Does the brain care about averages? A simple test

A. Tlaie,[1] K. A. Shapcott,[2] P. Tiesinga,[3] M. L. Schölvinck,[4] and M. N. Havenith[4]

[1] *Neural Computation Group, Italian Institute of Technology, 16131, Genoa, Italy*

[2] *Singer Lab, Ernst Strüngmann Institute for Neuroscience, 60528 Frankfurt am Main, Germany*

[3] *Department of Neuroinformatics, Donders Institute, Radboud University, Heyendaalseweg 135, 6525 AJ Nijmegen, The Netherlands*

[4] *Zero-Noise Lab, Ernst Strüngmann Institute for Neuroscience, 60528 Frankfurt am Main, Germany*

Trial-averaged metrics, e.g. in the form of tuning curves and population response vectors, are a basic and widely accepted way of characterizing neuronal activity. But how relevant are such trial-averaged responses to neuronal computation itself? Here we present a simple test to estimate whether average responses reflect aspects of neuronal activity that contribute to neuronal processing in a specific context. The test probes two assumptions inherent in the usage of average neuronal metrics:

1. Reliability: Neuronal responses repeat consistently enough across single stimulus instances that the average response template they relate to remains recognizable to downstream regions.
2. Behavioural relevance: If a single-trial response is more similar to the average template, this should make it easier for the animal to identify the correct stimulus or action.

We apply this test to a large publicly available data set featuring electrophysiological recordings from 42 cortical areas in behaving mice. In this data set, we show that single-trial responses were less correlated to the average response template than one would expect if they simply represented discrete versions of the template, down-sampled to a finite number of spikes. Moreover, single-trial responses were barely stimulus-specific – they could not be clearly assigned to the average response template of one stimulus. Most importantly, better-matched single-trial responses did not predict accurate behaviour for any of the recorded cortical areas. We conclude that in this data set, average responses do not seem particularly relevant to neuronal computation in a majority of brain areas, and we encourage other researchers to apply similar tests when using trial-averaged neuronal metrics.

Introduction

For decades, neuroscientists have recorded local neuronal populations and estimated how much information the brain can extract from their activity to guide perception and decision-making. While research has traditionally focused on the average response preferences of individual neurons [1–13], more recent studies have explored population patterns of activity, either in their raw form and/ or in the form of population vectors obtained by dimensionality reduction in higher-order (e.g. principal component) space [14–16]. This recent work has highlighted, for instance, the adaptation of neuronal responses to the statistics of the perceptual environment [17] and orthogonalized neuronal coding of stimulus information, behavioural choices and memory [15, 18, 19].

Irrespective of the specific approach, these studies have in common that trial-averaged population activity is implicitly treated as meaningful. For instance, upon finding that with repeated stimulus exposure, average population responses become more discriminative of behaviourally relevant stimuli [6, 8, 20], it is implicitly assumed that this will improve an animal's ability to perceive these stimuli correctly. Related to this assumption is the notion that deviations of single-trial neuronal responses from the average population response represent 'noise' of one form or another. The exact interpretation of such neuronal noise has been debated for decades [21], ranging from truly random and meaningless activity [22–26], to neuronal processes that are meaningful but irrelevant for the neuronal computation at hand [27–29], to an intrinsic ingredient of efficient neuronal coding [30–33]. Nevertheless, in all of these cases a clear distinction is being made between neuronal activity that is directly related to the cognitive process under study (e.g. perceiving a specific stimulus) – which is typically approximated by a trial-averaged neuronal response – and 'the rest'. While this framework has undoubtedly been useful for neuroscientists aiming to characterize the general response dynamics of neuronal networks, it remains an outstanding issue whether trial-averaged population activity does in fact reflect an aspect of neuronal responses that transmits information between neurons. In other words, neuroscientists care about average population responses, but does the brain?

There is some evidence in both directions: On the one hand, studies highlighting the large inter-trial variability of individual neuronal responses [28, 29, 34–38] would suggest that a fixed 'template response' averaged across many stimulus instances may not be very useful in order to represent ongoing neuronal processing. In addition, there is the simple fact that outside the lab, any stimulus is unlikely to appear repeatedly in the same way and in the same behavioural context, and therefore pooling responses across stimulus repetitions seems unlikely to be an ecologically valid strategy for reliable neuronal coding. On the other hand, the fact that perceptual decisions can be shifted e.g. by simply increasing or suppressing the activity of specific neuronal populations away from their average activity [39–44] indicates that at least for the clear-cut contexts - and limited time frames [45] - typically presented in lab experiments, average population responses can directly shape perceptual decision making and must therefore be computationally relevant.

In this paper, we formally test whether the implicit assumptions inherent in the computation of average population responses do actually hold for neuronal activity. Specifically, if the brain cares about averages, i.e. if neuronal coding relies fundamentally on average 'templates' of population activity, it should satisfy two assumptions (see Fig 1A):

1) The responses of task-relevant neuronal populations are reliable – they repeat consistently enough across single stimulus instances that the information they carry remains recognizable to downstream regions (i.e. responses can be matched to the 'population template' of a given percept or action). 2) Population responses guide decision-making and behaviour – if a single-trial response is more similar to the average population template, this should make it easier for the animal to identify the correct stimulus or action.

We test these two assumptions in a large data set containing neural activity from multiple brain areas recorded during a perceptual decision task. After identifying which areas are most relevant in this context, we use simple tests to compare single trial activity in these areas to the average response templates. Based on these tests, we show that in this data set, neither of the two assumptions set out above is fully met.

Results

To test our two assumptions, we use a large and publicly available data set provided by [44]. The data set contains high-density electrophysiological (Neuropixel) recordings across a large number of brain regions in mice performing a two-choice contrast discrimination task. In the task, animals are presented with two gratings of varying contrast (0, 25, 50 or 100%) appearing in their left and right hemifield, respectively. To receive reward, animals turn a small steering wheel to bring the higher-contrast grating into their central vision, or refrain from moving the wheel if no grating appears on either side (see Fig. 1B). The original task also featured trials in which both stimulus contrasts were equal. In those cases, animals were randomly rewarded for turning right or left. Those trials were discarded in the current analysis since it is impossible to define one 'correct' behavioural response in this context. Neuronal recordings were obtained from 42 brain regions including cortical and subcortical targets (Fig. 1C).

To first establish which cortical areas are relevant for this task, we used a data-driven approach to identify across all recorded areas to what extent neuronal population activity predicted the presented stimulus and/or the animal's target choice. To this end, we trained a decoder (Multinomial GLM; see Methods) to identify either target choice (left turn, right turn, no movement) or stimulus condition (higher contrast on left, higher on right, zero contrast on both sides) based on the single-trial population response vectors of each cortical area. For the response vectors, we took into account neuronal activity from stimulus onset to $200ms$ post-onset (see Figs. S1 and S2 for a rationale of this choice and examples of neuronal activity during this time window). Finally, we computed the mutual information between the decoder predictions and the real outcomes. Figure 2 shows the amount of mutual information about stimulus condition and target choice that was conveyed by the neuronal population activity in different cortical areas.

As one can see, many cortical areas contained little information on either stimulus identity or target choice, suggesting that they were not crucially engaged in the task. We therefore used an elbow criterion (see Methods) to determine a threshold for selecting cortical areas that provided the highest information on either stimulus ($I^{thr}_{stim} = 0.242$ bits;blue quadrant), choice ($I^{thr}_{choice} = 0.248$ bits;red quadrant), or both (i.e. both thresholds exceeded; purple quadrant). With this approach we identified five cortical areas that contained predominantly stimulus information, one area that contained mainly choice information and three areas that contained both. These results seem largely congruent with the literature: For instance, latero-intermediate visual cortex (VISl) and primary visual cortex (VISp) would be expected to contain visual stimulus information. Meanwhile, choice information is conveyed most strongly by the reticular part of the substantia nigra (SNr), which is pivotal in reward-seeking and learning [46–48]. The fact that the Red Nucleus (RN), which is involved in coordinated paw movement, contains information about both stimulus and choice is also in agreement with previous literature [49, 50]. More surprisingly, Inferior Colliculus (IC), which is classically regarded a hub of auditory processing [51], also contains stimulus and choice information, emphasizing the fact that information is widely distributed across cortical areas [14, 44, 52].

Having identified the most task-relevant cortical areas in a data-driven way, we used the neuronal recordings from these areas as well as from three comparison areas that contained the least relevant information – nucleus accumbens (ACB), dorsal endopiriform nucleus (EPd) and substantia innominate (SI) – as a benchmark to test the assumptions

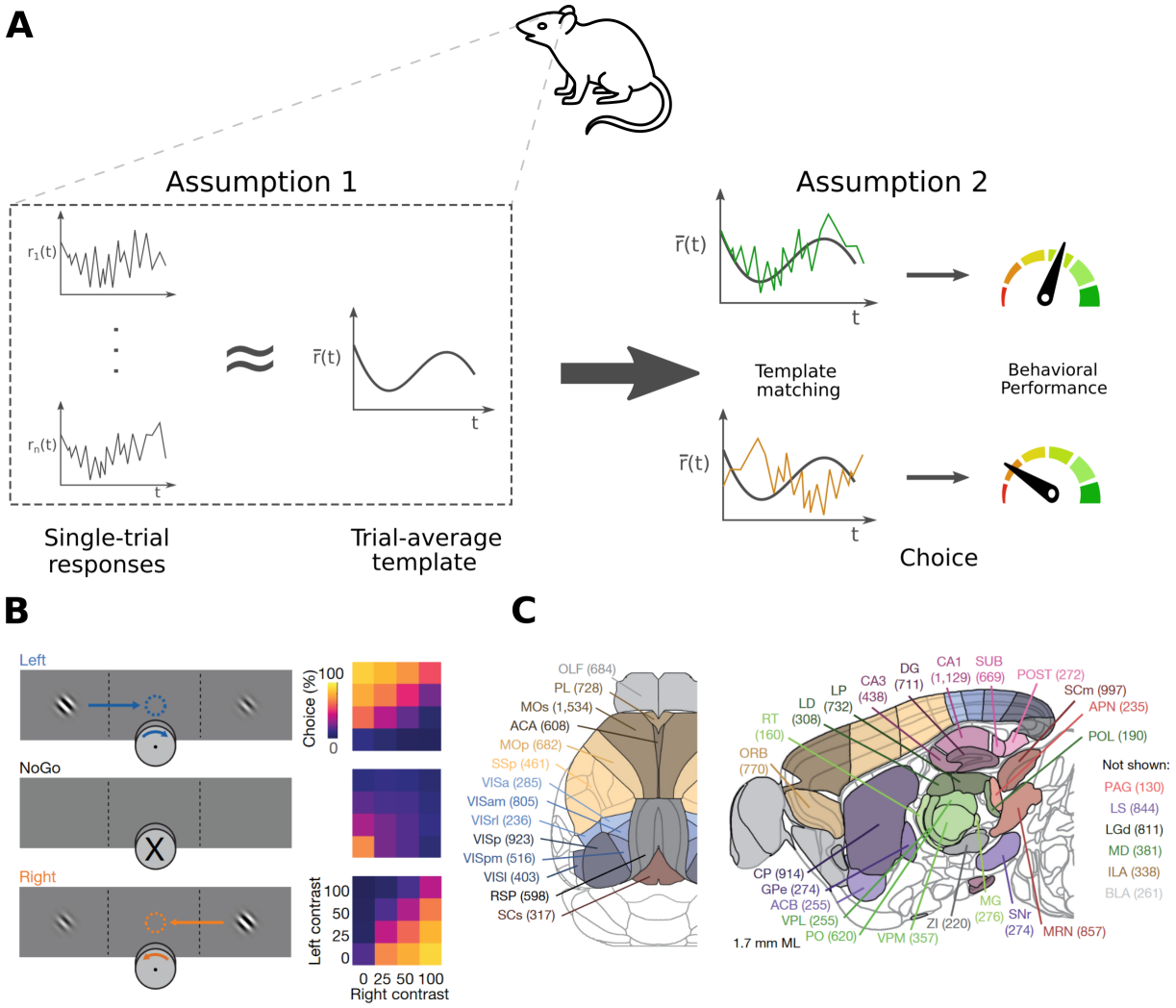

FIG. 1. (Color online) Overview of the analysis and data. A) Graphic summary of the two assumptions underlying the computation of average population responses: Single-trial responses correspond at least somewhat to the trial-averaged response template (left), and better matched single-trial responses lead to more efficient behaviour (right). B) Task structure: To obtain reward, animals need to move a steering wheel to bring the higher-contrast grating stimulus into the centre, or refrain from moving the wheel when no gratings are presented. Average behavioural performance on this task is shown on the right. C) Recording sites and, in parenthesis, total number of recorded neurons. B and C are reproduced with permission from [44].

set out above. As a first step, we computed average population responses ('templates') for two stimulus constellations: Target stimulus on the left or target stimulus on the right. Note that these average templates each summarize several different contrast levels (e.g. trials with a contrast difference of 50% right – 0% left and 100% right – 50% left will both be pooled in the 'Target right' template). We chose this approach to avoid working with $4 \times 4 = 16$ different contrast combinations with trial numbers as low as $n = 2$, which would have made averaging essentially meaningless. However, as Figure S3 shows, the average responses to contrast differences of the same 'direction' (e.g. 'Target right') were very comparable to each other, justifying the decision to pool them into the same average response template.

