Does the brain care about averages? A simple test

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

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

[2] *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*

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 trial that their average response template remains recognizable to downstream regions.
2. Behavioural relevance: If a single-trial response is more similar to the average template, it will be 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 71 brain areas in behaving mice. We show that single-trial responses were less correlated to the average response template than one would expect if they represented discrete versions of the template, down-sampled to the number of spikes in the single trial. 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 more accurately predict behaviour for any of the recorded brain 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.

* These authors contributed equally.

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 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 characterizing the general response dynamics of neuronal networks, it remains an outstanding issue whether trial-averaged population activity is used to transmit 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' 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 an unlikely 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] - of typical 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 (Fig 1A):

1) The responses of task-relevant neuronal populations are reliable – they repeat consistently enough across single trials that the information they carry remains recognizable to downstream regions. 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 71 brain areas recorded during a perceptual decision task. After identifying which areas are most relevant for this task, we use simple tests to compare single trial activity in these areas to the average response templates. We show that in this data set, neither of the

two assumptions set out above is fully met.

Results

The data set we use to test our two assumptions contains high-density electrophysiological (Neuropixel) recordings across 71 brain regions in mice performing a two-choice contrast discrimination task [44]. In the task, animals are presented with two gratings of varying contrast (0, 25, 50 or 100%) appearing in their left and right hemifield. 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 (Fig. 1B). When both stimulus contrasts were equal, animals were randomly rewarded for turning right or left. Those trials were discarded in the current analysis since there is no 'correct' behavioural response in this context. Neuronal recordings were obtained from 71 cortical and subcortical targets (Fig. 1C).

To first establish which brain 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) based on single-trial population vectors, 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). For the response vectors, we took into account neuronal activity $0 - 200ms$ post stimulus onset (see Figs. S1 and S2 for a rationale of this choice and examples of neuronal activity during this time window). We then 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 each area.

Many brain 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 (Methods) to determine a threshold for selecting brain 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 areas that contained predominantly stimulus information, one area that contained mainly choice information and three areas that contained both. These areas seem largely congruent with the literature: latero-intermediate visual cortex (VISl) and primary visual cortex (VISp) would be expected to contain visual stimulus information, while 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]; and information about both stimulus and choice is reflected in the Red Nucleus (RN), which is involved in coordinated paw movement [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 brain areas [14, 44, 52].

Having identified the most task-relevant brain 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 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 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,

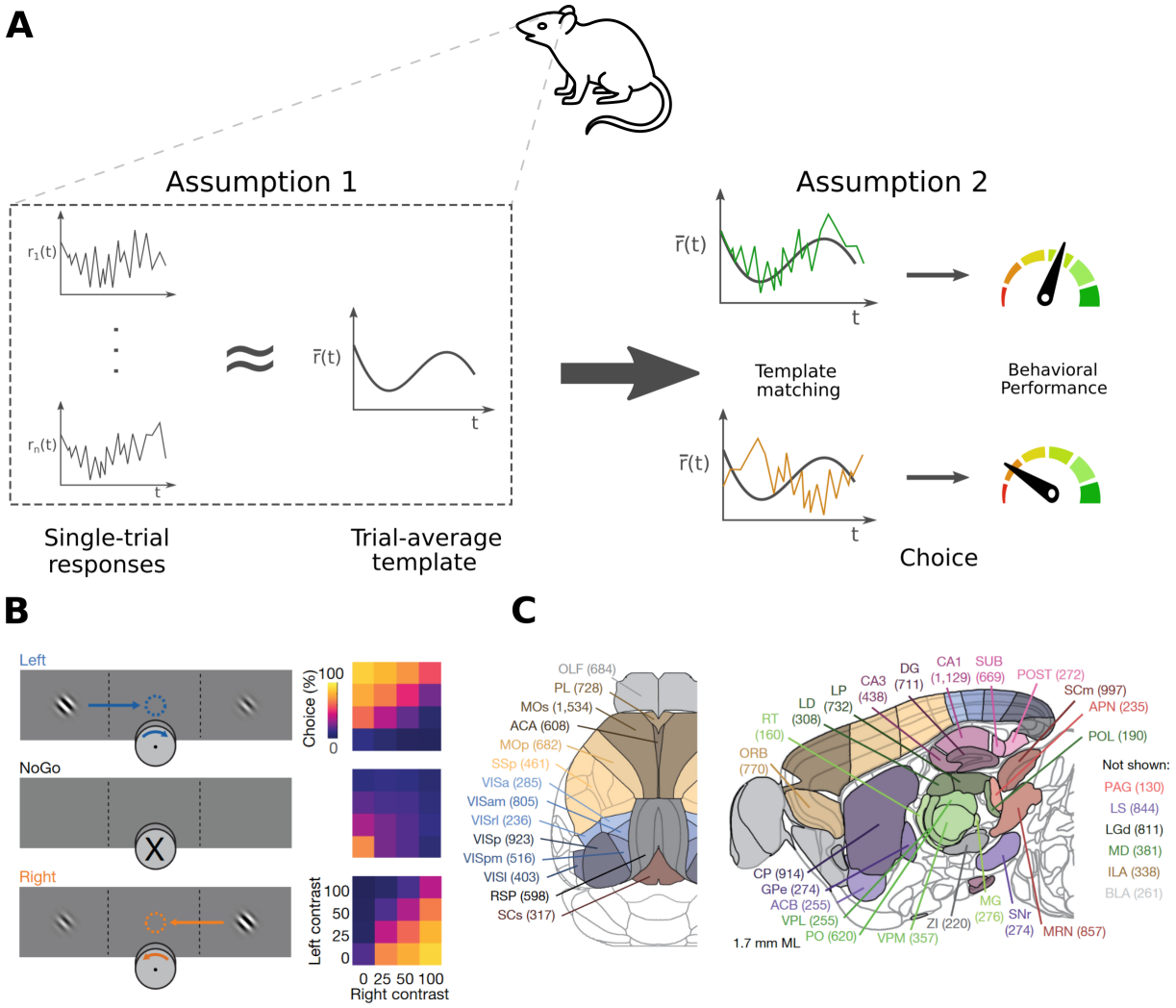

FIG. 1. Overview of the analysis and data. A) Graphic summary of the two assumptions underlying the computation of average population responses: Single-trial responses correspond at least somewhat to the trial-averaged response template (left), and better matched single-trial responses lead to more efficient behaviour (right). B) To obtain reward, animals needed 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. Note that in the original paper where the dataset was presented [44], 42 areas out of the 71 were included, as they excluded from the analyses those areas that were only recorded from one subject. B and C are reproduced with permission from [44].

as Figure S3 shows, the average responses to the individual contrast levels were very comparable to each other.

