Negative reward-prediction errors of climbing fiber inputs for cerebellar reinforcement learning algorithm

Huu Hoang[1*#], Shinichiro Tsutsumi[2*], Masanori Matsuzaki[3], Masanobu Kano[4,5], Keisuke Toyama[1], Kazuo Kitamura[6#], Mitsuo Kawato[7#]

[1] ATR Neural Information Analysis Laboratories, Japan

[2] RIKEN Center for Brain Science, Japan

[3] Department of Physiology, Graduate School of Medicine, The University of Tokyo, Japan

[4] Department of Neurophysiology, Graduate School of Medicine, The University of Tokyo, Japan

[5] International Research Center for Neurointelligence (WPI-IRCN), The University of Tokyo, Japan

[6] Department of Neurophysiology, Graduate School of Interdisciplinary Research, University of Yamanashi, Japan

[7] ATR Computational Neuroscience Laboratories, Japan

* These authors contributed equally to this work

# Correspondence: Huu Hoang (hoang@atr.jp), Kazuo Kitamura (kitamurak@yamanashi.ac.jp) and Mitsuo Kawato (kawato@atr.jp)

Abstract

Although the cerebellum is widely associated with supervised learning algorithm, abundant reward-related representations were found in the cerebellum. We ask the question whether the cerebellum also implements reinforcement learning algorithm, especially the essential reward-prediction error. By tensor component analysis on two-photon Ca^{2+} imaging data, we recently demonstrated that a component of climbing fiber inputs in the lateral zones of mouse cerebellum Crus II represents cognitive error signals for Go/No-go auditory discrimination task. Here, we applied the Q-learning model to quantitatively reproduce Go/No-go learning behaviors, as well as to compute reinforcement learning variables including reward, predicted reward and reward-prediction error within each learning trial. Climbing fiber inputs to the cognitive-error component are strongly correlated with the negative reward-prediction error, and decreased as learning progressed. Assuming parallel-fiber Purkinje-cell synaptic plasticity, Purkinje cells of this component could acquire necessary motor commands based on the negative reward-prediction error conveyed by their climbing fiber inputs, thus providing an actor of reinforcement learning.

Introduction

Reinforcement learning algorithms have made significant progress in recent years (Sutton and Barto, 2018; Barto et al., 1983; Lecun et al., 2015), with a number of notable successes in video games (Mnih et al., 2015; Wurman et al., 2022) and GO games (Silver et al., 2016). However, AI still struggles with learning control of robots (Atkeson et al., 2018). Studying reinforcement learning in the brain can inspire the development of new AI algorithms that can learn more efficiently and adapt to changing environments. Traditionally, the cerebellum has been thought to implement a supervised learning algorithm (Marr, 1969; Albus, 1971; Ito ,1969; Kawato et al., 1987; Kawato and Gomi, 1992; Wolpert et al. 1998; Kawato, 1999), with the dopamine neurons in the basal ganglia being the center for a reinforcement learning algorithm (Schultz et al. 1997; Bayer and Glimcher 2005; Haruno and Kawato, 2006; Pessiglione et al., 2006; Akiti et al. 2022; Amo et al., 2022). Despite a vast amount of evidence that supports these theories (Shidara et al., 1993; Medina et al., 2000; Medina and Lisberger, 2008; Haruno et al. 2004; Doya 2000; Kaplan et al., 2020; Dabney et al. 2020; Starkweather and Uchida, 2020; Kawato and Samejima, 2007; Raymond and Medina, 2018; Kawato et al., 2020), recent studies have uncovered the presence of reward-related variables in both mossy-fiber-granule cell pathway and climbing fibers, two major excitatory inputs to the cerebellar cortex. On one hand, it has been reported that subsets of granule cells are activated by either reward prediction, reward delivery or reward omission during conditioning behaviors (Wagner et al., 2017) and that Purkinje cell simple spikes convey error signals of reward outcome (Sendhilnathan et al., 2020). On the other hand, climbing fibers also encode those reward-related signals (Heffley et al., 2018; Heffley and Hull 2019; Kostandinov et al., 2019; Larry et al., 2019; Sendhilnathan et al., 2021). Interestingly, such association between climbing fibers and reward variables is found to be dependent on zonal organization of the cerebellar cortex (Kostandinov et al., 2019; Tsutsumi et al., 2019). Furthermore, recent anatomical studies revealed that the cerebellum has reciprocal connectivity with the reward circuitry (Bostan and Strick, 2018; Carta et al., 2019; Wagner et al., 2019; Chabrol et al., 2019). All of these studies suggested a potential role of the cerebellum in reward processing (for reviews, see Wagner and Luo, 2020; Kawato et al., 2020; Kostadinov and Häusser, 2022), and in several reinforcement learning tasks that are driven by reward and penalty. However, it is still unknown whether the reward-prediction error, which is essential for reinforcement learning algorithm (Dabney et al. 2020; Amo et al., 2022; Starkweather and Uchida, 2020), is represented within the cerebellum or the cerebellum also implements reinforcement learning algorithm.

The cerebellar cortex is divided into zones, each of which possesses expression patterns of specific molecules such as aldolase-C (Sugihara and Shinoda, 2004, 2007). In our recent study, we examined two-photon Ca^{2+} imaging data of climbing fibers in eight aldolase-C zones (7+ to 4b-) from Crus II of mice cerebellum during the Go/No-go auditory discrimination task (Tsutsumi et al., 2019). The task is one of the reinforcement-learning tasks driven by reward and penalty. Hyper-resolution spike timing estimation algorithm (Hoang et al., 2020) and tensor component analysis (Williams et al., 2019) revealed that a component of climbing fiber inputs in the lateral zones represent cognitive error signals of this learning (Hoang et al., 2022). Among 8 aldolase-C positive and negative zones, zones 5-, 6+ and 6- contain this component most densely. In the present study, we utilized a formal reinforcement learning algorithm, namely Q-learning (Watkins and Dayan, 1992), to reproduce learning behaviors as well as to compute reinforcement learning variables such as reward, predicted reward and reward-prediction error within each learning trial. This simple Q-learning model with minimal number of hyper-parameters well reproduced learning behaviors of individual mice during Go/No-go tasks. We found that climbing fiber inputs to zones 5-, 6+ and 6- as well as the component of cognitive errors in the lateral zones are strongly correlated with the negative reward-prediction error. Assuming the long-term depression (Ito and Kano, 1982; Hirano 1990; Ito 2001) and long-term potentiation (Linden 1999; Hirano, 2013) of parallel-fiber Purkinje cell synapses of the component, Purkinje cells of this component could provide an actor (Barto et al., 1983) of reinforcement learning. That is, they can acquire necessary motor commands based on the negative reward-prediction error conveyed by their climbing fiber inputs.

Results

Q-learning model of licking behavior

We employed a Q-learning algorithm to model licking behavior of mice in the Go/No-go experiment (Fig 1A). The Q-learning model was selected because it is one of the simplest reinforcement learning algorithms, which are based on state-action value functions rather than state value functions, with a small number of hyperparameters. Briefly, mice were trained to distinguish two auditory cues by either lick or no-lick within a response period of 1s after cue onset to obtain liquid reward. The reinforcement-learning-algorithm state *s* is determined by the two auditory cues (*s* = Go / No-go), and the reinforcement-learning-algorithm action *a* is determined by lick within the time window of [0, 1] s after the auditory cue (*a* = Lick / No-lick). A reward value R-P was assigned according to cue-response conditions, including HIT trials (lick after Go cue) in Go task (R-P = 1), false alarm trials (FA, lick after No-go cue, R-P $=-\xi$, $0 \leq \xi \leq 1$), correct rejection trials (CR, no lick after No-go cue, R-P = 0) and MISS trials (no lick after Go cue, R-P = 0). Note that the R-P value for FA trials was constrained to be negative because a lick after a No-go cue was punished with a timeout of 4.5 s, while that of CR and MISS trials were 0 because neither reward nor penalty was given in those trials. The Q-learning algorithm assumed that a reward prediction Q was computed as a function of the two variables *s* and *a; state and action*. At each trial, a reward prediction error δQ, a difference between reward R-P and Q, was used to update computation of Q in the next trial with a learning rate α. For action policy, we used the softmax function with a single temperature parameter τ to convert the Q values into a probability distribution over the two possible actions; Lick and No-lick. Since the mice had undergone pre-training sessions lasting 3 days, when they were always rewarded with a lick within 1 second for both cues, the initial Q values for both Go and No-go cues were positive ($0 \leq q_1 \leq 1$, $0 \leq q_2 \leq 1$, for Go and No-go cues respectively, see Methods for details). It is important to note that we did not explicitly model the time course within trial as in previous studies (e.g., Schultz et al., 1997). Instead, we assumed the single timing of cue representation for computation of Q and δQ, which precedes the timing of reward delivery by about 500 ms. We will later discuss implications of this assumption on timing for possible neural computations.

Behavioral data indicated that the lick rate in HIT trials were high early in time (average and standard deviation of lick latency, 0.25 ± 0.15 s, Fig 1B) and extended over the reward delivery period (liquid rewards were delivered three times at 0.41, 0.82, and 1.23 s after the first lick). By contrast, the lick rate in FA trials was also high early in time initially of learning (average and standard deviation of lick latency, 0.31 ± 0.23 s), but it gradually reduced to baseline because no reward was given for No-go cues. We fitted the Q-learning model to the licking behavior of individual mice (n = 17) in a total of 26,517 trials. The fitting performance was good for both the Go and No-go trials, which showed an increase in fraction correct and a decrease in fraction incorrect, respectively, at both the population (Fig 1C) and individual (Fig 1D) levels (average and standard deviation of coefficient of determination of 17 mice, 0.87 ± 0.15 for Go cue and 0.61 ± 0.18 for No-go cue, respectively, see Methods for details). The hyperparameters estimated for individual mice were broadly distributed (average and standard deviation; 0.002 ± 0.002 for α, 0.12 ± 0.07 for q_1, 0.24 ± 0.17 for q_2, 0.14 ± 0.07 for τ, and 0.84 ± 0.24 for ξ, Fig 1E), indicating that each animal utilized a distinct strategy for optimally learning to obtain the reward. As temporal evolution of the reward-related variables with discrimination learning, the Q values for both cues and lick (s=Go / No-go and a=Lick) were positive and intermediate initially, then increased for HIT trials (s=Go and a=Lick) and decreased toward negative values for FA trials (s=No-go and a=Lick) as learning progressed (Fig 1F). Note that the negative Q value for FA trials at the later stage of learning well reflected the negative R-P value assigned for those trials by the Q-learning model. As a consequence, the δQ values converge to zero for both cue-response conditions (Fig 1G). More specifically, the δQ values for HIT trials (s=Go and a=Lick) were positive initially and monotonically decreased during learning. By contrast, the initial δQ value for FA trials (s=No-go and a=Lick) was negatively large because of a large difference in negative reward R-P and the initial positive value of reward prediction q_2>0. Throughout the course of learning, this δQ value monotonically increased (decreased its magnitude) to zero, indicating a better agreement between negative R-P and Q values. For CR and MISS trials, both Q and δQ remained constant at zero.

