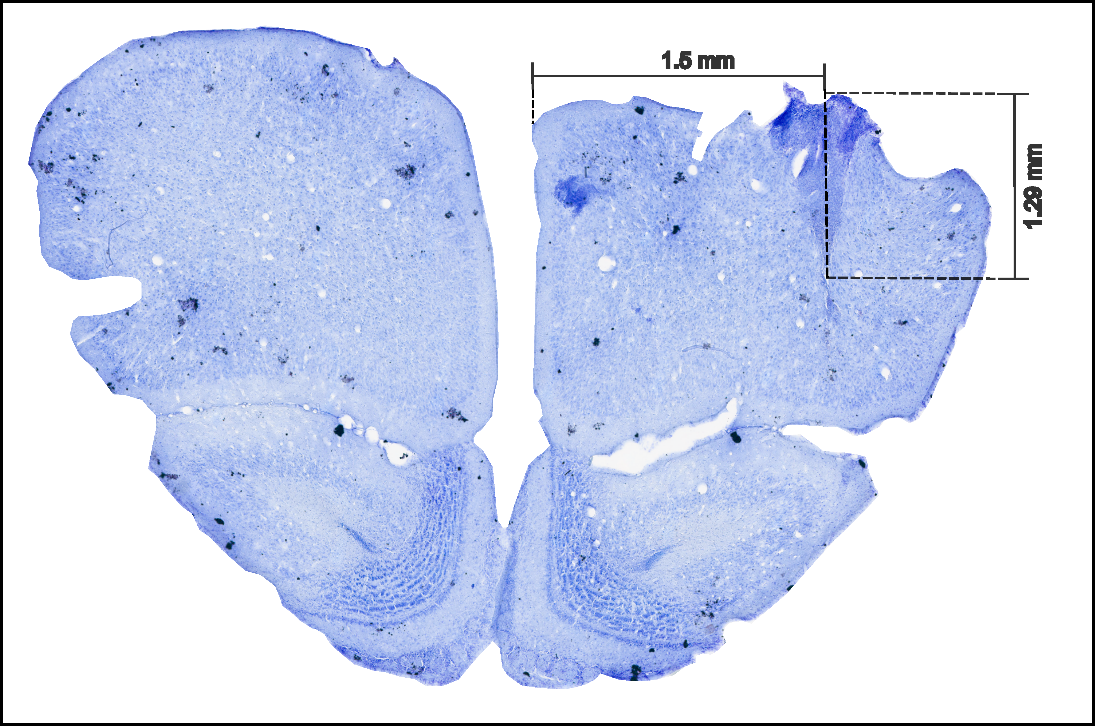
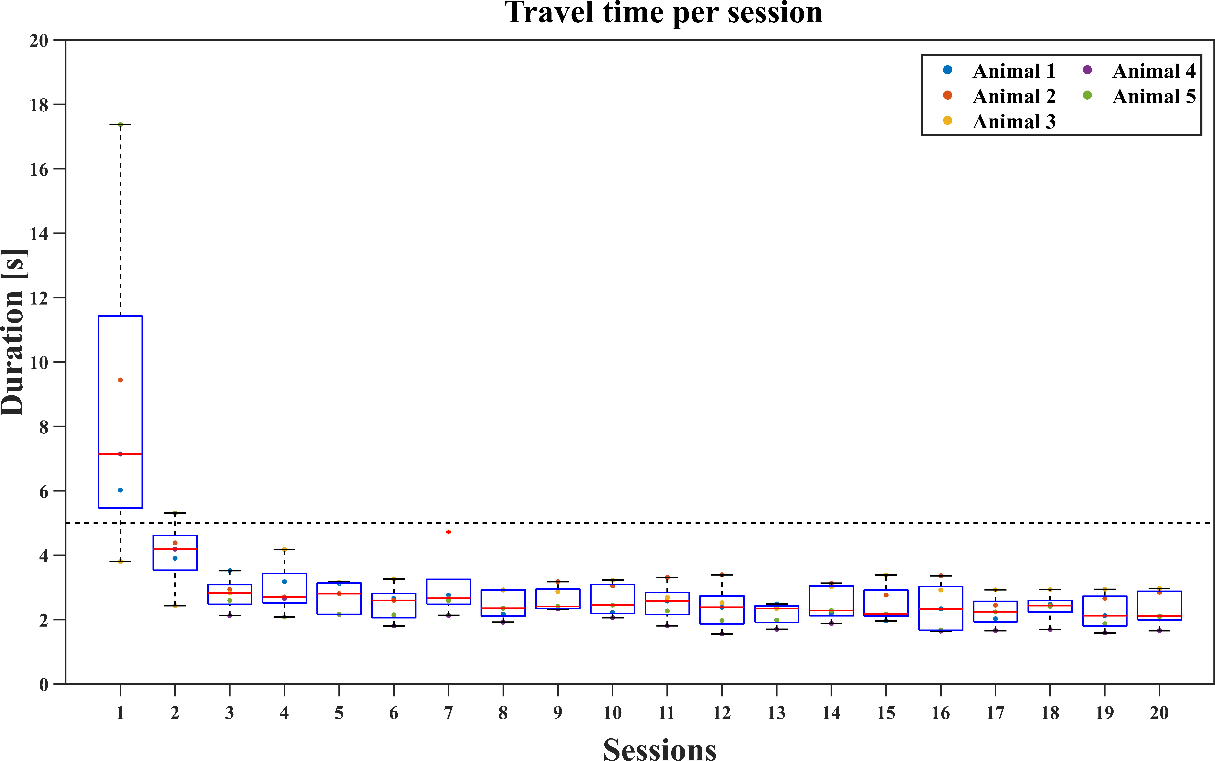
## Verification of the electrode implant