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 to 0.8 across all brain areas ($n = 89$ to 3560 trials per brain 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 computed a bootstrapped single-trial response vector by drawing the number of action potentials

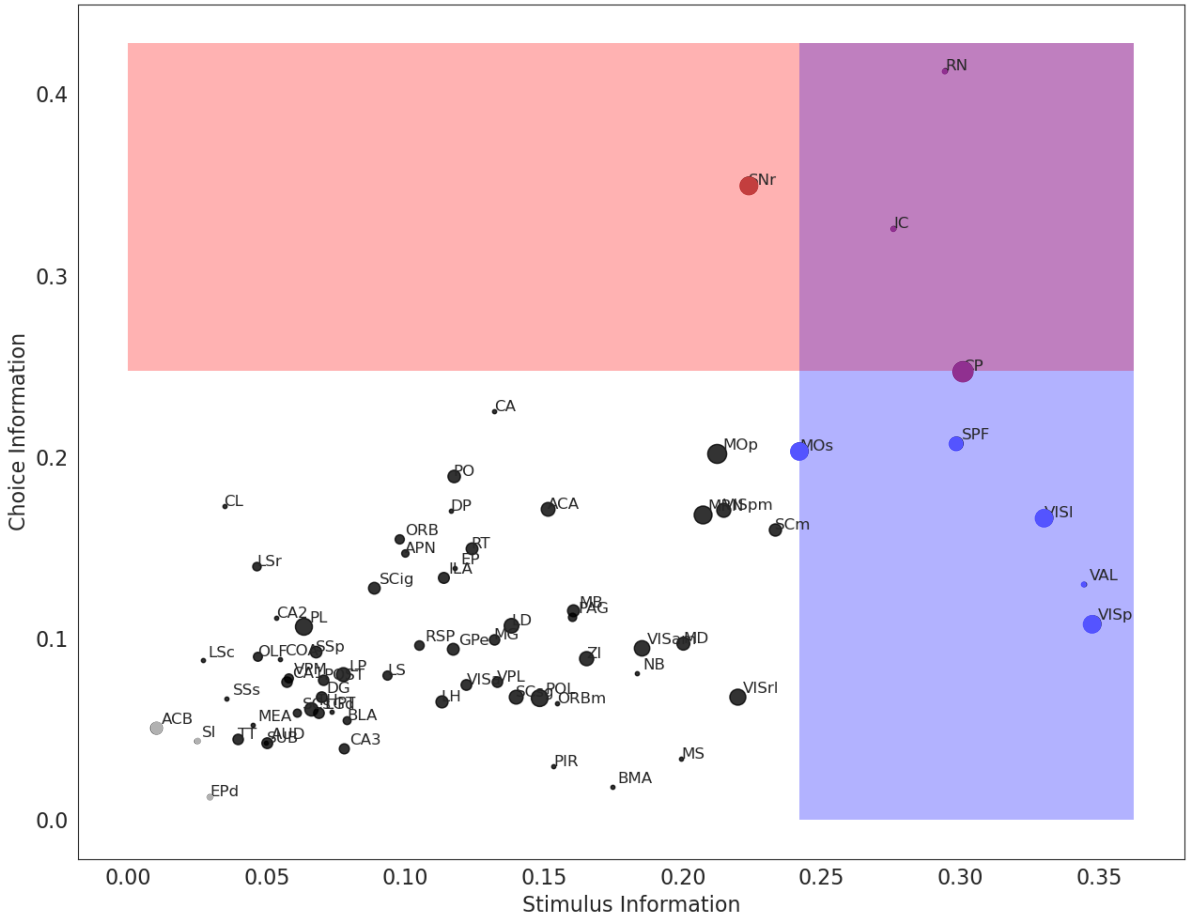

FIG. 2. Stimulus and target choice information decoded by a multinomial GLM decoder from the neuronal activity in various brain areas. Each data point represents the median (dot location) and standard deviation across sessions (dot size) of a brain area (see in-figure labels). The blue and red quadrants (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.

observed in that trial from the overall pool of spikes that make up the population template (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 the bootstrapped single-trial responses to the population template. 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 average response template 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. This notion is also supported by the observation that the three least informative, comparison brain areas showed comparable correlations to the population template as the brain 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 brain 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 brain areas) that correlations to the correct and incorrect templates were largely indistinguishable on a single-trial level. What's more, task-relevant brain areas did not show significantly more specific correlations than comparison brain areas (mean correlation difference for all task-relevant areas: 0.031; $n = 9$; for all comparison 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, 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 brain 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 brain 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 brain 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 the animal still needs to identify the presented stimulus and use it to make a correct perceptual choice. If trial-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 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 brain 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 the so-called *Common Language* effect size [60] (Ω; Fig. 4A, bottom). There were a few exceptions, such as the secondary motor area (MOs) ($\Omega = 0.66, p = 0.0142, n_{hit} = 2057, n_{miss} = 667$, 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

