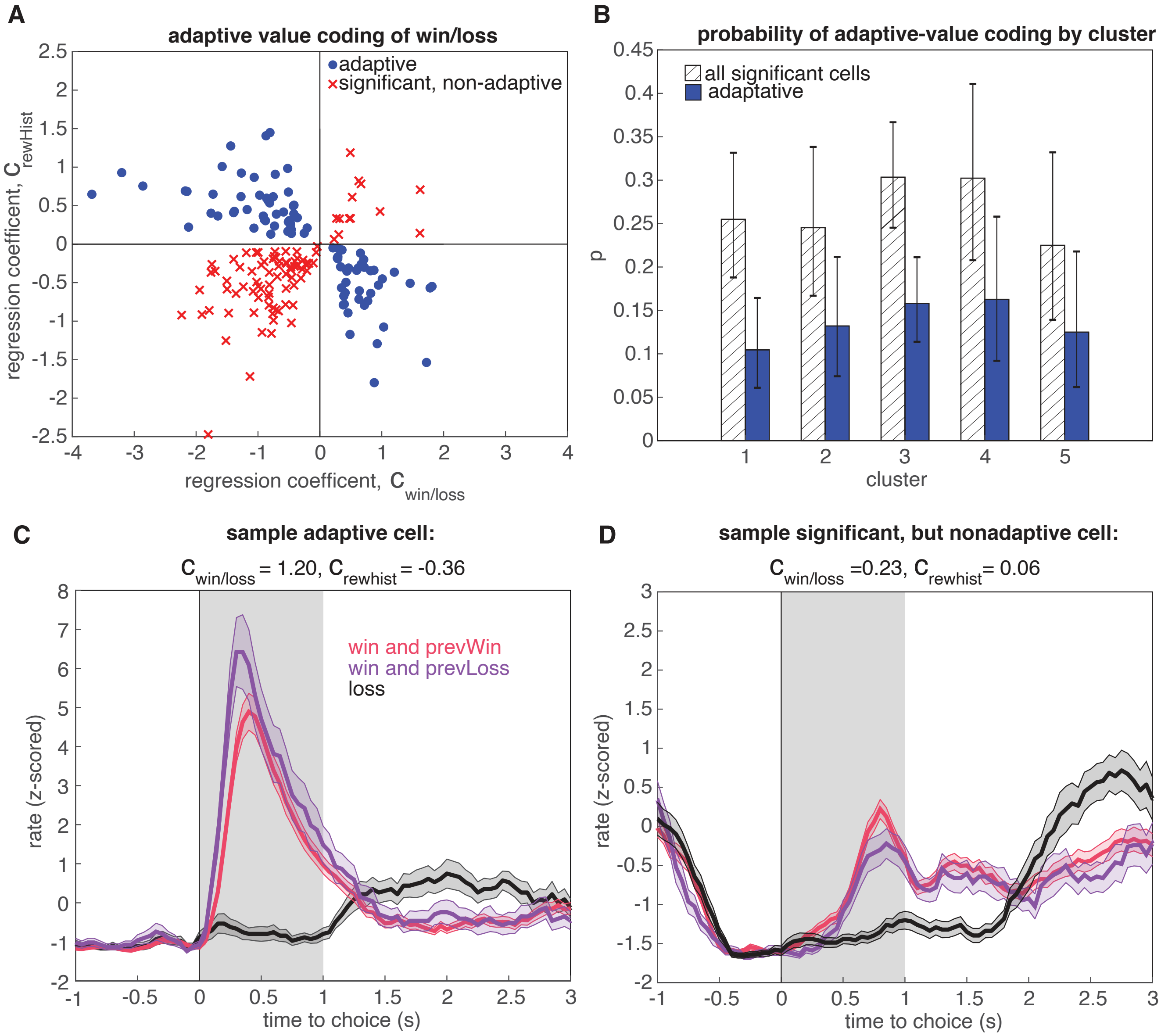


Figure 6: Adaptive value coding analysis using linear regression of firing rates for the 1s period following reward delivery against current and previous reward outcomes. A. Regression coefficients for the current reward outcome, $c_{\text{win/loss}}$ (parameterized as win=1, and loss=-1), and previous trial outcome, c_{rewhist}. 180 neurons had significant coefficients for both regressors ($p < 0.05$, t-test), and 81 neurons had coefficients with opposite signs, consistent with adaptive value coding (blue dots). The remaining neurons have differential responses due to reward history, but inconsistent with adaptive value coding (red crosses). B. Probability that a model with significant regressors for both current and past reward outcome would come from a given cluster. Shaded regions denote all models from panel A, and blue bars show the probability for adaptive neurons only. Error bars are the 95% confidence interval of the mean for a binomial distribution with observed counts from each cluster. C. Example cell demonstrating adaptive value coding. Shaded gray region denotes time window used to compute mean firing rate for the regression. D. Sample cell demonstrating significant modulation due to reward history, but with a relationship inconsistent with adaptive value coding.

trials. Consistent with this hypothesis, we identified a subset of neurons that had significant coefficients for rewards on current and previous trials, with opposite signs, and this fraction (∼15%) was comparable to the proportion of adaptive value coding neurons observed in central-lateral primate OFC [23]. Notably, we did not find any evidence of encoding of a reward prediction error during this epoch, or the difference between the expected value of the chosen lottery and the outcome. Therefore, the differential encoding of reward outcomes depending on reward history reflected a discrepancy between reward outcomes and recent experience, not a discrepancy between reward outcomes and expectations (as in a reward prediction error). These were dissociable in our task, due to the sensory cues that explicitly indicated expected value on each trial.

It is striking that we observed these correlates of adaptive value coding, as the present task did not vary reward statistics over trials. Therefore, OFC seems to dynamically adjust response gain on a trial-by-trial basis, even in contexts with stable reward contingencies. This activity was broadly distributed across all clusters in our dataset. Similar highly dynamic value coding has been observed in primate OFC [23].

4.3 Representations of reward history in OFC

In studies that require animals to learn and update the value of actions and outcomes from experience (*i.e.*, in accordance with reinforcement learning), values are often manipulated or changed over the course of the experiment to assess behavioral flexibility, value-updating, and goal-directed behavior. In rodents and primates, lesion and perturbation studies have shown that the OFC is critical for inferring value in these contexts, suggesting an important role in learning and dynamically updating value estimates for model-based reinforcement learning [27, 28, 41–46]. Neural recordings in these dynamic paradigms have revealed activity in OFC that reflects reward history and outcome values, which could subserve evaluative processing and learning [27, 29, 47].

Other studies, including this one, have used sensory cues (*e.g.*, static images, odors) to convey information about reward attributes. Once learned, the mapping between these cues and reward attributes is fixed, and the subject must choose between options with explicitly cued values. Neurons in the central-lateral primate OFC and rat lOFC have been shown to represent the values associated with these sensory cues in their firing rates (although in rats these representations tend to occur after the choice, if the task requires one), as well as reward outcomes and outcome values [7, 29, 33, 34, 40, 48–54]. However, the extent to which OFC is causally required for these tasks is a point of contention, and may differ across species [36, 55, 56]. Notably, even when reward contingencies are fixed over trials, animals often show sequential learning effects [5, 57], and reward history representations in OFC have been reported in tasks with stable reward contingencies [6, 23, 58].

In this study, we have described dynamic trial-by-trial changes in firing rates that reflected reward history just preceding the choice. In contrast to broadly distributed adaptive value coding, this activity was restricted to a particular subset of neurons that were identifiable by two independent clustering methods, and that exhibited the strongest encoding of reward outcomes. We hypothesize that this subpopulation of neurons, which were identifiable based on their temporal response profiles alone, may potentially mediate sequential effects, or the effect of previous outcomes on current choices. We have previously shown that optogenetic perturbations of lOFC in this task, triggered when rats exited the center port, disrupted sequential trial-by-trial learning effects [6]. The subpopulation of neurons in cluster 3 that specifically encodes (1) reward history preceding choice, and (2) reward outcomes immediately following choice are well-suited to mediate the influence of previous trials on the animal's choice. Future experiments should determine whether these neurons project to a common downstream target in service of trial-by-trial learning.

5 Methods

5.1 Animal subjects and behavior

The details for the animal subjects, behavioral task, and electrophysiological recordings have been described in detail elsewhere [5, 6]. Briefly, neural recordings from three male Long-Evans rats were used in this work. Animal use procedures were approved by the Princeton University Institutional Animal Care and Use Committee and carried out in accordance with National Institutes of Health standards.

Rats were trained in a high-throughput facility using a computerized training protocol. The task was performed in operant training boxes with three nose ports, each containing an LED. When the LED from the center port was illuminated, the animal was free to initiate a trial by poking his nose in that port (trial start epoch). While in the center port, rats were continuously presented with a train of randomly timed auditory clicks from a left and right speakers. The click trains were generated by Poisson processes with different underlying rates [59]; the rates from each speaker conveyed the water volume baited at each side port. Following a variable pre-flash interval ranging from 0 to 350 ms, rats were also presented with light flashes from the left and right side ports, where the number of flashes conveyed reward probability at each port. Each flash was 20 ms in duration, presented in fixed bins, and spaced every 250 ms to avoid perceptual fusion of consecutive flashes. After a variable post-flash delay period from

0 to 500 ms, there was an auditory "go" cue and the center LED turned back on. The animal was then free to leave the center port (exit center port epoch) and choose the left or right port to potentially collect reward (choice epoch).