We then quantified how well single-trial population responses correlated with the average template for that given stimulus constellation (Fig. 3; see also [53]). Correlations were generally high, typically ranging from r values of 0.4

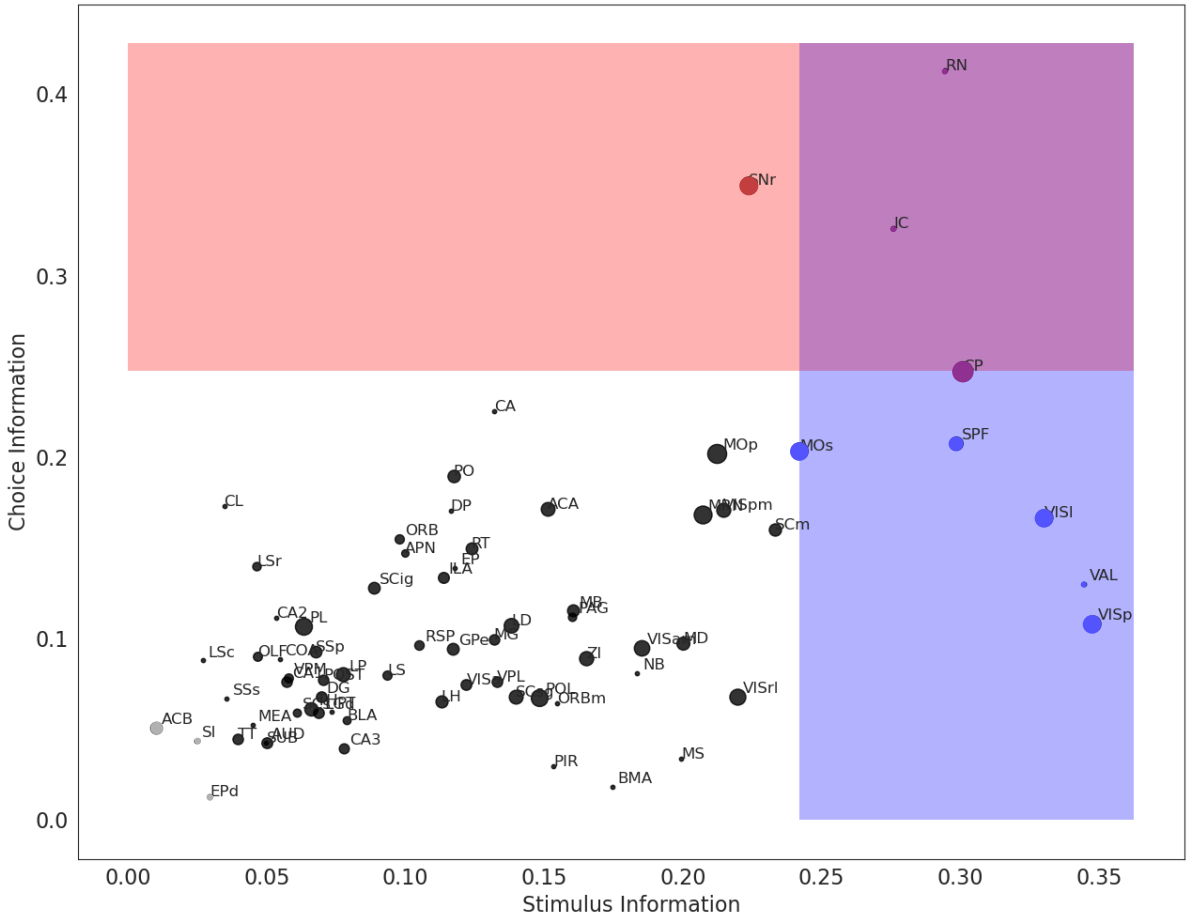

FIG. 2. (Color online) Stimulus and target choice information decoded from the neuronal activity in various cortical areas. Scatter plot of stimulus information versus target choice information extracted from different cortical areas by a multinomial GLM decoder. Each data point represents the median (dot location) and standard deviation across sessions (dot size) of one cortical area (see in-figure labels). The blue and red quadrants (determined by an elbow criterion, see Methods) represent those areas where either stimulus or choice information was high, respectively; the purple quadrant shows those areas where both choice and stimulus information was high. Comparison areas (in the white quadrant) are marked in grey.

to 0.8 across all cortical areas ($n = 89$ to 3560 trials per cortical area; all $p < 0.001$). This suggests that single-trial responses are quite a faithful representation of the trial-averaged template. To assess more precisely if single-trial responses can be regarded as a down-sampled version of the population template, we generated surrogate data. For each trial, we repeatedly drew the number of action potentials observed in that trial from the overall pool of spikes that make up the population template, and then computed a bootstrapped single-trial response vector based on the drawn spikes (Fig. 3A). Therefore, the bootstrapped responses contained the same number of spikes as the measured single trial, but the neurons that produced each of these spikes were chosen randomly according to the probability with which they contributed to the average template. We then correlated these bootstrapped single-trial responses to the population template. If single-trial responses are indeed simply the result of a Poisson process sampling from the average population template, the correlations resulting from our surrogate data should be comparable in the original data. The surrogate data uniformly showed higher correlations to the population template than the original data

(Fig. 3B). In fact, most correlations derived from the original data land far below the 50th percentile when compared to the distribution of correlations derived from surrogate data (Fig. S4). This suggests that single-trial responses exhibit more variation around the time-averaged mean than strictly explained by (Poissonian) down-sampling.

We suspected that the correlations between single-trial responses and the population template resulted at least partially from the basic firing properties of different neurons, which would not carry any specific information since this would not vary with the stimulus and/or task response. This notion is also supported by the observation that the three least informative cortical areas (ACB, EPd and SI) showed comparable correlations to the population template as the cortical areas carrying the most stimulus and/or choice information. To estimate more precisely what portion of the correlation was stimulus-specific, we also computed the correlation of each single-trial response to the population template for the stimulus that was not in fact shown at that particular time, i.e. the incorrect template. As one might expect, the resulting correlations are indeed lower than those with the population template for the correct stimulus (Fig. 3C; average difference between median correlations: 0.03 ± 0.02; t-test for dependent samples; $n = 89$ to 3560 trials per cortical area; $t = 3.0$ to 27.2; all $p < 0.01$, corrected for multiple comparisons using a FDR correction imposing a family-wise error rate of 0.05). While statistically significant, this correlation difference (0.03 ± 0.02) is so small compared to the typical spread of single-trial correlations (standard deviation: 0.04 to 0.39 across cortical areas) that correlations to the correct and incorrect templates were largely indistinguishable on a single-trial level. What is more, highly informative cortical areas did not show significantly more specific correlations than non-task-related cortical areas (mean correlation difference for all highly informative areas: 0.031; $n = 9$; for all unrelated areas: 0.022; $n = 62$; Welch's t-test: $t = 1.17$, $p = 0.27$). Thus, most of the correlation between single-trial responses and the time-averaged template is not explained by stimulus-specific response patterns. To quantify this more precisely, we computed a metric referred to as the specificity index, which represents for each single-trial response the correlation to the correct template minus the incorrect template. The distribution of single-trial specificity indices across cortical areas is shown in Fig. 3D. Most values are positive, indicating that single-trial responses were generally more related to the correct than incorrect template (t-test for difference from zero; $n = 89$ to 3560 trials per cortical area; $t = 3.0$ to 27.2; all $p < 0.01$, corrected for multiple comparisons using a FDR correction imposing a family-wise error rate of 0.05). However, the distributions also remain close to zero, with correlation differences rarely exceeding 0.1 (Fig. 3D). In addition, the specificity of the original data tended to be slightly lower than that of the bootstrapped data introduced above. This indicates that in this data set, across all examined cortical areas including visual ones, single-trial responses were barely more similar to the correct stimulus template than to the incorrect one – and deviated more from the template than necessary due to downsampling.

These results may not come entirely as a surprise since recent work has demonstrated how strongly non-task-related factors can drive neuronal responses even in primary sensory areas like visual cortex [28, 29, 54–59]. As a consequence, single-trial responses would be expected to vary strongly according to factors that are neither related to the perceived stimulus nor the target choice. However, that does not change the fact that in the presence of these expected single-trial variations, the animal still needs to identify the presented stimulus and make a correct perceptual choice based on that. If time-averaged response templates were relevant to this perceptual decision, we would expect that single-trial responses that for whatever reason fail to resemble the average response template should be more difficult to process for downstream areas, and hence lead to less efficient behavioural choices [44].

To directly test if the match between single-trial responses and the correct response template predicted target choices, we quantified single-trial correlations separately for hit and miss trials. In this context, miss trials are defined