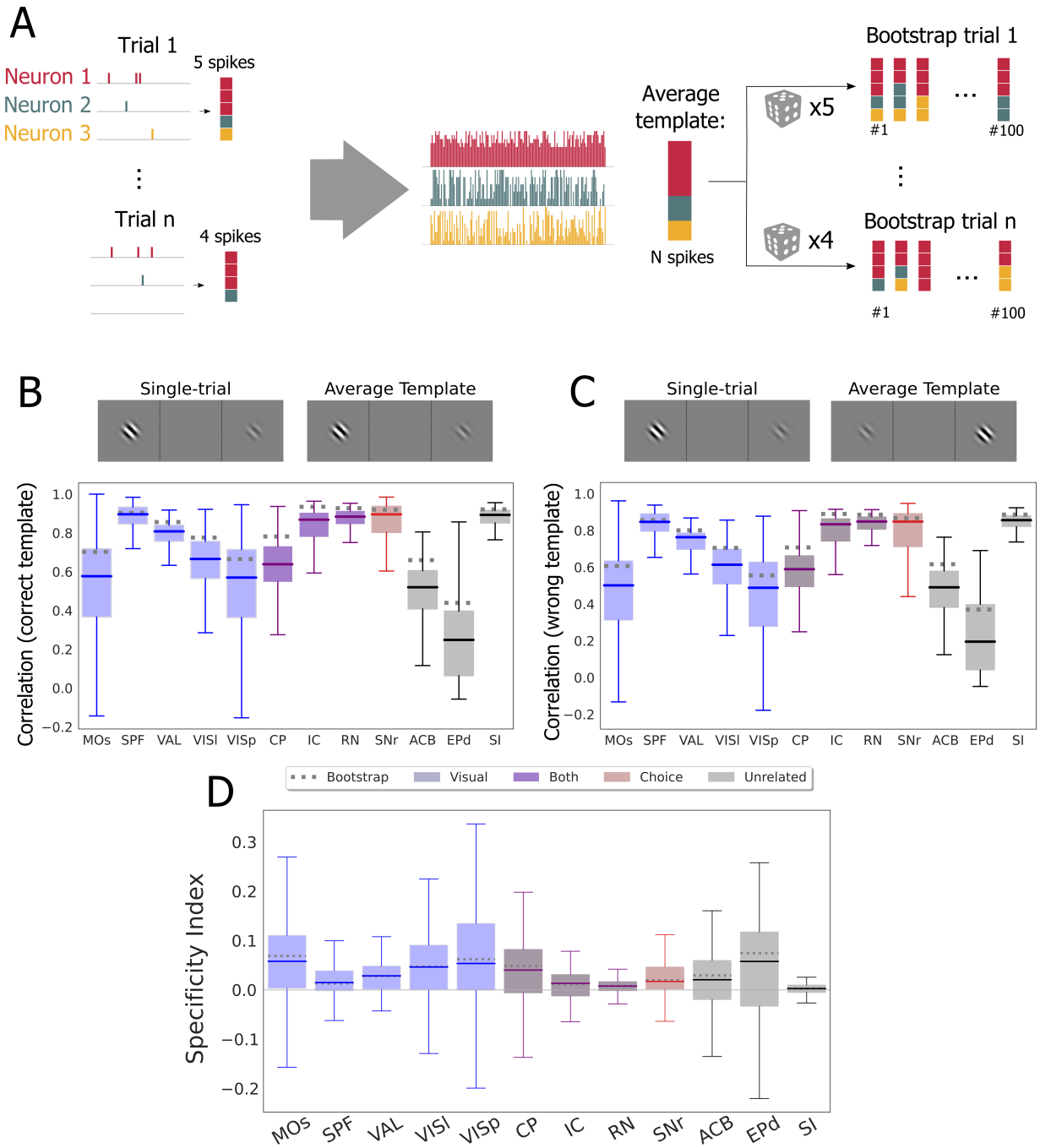

FIG. 3. Correspondence between average and single-trial responses. A) Representation of the bootstrapping procedure. The number of spikes stayed the same on each trial, but the neuron that produced each of these spikes was chosen randomly with a probability according to how often it spiked in the average 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 constellation. 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 brain 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 brain 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).

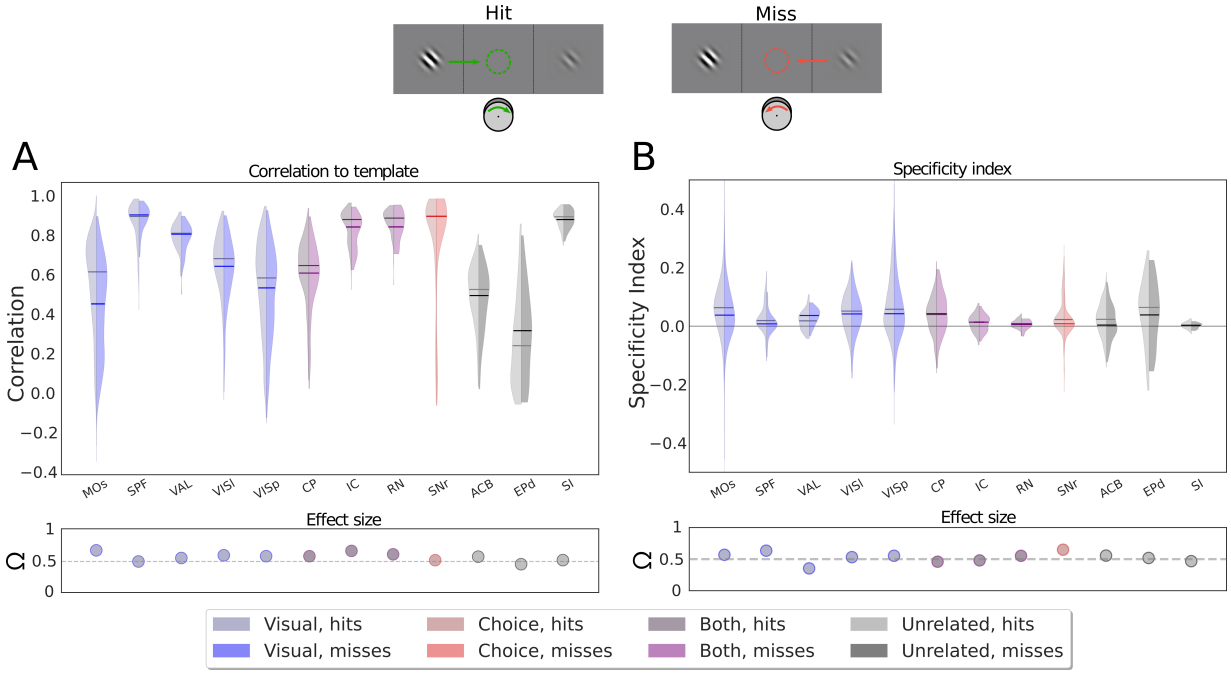

FIG. 4. (Color online) Match between single-trial responses and correct response template in hit and miss trials. A) Same as Figure 3B, but for single-trial correlations to the correct template, split by hit and miss trials. B) Same as Figure 3D, split by hit and miss trials. Common Language effect sizes (Ω) are shown below the plot and can take values between 0 and 1, with 0.5 indicating a complete overlap between distributions (see Methods).

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), indicating that single-trial responses that were more specific to the correct template did by and large not lead to improved target choices.

