we searched through IBL dataset and selected sessions and probes that includes channel(s) assigned to primary visual cortex. Along side the raw LFP data we extract two other dataset for this project: trials table containing event information such as stimulus contrast and the onset/offset timing for each trial; channel's location detailing the exact location of each channel and the associated brain region.

To compare the Inter-Trial Coherence (ITC) average and frequency power across different stimulus contrast levels, a series of non-parametric statistical tests were employed due to the non-normal distribution of the data. The Shapiro-Wilk test was initially used to assess the normality of the data. However, the results indicated that the data did not follow a normal distribution. Attempts to normalize the data by removing outliers (using the 95th percentile cutoff) were unsuccessful in achieving normality. Consequently, non-parametric methods were selected.

Given the non-normality and the repeated measures design of the experiment, the Friedman test was utilized as an alternative to repeated measures ANOVA. The Friedman test is appropriate for comparing the ITC average and frequency power across the different conditions, as it does not assume normal distribution and is suitable for dependent samples.

Following the Friedman test, which indicated significant differences among conditions, post-hoc pairwise comparisons were conducted using the Nemenyi test. The Nemenyi test is a suitable post-hoc method for the Friedman test, allowing for the identification of specific pairs of conditions that differed significantly in terms of ITC average and frequency power.

This approach ensured that the statistical analysis was robust and appropriately addressed the non-normality of the data while still allowing for meaningful comparisons across the different stimulus contrast levels.

Multiple Comparison for Time-Frequency ITC Estimate

To identify significant clusters in the time-frequency Inter-Trial Coherence (ITC) estimates, we employed the one-sample permutation cluster test as implemented in the MNE-Python library. The use of permutation tests is particularly crucial in this context due to the multiple

comparison problem inherent in time-frequency analyses, where statistical tests are conducted across many time points and frequency bands. Without proper correction, this can lead to a high rate of false positives (Maris & Oostenveld, 2007).

Why Permutation Tests Are Needed

The multiple comparison problem arises when numerous statistical tests are performed simultaneously, increasing the likelihood of incorrectly rejecting at least one null hypothesis (Type I error). In the context of time-frequency ITC estimates, testing for significance at each time-frequency point independently would require a large number of comparisons, leading to inflated false positive rates. Traditional methods of controlling for multiple comparisons, such as the Bonferroni correction, tend to be overly conservative, potentially reducing statistical power and increasing the likelihood of Type II errors (Nichols & Holmes, 2002).

Permutation tests provide a non-parametric solution to this problem. They make no assumptions about the distribution of the data, making them particularly suitable for neurophysiological data, which often do not follow normal distributions. Instead of relying on theoretical distributions to determine significance, permutation tests use the data itself to empirically construct a null distribution. This approach allows for accurate control of the family-wise error rate (FWER) while maintaining higher sensitivity in detecting true effects (Efron & Tibshirani, 1994).

How the Permutation Test Works

In our analysis, we used 1000 permutations to generate the null distribution. The permutation test works by randomly shuffling the data labels and recalculating the test statistic (in this case, the t-statistic) for each permutation. This process is repeated many times (1000 in our case), and for each permutation, the maximum t-value across all time-frequency points is recorded. This set of maximum t-values forms the null distribution.

The significance threshold is then defined based on this null distribution. Specifically, the threshold is set to a value that corresponds to the desired alpha level (e.g., 0.05), ensuring that the probability of observing a cluster of significant time-frequency points under the null

hypothesis is appropriately controlled. By setting the threshold to None, we allowed MNE-Python to automatically compute this threshold based on the distribution of maximum t-values from the permutations, ensuring that the threshold is data-driven and adapted to the observed data distribution.

Clusters of time-frequency points that exceed this threshold are considered statistically significant, indicating that the observed ITC values in these clusters are unlikely to have occurred by chance. This method not only controls for multiple comparisons but also capitalizes on the spatial and temporal structure of the data, increasing sensitivity to detect meaningful effects in the time-frequency domain (Blair & Karniski, 1993; Maris & Oostenveld, 2007).

Task detail

In the IBL task (Figure 1), head-fixed mice had to move a visual stimulus to the center by turning a wheel with their front paws. At the start of each trial, the mouse was required to refrain from moving the wheel for a quiescence period lasting between 400 and 700 milliseconds. After this period, a visual stimulus (Gabor patch) appeared on either the left or right side of the screen accompanied by a 100-millisecond tone (5 kHz sine wave). If the mouse correctly moved the stimulus to the center by turning the wheel over 35°, it received a 3 µL water reward. Incorrect responses or failing to respond within 60 seconds resulted in a 500-millisecond burst of white noise and a timeout (Benson et al. 2023). As shown in Figure 1c, mice typically responded quickly within 2 seconds. The stimulus is always presented for the first 1 second regardless of response time (RT). RT is defined as the time after stimulus when the wheel rotation exceeds the threshold.

The experiment began with 90 unbiased trials where the stimulus appeared equally on both sides. The stimulus contrast levels were presented in a ratio of [2:2:2:2:1] for contrasts [100%, 25%, 12.5%, 6%, 0%]. After this initial block, trials were organized into biased blocks where the likelihood of the stimulus appearing on one side was fixed at 20% for the left and 80% for the right in “right blocks” or vice versa in “left blocks.” These blocks consisted of 20 to