5.2 Electrophysiology and data pre-processing for spike train analyses and model inputs

Tetrodes were constructed from twisted wires that were either PtIr (18 µm, California Fine Wire) or NiCr (25 µm, Sandvik). Tetrode tips were platinum- or gold-plated to reduce impedances to 100–250 kΩ at 1 kHz using a nanoZ (White Matter LLC). Microdrive assemblies were custom-made as described previously [60]. Each drive contained eight independently movable tetrodes, plus an immobile PtIR reference electrode. Each animal was implanted over the right OFC. On the day of implantation, electrodes were lowered to 4.1 mm DV. Animals were allowed to recover for 2–3 weeks before recording. Shuttles were lowered 30–60 µm approximately every 2–4 days.

Data were acquired using a Neuralynx data acquisition system. Spikes were manually sorted using MClust software. Units with fewer than 1% inter-spike intervals less than 2 ms were deemed single units. For clustering and model fitting, we restricted our analysis to single units that had a mean firing rate greater that 1 Hz (659 units). To convert spikes to firing rates, spike counts were binned in 50 ms bins and smoothed using Matlab's smooth.m function with a 250ms moving window. Similarly, our neural response model fit spike counts in discretized bins of 50ms. When parsing data into cross-validated sets of trials balanced across conditions, a single trial was considered as the window of [-2,6]s around trial start. In all other analyses, conditional responses were calculated on trials with a window [-2,4]s for data aligned to trial start, and [-4,4]s for choice-aligned or leave centerpoke-aligned responses.

5.3 Clustering of lOFC responses

5.3.1 Feature space parametrization

To analyze the heterogeneity in time-dependent lOFC responses, our first clustering procedure utilized trial-averaged PSTHs from each neuron to construct the feature space for clustering. Specifically, for a set of N neurons and trials of T timepoints, we z-scored the PSTH of each neuron and combined all responses into a matrix $Z \in \mathbb{R}^{N\times T}$, then performed PCA to obtain principal components, W, and score, M, as $Z = MW^T$. We found that $k = 18$ components explained $> 95\%$ of the covariance in Z and used the first k columns of M, the PSTH projected onto the top k PC, as our feature space on which to perform clustering.

Our second clustering procedure used time-averaged, conditional neural responses to construct the feature space for k-means clustering. This is similar in form to the approach in [7]. The feature space consisted of its z-scored firing rate conditioned on choice, reward outcome, reward history, presented offer value (EV of left and right offers), and rewarded volume, in time bins that most often corresponded to differential encoding of each variable (Fig. 2A, Fig. 2B). Specifically, we used conditional PSTHs that depend on a single condition, and marginalized away all other conditions. The conditions, $X^{(M)} = x_j$, are grouped into three categories for our task. Choice and outcome information is $\mathbf{X}^{(\text{reward})} \in \{\text{win,loss}\}, \mathbf{X}^{(\text{rewardHistory})} \in \{\text{previous Win,previous loss}\}, \mathbf{X}^{(\text{choice})} \in \{\text{left,right}\}$. Presented offer attributes on left and right ports are the expected value of reward as $EV = pV$, where p is reward probability conveyed through flashes, and V is the volume offer conveyed through Poisson clicks. Values were binned on a log-2 scale: $\mathbf{X}^{(\mathbf{EVL})}, \mathbf{X}^{(\mathbf{EVR})} \in \{[0,6),[6,12),[12,24),[24,48]\mu l\}$. Rewarded value is $\mathbf{X}^{(\text{rewVol})} \in \{0,6,12,24,48\mu l\}$.

The conditional PSTH responses were z-scored, and then each conditional PSTH was averaged over the time window in which the behavioral variable was maximally encoded (dictated by peak location in CPD, see Methods below) to obtain a conditional firing rate as a single feature for clustering. Specifically, the time windows for reward and reward volume information were averaged over $[0,3]s$ after reward delivery, $[-1,2]s$ after trial start for reward history, $[0,1.5]s$ after exiting the center port for left/right choice, and $[-1,0]s$ before the animal's choice (*i.e.*, entering the side port) for expected value of presented offers. These 19 features were combined and pre-processed using PCA in the same way as the PSTH-based clustering to yield 11 features.

5.3.2 Evaluation of k-means cluster quality

For each clustering procedure, we utilized k-means clustering to locate groups of functionally distinct responses in lOFC, and used the gap statistic criterion to determine a principled choice of the best number of clusters (evalclusters.m in Matlab) [9]. Specifically, we locate the largest cluster size K for which there was a significant jump in gap score $Gap(K)$,

$$Gap(K) \geq Gap(K-1) + 2SE(K-1). \tag{2}$$

This is similar to a standard option in evalclusters.m ('SearchMethod'= firstMaxSE'), which finds the smallest instance in which a non-significant jump in cluster size is located. The two methods often agree. Finally, we used 5000 samples for the reference distribution to ensure convergence of results in the gap statistic.

5.3.3 Cluster labeling consistency and cluster similarity

To compare the consistency of results between the PSTH clustering and conditional clustering (Fig. 2E), we calculated $P(C_{\text{conditional}}|C_{\text{PSTH}})$, the conditional probability of a neuron being assigned to cluster $C_{\text{conditional}}$ from the conditional clustering procedure, given that it was assigned to cluster C_{PSTH} in the PSTH cluster procedure. We evaluated the similarity of clusters within a given clustering procedure in two ways. First, we performed TSNE embedding of the features space to visualize cluster similarity in two dimensions (sklearn.manifold.TSNE in python, perplexity=50, n_iter=5000) [61]. We then colored each sample in this 2-D space based on cluster identity (Figs. S3A-B). We quantified the distance amongst clusters by calculating the cluster averaged Mahalanobis distance to the other clusters [62]. The Mahalanobis distance $D_M(x)$ calculates the distance of a sample x_A to a distribution B with a known mean, μ_B, and covariance, S_B:

$$D_M(x_A, B) = \sqrt{(x_A - \mu_B)^T S_B^{-1} (x_A - \mu_B)}. \tag{3}$$

The cluster-averaged distances in Figures S3C-D average D_M over all samples from a given cluster A as

$$D_M(A, B) = \mathop{\mathbb{E}}_{x_A \in A} [D_M(x_A, B)]. \tag{4}$$

5.4 Generalized linear model of neural responses in lOFC

Our neural response model is a generalized linear model with an exponential link function that estimates the probability that spiking from a neuron will occur in time bin t with a rate λ_t, and with Poisson noise, given a set of time-dependent task parameters. Specifically, for a design matrix $\mathbf{X}$ with S columns of task variables $X_t^{(s)}$, the probability of observing y_t spikes from a neuron in time bin $t + \Delta t$ is modeled as

$$p(y_t|\mathbf{X}) \sim Poiss(\lambda_t) = \frac{1}{y_t!} (\lambda_t \Delta t)^{y_t} e^{-\lambda_t \Delta t} \tag{5}$$

$$\lambda_t = \exp\left[\sum_{s=1}^{S} \left(X_{t-\tau:t}^{(s)} * k_s(\tau) \right) + \theta_0 \right], \tag{6}$$

where λ_t is the rate parameter of the Poisson process. Task variables are linearly convolved with a set of response kernels $k_s(\tau)$ that allow for a time-dependent response from a neuron that may occur after (causal) or before (acausal) the event in the trial. The kernels are composed of a set of N_s basis functions, $\phi_s^{(k)}$, linearly weighted by model parameters $\theta_s^{(k)}$:

$$k_s(\tau) = \sum_{j=1}^{N_s} \theta_s^{(j)} \phi_s^{(j)}(\tau). \tag{7}$$

Additionally, we include a parameter θ_0 that captures the background firing rate of each neuron.

We used 15 time-dependent task variables in our model that indicate both the timing of an event in the task, as well as behavioral or conditional information. Parameterization of the model in this manner with a one-hot encoding of each condition per variable allows for asymmetric responses to different outcomes (*e.g.*, wins vs. losses), and also captures the variable timing of each event in the task. The task variables were the following: (4) stimulus variables of left and right clicks and flashes (aligned to trial start). (4) Choice variables for choosing either the left of the right port (aligned to exit center port). Similarly, we included an alternate parameterization of choosing either the safe port ($p = 1$) or the risky port ($p < 1$). (5) Outcome variables were wins or losses on the current trial (aligned to choice); and reward history variables of previous wins, losses, or a previous opt-opt (aligned to trial start). We included a previous reward rate task variable that was calculated as the average rewarded volume from all previous trials in the session (aligned to trial start). We also included a "session progress" variable as the normalized [0,1] trial number in the session (aligned to trial start). This covariate captures motivational or satiety effects on firing rate over the course of a session. Finally, model comparison of cross-validated log-likelihood indicated that an autoregressive spike-history kernel was not necessary for our model.