Even though average responses were overall not particularly informative about either stimulus identity or subsequent behaviour, there could be a 'supergroup' of highly informative neurons about either of these aspects, which would then drive neuronal processing in downstream areas and ultimately target choice [61, 62]. To examine this, 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 increasing and decreasing the match of the single-trial response to the correct template.

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 brain 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 involved in the task. To explore this possibility, we first sub-sampled the population of recorded neurons in

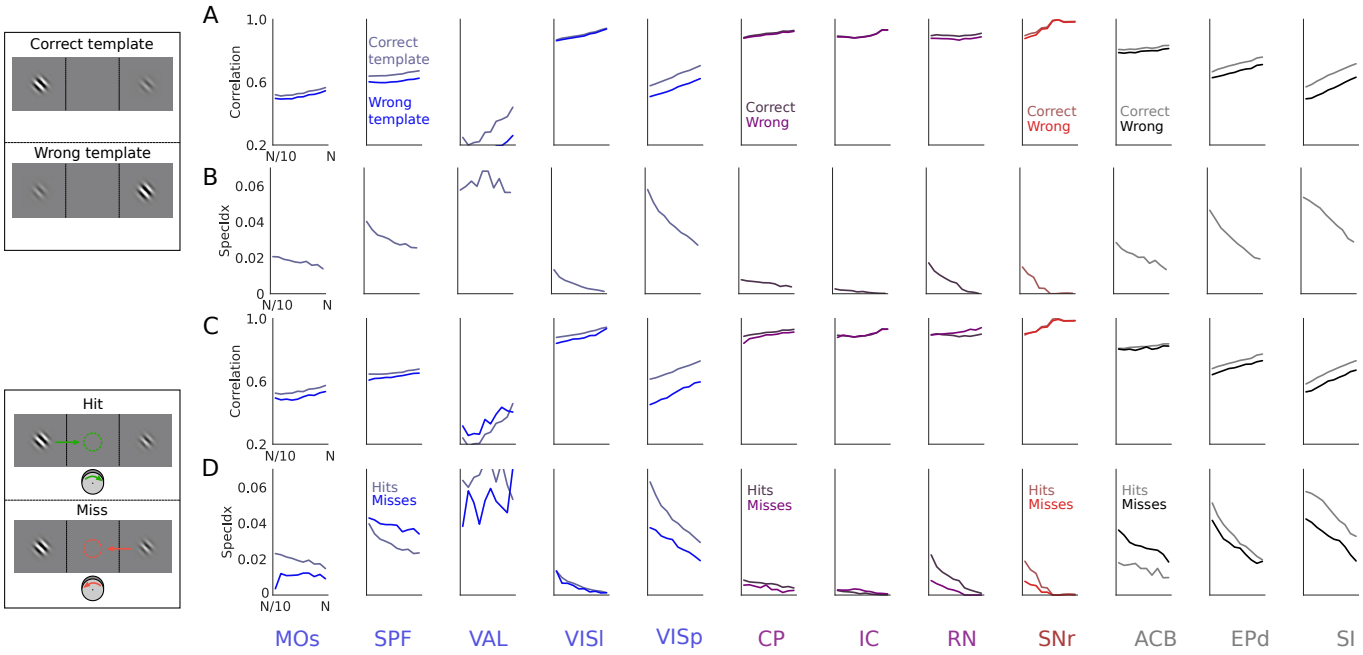

FIG. 5. Extrapolation of single-trial correlations and specificity index across different sample sizes. A) As the number of sampled neurons increases, single-trial responses do not get more stimulus-specific. B) The specificity index is always around 0 and actually tends to decrease with the number of sampled neurons C, D) Same as A), B) for the correct template and split over hits and misses.

each brain area at 10 different levels from $N/10$ to N. We then extrapolated how metrics like the Specificity Index would evolve as the number of available neurons grew. Single-trial correlations to the response template tended to grow with sample size (Fig. 5A), 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). In agreement with this, with growing N, specificity remains largely constant in some areas (e.g. MOs and SPF), and actually declines in many others (e.g. VISl, VISp, RN and SNr) (Fig. 5B). Moreover, with the exception of RN, the difference in single-trial correlations between hit and miss trials remained constant with growing n (Fig. 5C), and difference in response specificity between hit and miss trials tended to decrease (Fig. 5D). 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 from 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. One possibility is that 'average template matching' happens in a non-linear and/or multi-dimensional way. To take a first step at exploring this possibility, we repeated the analyses shown in Figs. 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 (Figs. 6-7).

Generally, the resulting single-trial response vectors did overall not represent their corresponding average vectors