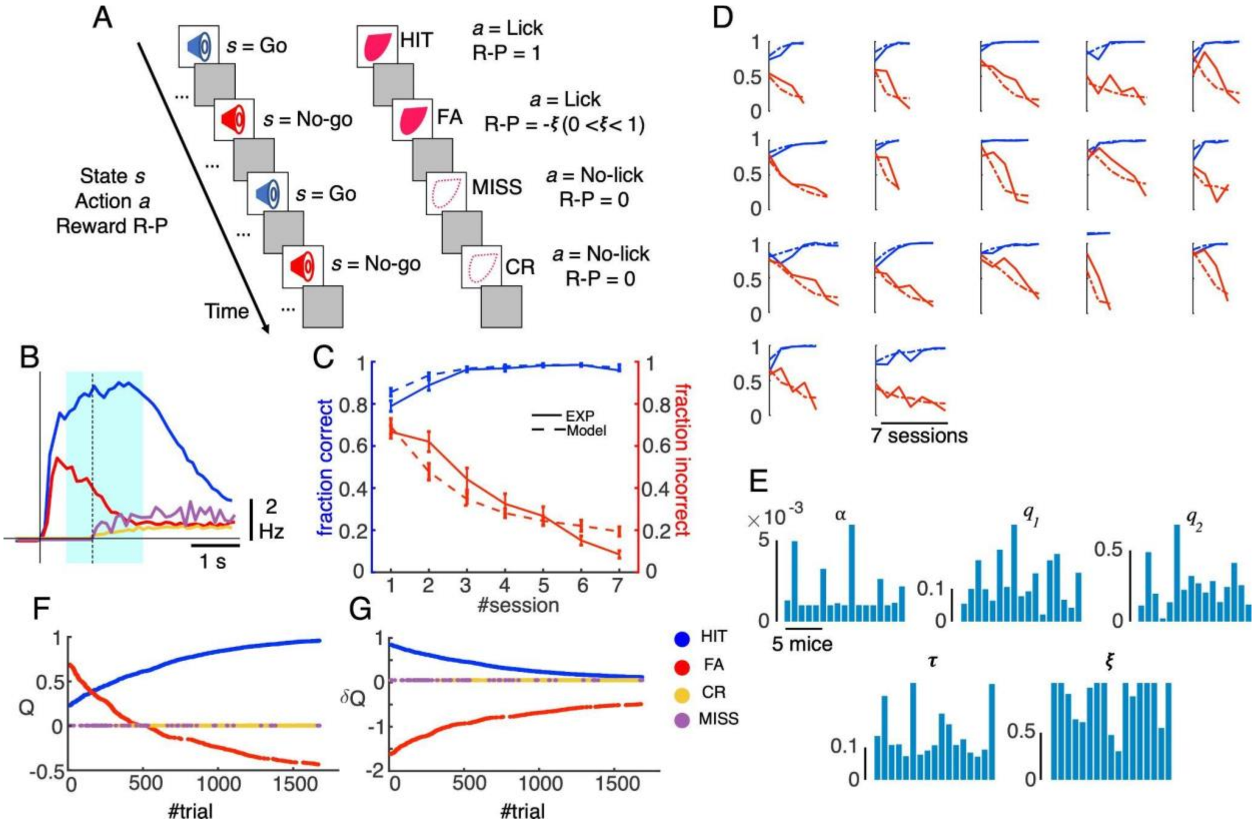

Figure 1: Q-learning model of Go/No-go experiment. A: schematic of the Q-learning model. B: the averaged lick rate for the four cue-response conditions (blue, red, orange and magenta for HIT, FA, CR and MISS trials, respectively: see also inset for color codes). Solid and dashed vertical lines indicate the cue onset and response window (1s after cue), respectively. Light cyan shadings represent the window for reward delivery (0.5 - 2 s after cue). C-D: fraction correct for Go cue (blue) and fraction incorrect for No-go cue (red) for experimental data (solid lines) and Q-learning model (dashed lines), averaged for 17 mice in 7 training sessions (C) and for individual mice (D). Vertical bars in C show standard errors. E: hyperparameter values of Q-learning model estimated for individual mice (1~17); and from left to right and top to bottom, learning rate α, initial Q values for Go and No-go cues, q_1 *and* q_2*, respectively,*

temperature τ, and punishment value for FA trials ξ. F-G: evolution of Q (F) and δQ (G) of a representative mouse for the four state-action combinations (HIT, FA, CR, MISS) during the course of learning.

Zonal climbing fiber responses and their correlations with reward prediction error

While mice learn Go/No-go discrimination tasks, we conducted two-photon recordings of climbing fiber responses (sampling rate, 7.8 Hz) from 6,445 Purkinje cells in eight cerebellar zones (from 7+ to 4b-, see Methods for details). A hyper-resolution algorithm (HA_time, Hoang et al. 2020) was applied to estimate the spike timings at the resolution of 100 Hz. Similar to the previous work (Hoang et al. 2022), we studied CF firing activity as population peri-stimulus time histograms (PSTHs) sampled in the three learning stages (from top to bottom, 1st, 2nd, and 3rd stages with fraction correct <0.6, 0.6 - 0.8, 0.8<, respectively, Fig 2A) for the four cue-response conditions, or the corresponding four state-action combinations. Briefly, CF responses in HIT trials (n = 3,788) were large and distributed across the entire medial hemisphere at the initial learning stage, but later on (2nd and 3rd stages), they became stronger and compartmentally focused on positive zones. By contrast, PSTHs in FA trials (n = 1,757) were distributed almost across the entire Crus II and gradually decreased and were more confined within lateral zones along with learning. PSTHs for CR trials (n = 2,229) were mainly distributed within zones 6- and 6+, and there was only spontaneous CS activity in MISS trials (n = 201).

Our previous study demonstrated that CF responses in the lateral hemisphere represent cognitive error signals related to No-go cues (Hoang et al., 2022). Therefore, in the present study, we analyzed trial-by-trial correlations between CF firings in zones 6-, 6+ and 5- with the reward prediction error δQ for FA trials. As expected, on a trial basis, we found that CF firings, defined as the mean firing rate in 0 - 0.5 s after cue onset, in these three zones were negatively correlated with δQ, with the strongest correlation found for zone 5- (slope = -0.53, p < 0.01 for 6-, slope = -0.60, p<0.001 for 6+ and slope = -0.76, p < 0.0001 for 5-, Fig 2B-D). The temporal window of 0 - 0.5 s for calculation of the mean firing rate was selected because

CF firings in FA trials of these three zones were mostly confined within 0 - 0.5 s after the cue onset (Fig 2A).

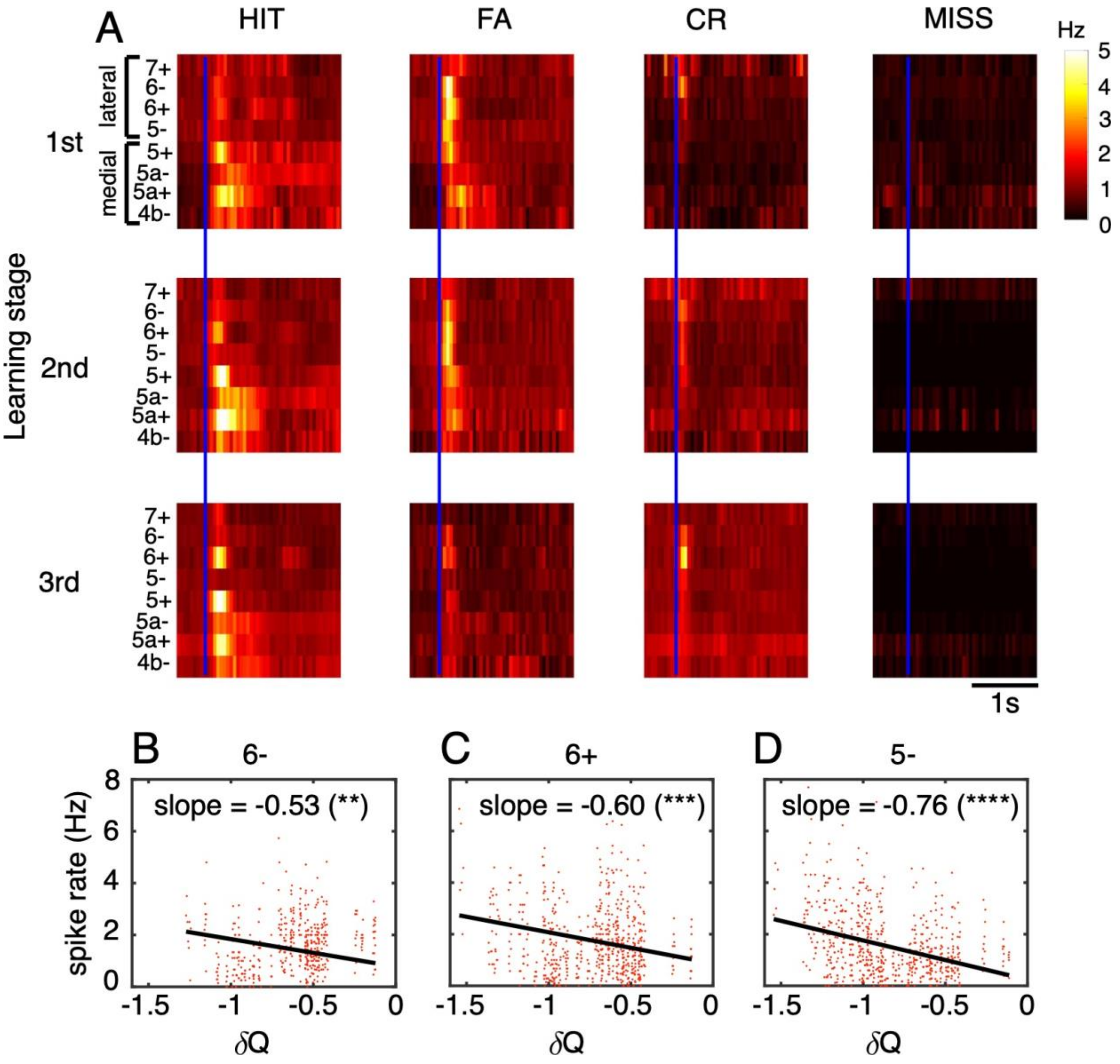

Figure 2: Climbing fiber responses to cues and their correlation with reward prediction error. A: Heat-maps showed PSTHs of climbing fiber firings in 8 Ald-C zones (7+ to 4b-) in the four cue-response conditions (columns) and three learning stages (rows, 1st, 2nd, and 3rd stages with fraction correct <0.6, 0.6 - 0.8, 0.8<, respectively). The blue line indicates cue onset. B-D: correlation of firing activity of Ald-C

zones 6- (B), 6+ (C) and 5- (D) with δQ in the FA trials. Each dot represents each trial of 17 mice. The Figure 2A was reproduced from (Hoang et al. 2022).