5.4.1 Parameter learning, hyperparameter choice, and model validation

The set of parameters θ of the model are the kernel weights $\theta_s^{(j)}$ and the background firing rate θ_0. θ were fit by minimizing $\mathcal{L}$, the negative-log likelihood $-\log[p(y|\theta)]$ with an additional L_2 penalty acting as a prior over model parameters [12]:

$$\begin{aligned}\mathcal{L} &= -\log p(y|\theta) - \log p(\theta) \\ &= \sum_t \lambda_t - y_t \log \lambda_t + \frac{\xi}{2}\|\theta\|^2. \end{aligned} \tag{8}$$

Model parameters were chosen through four-fold cross validation, and ξ was found through a grid search optimization on a held-out test set. Specifically, we split the data from each neuron into 5 equal parts that were balanced among trial contingencies (*i.e.*, equal amounts of wins/losses, previous wins/losses, and left/right choices per partition). One partition (test set) was not utilized in fitting θ, and was held out for later model comparison and hyperparameter choice. The remaining four partitions were used in cross validation to fit four models, with each model fit on three partitions and assessed on the fourth "validation" partition. The model with the lowest negative log-likelihood on the validation set was chosen for further analysis. This procedure was repeated iteratively on an increasingly smaller grid of initial hyperparameter values $\xi \in [10^{-5}, 10]$, and the hyperparameter yielding the lowest negative log-likelihood on the test partition was chosen. We chose this approach for hyperparameter optimization in lieu of approximations such as calculating evidence [63], as we found the underlying approximations to be limiting for our data. In general, when building our model we assessed the aspects of other hyperparameter choices (*i.e.*, kernel length, symmetric vs. asymmetric conditional kernels, number of task variables) with cross-validated negative log-likelihood on held-out test data. See the model comparison section of Supplementary Materials for further details. Finally, kernel covariances and standard deviations were estimated using the inverse of the Hessian of $\mathcal{L}$.

5.4.2 Basis functions

We utilized a log-scaled, raised cosine basis function set for our model [12]. These functions offer the ability to generate impulse-like responses shortly after stimulus presentation, as well as broader, longer time-scale effects. The form of the basis function is given as

$$\phi_j(\tau) = \frac{1}{2}\cos(a\log[\tau] - b_j) + \frac{1}{2}, \tag{9}$$

where a is parameter that controls breadth of support of each basis function, and b_j controls the location of its maxima. The parameters a, b_j were chosen to spread the set of basis functions $\{\phi_j\}$ in roughly a log-linear placement across their range of support. This gives a better coverage than linear spacing, as the basis functions increase in breadth as b_j increases. We used 9 such functions, and additionally augmented our basis set with two decaying exponential basis functions placed in the first two time bins to capture any impulse-like behavior at the onset of the task variable. This gives $M = 11$ basis functions for all kernels in the study. After a cursory model comparison of different kernel lengths we took a conservative approach and utilized a range of support of 4s for each kernel.

5.5 Coefficient of partial determination

The coefficient of partial determination (CPD) is a measure of the relative amount of % variance explained by a given covariate, and here we use it to quantify the encoding of behavioral information in single units. The CPD is defined as [27]

$$CPD(X^{(i)})_t = \frac{SSE(\mathbf{X}_{-i})_t - SSE(\mathbf{X}_{\text{all}})_t}{SSE(\mathbf{X}_{-i})_t} \times 100, \tag{10}$$

where SSE is the trial-summed, squared error between data and model. $\mathbf{X}_{\text{all}}$ refers to the full model with all covariates. $\mathbf{X}_{-i}$ implies a reduced model with the effect of $X^{(i)}$ omitted. Since our model utilized a ont-hot encoding for each condition, we calculated CPD by omitting the following groups of covariates: $\mathbf{X}^{(\text{reward})} \in \{x^{(\text{win})}, x^{(\text{loss})}\}$, $\mathbf{X}^{(\text{rewardHistory})} \in \{x^{(\text{prevWin})}, x^{(\text{prevLoss})}, x^{(\text{prevOptOut})}\}$, $\mathbf{X}^{(\text{choice})} \in \{x^{(\text{left})}, x^{(\text{right})}\}$, $\mathbf{X}^{(\text{clicks})} \in \{x^{(\text{Lclick})}, x^{(\text{Rclick})}\}$, $\mathbf{X}^{(\text{flashes})} \in \{x^{(\text{Lflash})}, x^{(\text{Rflash})}\}$.

Our measure of CPD assessed the encoding of the covariate of interest (*e.g.*, previous win or loss) separately from encoding of the general event-aligned response (*e.g.*, trial start). As such, our reduced model averaged the kernels that corresponded to the event-aligned response to create behaviorally irrelevant task variables. For example, for CPD of reward encoding: $X_t^{(\text{win})}, X_t^{(\text{loss})} \rightarrow \frac{1}{2}(X_t^{(\text{win})} + X_t^{(\text{loss})})$. For reward, reward history, and choice CPD calculations, we additionally weighted each trial type in SSE such that each condition contributed equally contributed to CPD, and omitted any bias due to an imbalance in trial statistics. Due to data limitations, we utilized the full set of training, testing, and validation data to calculate CPD. Units with a model fit of $R^2 > 0$ (590/659 units) were used in CPD calculation.

We assessed the significance of the CPD result by comparing it to a null distribution of 500 CPD values that were generated by shuffling the trial labels among the relevant covariates (*e.g.*, shuffle win and loss labels across trials, keeping timing of event and all other covariates fixed). CPD was deemed significant if it fell outside of the one-sided 95% confidence interval of the shuffle distribution, and plotted values in Figure 4 subtract off the mean of the shuffle distribution from the CPD.

5.6 Mutual information

Mutual information (MI) was used to calculate how much information about task variables is contained in lOFC firing rates, in different time windows throughout a trial. We calculated MI between spikes $\mathbf{Y_t}$ and a group of covariates $\mathbf{X^{(m)}} \in \{x_i\}$ (detailed in section 5.5) as:

$$\begin{aligned} MI_t &= H(\mathbf{X^m}) - H(\mathbf{Y_t}|\mathbf{X^{(m)}}) \\ &= -\sum_i p(x_i)\log(x_i) + \sum_{j,k} p(x_j)p(y_t|X^{(m)} = x_k)\log[p(y_t|X^{(m)} = x_k)]. \end{aligned} \tag{11}$$

The first term is the entropy of the stimulus, which is calculated from the empirical distribution. The second term is the conditional entropy of spiking, and requires calculation of the conditional distribution $p(y_t|X^{(m)} = x_k)$ by marginalizing over the fully conditional distribution that contains all other covariates from the model. We modeled the fully conditional distribution $p(y_t|X^{(1)}, X^{(2)}, ...X^{(S)}, ..X^{(m)} = x_k)$ via sampling of a doubly stochastic process, in which normal distributions of model parameters were propagated through our GLM and Poisson spiking. Specifically, for each time bin within each trial we sampled 500 parameter values from the normal distribution for each θ, where the covariance of θ was estimated as the inverse of the Hessian of $\mathcal{L}$. These samples were the passed through the exponential nonlinearity to generate a log-normal distribution of λ values, which were used as the rate parameter for a Poisson process that generated spikes. This spiking distribution of 500 samples was truncated at 10 spikes/50ms bin (a 200Hz cutoff), and the conditional distribution $p(y_t|X^{(m)} = x_k)$ was then taken as the average over trials in which $X^{(m)} = x_k$. Units with a model fit of $R^2 > 0$ (590/659 units) were used in MI calculations. We assessed significance of MI in each time bin by comparing to an analogously created distribution of 500 MI values in which trial labels were shuffled. Significant MI fell outside of the 95% confidence interval of this distribution.

5.7 Discriminability for reward history

The discriminability between previous wins and previous losses in Figure 5C was calculated in its unsigned form

$$d'_t = \frac{\left|\mu_{t,(\text{prevWin})} - \mu_{t,(\text{prevLoss})}\right|}{\sqrt{\frac{1}{2}(\sigma^2_{t,(\text{prevWin})} + \sigma^2_{t,(\text{prevLoss})})}}.$$

To account for an inflated range of nonsignificant d' values around zero due to the unsigned form of d', we subtracted from this quantity the mean d' of a trial-shuffled distribution. The shuffled distribution was generated by shuffling trial labels 1000 times. Significant d' values were identified as being outside of the one-sided 95% confidence interval of the shuffled distribution. Trial types were balanced when generating the shuffle distributions.