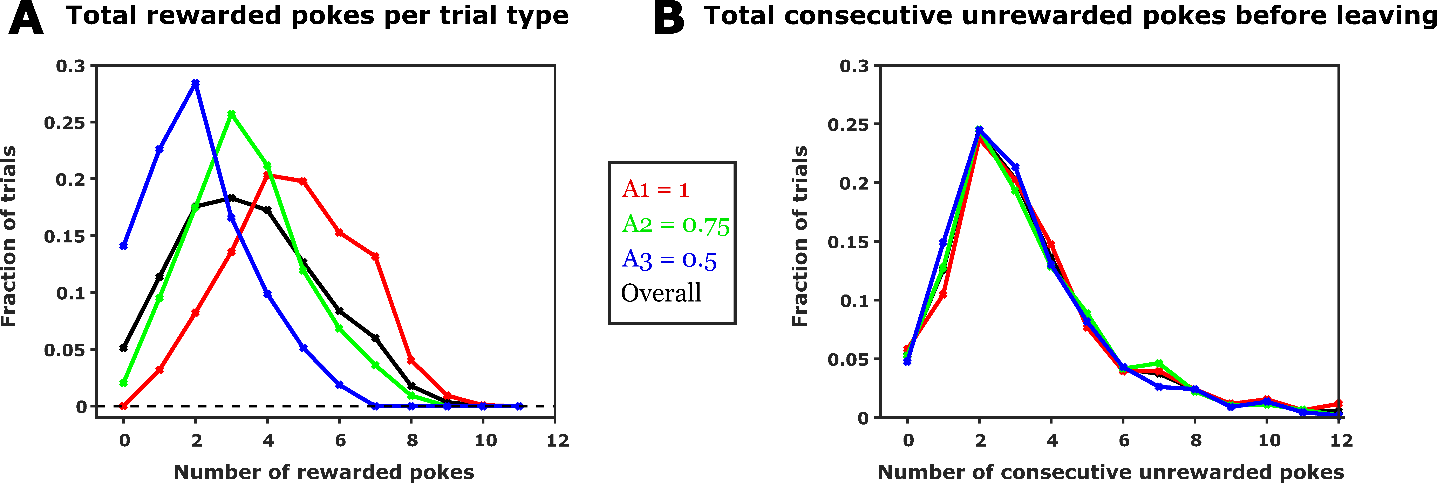


**Figure 1: Sample histology image of the frontal field A (FrA).** The histology slice is taken 4.85 mm anterior to Bregma. The electrode location from the staining can be seen 1.5 mm lateral with a cortical depth of 1.29 mm approximately.