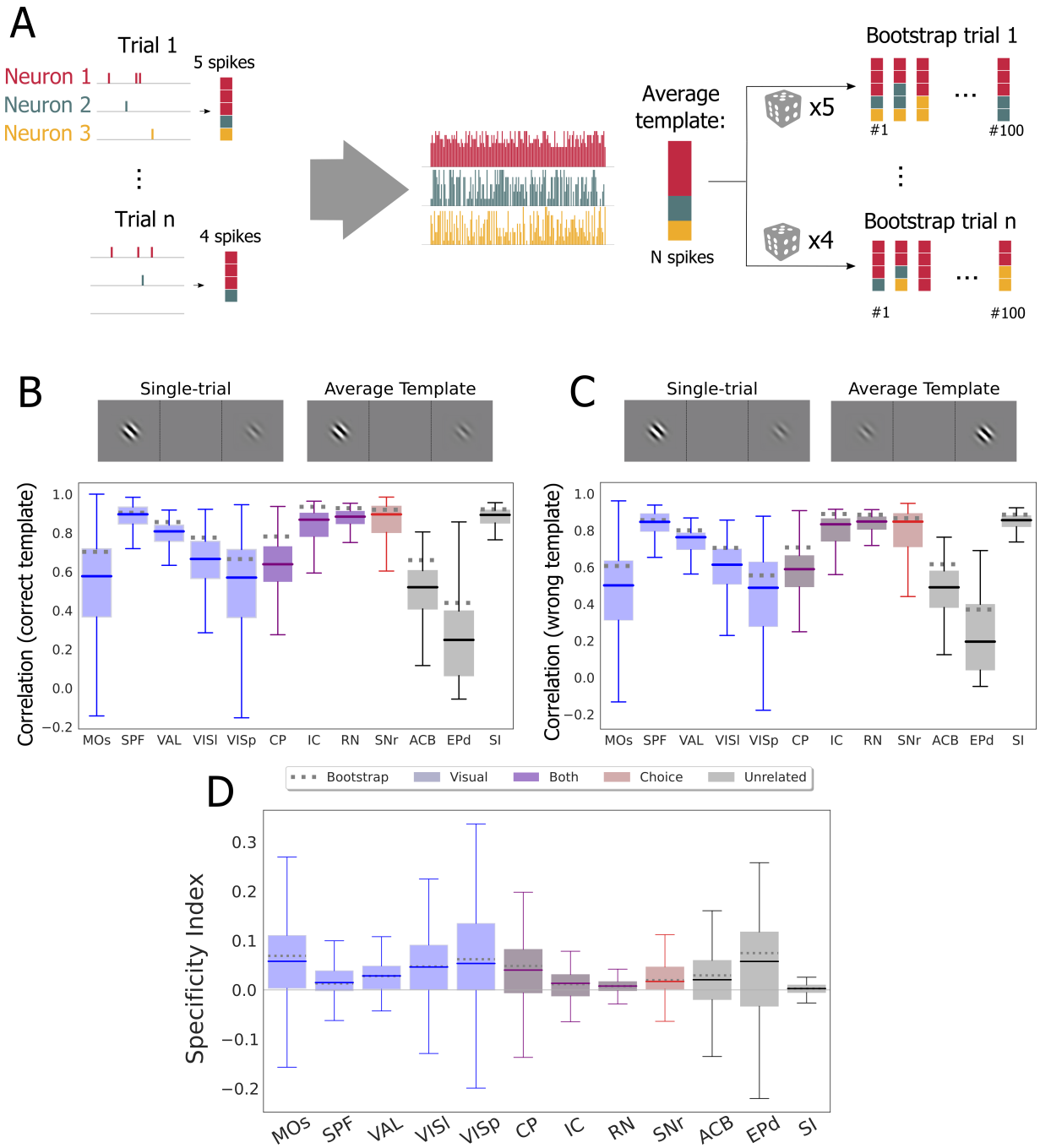

FIG. 3. Correspondence between average and single-trial responses. A) Representation of the bootstrapping procedure. The number of spikes stayed the same on each trial, but the neuron that produced each of these spikes was chosen randomly with a probability according to how often it spiked in the average template. We repeated this procedure 100 times. B) Distribution of the correlations between single-trial response vectors and the trial-averaged response template for the correct stimulus. Icons on top represent an example of a correct match between the stimulus constellations on a single trial, and that used to calculate the average response template. Box: 25th and 75th percentile. Centre line: median. Whiskers: 10th and 90th percentile. Colors: Classification of cortical areas (see in-figure legend). Dotted lines: Median correlation of bootstrapped data. Variance around this median value ranged from 0.02 (RN) to 0.23 (EPd). C) Same as B, but for correlations to the response template of the incorrect stimulus constellation. D) Specificity index of single-trial responses across cortical areas, defined as the difference between the correlations to the correct and incorrect template. Solid gray line highlights the Specificity Index of 0.0, which translates to exactly equal correlation to correct and incorrect template. Dotted lines represent the specificity index of the medians of the bootstrapped values for each recorded area, with variances around those median values ranging from 0.03 (VISl) to 0.33 (VISp).

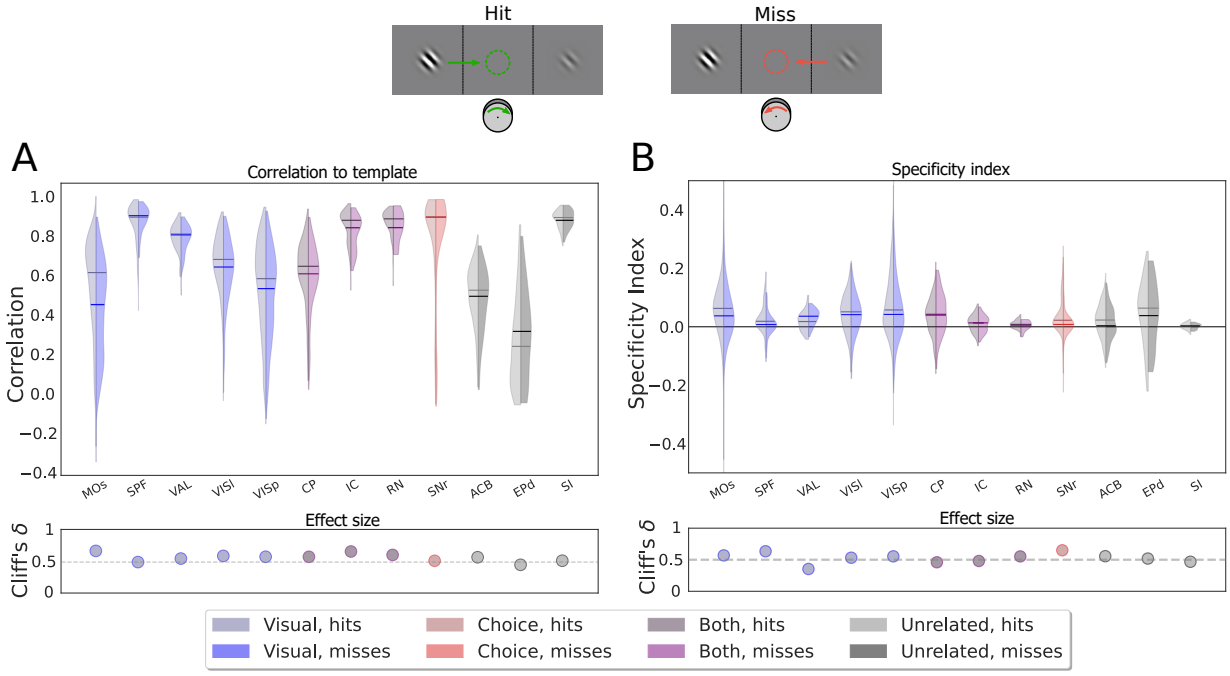

FIG. 4. (Color online) Match between single-trial responses and correct response template in hit and miss trials. A) Same as Figure 3A, but for single-trial correlations to the correct template, split by hit and miss trials. B) Effect size (Cliff's δ) of the difference in single-trial correlations for hit versus miss trials is shown below the plot. Cliff's δ can take values between 0 and 1, with 0.5 indicating a complete overlap between distributions (see Methods).

as trials where the animal either did not respond at all or chose the wrong target. Single-trial correlations were somewhat lower in miss trials than in hit trials across most cortical areas, suggesting that a worse match to the average template did indeed tend to produce miss trials more often (Fig. 4A, top). However, overall the difference between the correlations in hit versus miss trials was small, as quantified by Cliff's δ (Fig. 4A, bottom). There were a few exceptions, such as the secondary motor area (MOs) (Cliff's $\delta = 0.66, p = 0.0142, n_{hit} = 2057, n_{miss} = 667$, see Methods). This is potentially due to the fact that hit and miss trials are associated with fundamentally different motor responses, particularly since miss trials also include trials when the animal did not respond at all. Yet the overall pattern suggests that the absolute correlation between a single-trial response and the average response template has low - and inconsistent - predictive power regarding perceptual decision-making.

It is however possible that the important factor for perceptual decision making is not the overall correlation between the single-trial response and the correct response template, but whether it resembles the correct template more than the incorrect one. To explore this possibility, we compared the distribution of specificity indices (see Fig. 3D) between hit and miss trials. There was again no consistent difference (Fig. 4B), with many relevant areas (e.g. CP, IC and RN) showing no difference at all, others (like MOs, VISl and VISp) showing small differences with a large distribution overlap, and again others (VAL) showing an inverted difference. This indicates that single-trial responses that were more specific to the correct template did by and large not lead to improved target choices.

While in the neuronal populations recorded here, average responses were overall not particularly informative about either stimulus identity or subsequent behaviour, it is possible that there could be a 'supergroup' of highly informative neurons whose activity carries a larger amount of information about either of these aspects. These neurons would then

drive neuronal processing in downstream areas and ultimately target choice (for examples of such highly informative neurons, see e.g. [60, 61]). To examine whether such highly informative neurons existed in this context, we removed one neuron at a time from the data, and quantified whether this reduced single-trial correlations to and specificity for the correct stimulus template. The contributions of individual neurons to the overall template match in a given trial were typically equally low, and there was no distinct outlier group of neurons that boosted single-trial correlations or their stimulus specificity – in either hit or miss trials (Fig. S5). In a few areas (e.g. MOs and VISp), there seemed to be at least some neurons that contributed more substantially to the template correlations, however this did not translate to response specificity: For most areas' correlations and all areas' specificity indices, a roughly equal number of neurons were contributing to and subtracting from the match of the single-trial response to the correct template (as measured by the proportion of data points above and below zero).

It thus seems that single-trial responses are less correlated to the average than a bootstrapped version of it, and that they are only slightly predictive of subsequent behaviour, in only a few cortical areas. However, the available information might be more than enough to generate accurate perceptions and behaviour when scaled up to the number of neurons actually present in a local circuit. To explore this possibility, we first sub-sampled the population of recorded neurons in each cortical area at 10 different levels from $N/10$ to N. From these sub-samplings, we extrapolated how metrics like the Specificity Index would evolve as the number of available neurons grew. As shown in Figure 5A, single-trial correlations to the response template tended to grow with sample size, suggesting that in a realistic population sampled by a downstream neuron (e.g. 30.000 inputs), template matching would be quite strong. However, correlations to the correct and incorrect template appeared to grow at the same rate, so that the resulting correlations would be high but not stimulus-specific (Fig. 5A). This is also borne out by the development of the Specificity Index with sample size: With growing N, specificity remains largely constant in some areas (e.g. MOs and SPF), and actually declines in many others, including lateral and primary visual cortex (VISl and VISp), red nucleus (RN) and substantia nigra (SNr) (Fig 5B). Moreover, neither single-trial correlations nor specificity showed a tendency to become more predictive of behaviour with larger sample sizes (Fig. 5C,D). With the exception of RN, the difference in single-trial correlations between hit and miss trials remained constant with growing n. The difference in response specificity between hit and miss trials actually tended to decrease (e.g. in MOs, RN and SNr). In other words, the single-trial match to the average template did not become more indicative of subsequent behavioural choices with larger neuron numbers.

Together, these results suggest that the relation between single-trial population responses and their trial-averaged response templates is both less strong and less stimulus-specific than what one would expect if single-trial responses were simply a down-sampled representation of the average. Most importantly, single-trial responses that better resembled the correct time-averaged template did not evoke better target choices. This suggests that if 'average template matching' is part of neuronal processing in this context, it happens in a non-linear and/or multi-dimensional way that is not captured by simple correlations. To take a first step at exploring this possibility, we repeated the analyses shown in Figures 3-5 by characterizing population responses using Principal Component Analysis (PCA) via Singular Value Decomposition (SVD), and quantifying their resemblance to an average template in this dimensionally-reduced space rather than by linear correlation (Figs. 6-7).