5.8 Linear regression models of pre-choice and post-choice epochs

For the linear regression of pre-choice epoch in Figure S7, the model used the trial history of wins and losses on the previous five trials as regressors, $x^{(i)}_{\text{win/loss}}$, the choice on that trial, $x^{(n)}_{\text{choice}}$, and an offset term, c_0:

$$r_n = c_0 + x^{(n)}_{\text{choice}} + \sum_{i=n-1}^{n-5} c^{(i)}_{\text{win/loss}} x^{(i)}_{\text{win/loss}}. \tag{12}$$

The regressors were binary +1/-1 variables for win(+1)/loss(-1) outcomes and left(+1)/right(-1) choice. Model coefficients c were fit with Matlab's fitlm function, and significance was assessed via an F-test comparing to the baseline model containing only c_0. The significance of each coefficient was assessed via a t-test with a cutoff of $p < 0.05$ for significance.

Similarly, the results in Figure 6 and Figure S9 investigating the adaptation of reward representations regressed r_n, the average firing rate in the 1 second interval after reward onset on each trial n (post-choice epoch), against previous and current trial outcomes. The first model investigated adaptation of the rewarded outcome representation by including a binary win(+1)/loss(-1) regressor for current trial reward outcome, a binary win(+1)/loss(-1) regressor for previous reward outcome, the left(+1)/right(-1) choice on that trial, and an offset term :

$$r_n = c_{\text{win/loss}} x^{(n)}_{\text{win/loss}} + c_{\text{rewhist}} x^{(n-1)}_{\text{win/loss}} + c_{\text{choice}} x^{(n)}_{\text{choice}} + c_0. \tag{13}$$

Similarly, the other model investigating adaptation of reward volume representations instead used a regressor for rewarded volume, and a binary loss(1/0) regressor for outcome on the current trial:

$$r_n = c_{\text{vol}} x_{\text{vol}}^{(n)} + c_{\text{loss}} x_{\text{loss}}^{(n)} + c_{\text{rewhist}} x_{\text{win/loss}}^{(n-1)} + c_{\text{choice}} x_{\text{choice}}^{(n)} + c_0. \tag{14}$$

Finally, the regression in Figure S8 modeled the post-choice epoch using a reward prediction error as a regressor:

$$r_n = c_{\text{rpe}} x_{\text{rpe}}^{(n)} + c_{\text{choice}} x_{\text{choice}}^{(n)} + c_0, \tag{15}$$

where the RPE regressor was the difference in rewarded volume and the expected value of the chosen option on that trial: $x_{\text{rpe}}^{(n)} = V_{\text{reward}}^{(n)} - p^{(n)} V_{\text{choose}}^{(n)}$.

References

[1] Matthew FS Rushworth, MaryAnn P Noonan, Erie D Boorman, Mark E Walton, and Timothy E Behrens. Frontal cortex and reward-guided learning and decision-making. *Neuron*, 70(6):1054–1069, 2011.

[2] Thomas A Stalnaker, Nisha K Cooch, Michael A McDannald, Tzu-Lan Liu, Heather Wied, and Geoffrey Schoenbaum. Orbitofrontal neurons infer the value and identity of predicted outcomes. *Nature communications*, 5(1):1–13, 2014.

[3] Jonathan D Wallis. Orbitofrontal cortex and its contribution to decision-making. *Annu. Rev. Neurosci.*, 30:31–56, 2007.

[4] Robert C Wilson, Yuji K Takahashi, Geoffrey Schoenbaum, and Yael Niv. Orbitofrontal cortex as a cognitive map of task space. *Neuron*, 81(2):267–279, 2014.

[5] Christine M Constantinople, Alex T Piet, and Carlos D Brody. An analysis of decision under risk in rats. *Current Biology*, 29(12):2066–2074, 2019.

[6] Christine M Constantinople, Alex T Piet, Peter Bibawi, Athena Akrami, Charles Kopec, and Carlos D Brody. Lateral orbitofrontal cortex promotes trial-by-trial learning of risky, but not spatial, biases. *eLife*, 8, 2019.

[7] Junya Hirokawa, Alexander Vaughan, Paul Masset, Torben Ott, and Adam Kepecs. Frontal cortex neuron types categorically encode single decision variables. *Nature*, 576(7787):446–451, 2019.

[8] Vijay Mohan K Namboodiri, James M Otis, Kay van Heeswijk, Elisa S Voets, Rizk A Alghorazi, Jose Rodriguez-Romaguera, Stefan Mihalas, and Garret D Stuber. Single-cell activity tracking reveals that orbitofrontal neurons acquire and maintain a long-term memory to guide behavioral adaptation. *Nature neuroscience*, 22(7):1110–1121, 2019.

[9] Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.

[10] Jonathan W Pillow, Liam Paninski, Valerie J Uzzell, Eero P Simoncelli, and EJ Chichilnisky. Prediction and decoding of retinal ganglion cell responses with a probabilistic spiking model. *Journal of Neuroscience*, 25(47):11003–11013, 2005.

[11] Jonathan W Pillow, Jonathon Shlens, Liam Paninski, Alexander Sher, Alan M Litke, EJ Chichilnisky, and Eero P Simoncelli. Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, 454(7207):995–999, 2008.

[12] Il Memming Park, Miriam L R Meister, Alexander C Huk, and Jonathan W Pillow. Encoding and decoding in parietal cortex during sensorimotor decision-making. *Nature Neuroscience*, 17(10):1395–1403, 2014.

[13] Shunsuke Kobayashi, Ofelia Pinto de Carvalho, and Wolfram Schultz. Adaptation of reward sensitivity in orbitofrontal neurons. *Journal of Neuroscience*, 30(2):534–544, 2010.

[14] Ryan Webb, Paul W Glimcher, and Kenway Louie. Rationalizing context-dependent preferences: divisive normalization and neurobiological constraints on choice. *SSRN Electron. J*, 10, 2014.

[15] Hiroshi Yamada, Kenway Louie, Agnieszka Tymula, and Paul W Glimcher. Free choice shapes normalized value signals in medial orbitofrontal cortex. *Nature communications*, 9(1):1–11, 2018.

[16] Camillo Padoa-Schioppa. Range-adapting representation of economic value in the orbitofrontal cortex. *Journal of Neuroscience*, 29(44):14004–14014, 2009.

[17] Aldo Rustichini, Katherine E Conen, Xinying Cai, and Camillo Padoa-Schioppa. Optimal coding and neuronal adaptation in economic decisions. *Nature communications*, 8(1):1–14, 2017.

[18] Katherine E Conen and Camillo Padoa-Schioppa. Partial adaptation to the value range in the macaque orbitofrontal cortex. *Journal of Neuroscience*, 39(18):3498–3513, 2019.

[19] Eero P Simoncelli and Bruno A Olshausen. Natural image statistics and neural representation. *Annual review of neuroscience*, 24(1):1193–1216, 2001.

[20] Matteo Carandini and David J Heeger. Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, 13(1):51–62, 2012.

[21] Agnieszka Tymula and Paul Glimcher. Expected subjective value theory (esvt): A representation of decision under risk and certainty. *Available at SSRN 2783638*, 2020.

[22] Kenway Louie and Paul W Glimcher. Efficient coding and the neural representation of value. *Annals of the New York Academy of Sciences*, 1251(1):13–32, 2012.

[23] Steven W Kennerley, Timothy EJ Behrens, and Jonathan D Wallis. Double dissociation of value computations in orbitofrontal and anterior cingulate neurons. *Nature neuroscience*, 14(12):1581, 2011.

[24] Michael A McDannald, Federica Lucantonio, Kathryn A Burke, Yael Niv, and Geoffrey Schoenbaum. Ventral striatum and orbitofrontal cortex are both required for model-based, but not model-free, reinforcement learning. *Journal of Neuroscience*, 31(7):2700–2705, 2011.

[25] Yuji K Takahashi, Matthew R Roesch, Thomas A Stalnaker, Richard Z Haney, Donna J Calu, Adam R Taylor, Kathryn A Burke, and Geoffrey Schoenbaum. The orbitofrontal cortex and ventral tegmental area are necessary for learning from unexpected outcomes. *Neuron*, 62(2):269–280, 2009.

[26] Yuji K Takahashi, Chun Yun Chang, Federica Lucantonio, Richard Z Haney, Benjamin A Berg, Hau-Jie Yau, Antonello Bonci, and Geoffrey Schoenbaum. Neural estimates of imagined outcomes in the orbitofrontal cortex drive behavior and learning. *Neuron*, 80(2):507–518, 2013.

[27] Kevin J Miller, Matthew M Botvinick, and Carlos D Brody. Value representations in orbitofrontal cortex drive learning, but not choice. *bioRxiv*, page 245720, 2018.

[28] Kevin J Miller, Matthew M Botvinick, and Carlos D Brody. Dorsal hippocampus contributes to model-based planning. *Nature neuroscience*, 20(9):1269, 2017.

[29] Jung Hoon Sul, Hoseok Kim, Namjung Huh, Daeyeol Lee, and Min Whan Jung. Distinct roles of rodent orbitofrontal and medial prefrontal cortex in decision making. *Neuron*, 66(3):449–460, 2010.

[30] Adam P Steiner and A David Redish. Behavioral and neurophysiological correlates of regret in rat decision-making on a neuroeconomic task. *Nature neuroscience*, 17(7):995–1002, 2014.