Partial Least-Squares regression of zonal activity with reinforcement-learning and sensorimotor-control explanatory variables

The previous analysis was constrained to only three zones and δQ. Here, we systematically studied the correlations between the firing activity of neurons in all eight cerebellar zones with reward as well as sensorimotor variables by partial least-squares regression (PLS regression) to find meaningful correlations under the situation of multicollinearity between explanatory variables (range of correlations between explanatory variables, 0.12 ± 0.23 with the highest correlations found for R vs. Q and R vs. δQ as 0.82 and 0.78, respectively). For this purpose, the neuronal activity in each trial was defined as the mean firing rate in [-0.5, 2] s after cue onset of the neurons in the same Ald-C zone. For explanatory variables, we included R, Q, δQ as reward variables, and lick count in the three response windows ([0, 0.5] s for early lick - ELick, [0.5, 2] s for reward lick - RLick and [2, 4] s for late lick - LLick) for Go and No-go cues, separately, as sensorimotor variables. Here, we prepare 6 sensorimotor variables (2 × 3=6; Go vs No-go multiplied by three response windows of licks) as shown in the inset of Fig. 3. Note that these three response windows well correspond with licking behavior of mice as well as reward delivery period (Fig 1B). Because these sensorimotor variables span large temporal windows of 0 - 4 s after the cue, the mean firing rate was computed for the comparable large temporal window of [-0.5, 2] s after the cue onset. We further note that the physical reward R (R=1 for HIT trials, R=0 otherwise), which was used as one of explanatory variables of PLS regression, is different from the reward-punishment R-P used in the Q-learning model (see Methods for details).

We calculated the variable importance in projection scores (VIP scores) to quantitatively estimate the importance of a given explanatory variable to the PLS regression model (see Methods for details). We found that neurons in the lateral hemisphere were associated with Q and δQ, but not the reward R per se (Fig 3). More specifically, 6- and 6+ were strongly associated with Q (VIP score, 1.8 and 1.7 for 6- and 6+, respectively), while 6- and 5- was strongly associated with δQ (VIP score = 1.6 and 2.1 for 6- and 5-, respectively). We note that the VIP score does not provide the sign (positive or negative) of correlations, but the explanatory variables, whose VIP scores are larger than 1, are generally considered meaningfully important in PLS regression. Note that these results are consistent with the results of simple correlation analyses shown in Fig. 2B-D. By contrast, neurons in the medial hemisphere (5+ to 4b-) were strongly associated with the reward R (VIP score > 1.2). Notably, we also found a complicated association between zonal activity and sensorimotor variables (i.e., licking following two cues). That is, medial zones were associated with the lick number in the reward period (Go × RLick and No-go × RLick, VIP score > 1) while those in the lateral hemisphere (7+ and 5-) were associated with the lick number in the reward period following No-go cues (No-go × RLick, VIP score = 1.1 for both 7+ and 5-, respectively). Only zone 7+ was associated with late lick following Go cues (Go × LLick, VIP score = 1.1).

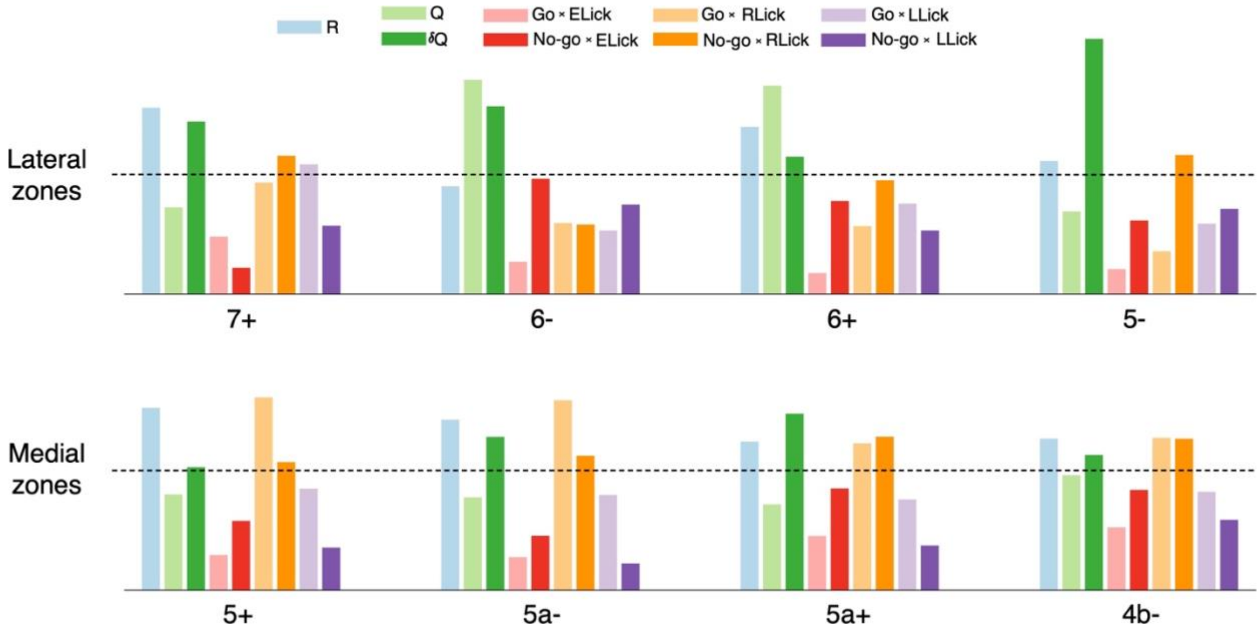

Figure 3: Partial least squares regression of spiking activity and reinforcement-learning and sensorimotor-control variables. Bars showed the variable-importance-in-prediction (VIP) scores of 9 reinforcement-

learning and sensorimotor-control variables (from left to right, R, Q, δQ, Go × ELick, No-go × ELick, Go × RLick, No-go × RLick, Go × LLick, No-go × LLick) for spiking activity of neurons in 8 Ald-C zones. Dashed lines indicated VIP score = 1, *which is considered a threshold of importance. See the inset for color codes of the 9 explanatory variables.*

The generative model of spiking activity at a trial basis by tensor component analysis

The PLS regression shown in Fig 3 suggested a functional organization of CF responses, moderately constrained by the zonal structure, with respect to sensorimotor and reward processing. In the previous study, we conducted the tensor component analysis (TCA) for CF PSTHs of >6,000 PCs and found four well separated components (TC1-4) that explained more than 50% of variance of PSTHs (Fig 4A, see Hoang et al. 2022 for details). One of the reasons why relatively moderate anatomical distributions of different functions are found in Fig. 3 is that each zone and even each neuron contain multiple functional components as demonstrated in Hoang et al. (2022) as well as in previous studies demonstrating multiplexed representations (Markanday et al. 2021; Ikezoe et al. 2022). In order to overcome this difficulty due to multiplexing for revealing precise functions of each component, we next examine functional representations of each tensor component utilizing Q-learning and trial-based analyses, while incorporating timing information of each spike instead of broadly computing the average spike rate over a wide temporal window.

In this study, we used TCs as a generative model of spiking activity at trial basis for elucidating the associations of functionally organized CF responses and reinforcement-learning and sensorimotor-control variables. Briefly, the TC score of a neuron in a particular trial was estimated by filtering the spike train of that neuron by the temporal profile of the corresponding TC, weighted by neuronal and cue-response condition coefficients (Fig 4B, see Methods for details). This computation incorporated trial-to-trial variability of spiking activity while maintaining fundamental properties of TCs. For example, the neurons, whose TC1 and TC2 coefficients were highest among all the neurons recorded, had high TC1 and TC2 activities in HIT and FA trials, respectively (Fig 4C-D). Note that TC1 and TC2 were selectively activated in HIT and FA trials, respectively (Fig 4A). Following such a computation, the averaged TC activity of all neurons in the same recording session shared a similar structure of zonal distribution and cue-response condition with those of the TCs (compare Fig 4A and Fig 4E, but note that abscissae are learning stages and cue-response conditions, respectively). Specifically, TC1 scores were high in HIT trials and for positive zones. By contrast, TC2 scores were high in FA trials and distributed in the lateral hemisphere. TC3 scores were high in HIT trials and distributed in the medial hemisphere. TC4 scores had similar zonal distribution with TC2 scores, except that they were non-zero only for CR trials.

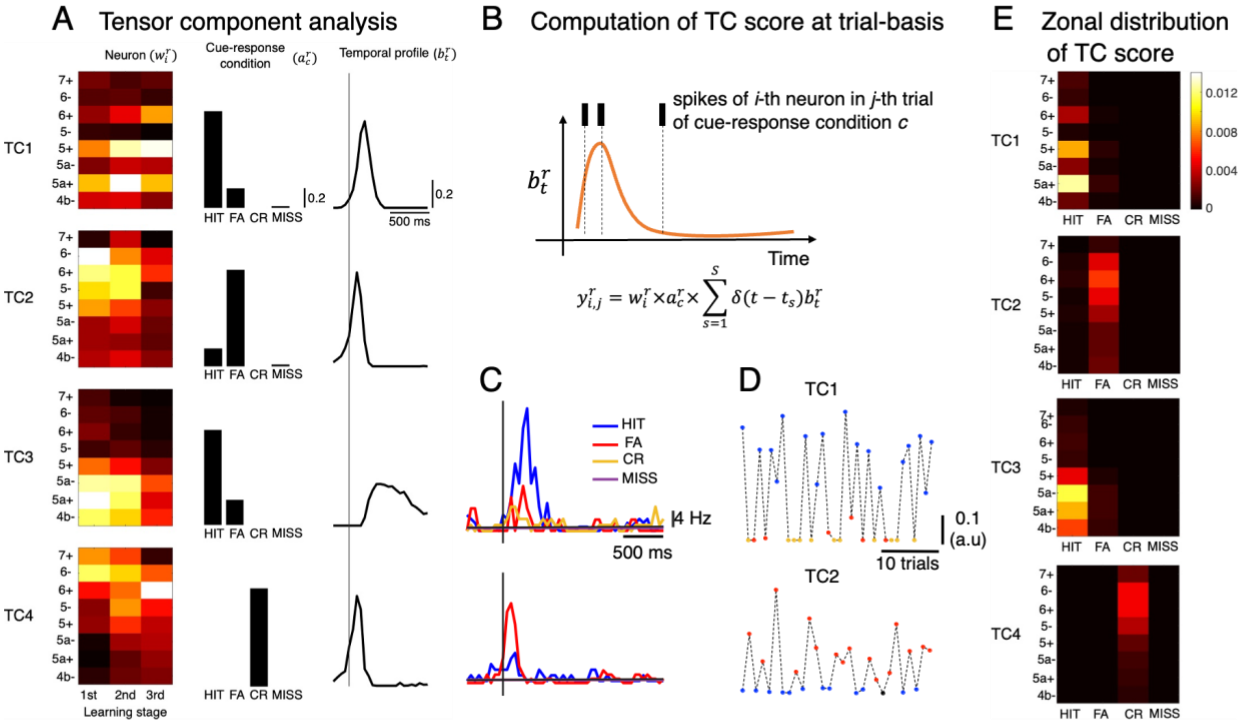

Figure 4: Tensor-component analysis (TCA) and computation of tensor score at a trial-basis. A: TCA was conducted for PSTHs in 4 cue-response conditions of n=6,775 neurons and the resulting four tensor components (TC1-4) explained more than 50% of variance. B: for the i-th single neuron, its activity for the r-th TC (y^r) in the particular j-th trial was computed by filtering spike timings with temporal profile of the r-th TC b_t^r, multiplying corresponding coefficients w_i^r of the i-th neuron and a_c^r of the cue-response

condition c. C-D: PSTHs (C) of two representative neurons, which have the highest coefficients of TC1 and TC2, respectively, and their TC1 and TC2 scores, respectively, computed for all trials in their corresponding sessions (D). E: Heatmaps showed TC1-4 scores averaged for all neurons in each of the eight zones distinctively for the four cue-response conditions