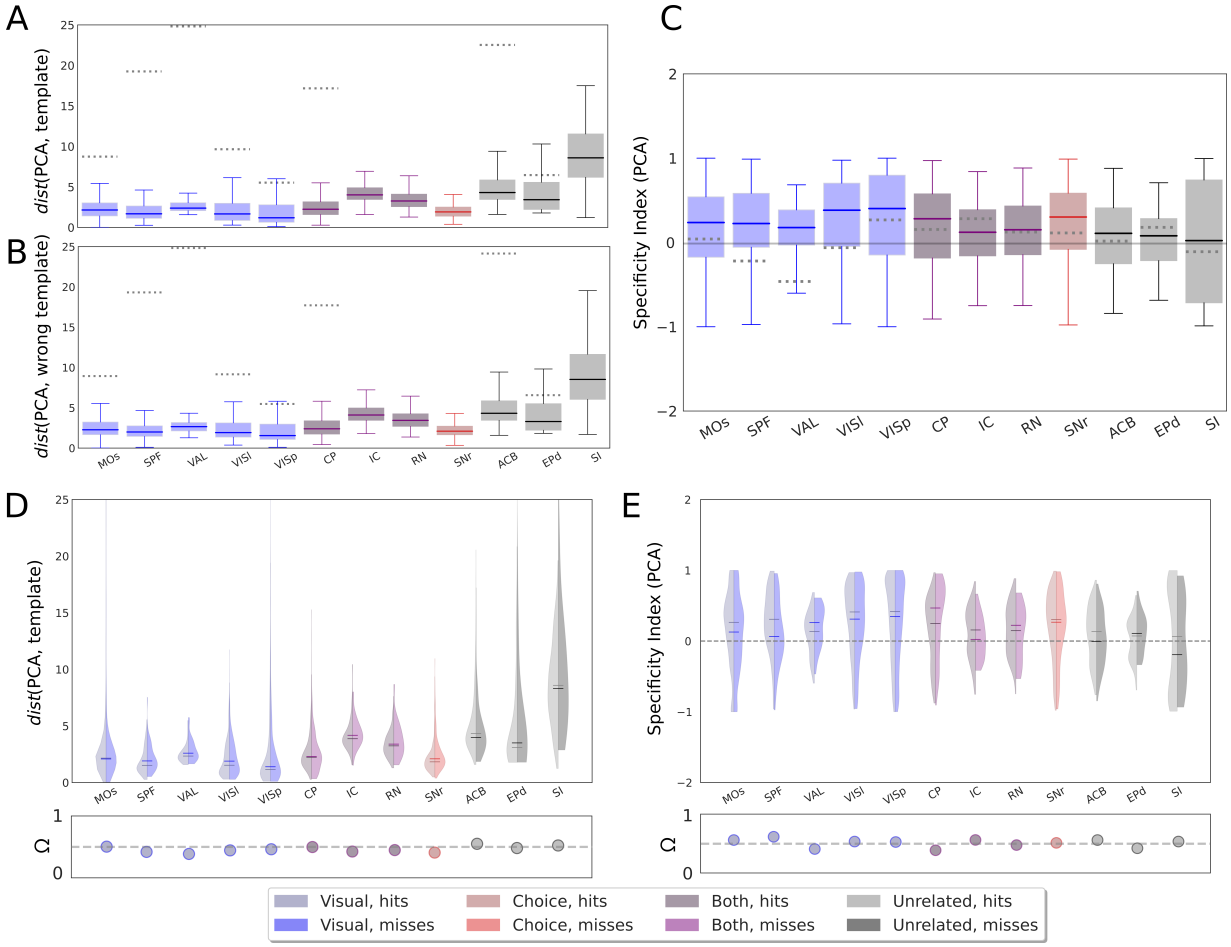

FIG. 6. Template matching using PCA instead of Pearson correlation. A) Distance of single-trial response vectors to the correct average template. For interpretability, distances are normalized by the distance between the two average templates. B) Same as A, but for the incorrect template. C) Specificity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

more effectively than the Pearson correlation 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 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, the specificity of single-trial responses for the correct average template was low, even more so 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. Given that PCA vector distances are not upper-bounded to 1 and often took on values between 5 and 10 (Fig. 6A-B), specificity indices < 1 imply negligible differences between the single-trial distances to correct and incorrect templates. Single-trial responses that were more similar and/or specific to the correct average vector also resulted in only slightly more 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, Ω between 0.38 and 0.55; all $p < 0.01$;

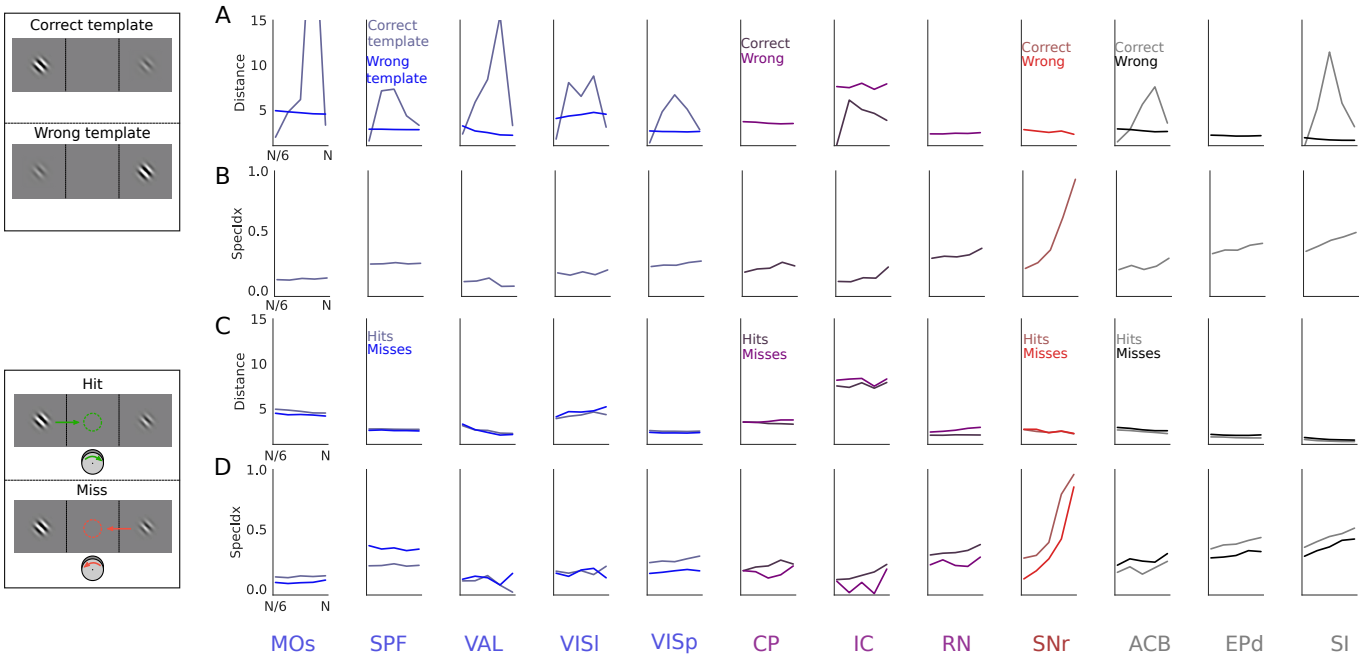

FIG. 7. Extrapolation of single-trial distances in PCA space and specificity indices across different sample sizes. A) Distance between single- trial vectors and correct and incorrect template over increasing sample sizes. B) Specificity index computed over increasing sample sizes. C,D) Same as A,B), but only for the correct template and split for hits and misses.

Mann-Whitney's U-test for differences in single-trial specificity in hit and miss trials: $n = 90$ to 3123, Ω between 0.38 and 0.61; all $p < 0.01$; corrected for multiple comparisons). Finally, unlike the Pearson correlation averages probed before, the similarity to the average template in PCA space (Fig. 7A) and the relation to the behaviour (Fig. 7C) seemed to not profit from increased neuron numbers, and improve only in very specific cases, such as the SNr, RN, IC, and VISp. Note however that even for these areas, the Specificity Index never exceeded 1, still pointing to low overall response specificity (see above). Overall, these results demonstrate that just like for Pearson correlations, resemblance of single-trial to average vectors in PCA space did not seem to drive neuronal processing in a decisive way across most brain areas. 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.