[31] Adam P Steiner and A David Redish. The road not taken: neural correlates of decision making in orbitofrontal cortex. *Frontiers in neuroscience*, 6:131, 2012.

[32] Zachary F Mainen and Adam Kepecs. Neural representation of behavioral outcomes in the orbitofrontal cortex. *Current opinion in neurobiology*, 19(1):84–91, 2009.

[33] Camillo Padoa-Schioppa and John A Assad. Neurons in the orbitofrontal cortex encode economic value. *Nature*, 441(7090):223–226, 2006.

[34] Zhongqiao Lin, Chechang Nie, Yuanfeng Zhang, Yang Chen, and Tianming Yang. Evidence accumulation for value computation in the prefrontal cortex during decision making. *Proceedings of the National Academy of Sciences*, 117(48):30728–30737, 2020.

[35] Erin L Rich and Jonathan D Wallis. Decoding subjective decisions from orbitofrontal cortex. *Nature neuroscience*, 19(7):973–980, 2016.

[36] Masaru Kuwabara, Ningdong Kang, Timothy E Holy, and Camillo Padoa-Schioppa. Neural mechanisms of economic choices in mice. *Elife*, 9:e49669, 2020.

[37] Wendy WP Tham, Richard J Stevenson, and Laurie A Miller. The functional role of the medio dorsal thalamic nucleus in olfaction. *Brain research reviews*, 62(1):109–126, 2009.

[38] Emmanuelle Courtiol and Donald A Wilson. The olfactory thalamus: unanswered questions about the role of the mediodorsal thalamic nucleus in olfaction. *Frontiers in neural circuits*, 9:49, 2015.

[39] Emmanuelle Courtiol and Donald A Wilson. Neural representation of odor-guided behavior in the rat olfactory thalamus. *Journal of Neuroscience*, 36(22):5946–5960, 2016.

[40] Léon Tremblay and Wolfram Schultz. Relative reward preference in primate orbitofrontal cortex. *Nature*, 398(6729):704–708, 1999.

[41] Michela Gallagher, Robert W McMahan, and Geoffrey Schoenbaum. Orbitofrontal cortex and representation of incentive value in associative learning. *Journal of Neuroscience*, 19(15):6610–6614, 1999.

[42] Alicia Izquierdo, Robin K Suda, and Elisabeth A Murray. Bilateral orbital prefrontal cortex lesions in rhesus monkeys disrupt choices guided by both reward value and reward contingency. *Journal of Neuroscience*, 24(34):7540–7548, 2004.

[43] Charles L Pickens, Michael P Saddoris, Michela Gallagher, and Peter C Holland. Orbitofrontal lesions impair use of cue-outcome associations in a devaluation task. *Behavioral neuroscience*, 119(1):317, 2005.

[44] MP Noonan, ME Walton, TEJ Behrens, J Sallet, MJ Buckley, and MFS Rushworth. Separate value comparison and learning mechanisms in macaque medial and lateral orbitofrontal cortex. *Proceedings of the National Academy of Sciences*, 107(47):20547–20552, 2010.

[45] Elizabeth A West, Jacqueline T DesJardin, Karen Gale, and Ludise Malkova. Transient inactivation of orbitofrontal cortex blocks reinforcer devaluation in macaques. *Journal of Neuroscience*, 31(42):15128–15135, 2011.

[46] Christina M Gremel and Rui M Costa. Orbitofrontal and striatal circuits dynamically encode the shift between goal-directed and habitual actions. *Nature communications*, 4(1):1–12, 2013.

[47] Ramon Nogueira Mañas, Juan M Abolafia, Jan Drugowitsch, Emili Balaguer-Ballester, Maria V Sanchez-Vives, and Rubén Moreno Bote. Lateral orbitofrontal cortex anticipates choices and integrates prior with current information. *Nature Communications. 2017 Mar 24; 8: 14823.*, 2017.

[48] SJ Thorpe, ET Rolls, and S Maddison. The orbitofrontal cortex: neuronal activity in the behaving monkey. *Experimental Brain Research*, 49(1):93–115, 1983.

[49] Geoffrey Schoenbaum, Andrea A Chiba, and Michela Gallagher. Orbitofrontal cortex and basolateral amygdala encode expected outcomes during learning. *Nature neuroscience*, 1(2):155–159, 1998.

[50] Jonathan D Wallis and Earl K Miller. Neuronal activity in primate dorsolateral and orbital prefrontal cortex during performance of a reward preference task. *European Journal of Neuroscience*, 18(7):2069–2081, 2003.

[51] Camillo Padoa-Schioppa. Neurobiology of economic choice: a good-based model. *Annual review of neuroscience*, 34:333–359, 2011.

[52] Dino J Levy and Paul W Glimcher. The root of all value: a neural common currency for choice. *Current opinion in neurobiology*, 22(6):1027–1038, 2012.

[53] Peter H Rudebeck and Elisabeth A Murray. The orbitofrontal oracle: cortical mechanisms for the prediction and evaluation of specific behavioral outcomes. *Neuron*, 84(6):1143–1156, 2014.

[54] Brian F Sadacca, Heather M Wied, Nina Lopatina, Gurpreet K Saini, Daniel Nemirovsky, and Geoffrey Schoenbaum. Orbitofrontal neurons signal sensory associations underlying model-based inference in a sensory preconditioning task. *Elife*, 7:e30373, 2018.

[55] Sébastien Ballesta, Weikang Shi, Katherine E Conen, and Camillo Padoa-Schioppa. Values encoded in orbitofrontal cortex are causally related to economic choices. *Nature*, 588(7838):450–453, 2020.

[56] Matthew PH Gardner, Davied Sanchez, Jessica C Conroy, Andrew M Wikenheiser, Jingfeng Zhou, and Geoffrey Schoenbaum. Processing in lateral orbitofrontal cortex is required to estimate subjective preference during initial, but not established, economic choice. *Neuron*, 108(3):526–537, 2020.

[57] Armin Lak, Emily Hueske, Junya Hirokawa, Paul Masset, Torben Ott, Anne E Urai, Tobias H Donner, Matteo Carandini, Susumu Tonegawa, Naoshige Uchida, et al. Reinforcement biases subsequent perceptual decisions when confidence is low, a widespread behavioral phenomenon. *ELife*, 9:e49834, 2020.

[58] Camillo Padoa-Schioppa. Neuronal origins of choice variability in economic decisions. *Neuron*, 80(5):1322–1336, 2013.

[59] Timothy D Hanks, Charles D Kopec, Bingni W Brunton, Chunyu A Duan, Jeffrey C Erlich, and Carlos D Brody. Distinct relationships of parietal and prefrontal cortices to evidence accumulation. *Nature*, 520(7546):220–223, 2015.

[60] Dmitriy Aronov and David W Tank. Engagement of neural circuits underlying 2d spatial navigation in a rodent virtual reality system. *Neuron*, 84(2):442–456, 2014.

[61] Geoffrey Hinton and Sam T Roweis. Stochastic neighbor embedding. In *NIPS*, volume 15, pages 833–840. Citeseer, 2002.

[62] Prasanta Chandra Mahalanobis. On the generalized distance in statistics. In *Proceedings of the National Institute of Sciences (Calcutta)*. National Institute of Science of India, 1936.

[63] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

6 Acknowledgements

We thank Carlos Brody, Carla Golden, and Alex Piet for helpful discussions and comments on the manuscript. This work was supported by NIMH grant 1R01MH125571-01 (CMC and CS).

7 Supplementary information

7.1 Excluded Data

The full data set consisted of 1881 single- and multi-units. Multi-unit recordings and neurons with a mean firing rate of less than 1 Hz were discarded, and the remaining 659 neurons were fit with the GLM, as well as utilized in the clustering procedures. Only neurons with a liberal threshold of $R^2 > 0$ were utilized in CPD and mutual information calculations (590 units). Additionally, we found that CPD measures for neurons can sometimes be noisy, and of the 659 neurons fit by the model, we further excluded neurons as outliers in the cluster-averaged CPD calculation if they had a CPD value > 0.5. This excluded a further 10 neurons.

7.2 GLM model comparison

To choose the best-fit model to our data we performed model comparison on held-out testing data. For comparison we considered models with additional task variables such as current and previously rewarded volume; as well as reduced models that omitted the variables relating to previous opt out trials, the previous reward rate, and the session progress. We also considered a reduced model that omitted the reward history contribution entirely. In each case, we assessed the population level change in model performance via a Wilcoxon signed rank test (Fig. S4). Our chosen model demonstrated a significantly lower population median in its negative log-likelihood than other models. We note that while our chosen model performed better than the reduced model at the population level, the median changes in neural responses were relatively slight (Fig. S4C, left panel). However, some neurons showed relatively strong improvement from introducing previousOptOut, previousRewardRate, and sessionProgress; which motivated us to keep them for the entire population (*i.e.*, Fig. S4C, middle panel). Additionally, we investigated a symmetrized form of our model that used a binary (±1) encoding of task variables, as opposed to a one-hot encoding. This model was rejected at the population level via model comparison (not shown).