Sparse canonical correlation analysis

To find the most contributing variables for each TC score, we conducted the sparse canonical correlation analysis (sCCA) of TC scores and the same 9 reinforcement-learning and sensorimotor-control variables used in PLS regression (see Methods for details). As a result, TC1 and TC3 were associated with only reward variables while TC2 and TC4 were associated with both reward and sensorimotor variables in No-go trials (Fig 5A). More specifically, TC1 and TC3 were positively correlated with reward R, reward prediction Q and also reward prediction error δQ, with high coefficients of R and Q for TC1 (0.66 and 0.62) and R for TC3 (0.78). We can safely state that TC1 is mainly related to reward and its prediction, and that TC3 is mainly related to reward. Remarkably, TC2 was negatively correlated with δQ (coefficient, -0.83) but it was positively correlated with the early lick count in No-go trials (0.42). Similarly, TC4 was negatively correlated with both R (coefficient, -0.65) and early lick count in No-go trials (-0.59). We note that those associations revealed by sCCA can also be found by PLS regression (Fig S1), thus they are not dependent on specific analysis methods. We further confirmed significant correlation for each of TC components by linear regression of TC1 vs. Q (slope = 0.46, $p < 0.0001$, Fig 5B), TC2 vs. δQ (slope = -0.42, $p < 0.0001$, Fig 5C), TC3 vs. R (slope = 0.30, $p < 0.0001$, Fig 5D) and TC4 vs. No-go × ELick (slope = -0.24, $p < 0.0001$, Fig 5E). Notably, these correlations were also significant ($p < 0.0001$) with comparable slopes even when using only the trials of the cue-response condition with which each of TCs is mostly associated (slope = 0.32 for TC1-HIT, slope = -1.24 for TC2-FA, and slope = -0.36 for TC4-CR trials, Figs 5BCE).

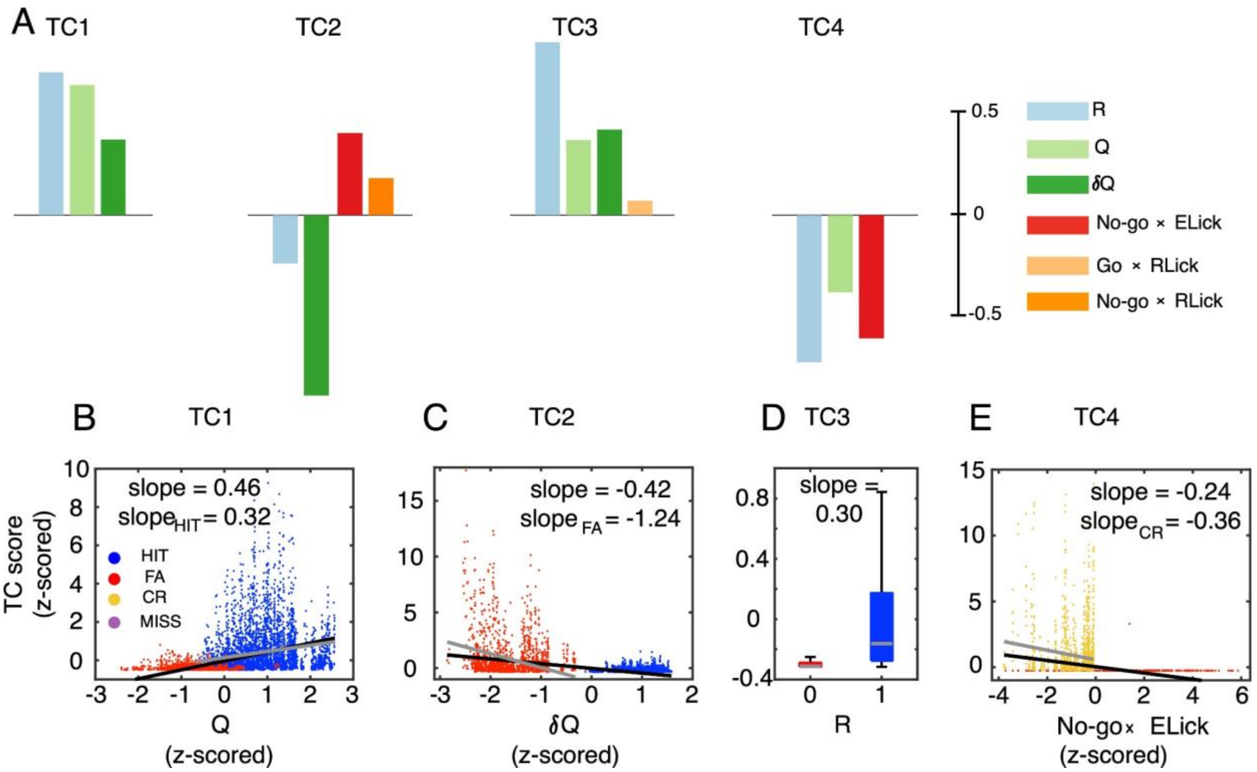

Figure 5: Sparse canonical-correlation analysis (sparse CCA) of TC scores and reinforcement-learning and sensorimotor-control variables. A: Bars showed the coefficients of reinforcement-learning and sensorimotor-control variables corresponding to TC1-4 scores. B-E: the scatter plots of trials showed correlations of TC1 with Q (B), TC2 with δQ (C), TC3 with R (D) and TC4 with No-go × ELick (E) at trial-basis. Black and gray lines indicated regression between variables when using all trials and the trials of the cue-response condition with which each of TCs is mostly associated (i.e., TC1-HIT, TC2-FA and TC4-CR), respectively. Panel D showed the boxplot with gray lines indicating the median and the bottom and top edges of the box the 25th and 75th percentiles, respectively. Note that the correlations in B-E are all significant (p<0.0001). Color convention of trials is the same as Figure 1. The inset of A shows color codes of the selected 6 reward and sensorimotor variables among 9 according to sCCA.

Discussions

In the present study, we conducted three different correlation analyses for elucidating contributions of CF inputs to behavior during learning of Go/No-go auditory discrimination tasks. In the first analysis, we showed associations between lateral zones and reward prediction error, which were suggested by our previous work demonstrating that the CF responses in zones 6-, 6+ and 5- convey cognitive error signals (Hoang et al. 2022). Since Tsutsumi et al. (2019) suggested distinct contributions of zonal-organized CF inputs to behavioral variables, in the second analysis, we systematically study the correlations between the firing rate of neurons in all eight cerebellar zones with 9 reinforcement-learning and sensorimotor-control variables by partial least-squares regression (PLS regression). PLS was necessary because reward and sensorimotor variables are considerably correlated with each other and we need to find meaningful correlations under this difficult situation of multicollinearity. PLS revealed relatively moderate zonal distributions of different variables, probably because each zone and even each neuron contain multiple functional components as demonstrated by previous studies (Markanday et al., 2021; Ikezoe et al., 2022; Hoang et al., 2022). Therefore, in the third analysis, we decomposed spiking activity into four tensor-components (TCs) by tensor component analysis (TCA) and then examined functional representations of each TC at a trial basis. First, these three analyses utilized distinct CF firings (mean firing rate in [0, 0.5] s and [-0.5, 2] s for the first two analyses, respectively, and precise spike timings in [-0.5, 2] s for the third one). Second, they started from selective (three lateral zones) and proceeded to more comprehensive (eight zones and functionally-organized CFs). Third, they started from a simple method (linear regression) and proceeded to sophisticated (PLS and sparse CCA) methods. However, the results were very consistent. We found that TC1 neurons, distributed most densely in 6+, were positively correlated with reward R and reward prediction Q (Figs 3 and 5B). By contrast, TC2 neurons in zones 6- and 5- were negatively correlated with reward prediction error δQ (Figs 2BD, 3 and 5C). TC3 neurons in medial zones were positively correlated with reward R and licking in reward delivery period (Figs 3 and 5D). Finally, TC4 neurons, distributed most densely in 6-, were negatively correlated with reward R and early lick following the No-go cues (Figs 3 and 5E). Note that we did not include the reward-penalty R-P into regression analyses due to extremely high correlations of this variable with other reward-related ones (e.g., correlations between R-P vs. δQ and R-P vs. R were 0.94 and 0.91, respectively, because we have the equation δQ= R-P - Q). Still, the aforementioned findings remained unchanged even when R-P was included (Fig S2).

Notably, these results were also in good agreement with our recent work. In Hoang et al. (2022), we found that TC1-4 corresponds to timing control of the first lick, cognitive error signals, reward-related signals and action inhibition, respectively. Here, we showed that TC2 were negatively correlated with reward prediction error δQ and thus cognitive error signals can be computed as sign-reversed reward prediction errors by reinforcement learning algorithms. That is, during learning, cognitive error signals decrease due to an increase (with the same magnitude) of negative reward prediction error. Similarly, in the previous study, we showed that TC4 neurons inhibit unwarranted licks specific to the No-go cue, which was further confirmed by sparse CCA showing negative correlation of TC4 with No-go × ELick. The negative correlations of TC4 with reward R and reward prediction Q can be explained that R and Q for No-go cues were less than the average level and decreased during learning, while TC4 activity was increased (Fig S3). For TC3, its correlations with reward R and reward lick were supported by both of our studies. The only difference between the two studies is that Hoang et al. (2022) showed that synchronized spikes of TC1 neurons within [0, 0.5] sec were positively correlated with precise timing of the first lick in HIT trials, but such a correlation cannot be found in this study. There are two possible reasons for this inconsistency. First, in this study, we investigated the correlations in all cue-response conditions and not specific for HIT trials, while our previous study examined only HIT trials for TC1. Second, we did not utilize synchronization of CF responses in this study, which has been shown to be important for timing control (Welsh et al., 1995; Tsutsumi et al., 2020; Wagner et al., 2021), while our previous study examined only synchronized spikes. Because R and Q for Go cues were higher than the average level and increased during learning, positive correlations of TC1 with these two reward-related variables are expected due to an increase of TC1 activity (Fig S3).