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 in their match 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 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 did not improve for progressively larger number of neurons, and only marginally improved for specific brain areas (Substantia Nigra and Red Nucleus) when applying dimensionality reduction techniques 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 analysis, one can envision several caveats, most of which pertain to the appropriateness of the average response template. First, in the current task, relevant stimulus information (the contrast difference between the two stimuli) was computed by comparing information across two brain hemispheres, but we only have access to neuronal recordings from one hemisphere. Thus, recording from both hemispheres might have yielded more stimulus-specific and informative population templates. However, the fact remains that the animal is able to make largely correct perceptual decisions in this task (Fig. 1B), which means that even if specific brain 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 there is not a single area that consistently represents the integrated stimulus information of the two hemifields.

Another limitation of our average population template may be that they were not specific enough, since the 16 different contrast combinations were pooled into only three stimulus categories. However, the fact that these three average response templates were highly correlated to each other (Fig. S3) suggests that the precision lost by pooling across stimulus pairs is largely negligible. Moreover, since the stimulus categories are congruent with the target choice of the animal (i.e. choose whether to turn the wheel to the left or right), the information on which hemifield contains the higher-contrast stimulus, irrespective of exact contrast should be reflected in at least some of the recorded brain areas, in order to drive the behavioural response – an assumption which is also borne out by the decoder analysis (Fig. 2).

A third potentially important factor is our choice to analyse neuronal responses $0 - 200ms$ post-stimulus onset. We chose this analysis window to focus on the decision process itself rather than on the licking response, which typically happened after $200ms$ (response peak at $520ms$, Fig. S1). It is possible that a different analysis window would highlight different and/or more task-related information across the recorded brain 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 trial-averaged response templates are much more relevant to neuronal computations in different behavioural contexts, stimulus structures or even species – or different metrics of neuronal activity that are being extracted and averaged over time [44]. We chose PCA as a benchmark of dimensionality-reduction techniques due to its relative simplicity and ubiquitous use, but other approaches might yield improved results. For instance, non-Negative Matrix Factorization [63] might outperform PCA because it defines neuronal ensembles in a sparser, and therefore more realistic, way [64].

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 computed here) that represents an estimate of how likely the information they convey is to be informative to the brain, rather than only

to the reader. 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 brain areas in which neuronal responses show more or less representational drift over time [34, 35, 65, 66].

To facilitate this, we have kept the computational tools employed here purposefully simple and general, by utilizing mainly linear correlations. 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; we could exclude those 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, at least in the particular context tested here, appear largely irrelevant to ongoing neuronal computations, 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 the many task-related and task-unrelated factors at any moment in time [12, 28, 29, 54, 56–59, 67–70], 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 into e.g. stimulus-related information and 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, potentially non-linearly interacting factors, might be more successful at teasing out reliable information from trial-averaged templates. If this were the case, we would suggest that the neuroscience community abandons 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, 71–73] throughout a wide range of experiments.

However, it is also possible that trial-averaged templates 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 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. However, 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 [73–79] or by explicitly including shared variability in their readout [80–82], do not always uncover consistent neuronal representations across all brain areas [34], yet they often provide vastly more informative and stable representations of neuronal activity despite seeming variability [74–76, 82]. Consistent with this, the decoder approach (Fig. 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 [83–85] or the relative timing of action potentials [86, 87] 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 brain 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 brain 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, 65, 76] showing that in many contexts, neuronal responses spontaneously shift over time; then a static average taken across time is then 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 analyses similar to the one presented here to 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, Viola Priesemann 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 [88]. 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 [89, 90]. 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 [91, 92].

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 [93]. After having extracted these, we used the method discussed in [94] to find the point of maximum curvature. We made use of the kneed Python package, implemented by the same authors [94].

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.

Common Language effect size

As a way to quantify the overlap between distributions (for example, correlation in hit vs miss trials) we relied on the Common Language effect size [60], that we labelled as Ω. This 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, Ω is especially interpretable. As it is related to the U statistic, 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 [96]. 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, Ω is given by normalizing it as

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

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) [97]. We have chosen PCA over non-Negative Matrix Factorization [98] or other more advanced dimensionality reduction techniques such as LFADS [80] or PSID [73] because we wanted to keep all analyses as general as we possibly could. Thus, we compute the truncated Singular Value Decomposition (tSVD) [99] 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] K. O. McGraw and S. P. Wong, Psychological bulletin **111**, 361 (1992).

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

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

[63] 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).

[64] 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).

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

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

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

[68] 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).

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

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

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

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

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

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

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

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

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

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

(2020).

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

[80] 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).

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

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

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

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

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

[86] 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).

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

[88] 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).

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

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

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

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

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

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

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

[96] D. S. Kerby, Comprehensive Psychology **3**, 11 (2014).