7.3 Supplementary Figures

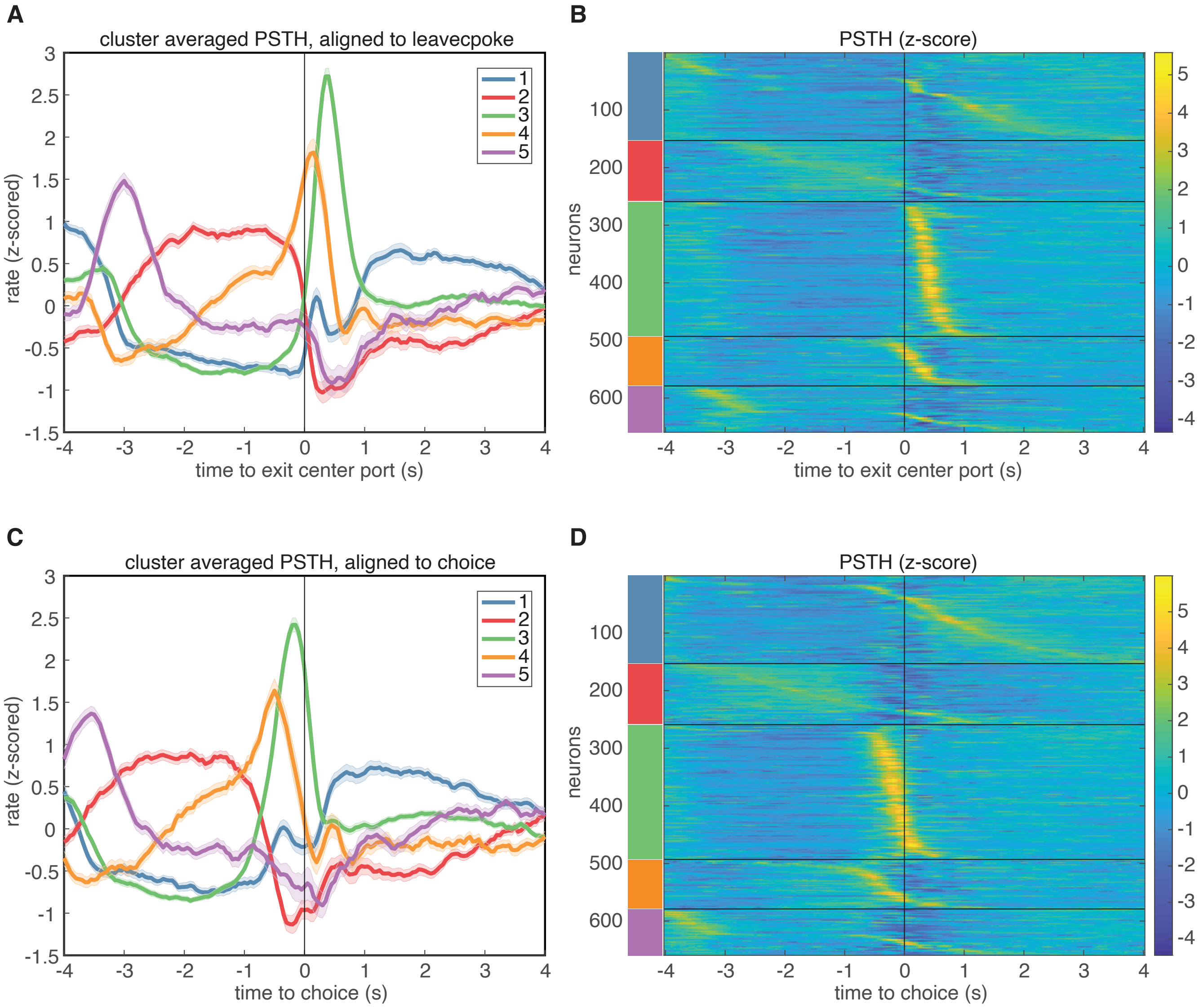


Figure S1: PSTH-based clustering of responses, aligned to different events in the task. A-B. Results aligned to exiting the center port. C-D. Results aligned to choice.

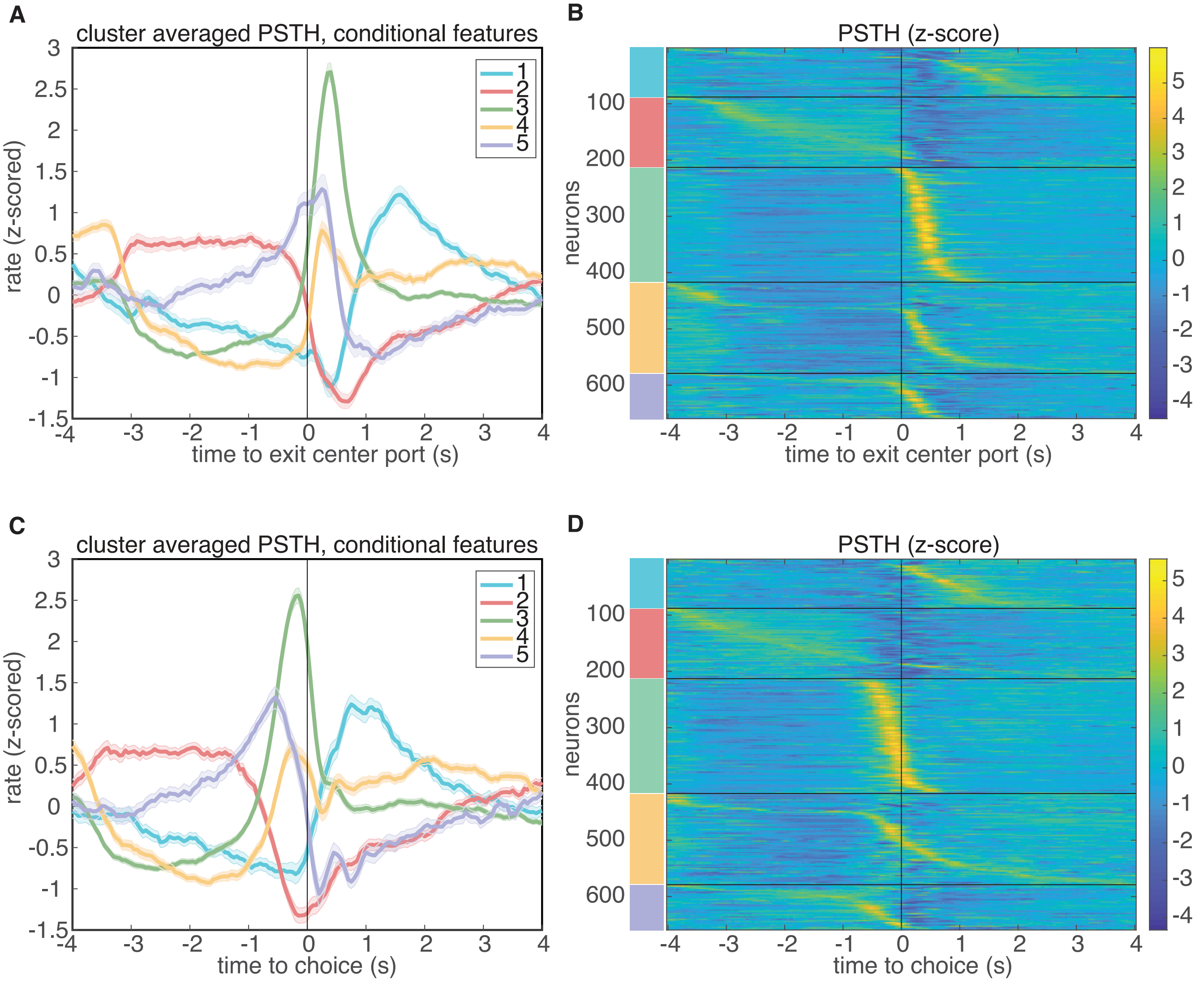


Figure S2: Results of clustering on conditional features space, aligned to different events in the task. A-B. Results aligned to exiting the center port. C-D. Results aligned to choice.