Once all the foraging sessions were completed, the animals were sacrificed and the brain slices were taken and Nissl stained to confirm the location of implanted electrode. According to the Gerbil brain atlas (ref), the target frontal region A (FrA) is at 4.65-5 mm anterior and 1.5 mm lateral to the bregma. Fig.1 shows a sample histology image from one animal taken at 4.85 mm anterior and 1.5 mm lateral to the bregma. The electrode location can be clearly identified with a maximum cortical depth of approximately 1.29mm.

## Performance of the foraging behaviour





**Figure 2: Performance of the foraging behaviour. A** shows the median ± SD of travel time per session remains consistent and less than 5 seconds. Travel time is calculated as the time between the end of last poke in a trial to the start of the first poke in the succeeding trial. **B** shows the total number of rewarded pokes across trials for different starting probabilities (A = 1, 0.75 and 0.5). **C** shows the total consecutive unrewarded pokes before leaving a spout in each trial having different starting probabilities.

### Does the animal learn what to do?

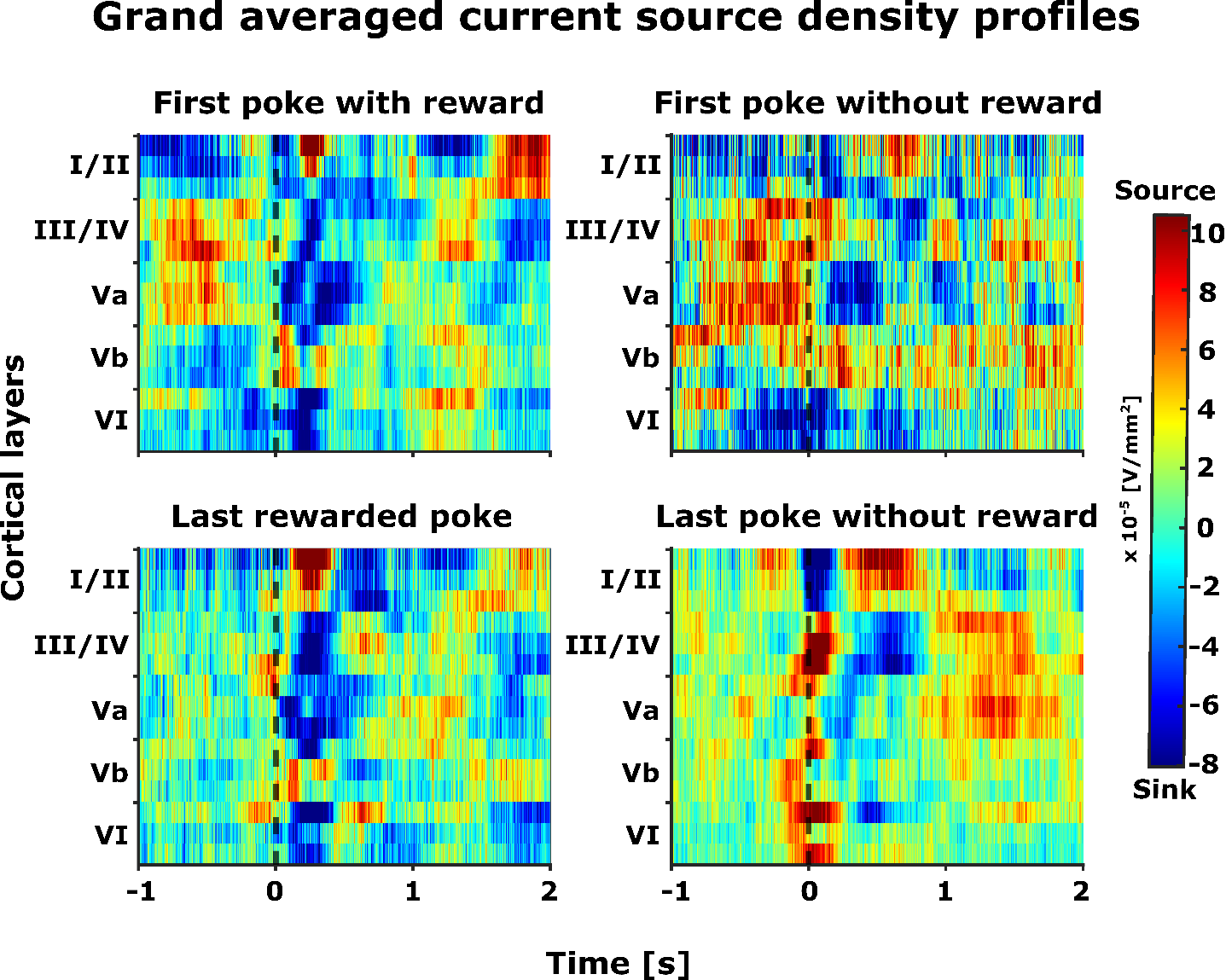
In order to see if the animal has learnt the foraging task and reduce the possibility of random exploration, we looked at their travel time. Here, travel time is defined as the time between the end of last poke in a trial and the start of the first poke in the succeeding trail (Methodology Fig). Lesser and consistent travel time indicate that the animal is goal-directed and not randomly exploring. Based on Fig.2A, the consistent and lower travel time indicates that after each trial, they were directed towards the other spout and not randomly exploring the arena. Hence, sessions 5-20 were considered for further behavioural and electrophysiological analyses.