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

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

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

SUPPLEMENTARY MATERIAL

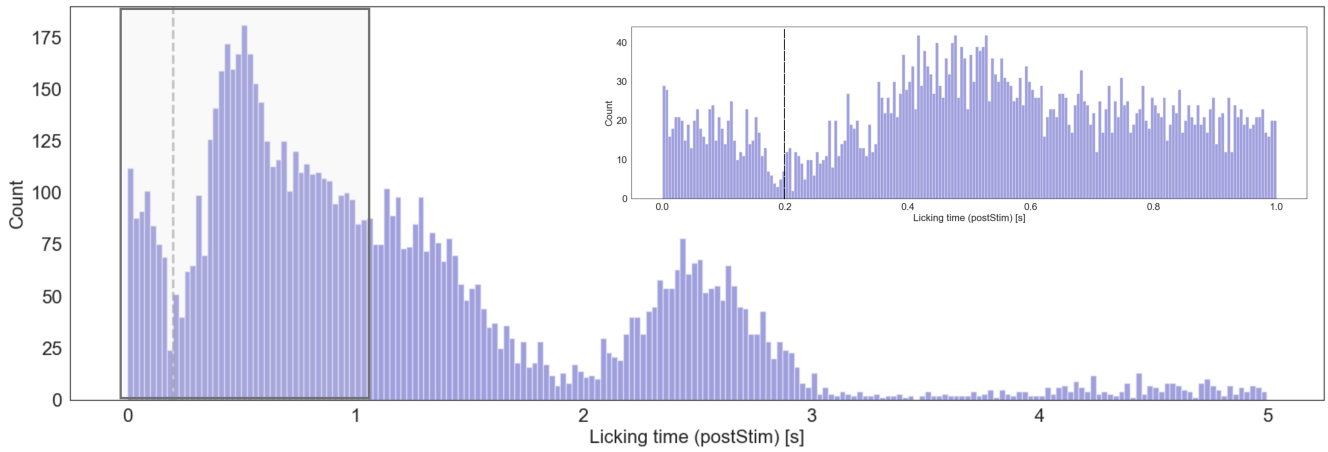

FIG. S1. (Color online) Post-stimulus licking time distribution. In the highlighted inset, we show licking times below a second after stimulus presentation. We have selected our analysis window of $200ms$ because of the relatively small number of lick events in that window (5% of all licks) and its likely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

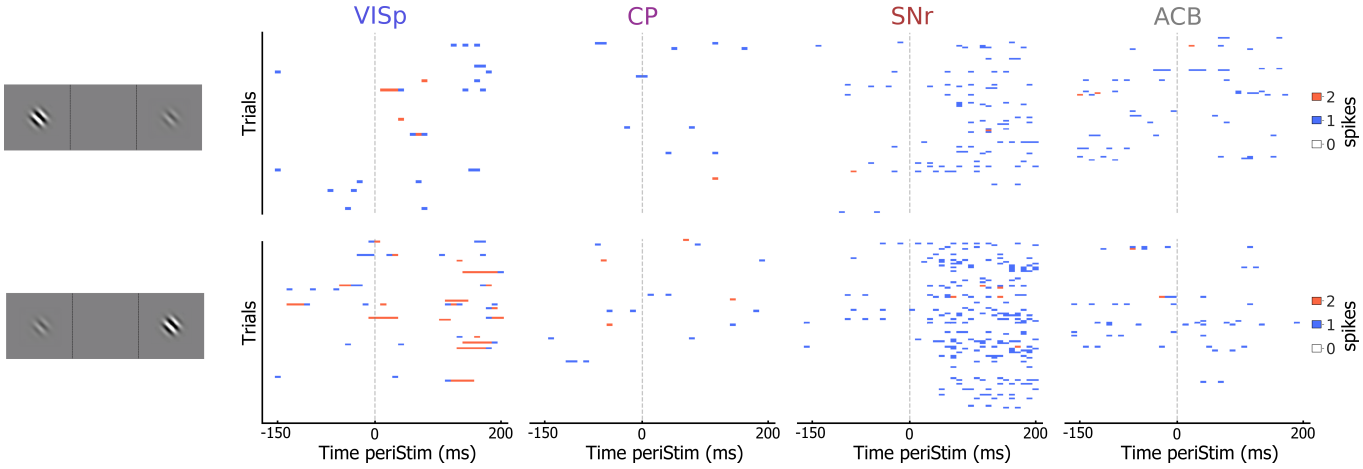

FIG. S2. (Color online) Example of neuronal responses to stimulus constellations with the target on the left and right, respectively. Responses are shown for one neuron each in four representative areas that are informative of the stimulus (VISp), the target choice (SNr), both (CP) and neither (ACB).

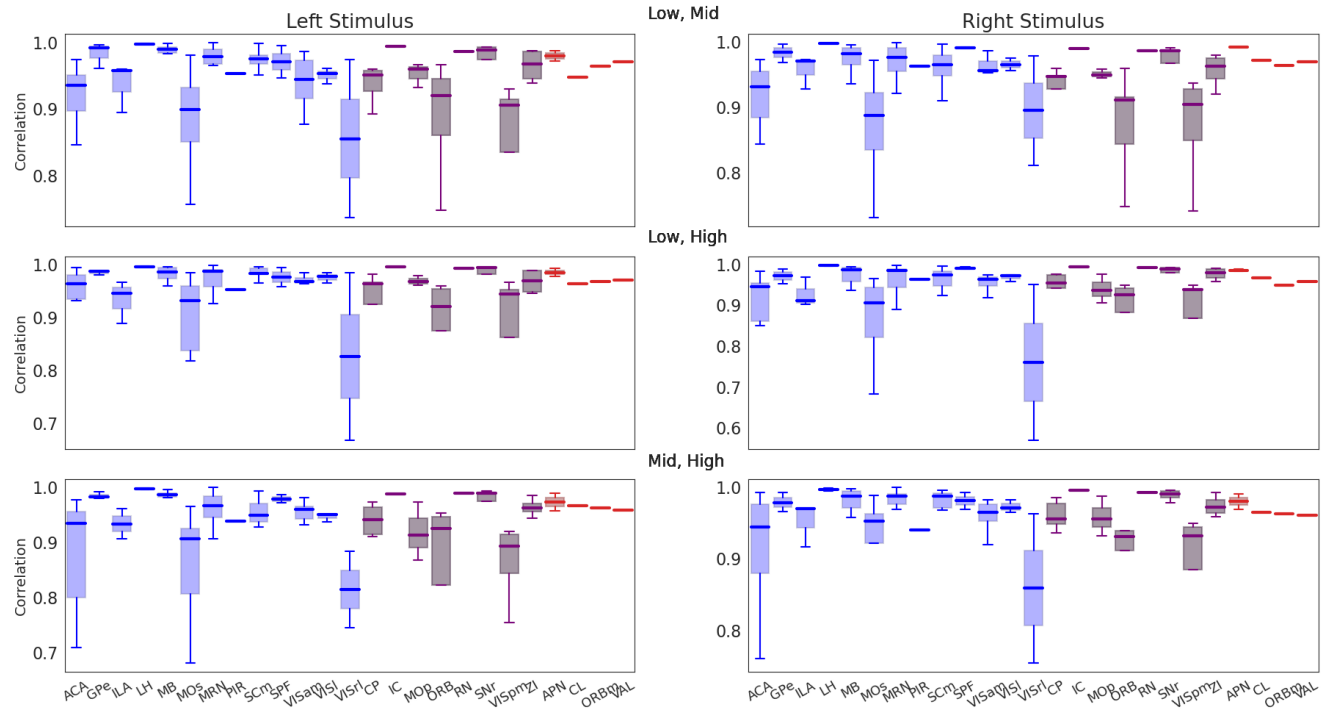

FIG. S3. (Color online) Correlations between response templates for different contrast combinations. These generally exceed 0.9, suggesting that one template should be sufficient to represent different contrast constellations.