The resulting single-trial response vectors did overall not represent the corresponding average vectors more effectively than the linear averages we had explored previously. In PCA space, single-trial vectors matched average vectors more closely than would be predicted from bootstrapping (Fig. 6A), but the match was nevertheless weak: The

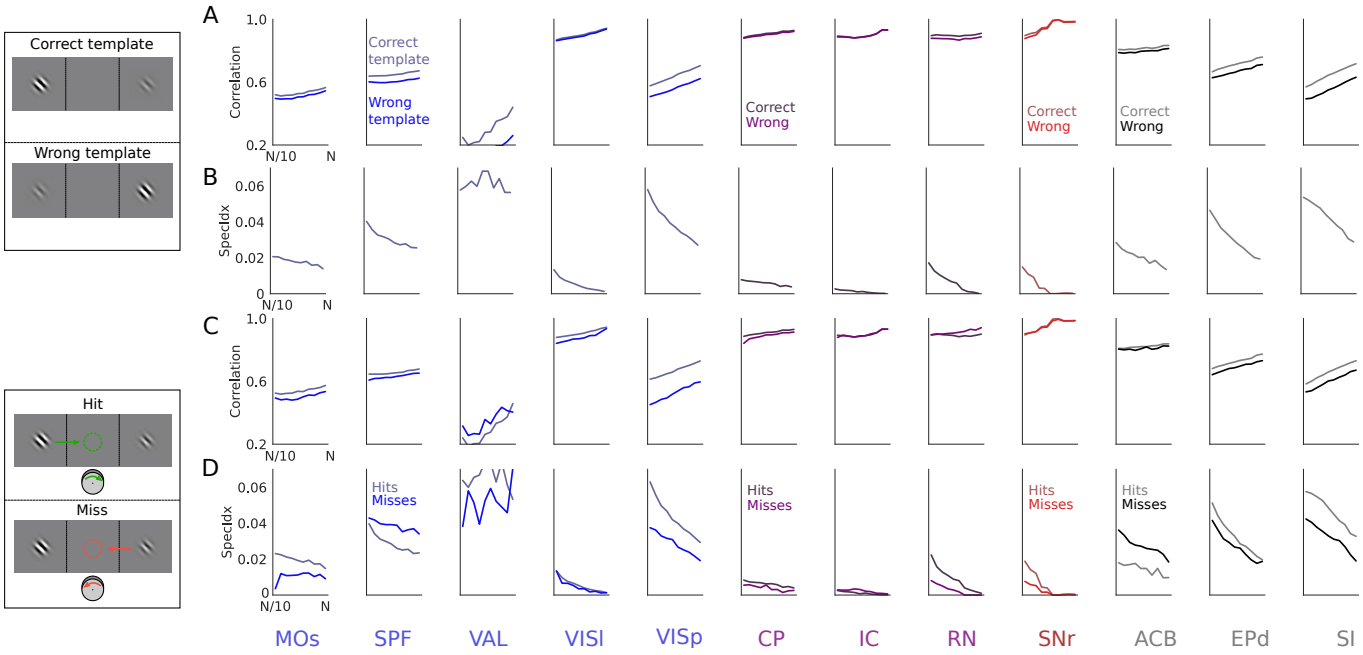

FIG. 5. (Color online) Extrapolation of single-trial correlations and specificity index across different sample sizes. A) As we increase the number of sampled neurons, single-trial responses do not get more stimulus-specific, given that the gap between curves does not increase. B) Specificity index. As expected from the previous panel, the specificity index is always around 0; moreover, when increasing the number of sampled neurons, we see a tendency towards less stimulus-specificity. C) Correlation, split over hits and misses. Sampling more neurons does not make single-trial responses differentially more correlated to their template in hit than in miss trials. D) Specificity index, split over hits and misses. The specificity index is always around 0, for both hits and misses, and the stimulus specificity of the single-trial responses seems to decline when the number of sampled neurons increases.

distance between a single-trial vector and its corresponding average template was typically $5-10$ times larger than the distance between correct and incorrect template. Consistently with this, single-trial responses were largely not stimulus-specific, in fact the specificity of single-trial responses for the correct average template was even lower in PCA space than for linear correlations (Fig. 6C; t-test for difference from Zero: $n = 90$ to 3123; $t = 0.6$ to 17.0; $p < 0.01$ except for $p(EPd) = 0.04$ and $p(SI) = 0.56$, corrected for multiple comparisons using a FDR procedure with a family-wise error rate of 0.05). Specificity indices were clustered tightly around zero, and never exceeded a value of 1. This is particularly remarkable because unlike correlation coefficients, which are bounded between 1 and -1, PCA vector distances are not upper-bounded and often took on values between 5 and 10 (Fig. 6A-B). Given these values, specificity indices < 1 imply negligible differences between the single-trial distances to correct and incorrect templates, respectively. Single-trial responses that were more similar and/or specific to the correct average vector also had only a marginally higher chance of resulting in correct behavioural choices (Fig. 6D-E; Mann-Whitney's U-test for differences in single-trial distances in hit and miss trials: $n = 90$ to 3123, Cliff's δ between 0.38 and 0.55; all $p < 0.01$; Mann-Whitney's U-test for differences in single-trial specificity in hit and miss trials: $n = 90$ to 3123, Cliff's δ between 0.38 and 0.61; all $p < 0.01$; corrected for multiple comparisons). Finally, unlike the linear averages probed before, the information conveyed by average vectors in PCA space seemed to at least somewhat profit from increased neuron numbers, but only in very specific cases. Single-trial vectors generally did not match average vectors

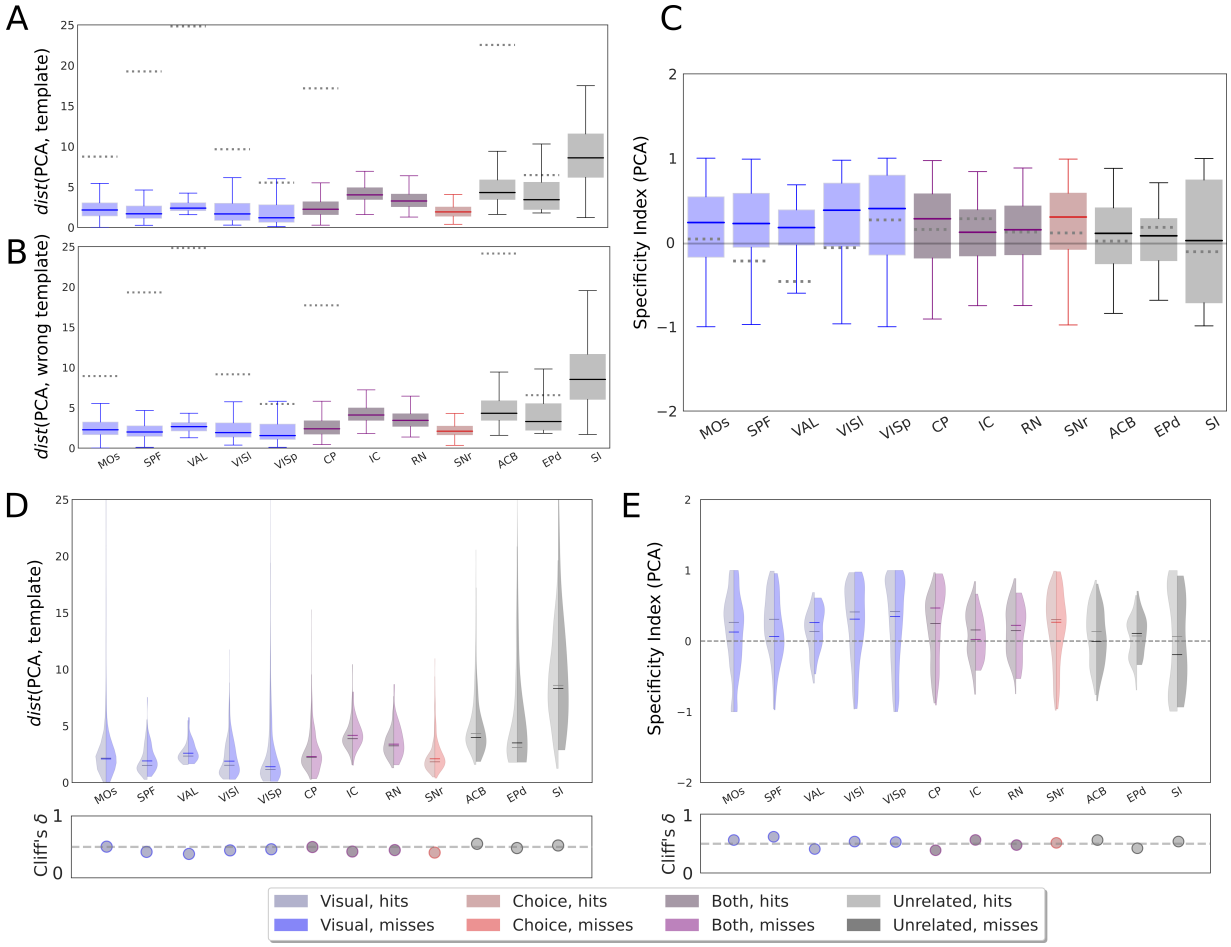

FIG. 6. (Color online) Template matching using PCA instead of Pearson correlation. A) Distance of single-trial response vectors to the correct average template. For interpretability, distances are normalized by the distance between the two average templates. B) Same as A, but for the incorrect template. C) Specificity index, computed for each trial as the difference between the distance to correct and incorrect template, respectively. D) Distribution of single-trial distances to the correct average template in hit and miss trials, respectively. Horizontal line: Median. Colors: see inset legend. E) Same as D for the Specificity index.

better with more neurons (Fig. 7A,C), but response specificity seemed to increase at least in some cortical areas, particularly Substantia Nigra (SNr) and to a lesser extent the Red Nucleus (RN), Inferior Colliculus (IC) and primary visual cortex (VISp). These areas showed not only increased specificity with larger neuronal populations (Fig. 7B), but also at least slightly higher response specificity in hit than miss trials (Fig. 7D). Note however that even with these improvements, the Specificity Index never exceeded 1, still pointing to low overall response specificity (see above). The remaining cortical areas did not seem to undergo any improvement in specificity with increased neuron numbers (Fig. 7B), and also showed no and/or inconsistent differences in specificity between hit and miss trials (Fig. 7D). Overall, these results demonstrate that just like for linear correlations, resemblance of single-trial to average vectors in PCA space did not seem to drive neuronal processing in a decisive way across most cortical areas, with the possible exception of Substantia Nigra and Red Nucleus. However, since PCA is a linear method too, this still leaves open the possibility that non-linear methods may reveal accurate template matching of single trials.

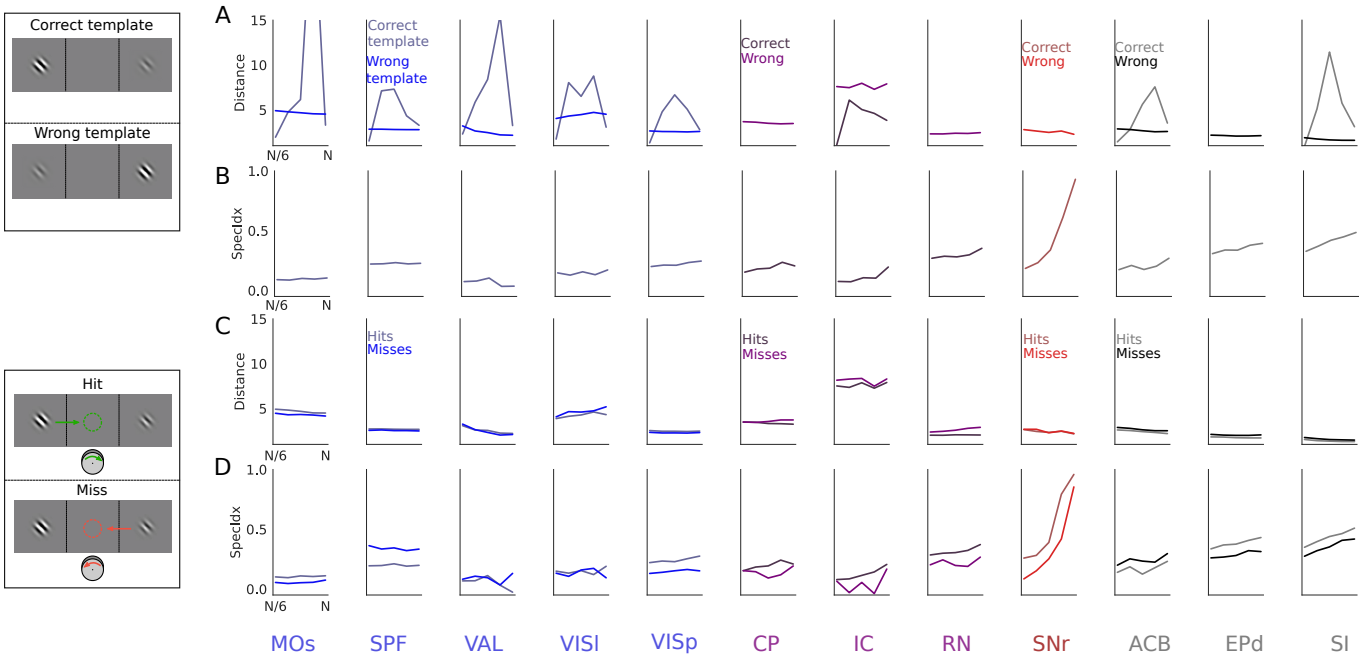

FIG. 7. (Color online) Extrapolation of single-trial distances in PCA space and specificity indices across different sample sizes. A) Distance between single- trial vectors and correct and incorrect template over increasing sample sizes. B) Specificity index computed over increasing sample sizes. C) Distance between single-trial response vectors and correct average template in PCA space, split for hit and miss trials. D) Same as C for specificity index.