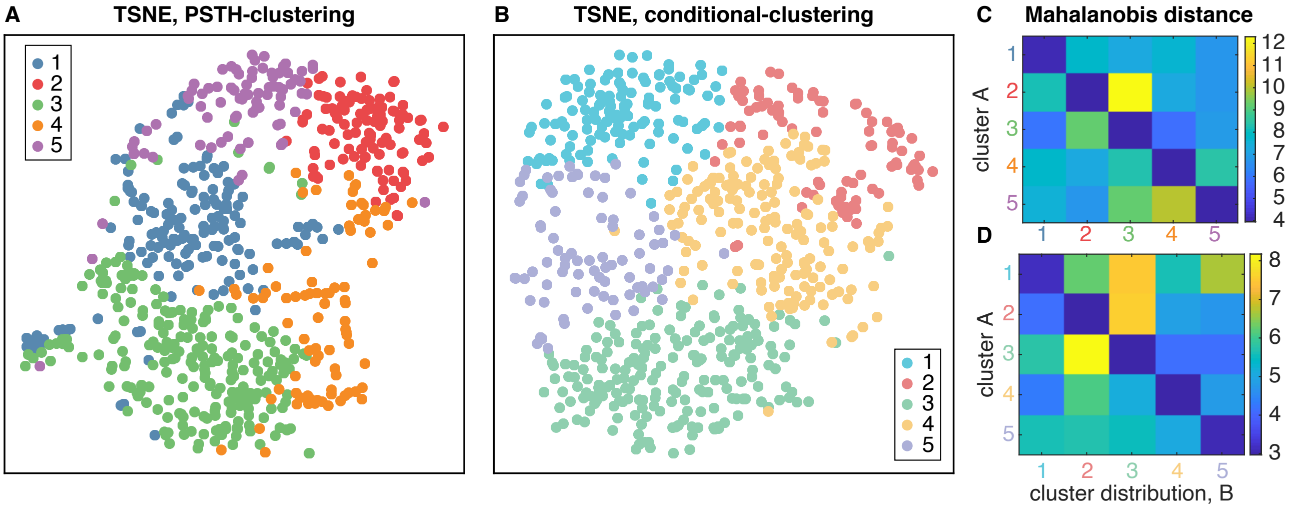


Figure S3: Similarity of clusters in feature space. A. t-SNE embedding of the PSTH feature space: Dots are individual neurons, and colors denote the cluster identity. B. Similar t-SNE embedding of the conditional feature space. Quantification of the cluster similarity through the cluster-averaged Mahalanobis distance of neurons from a given cluster distribution from all other clusters for C. the PSTH feature space clustering result and D. the conditional feature space clustering result.

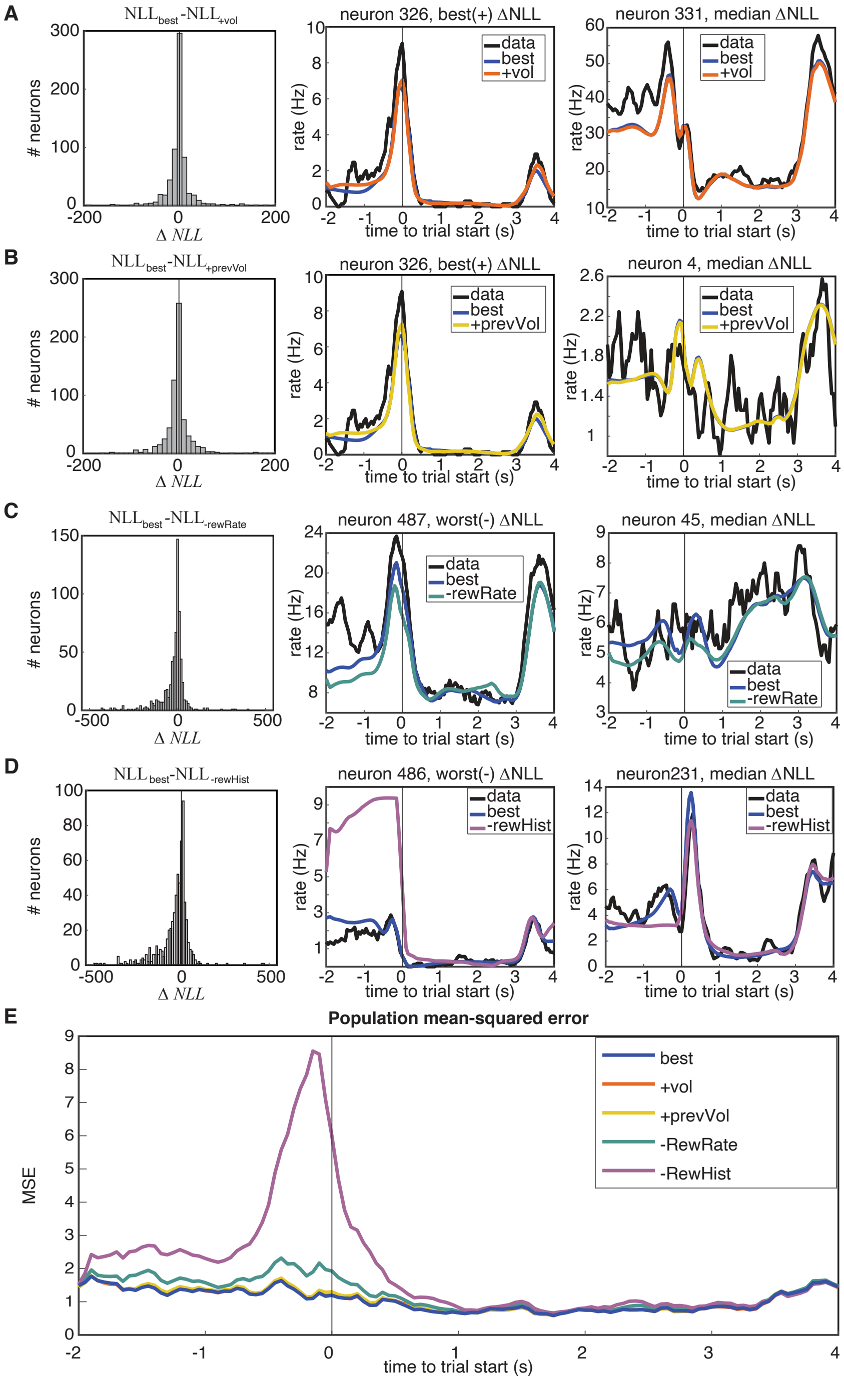


Figure S4: Model comparison. Model "best" was the model utilized in this work, and detailed in Figure 3. Two more complex models (+vol, +prevVol), as well as two simpler models (-RewRate,-RewHist) were compared through model comparison on held-out data log-likelihood. +prevVol substituted previous wins with rewarded volume on the previous trial, and +vol additionally substituted trial wins with rewarded volume. -RewRate omitted the sessionProgress, prevRewardRate, and previousOptOut kernels, and -RewHist further omitted the remaining reward history kernels. A-D. Individual model comparisons. (Left) Histograms of the population-level changes in held-out negative log-likelihood demonstrate when a model performs significantly better if the median ΔNLL is different than 0. Significance is assessed through Wilcoxon signed rank test. $p < 0.001$ in all comparisons. Median ΔNLL values were $\Delta NLL_{\text{best,+vol}} = -0.62$, $\Delta NLL_{\text{best,+prevVol}} = -1.77$, $\Delta NLL_{\text{best,-RewRate}} = -4.14$, $\Delta NLL_{\text{best,-RewHist}} = -15.13$. Sample neuron fits demonstrating the most extreme change in ΔNLL are shown in the middle columns, and neurons demonstrating the median change in ΔNLL are shown in right columns. E. Performance of models assessed through the mean-squared error on PSTHs from held-out test data as compared to model predictions.

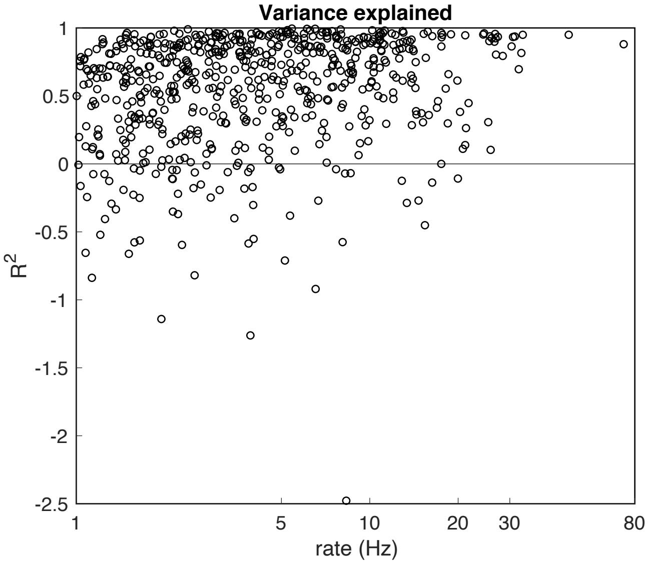


Figure S5: Variance explained, R^2, for all model fits. 69 units (10.5%) of the models had $R^2 < 0$, and were excluded from further analysis.