### Animals make inference-based decisions

To evaluate how well the animal learnt to make decisions during the exploitation-exploration dilemma, we focussed on the time point where the animal decided to shift from exploitation to exploration. This is because every trial begins with exploiting the current spout for rewards and slowly as the rewards gets exhausted, the important decision is to decide when to leave the current spout to explore the other. This can be interpreted by the consecutive number of unrewarded pokes the animal makes before leaving the current spout.

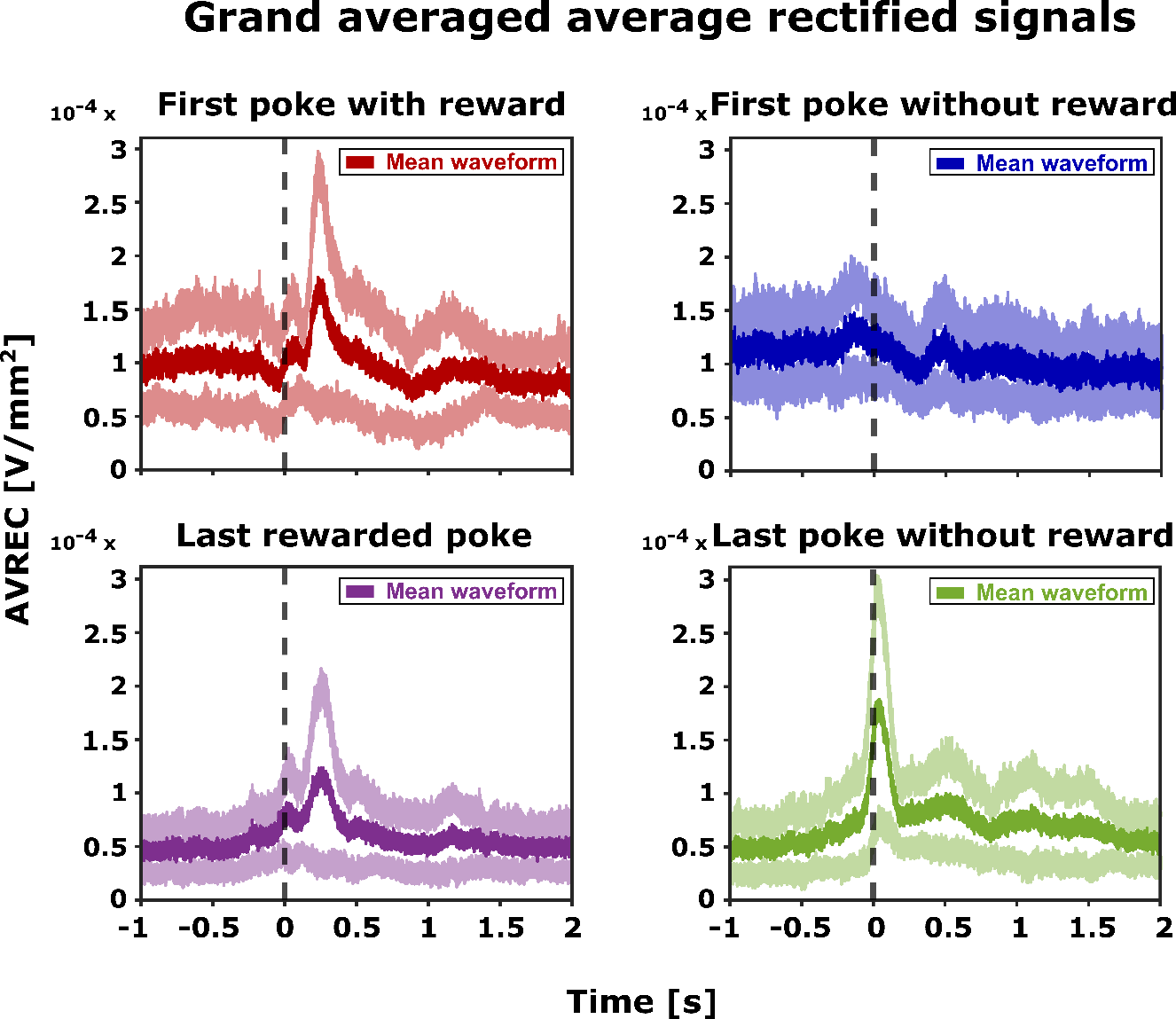
Following the experimental design (Methodology figure showing the exponential decrease of rewards for different starting probabilities), Fig.2B shows that the distribution of total rewards shifts towards right for trials starting with higher initial reward probability. This indicates that the if the animal makes same number of pokes in each trial, it will receive a greater number of rewards for trials starting with higher initial reward probability. On the contrary, this distinction between trials starting with different starting reward probabilities gets lost when it comes to the consecutive unrewarded pokes made before switching (Fig.2C) In other words, it shows that the animals irrespective of the starting reward probability and the total number of rewards received in a trial maintains a consistent number of consecutive unrewarded pokes before leaving a particular spout. This consistency in decision making may reflect the possibility that the animals form an inference about the hidden reward structure based on their learning of the task (explain stimulus bound and inference bound decision making in discussion).