We also attempted to elucidate the involvement of CF response in cue cognition and motor-related functions separately by introducing the auditory cue variable (Cue=1 for Go and Cue=0 for No-go cues) independent of motor-related variables (Fig S4). While the results of PLS regression for zonal activity were similar (Fig S4A), some notable insights were revealed by sparse CCA of TC activities (Fig S4B). First, negative correlation of TC4 with early lick following No-go cues was significantly weakened, and instead it showed a large negative correlation with auditory cue. This suggests that TC4 neurons conveyed motor commands for suppression of early licks for No-go cue, but they were not general motor commands irrespective of auditory cues, but specific motor commands only to No-go cue. Consequently, TC2 and TC4 neurons may altogether construct an actor of the reinforcement learning algorithm for the No-go cue, but not an actor for both Go and No-go cues. Second, while TC1 and TC3 were both positively correlated with reward R, their low and high positive correlations with the auditory cue, respectively, suggested that TC1 neurons were related only to reward and TC3 neurons may be related to motor control and sensory feedback of reward licks as shown in Hoang et al. (2022). Finally, strong and negative correlation of TC2 with reward prediction error δQ was retained. One important question to be answered in the future works is how TC2 neurons compute reward prediction error when TC2 activity was within 0.2 seconds (Fig S3) but rewards were only delivered between 0.5 and 2 seconds after the cue onset? By an observation that TC2 activity started increasing before the cue onset in FA trials (Fig. S3), we postulated that internal forward models (Wolpert et al., 1998; Kawato 1999), which may consist of a loop network of cerebral cortex, basal ganglia and the cerebellum (Bostan and Strick, 2018; Wagner and Luo, 2020; Watabe-Uchida et al., 2017; Kostadinov and Häusser, 2022), could be employed to compute the reward prediction error before actual rewards arrive.

CFs have been shown to be associated not only with sensorimotor variables but also a wide range of reward contingencies, suggesting that the cerebellum is involved in reinforcement learning of various tasks

(Wagner and Luo, 2020; Kostadinov and Häusser, 2022). However, previous studies could not indicate the algorithms (e.g. how the reward variables are computed) implemented in reinforcement learning. More importantly, it remains unclear how these reward-related CF inputs contribute to acquisition of behavior over the course of learning. To tackle these issues, we formally utilized the Q-learning algorithm to estimate reward variables from licking behavior of mice that underwent Go/No-go tasks for the following reason. From a computational view-point, we have a clear dichotomy about value functions, that is, expectation of summation of discounted future rewards as follows. The state value function V(s) is dependent only on the state s, while the state-action value function Q(s,a) is dependent both on the state s and action a. The two-photon Ca^{2+} imaging data clearly showed state-action value function rather than state value function because the imaging data is very different between different actions (lick) for HIT vs MISS or FA vs CR (Fig 1D). Thus we utilized a state-action Q value function to estimate the expected reward starting from a particular state and taking a particular action. For action policy, we used the softmax function for selecting actions based on the learned Q values. This simple Q-learning model with only five hyper-parameters fitted the licking behavior of individual mice very well (Figs 1CD), statistically better than the models with a reduced number of hyper-parameters (Fig S5) indicating that all these five hyper-parameters are important, at least for the present Go/No-go learning task. Interestingly, the estimated hyper-parameters revealed that each animal learns the task in a unique way (i.e., difference in learning rate, penalty value and initial Q values across animals, Fig 1E) while utilizing its own strategy (temperature) for optimally getting the rewards.

One of the technical advances of our study is to utilize tensor component analysis (TCA) as a generative model of spiking activity in single trials, under a challenging condition that single CFs multiplex various information (Markanday et al. 2021; Ikezoe et al., 2022; Hoang et al., 2022). Unsupervised statistical methods like TCA are effective in decomposing spiking activity into biologically-meaningful components, but they are sensitive to noise and thus not suitable for single trials. Therefore, we applied TCA for PSTH of CF firings to diminish the trial-by-trial variability as well as to find the low-dimensional tensor components (TCs) underlying spiking dynamics. The temporal profiles of TCs were then used for filtering spike activity in each trial (Fig 4B). This process resulted in TC scores that reliably captures spiking dynamics with respect to the four TCs found by TCA (Figs 4CDE).

Theoretically, the most exciting finding of this study is that TC2 was strongly and negatively correlated with reward prediction error δQ, and at the same time, it was positively correlated with early lick following No-go cues (Fig 5A). Note that TC2 firings in the FA condition decreased dramatically as learning proceeds (Fig S3). This result supports the integration of internal model theory and reinforcement learning in the cerebellum. According to this hypothesis, TC2 neurons acquire necessary motor commands from reward prediction error conveyed by climbing fibers as follows. In the beginning of learning, negative reward prediction error is large in amplitude, thus CFs of TC2 neurons are strongly activated. If we assume that SSs of those neurons are also activated in 0-500 ms with a positive baseline firing rate of 50-100 Hz, many erroneous (unwarranted) licks are generated within this temporal window. Following co-activation of parallel fiber and climbing fiber inputs, long-term depression (LTD) at the parallel-fiber-to-Purkinje-cell synapses occurs, and consequently, the SS modulation turns negative and becomes an approximate mirror image of CS modulation (similar phenomena were simulated by Yamamoto et al., 2002 in connection to adaptive control of ocular following responses). At later stages of learning, negative reward prediction error is close to zero, CFs are only spontaneous, thus SSs of these neurons contribute to suppression of the lick following No-go cues. According to this mechanism, TC2 may turn into TC4 as also suggested in Hoang et al. (2022). Future simultaneous recordings of SSs and CSs may reveal neural mechanisms of these neurons. In conclusion, our paper proposed a computational framework - combining Q-learning model, hyper-resolution spike detection algorithm and tensor component analysis - for elucidating reward processing in the cerebellum and we found that the climbing fibers in the lateral Crus II encode negative reward prediction error in reinforcement learning algorithm.

Methods

Q-learning model

We adopted a Q-learning model, a reinforcement learning algorithm, to model the licking behavior of mice in the Go/No-go experiment. The algorithm works by learning a state-action value function, commonly referred to as the Q-function. The Q-function is a mapping from a state-action pair to a scalar value that represents the expected reward for taking a particular action in a particular state. The Q-function is updated over time based on the observed rewards and transitions to new states.

In the Go/No-go auditory discrimination task, the Q-function of a state *s* (*s* = Go / No-go cue) and an action *a* (*a* = Lick / No-lick) was updated at the trial *t* as follows

$\delta Q_t = \text{R-P}_t - Q_t$

$Q_{t+1} = Q_t + \alpha \times \delta Q_t$

where δQ is the reward prediction error, α is the learning rate and R-P is the reward-penalty function

R-P = 1 for *s* = Go, *a* = Lick (HIT)

R-P = $-\xi$ for *s* = No-go, *a* = Lick (FA)

R-P = 0, otherwise (CR and MISS)

The probability of selecting a given action *a* in state *s* is determined by the softmax function comparing the Q-function values of an action to all others:

$$Prob(s,a) = \frac{exp\,\{Q(s,a)/\tau\}}{\sum_a exp\,\{Q(s,a)/\tau\}}$$

where τ is the temperature parameter, representing the trade-off between exploitation and exploration. Because mice underwent pre-training sessions of 3 days, during which they were rewarded by licking after 1 second of both cues, initial Q values for Go and No-go cues should be positive: Q_0(*s*=Go, *a*=Lick) = q_1 and Q_0(*s*=No-go, *a*=Lick) = q_2. In total, the Q-learning model contains 5 hyper-parameters: learning rate α (range, $0.001 \le \alpha \le 0.1$), ξ ($0 \le \xi \le 1$), temperature τ ($0.01 \le \tau \le 0.5$), initial Q values for Go and No-go cues, q_1 and q_2 ($0 \leq q_1 \leq 1$, $0 \leq q_2 \leq 1$). We estimated these parameters for individual animals by maximizing the likelihood defined as the sum of Prob(s,a) for all trials.

It is important to emphasize that we fitted the Q-learning model to behavioral data at a trial basis with the trials from different sessions concatenated. To evaluate the fitting, for each session, we computed the fraction correct for Go cues (number of HIT trials / total number of Go trials) and the fraction incorrect for No-go cues (number of FA trials / total number of No-go trials) from the data, while the probability Prob(s,a) of the last HIT and FA trials were selected from the model (Figs 1CD). The coefficient of determination (R^2) was calculated to measure goodness-of-fit between the data and model probabilities across sessions of individual animals.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

$$SS_{res} = \sum_i (h_i - f_i)^2$$

$$SS_{tot} = \sum_i (h_i - \bar{h})^2$$

, where h_i and f_i are the probabilities of the data and the model at the i-th session, respectively, and $\bar{h}$ is the mean of the data probability across sessions.

Estimation of CF firing activity from two-photon recordings

Ca signals in Purkinje cell dendrites were evaluated for regions-of-interest (ROIs) of the two-photon images extracted by Suite2p software and manually selected. Spike trains were reconstructed for 6,445 Purkinje cells sampled in the 17 mice, using hyperacuity software (HA_time, Hoang et al. 2020) that detected spike activities for Ca signals of two-photon imaging with a temporal resolution of 100 Hz (see Hoang et al. 2022 for details).

CF response to cue stimulus

For evaluating the CF response of a single neuron to the cue stimulus, we constructed the peri-stimulus time histogram (PSTH) of CF firings in [-0.5, 2] s with a time bin of 50 ms for the four cue-response conditions. They include HIT, FA, CR, or MISS, according to licking behavior within a response period of 1 s to the two cues, i.e., correct lick in response to the go cue, unwarranted lick in response to the No-go cue, correct response rejection to the No-go cue, or response failure to the Go cue, respectively. Each PSTH was subtracted from its baseline activity, defined as the mean value of the firing rate in the range of [-2, -1] s before cue onset.

Partial least-squares regression analysis

We aimed at revealing correlations between zonal activity and behavior variables at a single trial basis (Fig 2). For each trial, the neuronal activity was calculated as the mean firing rate in [-0.5, 2] s after cue onset of the neurons in the same Ald-C zone. Behavior variables include physical reward R (R = 1 for HIT trials, = 0 otherwise), Q, δQ, number of licks in the early period of 0-0.5 s for Go and No-go cues (Go × ELick and No-go × ELick), number of licks in the reward period (0.5 - 2 s after cue, Go × RLick and No-go × RLick) and number of licks in the late period (2 - 4 s after cue, Go × LLick and No-go × LLick).

We conducted partial least-squares (PLS) regression to resolve the multi-collinearity of behavior variables, e.g. between R, Q and δQ. PLS regression searches for a set of low-dimensional components that performs a simultaneous decomposition of dependent and explanatory variables with the constraint that these components explain as much as possible the covariance between the variables. The VIP (Variable Importance in Projection) score is a measure of the importance of each predictor variable in a PLS regression model, with higher scores indicating greater importance. The VIP score for the *j*-th variable is given as:

$$VIP_j = \sqrt{\frac{\sum_{f=1}^{F} w_{jf}^2 \times SSY_f \times J}{SSY_{total} \times F}}$$

, where w_{jf} is the weight for *j*-th variable and *f*-th PLS component, SSY_f is the sum of squares of explained variance for the *f*-th PLS component and *J* number of explanatory variables (J=9). SSY_{total} is the total sum of squares explained for the dependent variable, and *F* is the total number of PLS components. In our analysis, the number of PLS components was optimized by 10-fold cross-validation.

Note that the VIP score does not provide sign information (positive or negative) of correlations between explanatory and dependent variables. Although there was no known threshold for a systematic evaluation, VIP score > 1 was typically used as an indicator for a variable to be considered significant. PLS was conducted using the MATLAB functions *plsregress*.