Discussion

The present study set out to formally test the implicit assumptions we make when computing average population responses. Specifically, if average population responses are informative to the brain, single-trial responses should be sufficiently reliable and specific to be matched to the correct percept's population template, and single-trial responses more similar to the template should evoke more efficient behaviour. We find that these two assumptions are only fulfilled to a very limited extent in the data set examined here: Single-trial responses were reliably correlated to the average template - although less so than would be expected if they represented a discretized version of the population template. However, these correlations showed very low stimulus specificity, meaning that a large part of their reliability was likely based on stimulus-independent factors such as the baseline firing rate of different neuron types. Correcting for such differences in firing rate, as is done automatically by the PCA, did not improve the stimulus specificity of the correlations. In addition, single-trial responses that better resembled the correct population template hardly increased an animal's chance of making the correct target choice. Further analyses indicated that these results would not improve for a larger number of neurons, and were only marginally improved for specific brain areas (Substantia Nigra and Red Nucleus) when applying dimensionality reduction techniques (specifically PCA) before quantifying the match between single trials and average response vectors. This suggests that if the brain uses average population responses as a template, at least in the context of the given data set, this is not the central mechanism driving perceptual decision making.

Given that these conclusions are based on one specific data set and one specific set of analyses, one can envision several caveats, most of which pertain to the appropriateness of the average response template. First, given the nature of the behavioural task employed here, the most relevant stimulus information (the contrast difference between the two

presented stimuli) is computed by comparing information across two cortical hemispheres, but we only have access to neuronal recordings from one hemisphere. Thus, if we recorded neuronal responses from both hemispheres, population templates might become more stimulus-specific and informative. While this is likely true for some cortical areas, the fact remains that the animal is able to make largely correct perceptual decisions in this task, which means that even if specific cortical areas (e.g. primary visual cortex) did not contain information about the stimulus comparison between the two hemifields, some downstream area should receive the result of this cross-hemisphere computation in order to initiate the correct behavioural response. Since this data set is arguably the most complete set of neuronal recordings to date regarding the number of cortical and subcortical areas covered, it seems unlikely that across all these recorded areas, there is not a single one that consistently represents the integrated stimulus information of the two hemifields.

Another limitation of our average population template may be that animals were presented with 16 different contrast combinations, which we have pooled into only three stimulus categories (higher-contrast stimulus on the right or the left, and no stimulus on either side). Thus, if there were enough trials to compute average response templates for each specific stimulus pair, the measured single-trial correlations to the template as well as their stimulus specificity might be higher. While this is certainly possible, the fact that the average responses for each of the pooled stimulus pairs were highly correlated to each other (see Fig. S3) would suggest that the precision lost by pooling across stimulus pairs is largely negligible. In addition, since the stimulus categories we applied are congruent with the target choice the animal needs to make (i.e. choose whether to turn the wheel to the left or right), this global information (which hemifield contains the higher-contrast stimulus, irrespective of exact contrast) should be reflected in at least some of the recorded cortical areas in order to drive the behavioural response – an assumption which is also borne out by the decoder analysis shown in Figure 2.

A third potentially important factor is our choice to analyse neuronal responses within a time window of $200ms$ post stimulus onset. We chose this analysis window to largely exclude neuronal signals directly related to licking activity, since the majority of licks typically happened after $200ms$, with the response peak occurring at a delay of $520ms$ (see Fig. S1). In this way, our aim was to focus on the decision process that leads up to the behavioural response, rather than the response itself. Nevertheless, it is possible that a different analysis window would highlight different and/or more task-related information across the recorded cortical areas.

Most importantly, while the results obtained in this data set suggest a very limited utility of trial-averaged population responses for neuronal processing, these results may not hold for other cases. It is very possible that time-averaged response templates are much more relevant to neuronal computations given different behavioural contexts, stimulus structures or even species – or different metrics of neuronal activity that are being extracted and averaged over time [44]. Similarly, in this context we chose PCA as a benchmark of dimensionality-reduction techniques due to its relative simplicity and ubiquitous use, but other approaches might in principle yield improved results. For instance, non-Negative Matrix Factorization [62], though computationally more demanding and less widely used, might outperform PCA because it defines neuronal ensembles in a sparser, and therefore more realistic, way [63].

We would therefore encourage other researchers to run a simple 'rule-of-thumb' test like the one presented here on their data in order to gain an estimate of how crucial average population templates might be to the neuronal computations they are studying – and ideally choose their analysis approach based on that estimate. Over time, this might allow the neuroscience community to put together a 'map' of contexts in which averaged responses are more or less informative to the brain, similar to the emerging map of cortical areas in which neuronal responses show more or less representational drift over time [34, 35, 64, 65].

In statistics, it is common practice to explicitly test whether the assumptions (e.g. normal distribution) of a particular analysis (e.g. ANOVA) are fulfilled in a data set. We would argue in favour of a similar approach when it comes to the average metrics of neuronal activity typically applied in neuroscience: Population response vectors, tuning curves, PSTHs etc. should ideally come with a simple metric (like the Specificity Index) that represents an estimate of how likely the information they convey is to be informative to the brain, rather than only to the reader.

To facilitate this, we have kept the computational tools employed here purposefully simple and general, by utilizing mainly linear correlations. This aims to ensure that the analyses presented here provide an easy-to-use and intuitively interpretable way of estimating the relation between single-trial responses and time-averaged response templates. Second, since typical metrics of neuronal activity such as tuning curves, receptive fields and PSTHs do in fact rely on simple averaging, our test is designed to directly determine if these common metrics can be meaningfully applied to a specific data set. This does not exclude the possibility that our estimate is missing out on higher-order relations between single-trial and average population responses, which cannot be captured in simple correlations. As we have shown, this possibility can at least be excluded for relations that can be revealed by PCA, but other analyses may uncover strong and behaviourally meaningful links between individual and averaged responses on a more complex level.

Since the classical trial-averaged responses tested here appear largely irrelevant to ongoing neuronal computations at least in this particular context, how then could stimulus and target choice information be encoded? First, the stimulus-related response profiles explored here may underestimate the computational power of average responses by ignoring modulating factors: Neuronal responses in every cortical area are likely shaped by many task-related and task-unrelated factors at any moment in time [12, 28, 29, 54, 56–59, 66–69], only some of which will be accessible to the experimenter. This can make neuronal responses appear highly unpredictable, when they are in fact shaped systematically and reproducibly by a set of unmeasured, or 'latent', variables. In principle, downstream neurons may be able to disentangle these factors and dissect out e.g. stimulus-related information from the representation of other variables. Thus, while the simple population averages tested here may not appear particularly informative, other approaches considering joint neuronal response profiles for multiple factors, potentially including non-linear interactions between them, might be more successful at teasing out reliable information from trial-averaged templates. If this were the case, then we would suggest that the neuroscience community should abandon single-feature response averages in favour of multi-feature response averages. This would likely involve finding routine metrics to track ubiquitous latent variables like behavioural state [56, 58, 70–72] throughout a wide range of experiments.

However, it is also possible that trial-averaged templates, whether single- or multi-feature, are simply not the best way to represent neuronal information. Several recent papers have argued that factors such as stimulus properties, behavioural choices, and retrieved memories are encoded along largely orthogonal dimensions in neuronal response space [15, 18, 19]. If this is true and trial-averaged responses are informative along these different dimensions, then our PCA approach would be expected to more successfully retrieve e.g. stimulus identity from averaged neuronal population vectors by dissociating it from the response profiles related to other, orthogonally coded, factors. We show here that this was largely not the case.

This leaves several alternatives. First, information may be encoded mostly in joint neuronal dynamics that are only captured very imperfectly by static (single-or multi-feature) response preferences. Analysis approaches that take into account such dynamics, e.g. by tracking and/or tolerating ongoing rotations and translations in neuronal space [72–78] or by explicitly including shared variability in their readout [79–81] seem to generally fare better in capturing

robust features of neuronal coding. Even though such non-static approaches do not always uncover consistent neuronal representations across all cortical areas [34], they often provide vastly more informative and stable representations of neuronal activity despite seeming variability [73–75, 81]. Consistently with this, the decoder approach used in Figure 2 extracted information more successfully than the average templates derived from the same data – most likely because decoders build their predictive power on co-variability and co-dependences between the input data and the class labels, which are smoothed over when averaging across trials.

Finally, it is also possible that highly informative aspects of neuronal activity might not be captured by population response vectors at all, whether single-trial or trial-averaged. For instance, transient phase relationships between neuronal sub-populations [82–84] or the relative timing of action potentials [85, 86] will not be reflected in overall population responses. No matter which of these approaches turns out to be most successful, it is important to recognize that time-averaged population responses may at least in some contexts not be a fitting way to describe how information is represented in the brain.

Conclusion In this study, we present a simple analysis that can be used to determine whether trial-averaged population responses are likely to be relevant to the neuronal computations under study - or not. We apply this analysis to a publicly available data set containing electrophysiological recordings from a large number of cortical areas in behaving mice [44] and show that in this data set, average population responses seem to be largely irrelevant to perceptual decision making. Even in cortical areas that carry stimulus and/or target choice information, the relation between single-trial and trial-averaged population responses reflected neither stimulus nor target choice reliably. This fits with studies [34, 54, 64, 75] showing that in many contexts, neuronal responses spontaneously shift over time. In such (and other) instances, a static average taken across time is a very imprecise way of representing the ongoing neuronal computations. In other contexts, trial-averaged responses may be a much more meaningful representation of ongoing neuronal responses. We encourage other researchers to apply the analysis presented here or similar analyses on their own data sets. While trial-averaged metrics such as receptive fields, orientation preferences, or PSTHs can be a useful tool to summarize neuronal responses in a clear-cut way, it is important to know whether these metrics are mainly a shortcut for us neuroscientists, or whether we also expect the brain to make use of them to convey information.

Acknowledgements

We thank Jonathan Pillow and Mike X Cohen for valuable input on earlier versions of the manuscript.

METHODS

We have released all the scripts and data files to reproduce these analyses, they can be found at the following URL: https://github.com/atlaie/BrainAveraging. They are written in Python 3 and leverage on several libraries.