## Distinct motor and reward related activity patterns in the frontal field A



**Figure 3: Grand averaged current source density (CSD) profiles (n=5)** **– Distinct motor and reward related spatiotemporal neural activity in frontal field A**. The selected epochsrepresent -1 to +2 seconds from the end of the poke (black dashed line, t=0). The selected time interval was taken for four different events (pokes) and its corresponding consequence (reward): (top left) first poke with reward, (top right) first poke without reward, (bottom left) last rewarded poke, and (bottom left) last poke without reward.

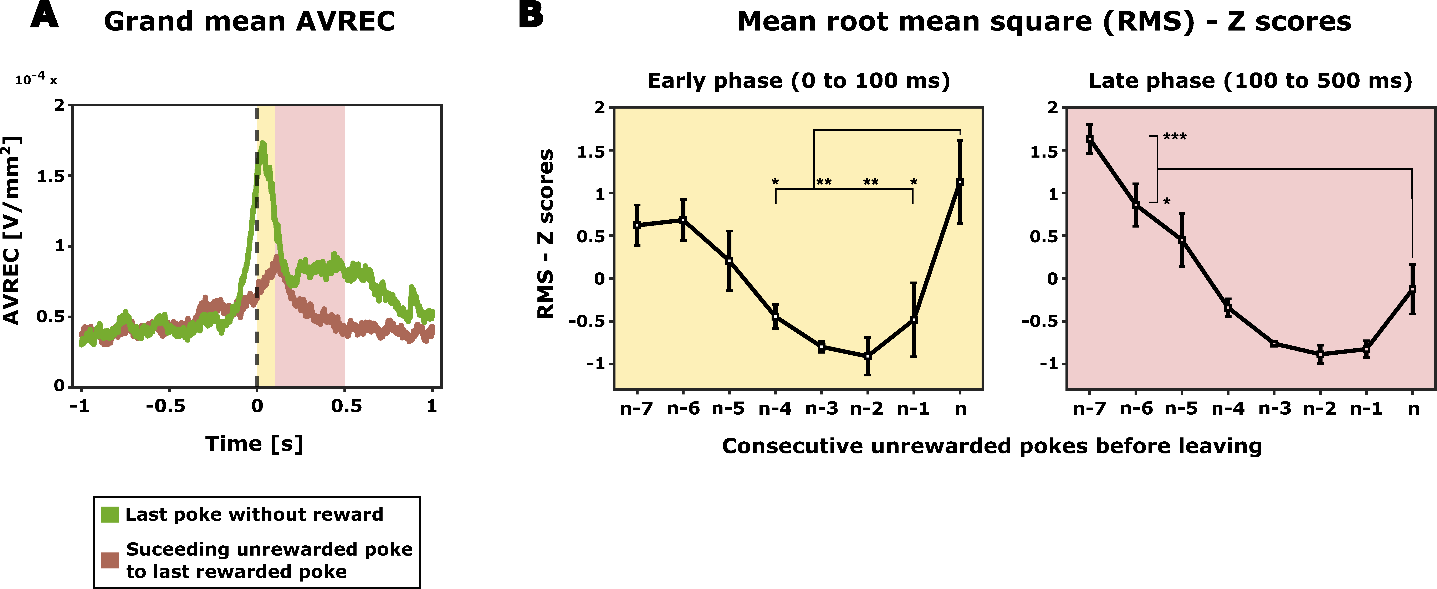
Fig.3 shows distinct spatiotemporal neural activity within frontal field A that encode both the poke and the ensuing reward. The selected epochs represent -1 to +2 seconds from the end of the poke (black dashed line, t=0) to compare the neural activity during decision-making phase at different time points. Four different time points were chosen to represent the exploitation and exploration phases in a trial. Based on the experimental design and animal performance (Fig), the time between first poke until last rewarded poke is considered as the exploitation phase because during this phase, even after experiencing unrewarded pokes, the animal insists on staying on the same side expecting more rewards. The last rewarded poke represents the end of exploitation phase and beginning of transition towards exploration. This is because after the last rewarded poke, the animal slowly starts to alter its expectation and move towards exploration phase which is best represented by the last poke in a trial. Further, the differential reward related activity patterns helped identify and distinguish deeper layers from superficial layers in the laminar recordings and perform channel-layer specification (Fig.3).