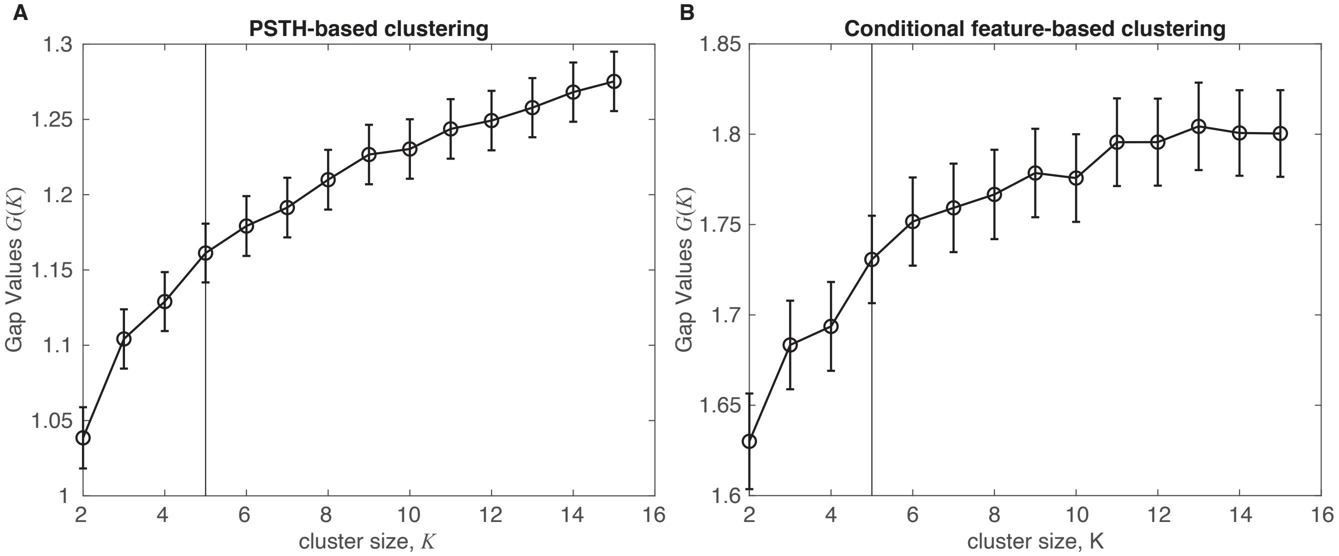


Figure S6: Gap statistic evaluation for both clustering approaches (see Methods for details). Error bars denote ± 2 s.e.m. A. PSTH-based clustering results. B. Alternative clustering approach based upon a conditional feature space. Vertical black lines denote largest significant number of clusters.

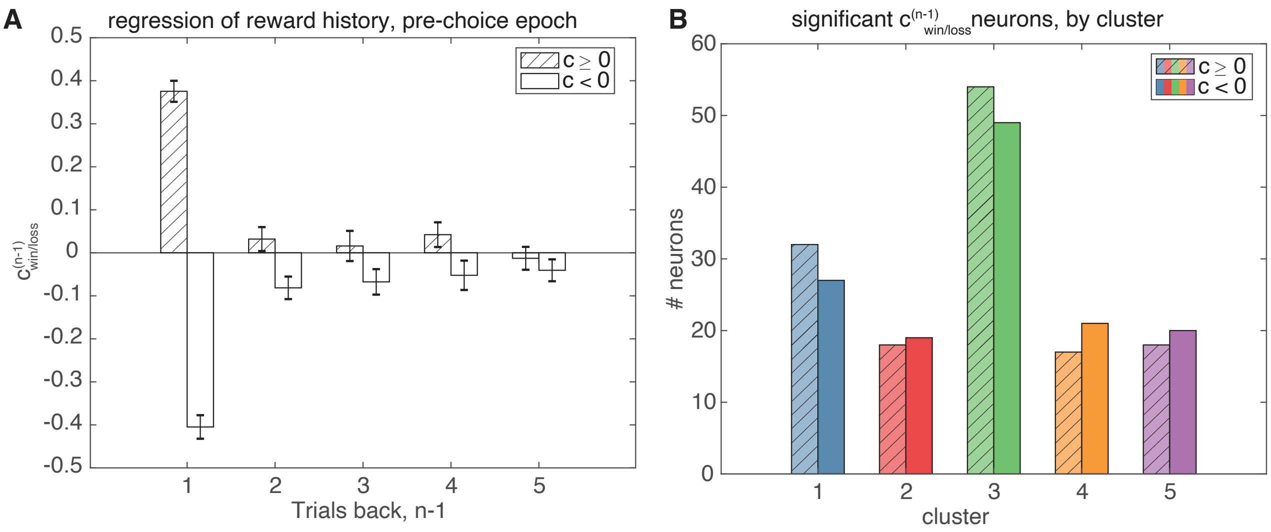


Figure S7: Linear regression of the pre-choice epoch based on recent trial history. A. Average regression coefficients for all statistically significant models with significant $c^{(n-1)}_{\text{win/loss}}$ ($p < 0.05$ for F-test comparing full model to mean-rate baseline model, $p < 0.05$ for t-test of $c^{(n-1)}_{\text{win/loss}}$). Regression coefficients one trial back have both positive and negative coefficients for reward outcome one trial back that convey different coding schemes for rewarded value, but coefficients for increasingly further trials back are not significant. B. Number of neurons with significant models containing either positive (shaded) or negative (not shaded) $c^{(n-1)}_{\text{win/loss}}$ coefficients, separated by cluster. Cluster 3 contains the majority of statistically significant neurons.

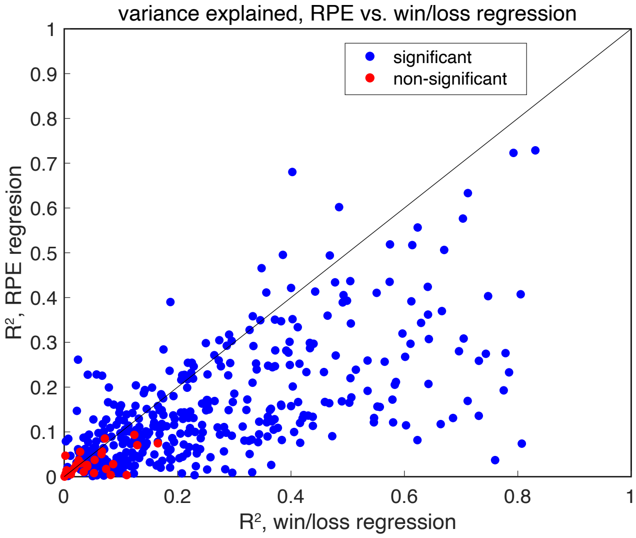


Figure S8: Linear regression of RPE in the post-choice epoch indicates that no RPE is present. Plotted is the variance explained for each neuron for two separate models: the binary win/loss model analyzed in Figure 6, and an equivalent model that replaces the current trial win/loss and past trial win/loss regressors with an RPE regressor. Blue dots indicate models with significant linear models, while red dots indicate non-significant models (threshold $p < 0.05$, F-test). The binary win/loss model captures more variance than the RPE model, indicating that wins and losses better explain that data. Further, model comparison between the two models by held out data log-likelihood reveals that the win/loss model is a better model (median $\Delta NLL = -5.9$, $p < 10^{-4}$, Wilcoxon signed rank test). The large proportion of RPE models that are significant is likely a reflection of the high correlation between RPE and win/loss regressors ($\rho = 0.80$).

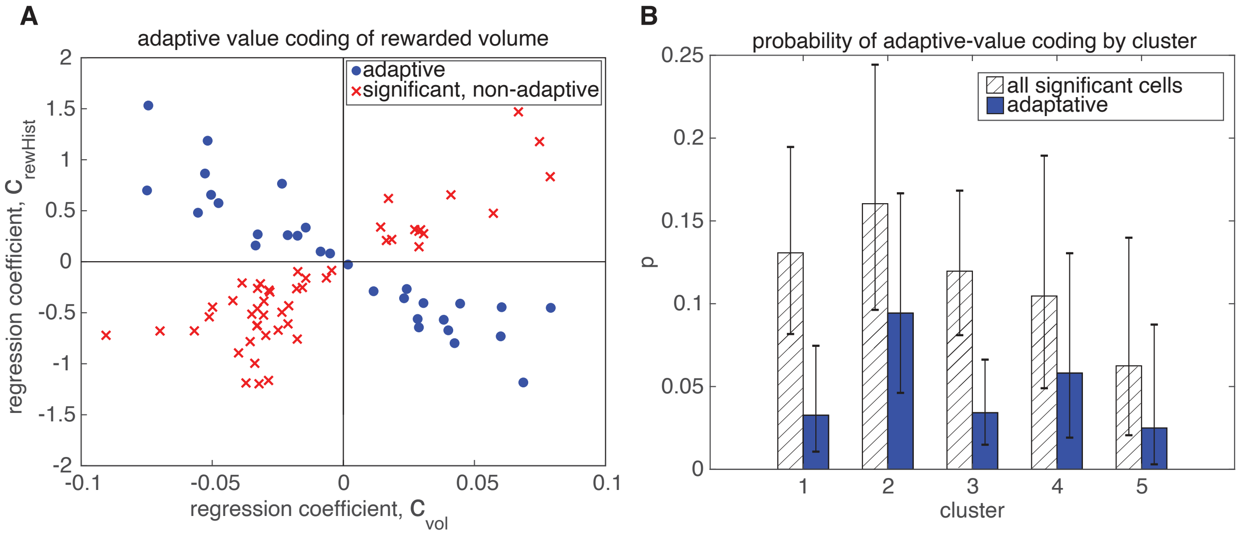


Figure S9: Adaptation of rewarded volume representation is present in fewer neurons than the reward outcome representation, but is still distributed across all clusters. Figure convention is similar to Fig. 6.