Tensor component analysis

Let x_{ntk} denote the PSTH of neuron *n* at time step *t* within cue-response condition *k*. TCA yields the decomposition

$$x_{ntk} \approx \hat{x}_{ntk} = \sum_{r=1}^{R} \lambda_r w_n^r b_t^r a_k^r$$

, where *R* is the number of tensor components, w_n^r, b_t^r and a_k^r are the coefficients of the neuron, temporal, and response condition factors, respectively. Those coefficients were scaled to be unit length with the rescaling value λ_r for each component *r*. We introduced a non-negative constraint of those coefficients ($w_n^r \geq 0$, $b_t^r \geq 0$ and $a_k^r \geq 0$ for all *r*, *n*, *t* and *k*). In the previous study, we optimized the number of components R = 4 for which the solutions were most stable and the fitting scores were high (Hoang et al., 2022).

Computation of tensor component score for a single trial

TCA was efficient for decomposing PSTHs into biologically-meaningful components (Hoang et al. 2022), but for systematic analysis of associations of CF responses and variables at a trial basis, it is crucial to estimate activity of such components for individual neurons in a single trial. Note that TCA was carried out for PSTHs computed over multiple trials within a session, but we need some firing index for each trial for Q-learning analysis. For that purpose, we proposed a novel approach to utilize the components decomposed by TCA as generative models of CF firings.

The tensor-based activity of the *i*-th neuron in the *j*-th trial (corresponding to the cue-response condition *c*) with respect to the component *r*-th (r = 1, .. , 4) was evaluated as:

$y_{i,j}^{r} = w_{i}^{r} \times a_{c}^{r} \times \sum_{s=1}^{S} \delta(t - t_{s}) b_{t}^{r}.$

, where $\sum_{s=1}^{S} \delta(t - t_{s})$ is the firings sampled from [-0.5, 2] s after cue onset as the summation of Dirac delta functions δ.

For the *j*-th trial, the TC score of the *r*-th component was calculated as the averaged $y_{i,j}^{r}$ across all neurons in the same recording session. As a result, each trial has four TC scores corresponding to the four TCs.

$$TC\ score_{j}^{r} = \frac{1}{N} \sum_{i=1}^{N} y_{i,j}^{r}$$

Sparse canonical correlation analysis

Because CFs may multiplex different information, we conducted sparse Canonical Correlation Analysis (sCCA) for analyzing the relationship between TC scores and behavior variables. The goal of sCCA is to find a set of linear combinations (known as "canonical variates") of the variables in each set, such that the correlation between the two sets of canonical variates is maximized. The sCCA includes a sparsity constraint, which promotes solutions in which only a small number of variables are considered in the calculation of the canonical variates. This can result in more interpretable and biologically relevant solutions, as it reduces the amount of noise in the data.

In our analysis, sCCA was conducted using the *PMA* package of R, with a LASSO penalty applied to enforce sparsity. L1 bounds were set 0.4 and 0.6 for TC scores and behavior variables, respectively (larger L1 bound corresponds to less penalization). The explanatory and dependent variables were standardized to have mean zero and standard deviation 1 before performing the analysis. The number of canonical variables was 4. The resulting coefficient of TC scores was either 1 or -1. For easier interpretations, coefficient vectors of behavior variables (reported in Fig 5A) were re-signed so that the coefficient of TC scores was all 1.

Statistics

All statistical analyses were performed using MATLAB software. Unless otherwise stated, data are presented as means ± SD. For evaluation of correlations between neuronal response and single behavior variables (Fig 2B-D and Fig 5B-E), we fitted a linear mixed-effects model with fixed effect for behavior variables and mouse index as random intercept. The analysis was conducted using MATLAB function *fitlme,* with the slope and its significance p-value were reported. Significance level: n.s, $p > 0.05$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; **** $p < 0.0001$.

Acknowledgments

H.H., K.K. and K.T. were supported by Grants-in-Aid for Scientific Research in Innovative Areas (17H06313) and for Transformative Research Areas (22H05160 to M.M., 22H05156 to M.Kawato, and 22H05161 to K.K.). H.H. and K.T. were partially supported by JST ERATO (JPMJER1801, "Brain-AI hybrid"). H.H., M.Kawato, and K.K. were partially supported by the Grant Number JP21dm0307002, JP21dm0307008, and JP19dm0207080, respectively, Japan Agency for Medical Research and

Development (AMED). M.Kawato was partially supported by Innovative Science and Technology Initiative for Security Grant Number JP004596, Acquisition, Technology & Logistics Agency (ATLA), Japan. M.Kano and K.K. was partially supported by Grants-in-Aid for Scientific Research (JP18H04012, JP20H05915, JP21H04785 to M.Kano and JP22H00460 to K.K.) from Japan Society for the Promotion of Science (JSPS).

References

Akiti K, Tsutsui-Kimura I, Xie Y, Mathis A, Markowitz JE, Anyoha R, Datta SR, Mathis MW, Uchida N, Watabe-Uchida M. Striatal dopamine explains novelty-induced behavioral dynamics and individual variability in threat prediction. *Neuron* **110**(22), 3789–3804 (2022).

Albus, J. S. A theory of cerebellar function. *Math Biosci* **10**, 25–61 (1971).

Amo R, Matias S, Yamanaka A, Tanaka KF, Uchida N, Watabe-Uchida M. A gradual temporal shift of dopamine responses mirrors the progression of temporal difference error in machine learning. *Nat Neurosci.* **25**(8),1082–1092 (2022).

Atkeson, C.G. et al. (2018). What Happened at the DARPA Robotics Challenge Finals. In: Spenko, M., Buerger, S., Iagnemma, K. (eds) The DARPA Robotics Challenge Finals: Humanoid Robots To The Rescue. Springer Tracts in Advanced Robotics, vol 121. Springer, Cham.

Barto A., Sutton R., Anderson C. Neuron-like elements that can solve difficult learning control problems *IEEE Trans Syst Man Cybern* **13**(5), 835–846 (1983).

Bayer H., Glimcher P. Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron* **47**(1), 129–141 (2005).

Bostan, A. C. & Strick, P. L. The basal ganglia and the cerebellum: nodes in an integrated network. *Nat Rev Neurosci* **19**, 338–350 (2018).

Carta I., Chen C., Schott A., Dorizan S., Khodakhah K. Cerebellar modulation of the reward circuitry and social behavior. *Science* **363**(6424), eaav0581 (2019).

Chabrol F., Blot A., Mrsic-Flogel T. Cerebellar contribution to preparatory activity in motor neocortex. *Neuron* **103**, 506–519 (2019).

Dabney W, Kurth-Nelson Z, Uchida N, Starkweather CK, Hassabis D, Munos R, Botvinick M. A distributional code for value in dopamine-based reinforcement learning. *Nature* **577**(7792), 671–675 (2020).

Doya K. Complementary roles of basal ganglia and cerebellum in learning and motor control. *Curr Opin Neurobiol.* **10**(6), 732–739 (2000).

Haruno M., Kawato M. Different neural correlates of reward expectation and reward expectation error in the putamen and caudate nucleus during stimulus-action-reward association learning. *J Neurophysiol.* **95**(2), 948–959 (2006).

Haruno M, Kuroda T, Doya K, Toyama K, Kimura M, Samejima K, Imamizu H, Kawato M. A neural correlate of reward-based behavioral learning in caudate nucleus: a functional magnetic resonance imaging study of a stochastic decision task. *J Neurosci.* **24**(7), 1660–1665 (2004).

Heffley, W. & Hull, C. Classical conditioning drives learned reward prediction signals in climbing fibers across the lateral cerebellum. *Elife* **8**, (2019).

Heffley, W. *et al.* Coordinated cerebellar climbing fiber activity signals learned sensorimotor predictions. *Nat Neurosci* **21**, 1431–1441 (2018).

Hirano T. Depression and potentiation of the synaptic transmission between a granule cell and a Purkinje cell in rat cerebellar culture. *Neurosci Lett* **119**, 141–144. (1990).

Hirano T. Long-term depression and other synaptic plasticity in the cerebellum. *Proc Jpn Acad Ser B Phys Biol Sci* **89**(5), 183–195 (2013).

Hoang, H. *et al.* Improved hyperacuity estimation of spike timing from calcium imaging. *Sci Rep* **10**, 17844 (2020).

Hoang, H., Tsutsumi, S., Matsuzaki, M., Kano, M., Kawato, M., Kitamura, K., Toyama, K. Dynamic organization of cerebellar climbing fiber response and synchrony in multiple functional modules reduces dimensions for reinforcement learning, bioRxiv, doi: https://doi.org/10.1101/2022.12.05.518634 (2022).

Ikezoe, K. *et al.* Cerebellar climbing fibers convey behavioral information of multiplex modalities and form functional modules. *bioRxiv* 2022.08.24.505210 (2022)

Ito M. Cerebellar long-term depression: characterization, signal transduction, and functional roles. *Physiol Rev* **81**(3), 1143–1195 (2001).

Ito, M. Neurophysiological aspects of the cerebellar motor control system. *Int J Neurol Neurother* **7**, 162–176 (1970).

Ito M., Kano M. Long-lasting depression of parallel fiber-Purkinje cell transmission induced by conjunctive stimulation of parallel fibers and climbing fibers in the cerebellar cortex. *Neurosci Lett* **33**(3), 253–258 (1982).

Kaplan A, Mizrahi-Kliger AD, Israel Z, Adler A, Bergman H. Dissociable roles of ventral pallidum neurons in the basal ganglia reinforcement learning network. *Nat Neurosci.* **23**(4), 556–564 (2020).

Kawato M. Internal models for motor control and trajectory planning. *Curr Opin Neurobiol* **9**(6), 718–727 (1999).

Kawato M., Furukawa K., Suzuki R. A hierarchical neural-network model for control and learning of voluntary movement. *Biol Cybern* **57**(3), 169–85 (1987).

Kawato M., Gomi H. A computational model of four regions of the cerebellum based on feedback-error learning. *Biol Cybern* **68**, 95–103 (1992).

Kawato, M., Ohmae, S., Hoang, H., Sanger, T. 50 Years Since the Marr, Ito, and Albus Models of the Cerebellum. *Neurosci.*, **462**, 151–174 (2021).

Kawato M, Samejima K. Efficient reinforcement learning: computational theories, neuroscience and robotics. *Curr Opin Neurobiol.* **17**(2), 205–212 (2007).

Kostadinov, D., Beau, M., Blanco-Pozo, M. & Häusser, M. Predictive and reactive reward signals conveyed by climbing fiber inputs to cerebellar Purkinje cells. *Nat Neurosci* **22**, 950–962 (2019).

Kostadinov, D. & Häusser, M. Reward signals in the cerebellum: Origins, targets, and functional implications. *Neuron* **110**, 1290–1303 (2022).

Larry, N., Yarkoni, M., Lixenberg, A., Joshua, M. Cerebellar climbing fibers encode expected reward size. *Elife* 8:e46870 (2019).

LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).

Linden D. The return of the spike: postsynaptic action potentials and the induction of LTP and LTD. *Neuron* **22**, 661–666 (1999).