**Figure 4: Grand averaged AVREC (n=5)** **– Overall frontal cortical activity shows distinct motor and reward related signal**. Mean average rectified waveform (dark) along with its standard error (light) is plotted for selected time intervals (epochs). The selected epochsrepresent -1 to +2 seconds from the end of the poke (t=0). AVREC were taken for four different events (pokes) and its corresponding consequence (reward): (top left) first poke with reward, (top right) first poke without reward, (bottom left) last rewarded poke, and (bottom left) last poke without reward.

In order to visualize the overall frontal cortical activity, the CSD signals were rectified and averaged across the laminar electrodes to lose the spatial information. The average rectified signals (AVREC) from the CSD profiles also show distinct motor and reward related signals (Fig.4). Further, it can be seen that in frontal field A, the encoding of the expected (towards the end of pokes) and received reward (subsequently) shows different activity patterns during the exploitation (first poke until last rewarded poke) and exploration (last poke) phases of the trial. Both the rewarded pokes (first rewarded and last rewarded pokes) show a similar waveform where there is an early peak in amplitude immediately after the end of the poke that may encode for the expectation of a reward followed by a slightly increased peak occurring in less than 250 ms from the end of the pokes that may encode for the reward evaluation. On the other hand, in the unrewarded pokes (first unrewarded and last poke), the reward evaluation shows a dip followed by a later peak around 500 ms that may encode for a prediction error from the deeper layers. Compared to all other pokes, the last poke which is the starting point of exploration phase where the animal decides to switch to the other spout, shows a heightened frontal activity immediately after the end of the poke (<100 ms). These distinct activation patterns for different time points seen in Fig.3 and Fig.4 i.e., first rewarded poke, first unrewarded poke, last rewarded poke and last poke show that the frontal field A encodes not just the motor activity but also the reward related activity (expectation and prediction error).