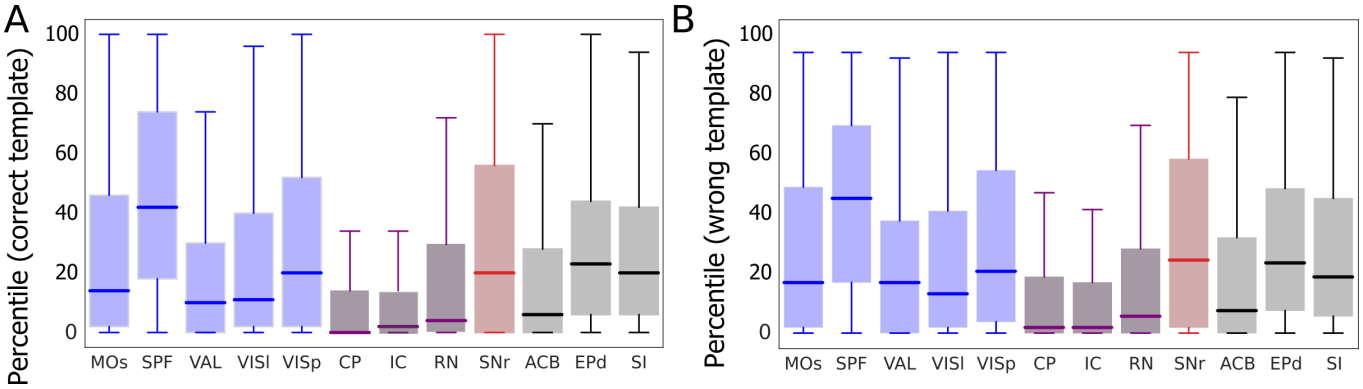

FIG. S4. (Color online) Percentile distributions for the correlation between single-trial responses and the correct (A) and incorrect templates (B). Box: 25th and 75th percentile. Solid line: Median. Whisker bars: 10th and 90th percentile. In both cases, these correlations lie well below the 95th percentile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

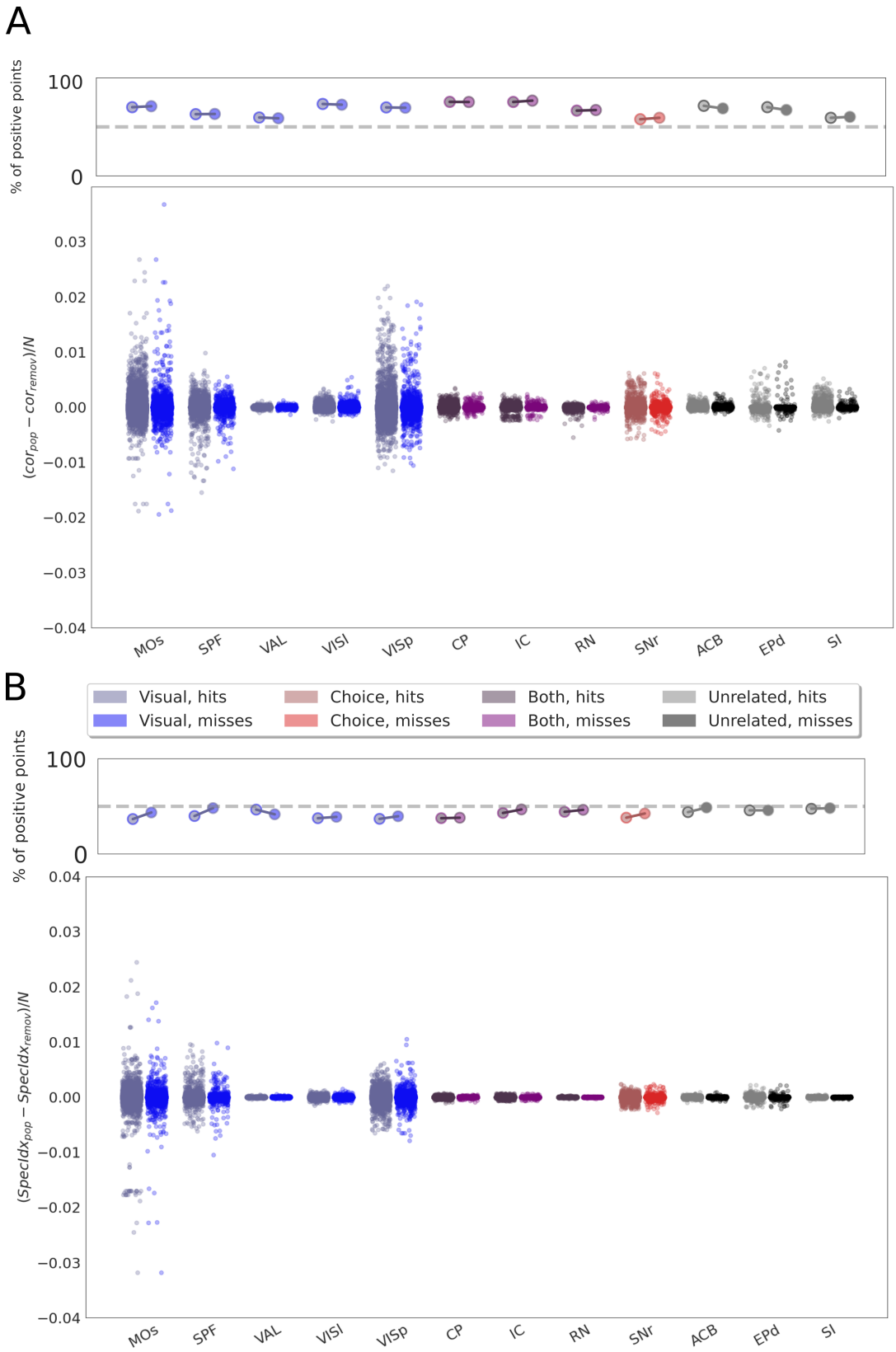

FIG. S5. (Color online) Neuron removal analyses. A) Change in single-trial correlations to the correct average template when one individual neuron was removed. Data points: Trials. Colors: see inset legend. B) Same for single-trial specificity.