Markanday, A., Inoue, J., Dicke, P. W. & Thier, P. Cerebellar complex spikes multiplex complementary behavioral information. *PLoS Biol* **19**, e3001400 (2021).

Marr, D. A Theory of Cerebellar Cortex. *J. Physiol* **202**, 437–470 (1969).

Medina JF, Lisberger SG. Links from complex spikes to local plasticity and motor learning in the cerebellum of awake-behaving monkeys. *Nat Neurosci.* **11**(10):1185–1192 (2008).

Medina JF, Nores WL, Ohyama T, Mauk MD. Mechanisms of cerebellar learning suggested by eyelid conditioning. *Curr Opin Neurobiol.* **10**(6), 717–724 (2000).

Mnih, V. et al. Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015).

Pessiglione M, Seymour B, Flandin G, Dolan RJ, Frith CD. Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature* **442**(7106), 1042–1045 (2006).

Raymond JL, Medina JF. Computational Principles of Supervised Learning in the Cerebellum. *Annu Rev Neurosci.* **41**, 233–253 (2018).

Schultz W., Dayan P., Montague P. A neural substrate of prediction and reward. *Science*. **275**(5306), 1593–1599 (1997).

Sendhilnathan, N., Ipata, A. & Goldberg, M. E. Mid-lateral cerebellar complex spikes encode multiple independent reward-related signals during reinforcement learning. *Nat Commun* **12**, 6475 (2021).

Sendhilnathan, N., Semework, M., Goldberg, M., Ipata, A. Neural Correlates of Reinforcement Learning in Mid-lateral Cerebellum. *Neuron* 106(1):188–198.e5 (2020).

Shidara M., Kawano K., Gomi H., Kawato M. Inverse-dynamics model eye movement control by Purkinje cells in the cerebellum. *Nature* **365**, 50–52 (1993).

Silver, D. et al. Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484–489 (2016).

Starkweather CK, Uchida N. Dopamine signals as temporal difference errors: recent advances. *Curr Opin Neurobiol.* **67**, 95–105 (2021).

Sugihara I, Shinoda Y. Molecular, topographic, and functional organization of the cerebellar cortex: a study with combined aldolase C and olivocerebellar labeling. *J Neurosci.* **24**(40), 8771–8785 (2004).

Sugihara I, Shinoda Y. Molecular, topographic, and functional organization of the cerebellar nuclei: analysis by three-dimensional mapping of the olivo-nuclear projection and aldolase C labeling. *J Neurosci.* **27**(36), 9696–9710 (2007).

Sutton, R. S. & Barto, A. G. Reinforcement Learning: An Introduction 2nd edn, *MIT Press* (2018).

Tsutsumi, S. *et al.* Modular organization of cerebellar climbing fiber inputs during goal-directed behavior. *Elife* **8**, (2019).

Tsutsumi, S. et al. Purkinje Cell Activity Determines the Timing of Sensory-Evoked Motor Initiation. *Cell Rep* **33**, 108537 (2020).

Wagner, M. J., Kim, T. H., Savall, J., Schnitzer, M. J. & Luo, L. Cerebellar granule cells encode the expectation of reward. *Nature* **544**, 96–100 (2017).

Wagner, M. J. & Luo, L. Neocortex-Cerebellum Circuits for Cognitive Processing. *Trends Neurosci* **43**, 42–54 (2020).

Wagner, M. J. *et al.* Shared Cortex-Cerebellum Dynamics in the Execution and Learning of a Motor Task. *Cell* **177**, 669–682.e24 (2019).

Wagner, M. J. et al. A neural circuit state change underlying skilled movements. *Cell* **184**, 3731–3747.e21 (2021).

Watabe-Uchida, M., Eshel, N. & Uchida, N. Neural Circuitry of Reward Prediction Error. *Annu Rev Neurosci* **40**, 373–394 (2017).

Watkins, C., Dayan, P. Q-learning. *Mach. Learn.* **8**, 279–292 (1992).

Welsh, J. P., Lang, E. J., Suglhara, I. & Llinás, R. Dynamic organization of motor control within the olivocerebellar system. *Nature* **374**, 453–457 (1995).

Williams, A. H. *et al.* Unsupervised Discovery of Demixed, Low-Dimensional Neural Dynamics across Multiple Timescales through Tensor Component Analysis. *Neuron* **98**, 1099–1115.e8 (2018).

Wolpert D., Miall C., Kawato M. Internal models in the cerebellum. *Trends in Cognitive Sciences* **2**, 338–347 (1998).

Wurman, P. R. et al. Outracing champion Gran Turismo drivers with deep reinforcement learning. *Nature* **602**, 223–228 (2022).

Supplemental Figures

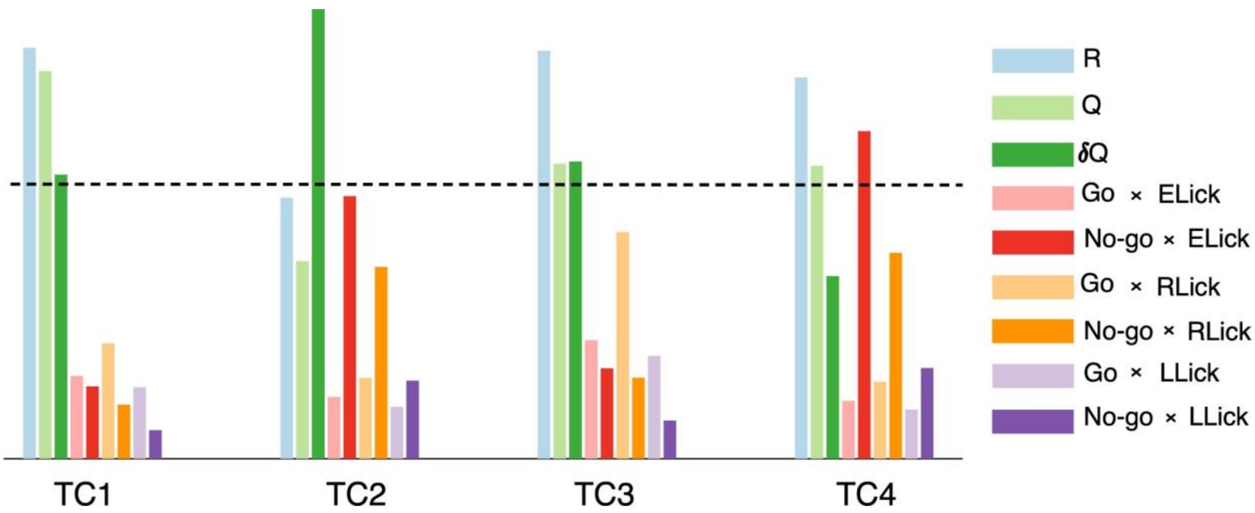

Figure S1: PLSR of TC activity and behavior variables. Color convention is the same as Figure 3.

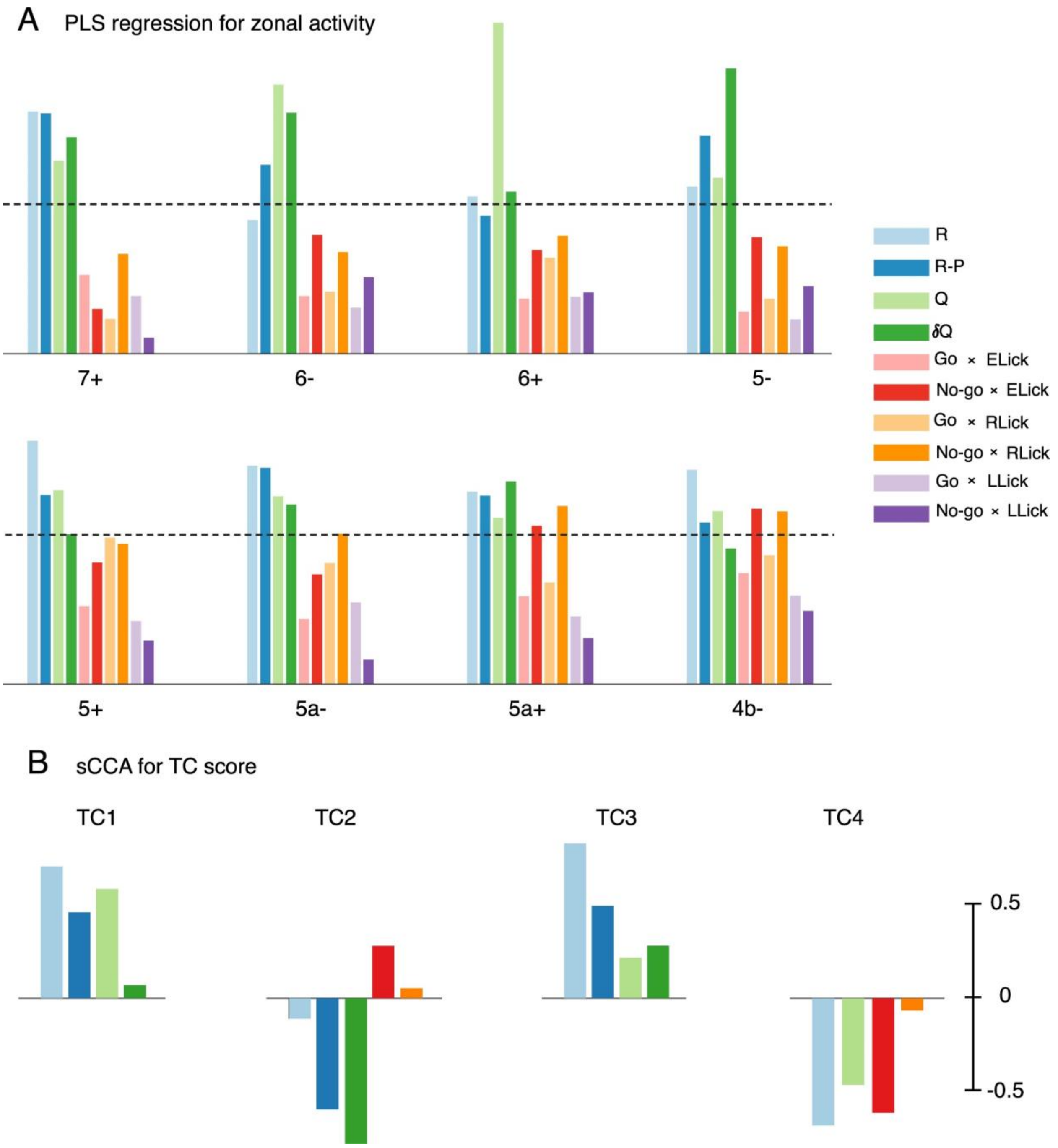

Figure S2: PLS regression for zonal activity (A) and sparse CCA for TC1-4 scores (B) with an additional reward-penalty R-P variable. Color convention is the same as Figure 3 with a new dark blue column for R-P.