## Shifts in frontal activity patterns: Exploitation to exploration



**Figure 5: Shift from exploitation to exploration**. **A** – shows the distinct activation pattern of grand AVREC of first unrewarded poke after last rewarded poke (brown) and the last poke without reward (green). Based on the grand AVREC data (A), two distinct time intervals (epochs) were chosen for RMS computation: early phase (0 – 100 ms, yellow), and late phase (100 – 500 ms, light pink). **B** – The average root mean square (RMS) Z score for unrewarded pokes between last rewarded poke and the last poke (n) before disengaging from the current spout (in this figure, we show a scenario of 7 consecutive unrewarded pokes where n-7 is the first unrewarded poke after last rewarded poke while n represents the last poke (unrewarded)). How to write the statistics part?

### What makes the last poke the last one?

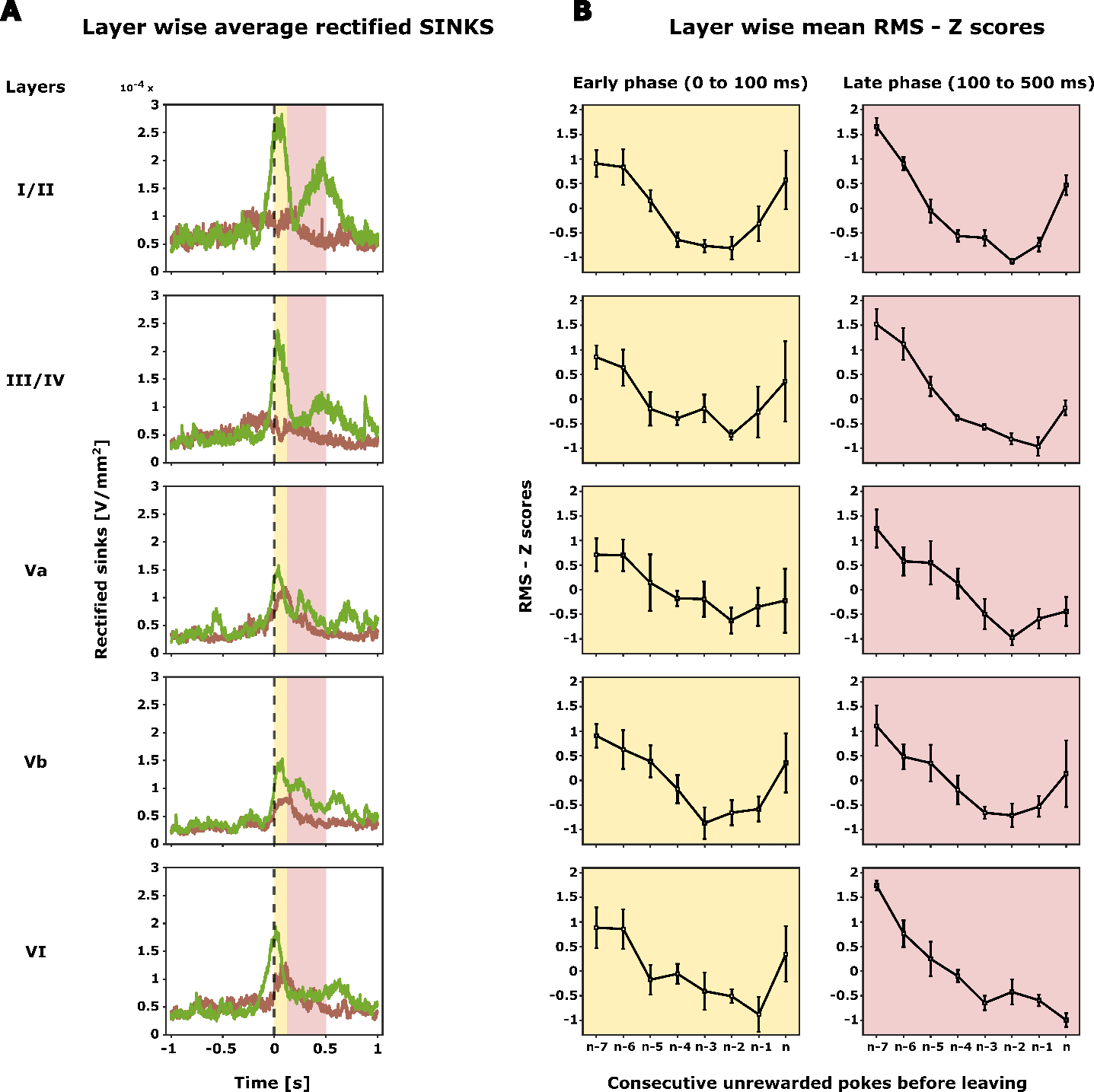
Now that we have confirmed that the frontal field A encodes distinct activity patterns for different phases (Fig.3 and Fig.4), it is essential to understand how the animal decides the last poke in a trial. This is essential as the animals may encounter multiple consecutive unrewarded pokes after the last rewarded poke (Fig.2C). Hence, the last poke also being an unrewarded poke needs to be distinguished from its preceding unrewarded pokes.