Decoder

We trained a multinomial Generalized Linear Model (GLM) using the SciKit-Learn package in Python [87]. In order to avoid overfitting, we introduced a L2-regularization. If we have K classes in which we want to classify our label data, this model states that the probability of a particular data point y_i belonging to class c is dependent of the input data x_i and the bias for that class (b_c) takes the form of:

$$p\left(y_i = c \mid x_i\right) = \frac{e^{w_c \cdot x_i + b_c}}{\sum_{j=1}^{K} e^{w_j \cdot x_i + b_j}} \tag{1}$$

After having the probabilities of y_i belonging to each class c, the highest one will be taken to be 1 and the rest will be set to 0. Therefore, the objective is to find the weight vector w_c that minimizes the distance between the predicted ($\hat{y}_i$) and the actual (y_i) class labels by optimizing (in this case, minimizing) the following loss function:

$$L\left(\hat{y}_i, y_i\right) = -\log\left(\frac{e^{w_c \cdot x_i + b_c}}{\sum_{j=1}^{K} e^{w_j \cdot x_i + b_j}}\right) + \lambda \left\|w_c\right\|_2^2 \tag{2}$$

where $\|w_c\|_2^2$ is the L2-norm of the weight vector for class c , accounting for the L2-regularization term – with the hyperparameter λ modulating its strength.

For each experimental session, there are several recorded regions. Thus, we trained independent decoders using the single-trial population vector for each region. The labels to be predicted would be either choice (left wheel turn, right wheel turn or no movement) or stimulus (right-higher contrast, left-higher contrast, both equal). We split the data following a 80-20 ratio (train-test) and, given the imbalanced nature of the dataset, we used an stratified 10-repeated 5-fold Cross-Validation approach. We then performed hyperparameter optimization via a greedy algorithm (grid search) and checked that the model performance (Accuracy and LogLoss score) was above chance and above majority class (i.e., always predicting the most abundant label) and random models.

We then computed the Mutual Information between the predicted and the test class labels, as a proxy of the amount of stimulus – or choice – information there was in the population vector.

Mutual Information

This quantity is defined in the context of classical Information Theory [88, 89]. We can compute it for two discrete stochastic variables X and Y. Assuming these have a joint probability mass function given by $p_{X,Y}(x, y) = P(Y = y \mid X = x) \cdot P(X = x)$ and that each of them follows a marginal probability distribution given by $p_X = \sum_{y \in Y} p_{X,Y}(x, y)$, one can mathematically define the Mutual Information between X and Y as:

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(X,Y) \log\left(\frac{p_{X,Y}(x,y)}{p_X p_Y}\right) \tag{3}$$

Intuitively, one can understand $I(X;Y)$ as the uncertainty reduction in X that follows if Y is measured (or vice versa, as $I(X;Y)$ is invariant when swapping X and Y). If (and only if) they are independent of each other, then $I(X;Y) = 0$. Therefore, this is a strictly non-negative quantity. It is noteworthy that $I(X;Y)$ captures all linear and nonlinear dependencies between X and Y, thus generalizing the notion of correlation measures. For further discussion of this measure, see [90, 91].

Elbow method

In order to select a threshold when selecting the task-related areas based on their stimulus and choice information (Figure 2 in the main text), we used the data to compute the Kernel Density Estimate, via Gaussian kernels [92]. After having extracted these, we used the method discussed in [93] to find the point of maximum curvature. We made use of the kneed Python package, implemented by the same authors [93].

Surrogate models

We were interested in comparing the experimental neuronal population response with a downsampled version of the trial-averaged template. To do that, we built our surrogate models by constructing $N(=100)$ random vector with the following constraints:

1. Its size is equal to the number of neurons comprising the neural population for that area and that session.
2. The probability that at n spikes are allocated at a particular location m (i.e., that neuron m has spiked n times) is given by $P_{m,n} = \left(\frac{\lambda_m}{\sum_m \lambda_m}\right)^n$, where λ_m is the mth element of the template vector.
3. The total number of spikes is constant and equal to the total recorded number of spikes for that area and that session.

By imposing these constraints, we are testing the alternative hypothesis that neurons are independent from each other (uncorrelated) and it is therefore equivalent to keeping the single-trial population statistical response, while scrambling across trials. This is also the same as drawing single-neuron responses from the underlying template distribution following a Poisson process.

Specificity index

With the intent of characterizing whether the neural response is more similar to the appropriate template (i.e., the one corresponding to the stimulus that was actually presented in that trial) or the other one, we introduced a simple quantity we termed specificity index. It is defined as:

$$\rho_i = cor(\lambda_{appropriate}, r_i) - cor(\lambda_{wrong}, r_i) \tag{4}$$

where *cor* is the Pearson correlation, λ denotes a given neural template and r_i is the population vector of the ith trial. Thus, the specificity index captures the differential similarity of a given neural response to each of the templates. It is key to note that, given that the Pearson correlation is bounded between -1 and 1, the specificity index can attain values between -2 and 2 and, as we were just interested in its sign and global tendencies, we did not introduce any normalization factor.

Cliff's δ

As a way to quantify the overlap between distributions (for example, correlation in hit vs miss trials) we relied on Cliff's δ [94]. Cliff's δ is an effect size derived from the Mann-Whitney U-test –a non-parametric statistical test that is particularly useful when distributions are not Gaussian [95]. Furthermore, Cliff's δ is especially interpretable. It can be thought of as the probability of a randomly selected point from one distribution being higher than another randomly selected point from the other one. Mathematically, if we have two distributions A and B, the U-statistic is given by:

$$U = \sum_{i=1}^{a} \sum_{j=1}^{b} S\left(A_i, B_j\right) \tag{5}$$

with a and b being the number of elements of A and B, respectively; and

$$S\left(A_i, B_j\right) = \begin{cases} 1, & \text{if } A_i < B_j \\ 1/2, & \text{if } A_i = B_j \\ 0, & \text{if } A_i > B_j \end{cases} \tag{6}$$

Having computed the U-statistic, Cliff's δ is given by normalizing it as

$$\delta = \frac{U}{ab} \tag{7}$$

Thus, it is bounded between 0 and 1. If A and B are maximally overlapping, $\delta = 0.5$; if there is no overlap, $\delta = 1$ (or 0 if we take the U-test in the reverse direction). Therefore, the more its value deviates from 0.5, the less overlapping the distributions are.

Templates and distances in PCA space

As an alternative to Pearson's correlation, we applied Principal Component Analysis (PCA) [96]. We have chosen PCA over non-Negative Matrix Factorization [97] or other more advanced dimensionality reduction techniques such as LFADS [79] or PSID [72] because we wanted to keep all analyses as general as we possibly could. Thus, we compute the truncated Singular Value Decomposition (tSVD) [98] for the matrix consisting on Z-scored single-trial population vectors, for a given area and session. Then, we extract the knee (elbow) using the aforementioned method, to select the number of components based on the variance explained. After the number of components has been selected, we project each single-trial into this (dimensionally-reduced) space and compute the Euclidean distance between this new vector and the template (also projected into this space). We normalize by the distance between the projection of the two templates in this new space. For the subsampling analyses, the only thing we require is that the dimensionality of the extracted subspace should be larger than one. That explains the discrepancy between the correlation subsampling range (from $N/10$ to N) and this other one (from $N/6$ to N).

Specificity Index for PCA distances

Since in PCA analyses we dealt with distances rather than correlations (i.e., differences rather than similarities), we inverted the computation of the Specificity Index in this context so that positive values continued to signify a stronger relation between single trial response and correct template than incorrect template. The corresponding formula is:

$$\rho_i^{PCA} = d(\lambda_{wrong}^{PCA}, r_i^{PCA}) - d(\lambda_{appropriate}^{PCA}, r_i^{PCA}) \tag{8}$$

where d stands for Euclidean distance and r_i^{PCA} is the PCA-projected version of the population vector measured in the ith trial; λ_{wrong}^{PCA} and $\lambda_{appropriate}^{PCA}$ are the PCA-projected version of the trial-average templates (wrong and appropriate, respectively).

[1] K. H. Britten, M. N. Shadlen, W. T. Newsome, and J. A. Movshon, J. Neurosci **12** (1992).

[2] W. T. Newsome, K. H. Britten, and J. A. Movshon, Nature **341**, 52 (1989).

[3] G. B. Keller, T. Bonhoeffer, and M. Hübener, Neuron **74**, 809 (2012).

[4] L. Busse, A. R. Wade, and M. Carandini, Neuron **64**, 931 (2009).

[5] I. Nauhaus, A. Benucci, M. Carandini, and D. L. Ringach, Neuron **57**, 673 (2008).

[6] J. Poort, A. G. Khan, M. Pachitariu, A. Nemri, I. Orsolic, J. Krupic, M. Bauza, M. Sahani, G. B. Keller, T. D. Mrsic-Flogel, *et al.*, Neuron **86**, 1478 (2015).

[7] D. Kvitsiani, S. Ranade, B. Hangya, H. Taniguchi, J. Huang, and A. Kepecs, Nature **498**, 363 (2013).

[8] V. Dragoi, J. Sharma, and M. Sur, Neuron **28**, 287 (2000).

[9] L. Pinto and Y. Dan, Neuron **87**, 437 (2015).

[10] S. El-Boustani and M. Sur, Nature communications **5**, 1 (2014).

[11] E. M. Diamanti, C. B. Reddy, S. Schröder, T. Muzzu, K. D. Harris, A. B. Saleem, and M. Carandini, Elife **10**, e63705 (2021).

[12] R. N. Ramesh, C. R. Burgess, A. U. Sugden, M. Gyetvan, and M. L. Andermann, Neuron **100**, 900 (2018).

[13] P. Bao, L. She, M. McGill, and D. Y. Tsao, Nature **583**, 103 (2020).

[14] M. J. Goard, G. N. Pho, J. Woodson, and M. Sur, elife **5**, e13764 (2016).

[15] S. W. Failor, M. Carandini, and K. D. Harris, bioRxiv (2021).

[16] M. C. Aoi, V. Mante, and J. W. Pillow, Nature neuroscience **23**, 1410 (2020).

[17] H. J. Ladret, N. Cortes, L. Ikan, F. Chavane, C. Casanova, and L. U. Perrinet, bioRxiv (2021).

[18] A. Libby and T. J. Buschman, Nature neuroscience **24**, 715 (2021).

[19] S. B. M. Yoo and B. Y. Hayden, Neuron **105**, 712 (2020).

[20] A. G. Khan, J. Poort, A. Chadwick, A. Blot, M. Sahani, T. D. Mrsic-Flogel, and S. B. Hofer, Nature neuroscience **21**, 851 (2018).

[21] L. Q. Uddin, Trends in Cognitive Sciences **24**, 734 (2020).

[22] M. Kafashan, A. W. Jaffe, S. N. Chettih, R. Nogueira, I. Arandia-Romero, C. D. Harvey, R. Moreno-Bote, and J. Drugowitsch, Nature communications **12**, 1 (2021).

[23] M. M. Churchland, M. Y. Byron, J. P. Cunningham, L. P. Sugrue, M. R. Cohen, G. S. Corrado, W. T. Newsome, A. M. Clark, P. Hosseini, B. B. Scott, *et al.*, Nature neuroscience **13**, 369 (2010).

[24] M. R. Cohen and J. H. Maunsell, Nature neuroscience **12**, 1594 (2009).

[25] E. Zohary, M. N. Shadlen, and W. T. Newsome, Nature **370**, 140 (1994).

[26] M. Gur and D. M. Snodderly, Cerebral cortex **16**, 888 (2006).

[27] N. Roth and N. C. Rust, Journal of neurophysiology **121**, 115 (2019).

[28] S. Musall, M. T. Kaufman, A. L. Juavinett, S. Gluf, and A. K. Churchland, Nature neuroscience **22**, 1677 (2019).

[29] C. Stringer, M. Pachitariu, N. Steinmetz, C. B. Reddy, M. Carandini, and K. D. Harris, Science **364** (2019).

[30] E. Y. Walker, R. J. Cotton, W. J. Ma, and A. S. Tolias, Nature Neuroscience **23**, 122 (2020).

[31] L. Waschke, N. A. Kloosterman, J. Obleser, and D. D. Garrett, Neuron (2021).

[32] M. Valente, G. Pica, G. Bondanelli, M. Moroni, C. A. Runyan, A. S. Morcos, C. D. Harvey, and S. Panzeri, Nature Neuroscience , 1 (2021).