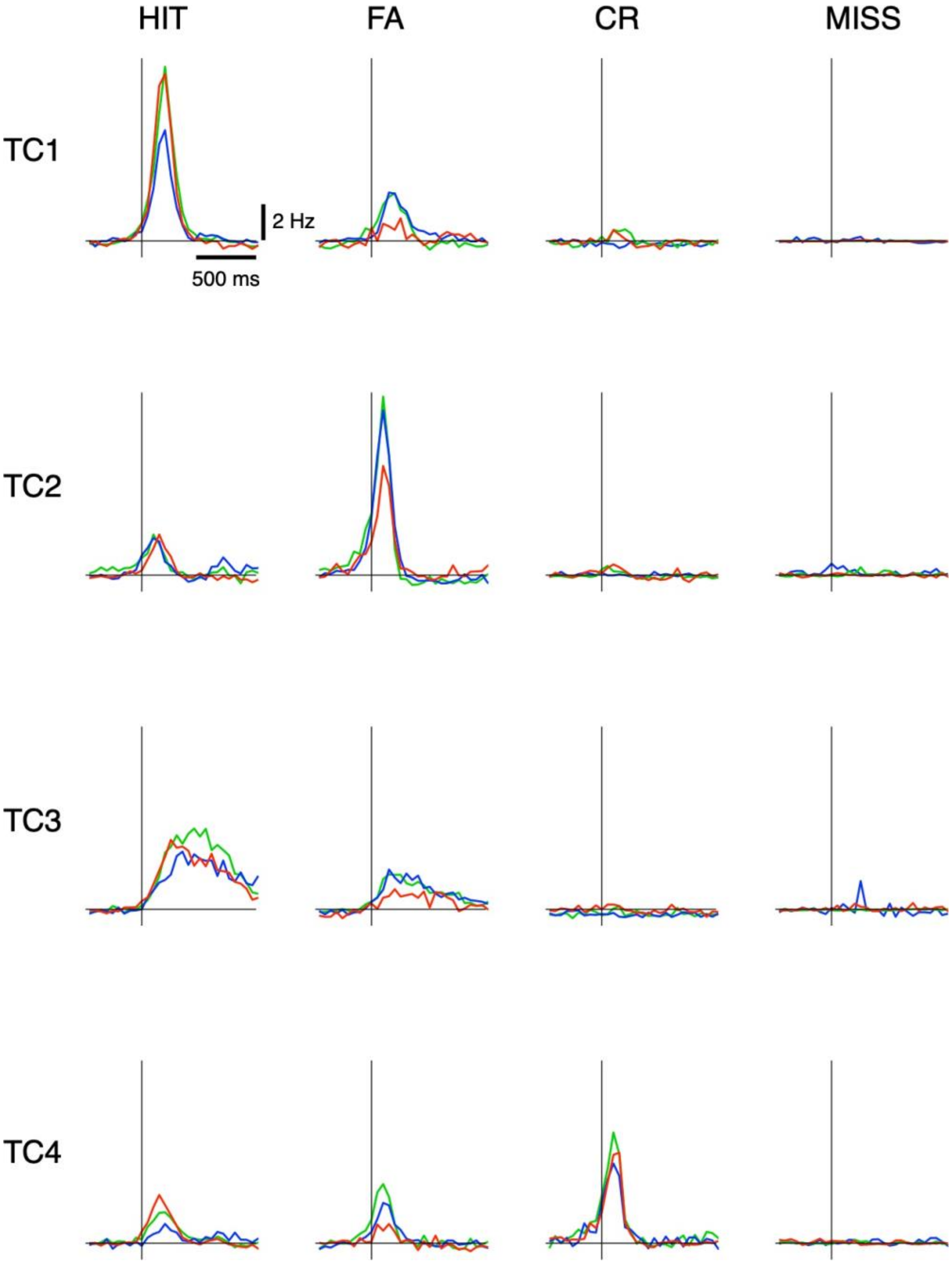

Figure S3: At each learning stage, we first sampled 300 neurons whose coefficients were highest for each of the tensor components, TC1-4, then excluded those that appeared in at least two samples (e.g TC1 and TC3). This figure plots PSTHs in the four cue-response conditions of selected neurons for the four tensor components. Blue, green and red traces are for 1st, 2nd and 3rd learning stages, respectively.

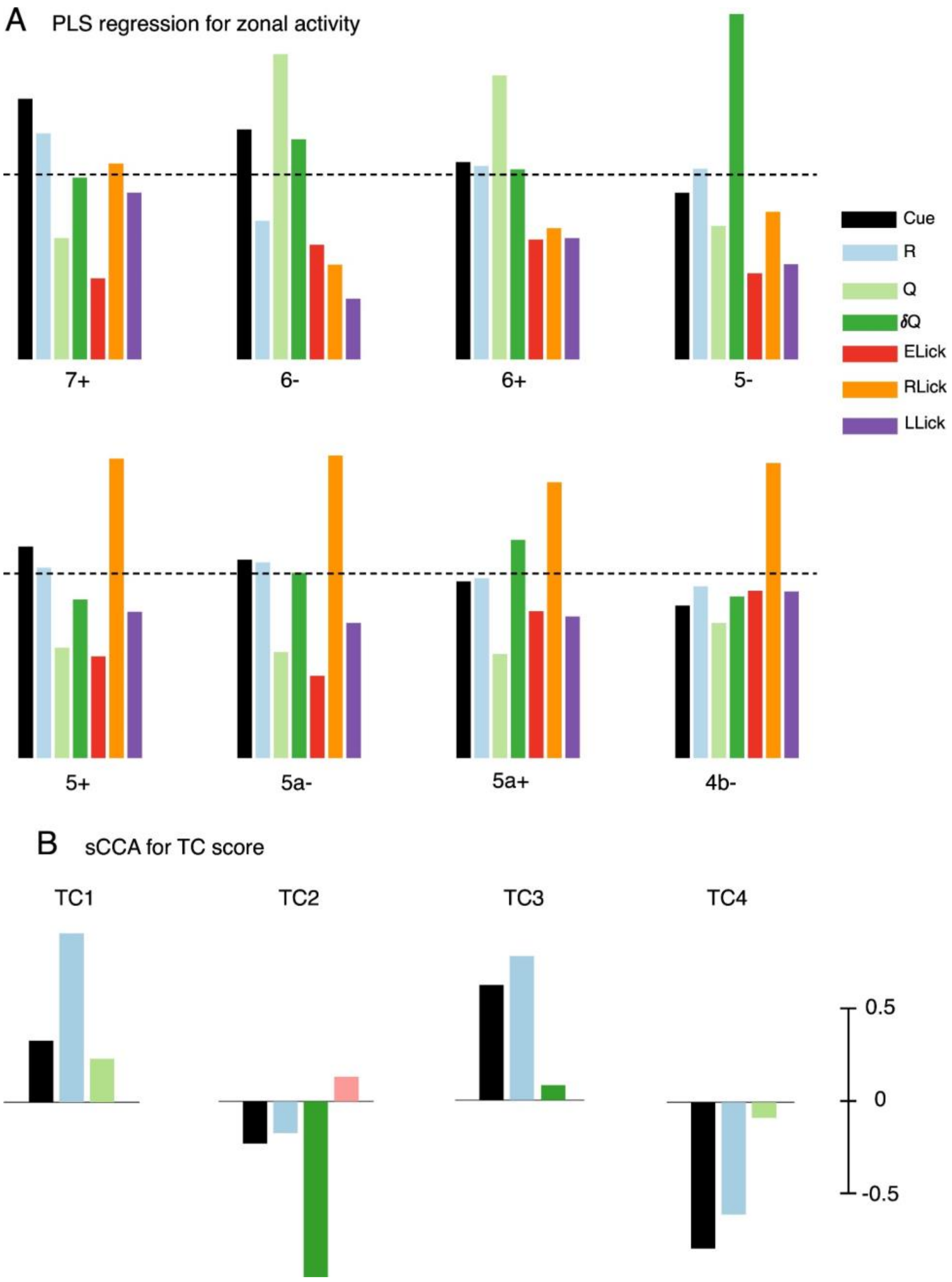

Figure S4: PLS regression for zonal activity (A) and sparse CCA for TC1-4 scores (B) with auditory cue and motor variables independent. In these analyses, the auditory cue (Cue=1 for Go and Cue=0 for No-go cues) was made independent from the motor (licking) variables, resulting in a total of 7 exploratory variables (Cue, R, Q, δQ, ELick, RLick, and LLick). Color convention is the same as Figure 3 with a new black column corresponding to the auditory cue.

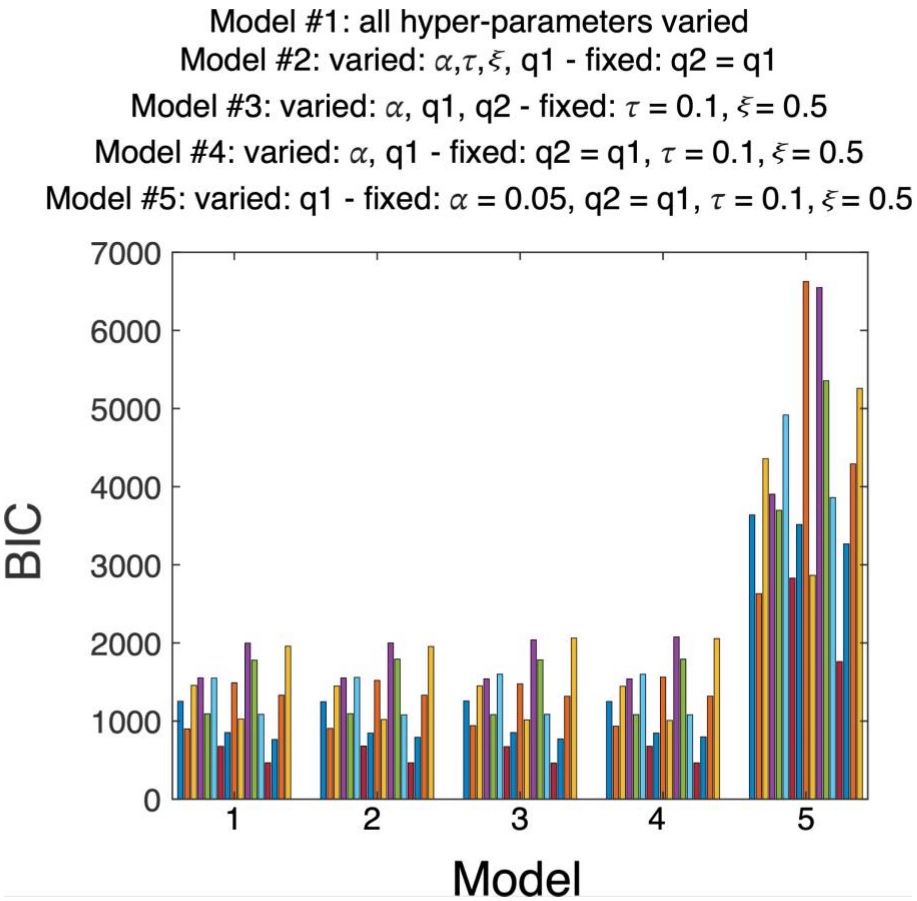

Figure S5: BIC score, estimated for 17 mice, of the five Q-learning models with different numbers of hyper-parameters. The total BIC score was 21,235; 21,291; 21,414; 21,532 and 69,325 for Model #1-5, respectively.