Fig.5A evidently shows that the last unrewarded poke has a distinct activity pattern compared to that of first unrewarded poke succeeding the last rewarded poke. Further, during the exploitation phase (first unrewarded poke after last rewarded poke), extended persistent activity patterns are observed after 100 ms following the nose poke (reward evaluation). Conversely, at the onset of the exploration phase (last poke), heightened frontal activity is evident during the animal's decision-making, occurring within 100 ms after the nose poke.

### The evolution of frontal activity from exploitation to exploration

Further, in order to look at the evolution of change in frontal activity patterns towards the last poke, we computed the z-scores of root mean square (RMS) of the AVREC signal for all the unrewarded pokes between last rewarded poke and last poke (Fig.5B and 5C). Based on the grand AVREC (Fig.5A), two different phases were selected i.e., early phase (0-100 ms from the end of the poke, Fig.5A yellow phase) and late phase (100-500 ms from the end of the poke, Fig.5A light pink phase) in which the RMS was computed for each of the pokes. This is to specifically narrow down the animal’s decision-making time point (<100 ms from the end of the poke) and separate it from the reward evaluation time point (> 100 ms from the end of the poke). Over the transition from exploitation to exploration (Fig.5B, n-7 to nth poke), during the early phase, there is an initial phase where the overall frontal activity decreased (n-6 to n-2) and then increased just before the animal decides to leave the spout (n-2 to nth poke) (One way ANOVA with Bonferroni correction, \* p < 0.05, \*\* p < 0.01, \*\* p<0.001). However, during the late phase, as it mainly encodes for reward evaluation, it constantly decreases as the unrewarded pokes keeps increasing.

## Layer specific motor and reward related activity patterns in the frontal field A



**Figure 6: Layer specific frontal motor and reward related activity**. **A** – The grand averaged rectified sinks (n=5) was computed for all the identified layers from the grand CSD profile (Fig.3). The selected epochsrepresent -1 to +2 seconds from the end of the poke (t = 0). The selected time interval was taken for first unrewarded poke after last rewarded poke (brown) and last poke without reward (green). Based on the averaged sinks (A), two distinct time intervals (epochs) were chosen for RMS computation: early phase (0 – 100 ms, yellow), and late phase (100 – 500 ms, light pink). **B** – The layer wise average root mean square (RMS) Z score was calculated from the average rectified sinks for unrewarded pokes between last rewarded poke and the last poke (n) before disengaging from the current spout. (Similar to Fig.5B, in this figure, we show a scenario of 7 consecutive unrewarded pokes where n-7 is the first unrewarded poke after last rewarded poke while n represents the last poke (unrewarded)).

Five distinct cortical layers were identified from the laminar recordings (Fig.3). The source signal was removed and only the sinks were considered to ensure that the signal is contributed only by layer specific local excitatory synaptic populations. Comparison of layer-wise averaged sink activity indicates a transition of activity from deeper to upper cortical layers once the reward expectation at the current spout decreased (Fig.6A, add last rewarded poke). Shortly before the decision of the animal to explore the other spout, particularly layers I/II and III/IV showed two prominent activity peaks in amplitude directly at the retraction of the spout (0-100 ms) and afterwards (100-500 ms) (Fig.6A). During the phase of switching from exploitation to exploration, activity decreases in all cortical layers. At the beginning of the exploration phase, the upper layers I/II and III/IV, as well as Vb, are active. The increase is mainly observed in the early phase following the nose poke (expectation), and then with regard to the evaluation of the absence of rewards (Fig.6B).