[33] D. Festa, A. Aschner, A. Davila, A. Kohn, and R. Coen-Cagli, Nature Communications **12**, 1 (2021).

[34] C. E. Schoonover, S. N. Ohashi, R. Axel, and A. J. Fink, Nature , 1 (2021).

[35] M. E. Rule, T. O'Leary, and C. D. Harvey, Current opinion in neurobiology **58**, 141 (2019).

[36] D. Shimaoka, N. A. Steinmetz, K. D. Harris, and M. Carandini, Elife **8**, e43533 (2019).

[37] W. R. Softky and C. Koch, Journal of neuroscience **13**, 334 (1993).

[38] M. Carandini and C. Stevens, PLoS biology **2**, e264 (2004).

[39] N. T. Robinson, L. A. Descamps, L. E. Russell, M. O. Buchholz, B. A. Bicknell, G. K. Antonov, J. Y. Lau, R. Nutbrown,

C. Schmidt-Hieber, and M. Häusser, Cell **183**, 1586 (2020).

[40] C. D. Salzman, K. H. Britten, and W. T. Newsome, Nature **346**, 174 (1990).

[41] M. Jin and L. L. Glickfeld, Current Biology **30**, 4682 (2020).

[42] P. Zatka-Haas, N. A. Steinmetz, M. Carandini, and K. D. Harris, bioRxiv , 501627 (2021).

[43] K. H. Britten, W. T. Newsome, M. N. Shadlen, S. Celebrini, and J. A. Movshon, Visual neuroscience **13**, 87 (1996).

[44] N. A. Steinmetz, P. Zatka-Haas, M. Carandini, and K. D. Harris, Nature **576**, 266 (2019).

[45] C. R. Fetsch, N. N. Odean, D. Jeurissen, Y. El-Shamayleh, G. D. Horwitz, and M. N. Shadlen, Elife **7**, e36523 (2018).

[46] A. G. Ramayya, A. Misra, G. H. Baltuch, and M. J. Kahana, Journal of Neuroscience **34**, 6887 (2014).

[47] J. N. Reynolds, B. I. Hyland, and J. R. Wickens, Nature **413**, 67 (2001).

[48] K. A. Zaghloul, J. A. Blanco, C. T. Weidemann, K. McGill, J. L. Jaggi, G. H. Baltuch, and M. J. Kahana, Science **323**, 1496 (2009).

[49] G. A. Basile, M. Quartu, S. Bertino, M. P. Serra, M. Boi, A. Bramanti, G. P. Anastasi, D. Milardi, and A. Cacciola, Brain Structure and Function **226**, 69 (2021).

[50] R. Pacheco-Calderón, A. Carretero-Guillén, J. M. Delgado-García, and A. Gruart, Journal of Neuroscience **32**, 12129 (2012).

[51] J. A. Winer and C. E. Schreiner, in *The inferior colliculus* (Springer, 2005) pp. 1–68.

[52] W. E. Allen, I. V. Kauvar, M. Z. Chen, E. B. Richman, S. J. Yang, K. Chan, V. Gradinaru, B. E. Deverman, L. Luo, and K. Deisseroth, Neuron **94**, 891 (2017).

[53] M. Carandini, D. Shimaoka, L. F. Rossi, T. K. Sato, A. Benucci, and T. Knöpfel, Journal of Neuroscience **35**, 53 (2015).

[54] M. L. Schölvinck, A. B. Saleem, A. Benucci, K. D. Harris, and M. Carandini, Journal of Neuroscience **35**, 170 (2015).

[55] C. M. Niell and M. P. Stryker, Neuron **65**, 472 (2010).

[56] M. Vinck, R. Batista-Brito, U. Knoblich, and J. A. Cardin, Neuron **86**, 740 (2015).

[57] J. Fournier, A. B. Saleem, E. M. Diamanti, M. J. Wells, K. D. Harris, and M. Carandini, Current Biology **30**, 3811 (2020).

[58] W. E. Allen, M. Z. Chen, N. Pichamoorthy, R. H. Tien, M. Pachitariu, L. Luo, and K. Deisseroth, Science **364** (2019).

[59] A. Arieli, A. Sterkin, A. Grinvald, and A. Aertsen, Science **273**, 1868 (1996).

[60] C. Stringer, M. Michaelos, D. Tsyboulski, S. E. Lindo, and M. Pachitariu, Cell **184**, 2767 (2021).

[61] J. Pérez-Ortega, T. Alejandre-García, and R. Yuste, Elife **10**, e64449 (2021).

[62] J. K. Liu, H. M. Schreyer, A. Onken, F. Rozenblit, M. H. Khani, V. Krishnamoorthy, S. Panzeri, and T. Gollisch, Nature communications **8**, 1 (2017).

[63] E. A. Pnevmatikakis, D. Soudry, Y. Gao, T. A. Machado, J. Merel, D. Pfau, T. Reardon, Y. Mu, C. Lacefield, W. Yang, *et al.*, Neuron **89**, 285 (2016).

[64] T. D. Marks and M. J. Goard, Nature communications **12**, 1 (2021).

[65] J. Bauer and T. Rose, Current Biology **31**, R1129 (2021).

[66] A. Basole, L. E. White, and D. Fitzpatrick, Nature **423**, 986 (2003).

[67] A. Lak, M. Okun, M. M. Moss, H. Gurnani, K. Farrell, M. J. Wells, C. B. Reddy, A. Kepecs, K. D. Harris, and M. Carandini, Neuron **105**, 700 (2020).

[68] A. Fiser, D. Mahringer, H. K. Oyibo, A. V. Petersen, M. Leinweber, and G. B. Keller, Nature neuroscience **19**, 1658 (2016).

[69] J. M. Beck, W. J. Ma, X. Pitkow, P. E. Latham, and A. Pouget, Neuron **74**, 30 (2012).

[70] M. N. Havenith, P. M. Zijderveld, S. van Heukelum, S. Abghari, J. C. Glennon, and P. Tiesinga, Scientific reports **8**, 1 (2018).

[71] M. N. Havenith, P. M. Zijderveld, S. van Heukelum, S. Abghari, P. Tiesinga, and J. C. Glennon, Scientific reports **9**, 1 (2019).

[72] O. G. Sani, H. Abbaspourazad, Y. T. Wong, B. Pesaran, and M. M. Shanechi, Nature Neuroscience **24**, 140 (2021).

[73] G. Okazawa, C. E. Hatch, A. Mancoo, C. K. Machens, and R. Kiani, Cell **184**, 3748 (2021).

[74] J. Xia, T. D. Marks, M. J. Goard, and R. Wessel, Nat. Commun. 12 **12** (2021).

[75] J. A. Gallego, M. G. Perich, S. N. Naufel, C. Ethier, S. A. Solla, and L. E. Miller, Nature communications **9**, 1 (2018).

[76] D. Deitch, A. Rubin, and Y. Ziv, Current Biology **31**, 4327 (2021).

[77] E. Froudarakis, U. Cohen, M. Diamantaki, E. Y. Walker, J. Reimer, P. Berens, H. Sompolinsky, and A. S. Tolias, bioRxiv (2020).

[78] R. J. Low, S. Lewallen, D. Aronov, R. Nevers, and D. W. Tank, BioRxiv , 418939 (2018).

[79] C. Pandarinath, D. J. O'Shea, J. Collins, R. Jozefowicz, S. D. Stavisky, J. C. Kao, E. M. Trautmann, M. T. Kaufman, S. I. Ryu, L. R. Hochberg, *et al.*, Nature methods **15**, 805 (2018).

[80] J. S. Montijn, G. T. Meijer, C. S. Lansink, and C. M. Pennartz, Cell reports **16**, 2486 (2016).

[81] J. W. Pillow, J. Shlens, L. Paninski, A. Sher, A. M. Litke, E. Chichilnisky, and E. P. Simoncelli, Nature **454**, 995 (2008).

[82] T. Womelsdorf, J.-M. Schoffelen, R. Oostenveld, W. Singer, R. Desimone, A. K. Engel, and P. Fries, science **316**, 1609 (2007).

[83] U. Rutishauser, I. B. Ross, A. N. Mamelak, and E. M. Schuman, Nature **464**, 903 (2010).

[84] J. Duprez, R. Gulbinaite, and M. X. Cohen, NeuroImage **207**, 116340 (2020).

[85] M. N. Insanally, I. Carcea, R. E. Field, C. C. Rodgers, B. DePasquale, K. Rajan, M. R. DeWeese, B. F. Albanna, and R. C. Froemke, Elife **8**, e42409 (2019).

[86] M. N. Havenith, S. Yu, J. Biederlack, N.-H. Chen, W. Singer, and D. Nikolić, Journal of neuroscience **31**, 8570 (2011).

[87] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, the Journal of machine Learning research **12**, 2825 (2011).

[88] R. Q. Quiroga and S. Panzeri, Nature Reviews Neuroscience **10**, 173 (2009).

[89] C. E. Shannon, The Bell system technical journal **27**, 379 (1948).

[90] N. M. Timme and C. Lapish, eneuro **5** (2018).

[91] T. M. Cover and J. A. Thomas, Elements of Information Theory **1**, 279 (1991).

[92] B. W. Silverman, **26** (1986).

[93] V. Satopaa, J. Albrecht, D. Irwin, and B. Raghavan, in *2011 31st international conference on distributed computing systems workshops* (IEEE, 2011) pp. 166–171.

[94] N. Cliff, Psychological bulletin **114**, 494 (1993).

[95] H. B. Mann and D. R. Whitney, The annals of mathematical statistics , 50 (1947).

[96] K. Pearson, The London, Edinburgh, and Dublin philosophical magazine and journal of science **2**, 559 (1901).

[97] D. D. Lee and H. S. Seung, Nature **401**, 788 (1999).

[98] P. C. Hansen, BIT Numerical Mathematics **27**, 534 (1987).

SUPPLEMENTARY MATERIAL

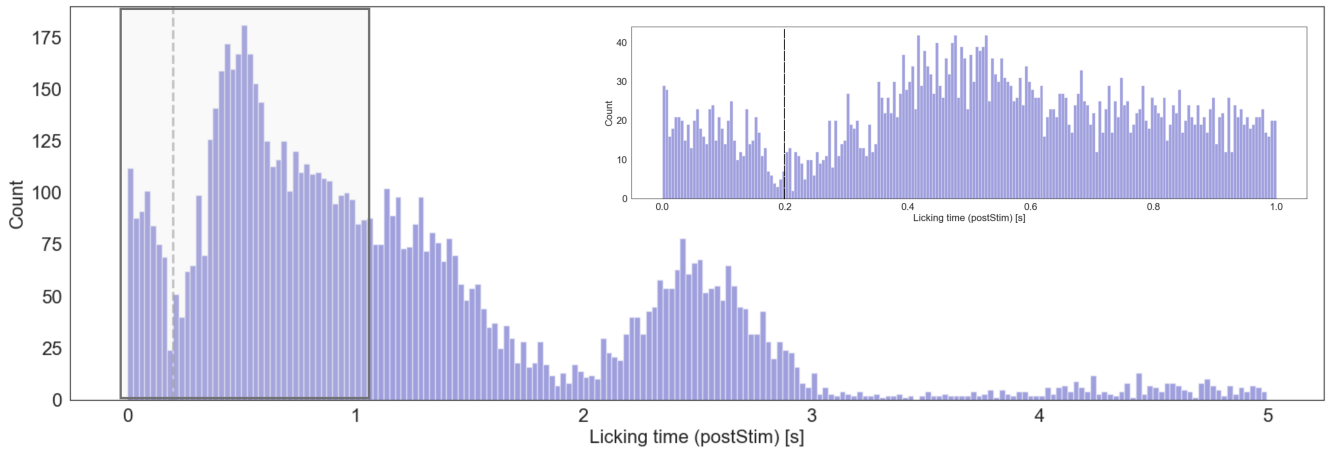

FIG. S1. (Color online) Post-stimulus licking time distribution. In the highlighted inset, we show licking times below a second after stimulus presentation. We have selected our analysis window of $200ms$ because of the relatively small number of lick events in that window (5% of all licks) and its likely relevance to stimulus processing and behavioural decision making.

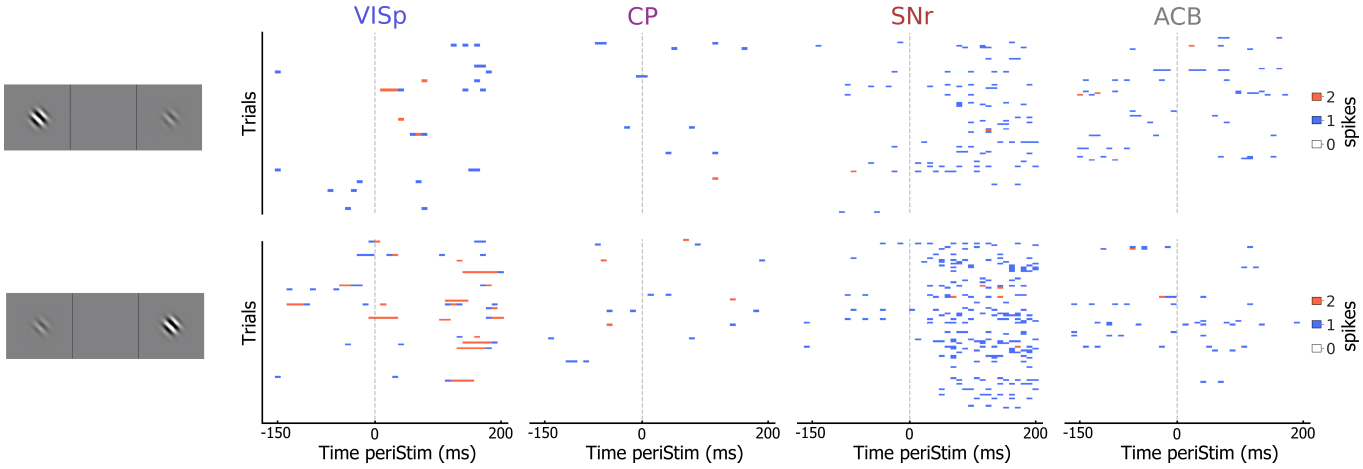

FIG. S2. (Color online) Example of neuronal responses to stimulus constellations with the target on the left and right, respectively. Responses are shown for one neuron each in four representative areas that are informative of the stimulus (VISp), the target choice (SNr), both (CP) and neither (ACB).

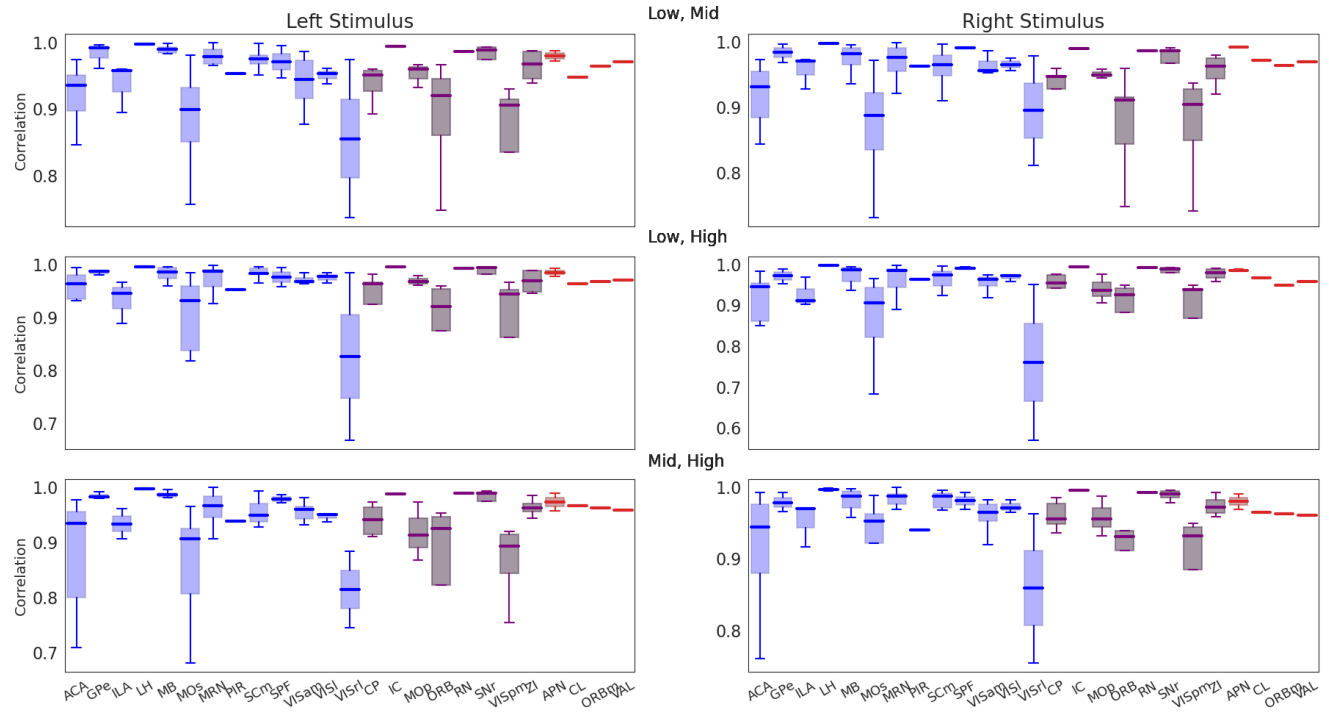

FIG. S3. (Color online) Correlations between response templates for different contrast combinations. These generally exceed 0.9, suggesting that one template should be sufficient to represent different contrast constellations.

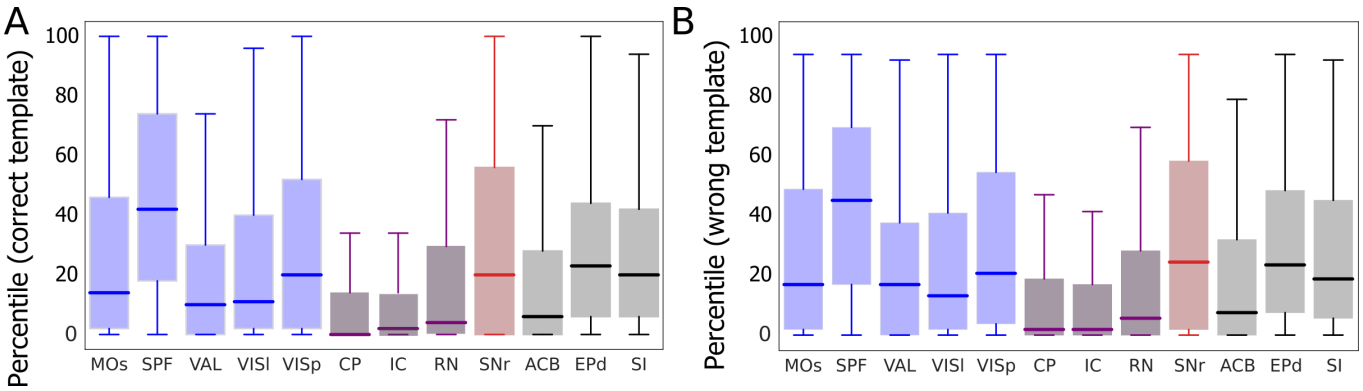

FIG. S4. (Color online) Percentile distributions for the correlation between single-trial responses and the correct (A) and incorrect templates (B). Box: 25th and 75th percentile. Solid line: Median. Whisker bars: 10th and 90th percentile. In both cases, these correlations lie well below the 95th percentile when compared to the surrogate distributions we computed (main text, Fig. 3A).

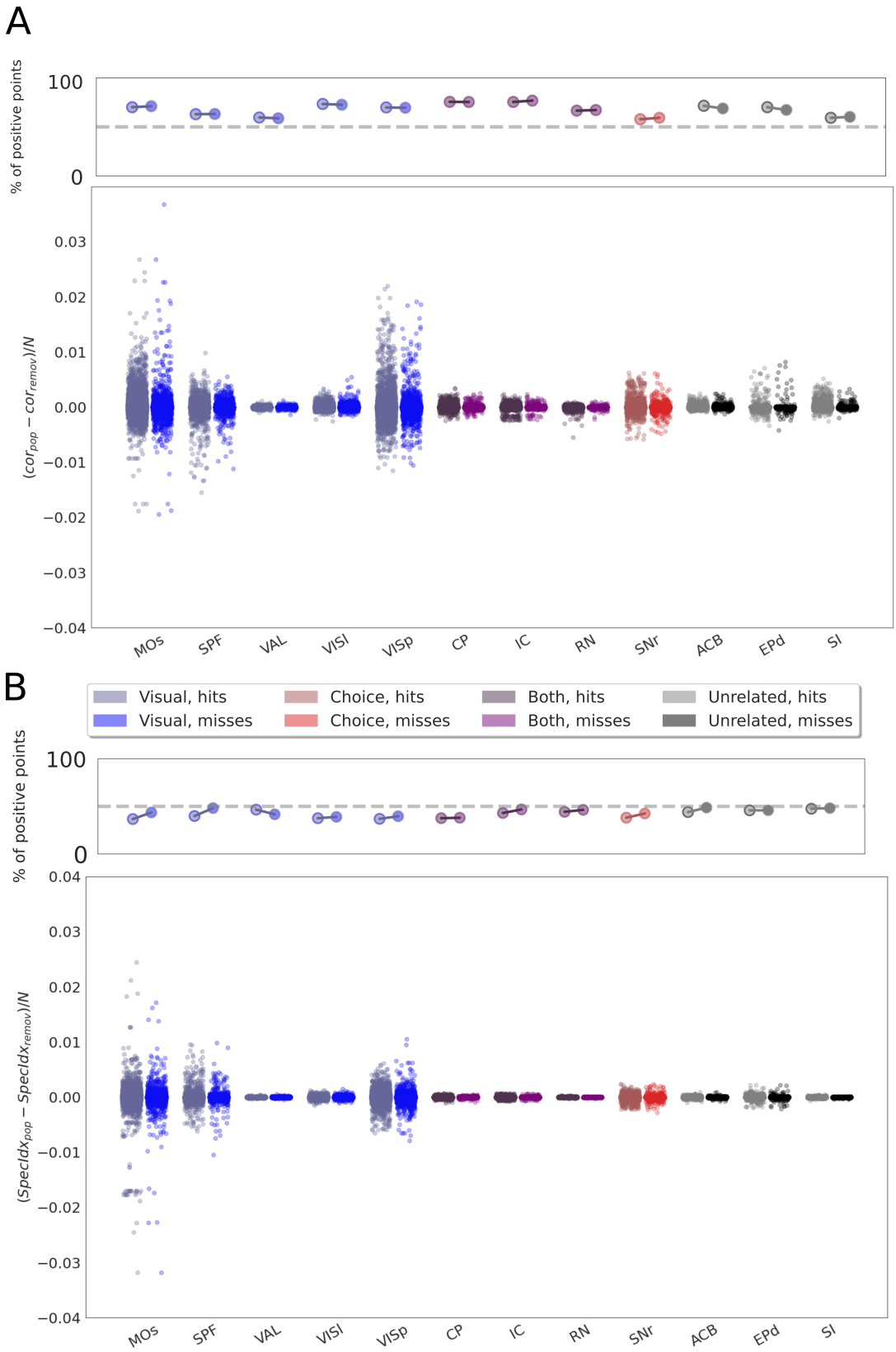

FIG. S5. (Color online) Neuron removal analyses. A) Change in single-trial correlations to the correct average template when one individual neuron was removed. Data points: Trials. Colors: see inset legend. B) Same for single-trial specificity.