Modelling the stable subspace of distributed brain responses in stimulus processing

Diego Vidaurre[1,2]

1. Center for Functionally Integrative Neuroscience, Department of Clinical Medicine, Aarhus University (Denmark)
2. Department of Psychiatry, Oxford University (UK)

Abstract: Each brain response to a stimulus is, to a large extent, unique. However this variability, our perceptual experience feels stable. Standard decoding models, which utilise information across several areas to tap into stimuli distributed representations and processing, are fundamentally based on averages and, therefore, can focus precisely on the features that are most stable across stimulus presentations. But what are these stable features that may sustain stable perception across experiences is still poorly understood. This question is indeed difficult to address in the absence of a generative model of the signal. Here, I introduce *genephys*, a generative model of brain responses to stimulation that, when confronted with a decoding algorithm, can reproduce the structured patterns of decoding accuracy observed in real data. Using this approach, I characterise how these patterns can be brought about by the different aspects of the signal, which in turn may translate into distinct putative neural mechanisms. In particular, the model shows that the features in the data that support successful decoding — and, therefore, likely reflect stable mechanisms of stimulus representation— lie in a high-dimensional, multi-frequency space of phase resets and frequency entrainments, combined with an additive non-oscillatory response. At the individual trial level, still, responses are found to be highly variable, which can be due to various factors including phase noise and probabilistic activations.

Introduction

There are virtually infinite manners by which a constant stimulus can impinge into the sensorium of an animal. For instance, our noses have many receptors that can sense a given odorant molecule, but only a small subset of those are excited each time the odour is perceived (Axel, 1995). Similarly, photons hit and excite photoreceptors in the retina randomly and sparsely for a given presented image (Dowling, 1987). An important aspect of sensation and perception is that never the exact same receptors are excited every time we perceive and, still, our perceptual experiences are quite stable. It is therefore sensible to assume that the brain must have a way to transit from lack of invariance at the microscopic sensory level toward invariance at the macroscopic level, which ultimately supports the invariant aspects of conscious perception and behaviour. However, we observe significant variability in the brain responses at the trial level (Stein, Gossen, & Jones , 2005; McIntosh, Kovacevic, & Itier, 2008; Garrett, et al., 2013), including at the earliest layers of the perceptual hierarchy (Croner, Purpura, & Kaplan, 1993; Freeman W. J., 1978) —that is, each perceptual experience is associated with a unique neural trajectory that does not repeat. The gap between stability in subjective perception and the changing nature of brain responses in an important question in neuroscience. Here, I investigate the distributed aspects of brain activity that are most stable across experimental repetitions, and are therefore most likely to support stable perceptual experiences.

One possible approach to characterise the stable aspects of brain responses to stimulation is the analysis of average evoked responses potentials or fields (ERP/F) (Dawson, 1954; Steven & Kappenman, 2011; Pfurtscheller & Lopes da Silva, 1999). However, this is limited because perceptual experiences emerge from activity patterns across multiple brain areas acting in a distributed fashion, whereas ERP/Fs are separately evaluated channel by channel. Also, ERP/F

analyses are not concerned with what aspects of the signal carry specific information about the stimulus —i.e. they are not predictive of the stimulus (Kragel, Koban, Feldman Barrett, & Wager, 2018). In contrast, decoding analysis uses a multivariate machine learning algorithm to predict the identity of the observed stimulus from brain data (Haxby, Connolly, & Guntupalli, 2014; Stokes, Wolff, & Spaak, 2015), producing a set of algebraic subspaces (one per time point) where the brain signals maximally discriminate between conditions, and offering a measure of accuracy across time that reflects how much information the data conveys about the stimuli. The assessment, via decoding accuracy, of how the discriminative space changes throughout the trial offers a view of the properties of stimulus representation and processing. However, it is not straightforward to know what specific aspects of the signal cause the patterns of decoding accuracy that we observe in perceptual experiments. Without this capacity, it is hard to link these patterns to actual neural mechanisms.

To fill this gap, I introduce a generative model of multichannel electrophysiological activity (e.g. EEG or MEG) that, under no stimulation, exhibits chaotic phasic and amplitude fluctuations; and that, when stimulated, responds by manipulating certain aspects of the data, such as ongoing phase or signal amplitude, in a stimulus-specific manner. Specifically, in every trial, each channel may or may not respond to stimulation, according to a certain probability. When a channel responds, it can do it in different ways: (i) by phase resetting of the ongoing oscillation and then entraining to a given frequency, (ii) by modulating the amplitude of the ongoing oscillation, (iii) by an additive non-oscillatory response, or (iv) by an additive oscillatory response. Each of these effects can arguably be underpinned by distinct neural mechanisms. I named this model *genephys* by *gen*erative model of *e*mpirical electro*phys*iological signals. This model is empirical in the sense that it purely accounts for the features in the signal that are observable, without making any assumption about the underlying neurobiological causes; that is, it can generate signals that, depending on its parametrisation, can share empirical properties with real data. In particular, when confronted with a decoding algorithm, the data generated by this model can yield, depending on its configuration, patterns of decoding accuracy with the same main characteristics that we observe in real data. Specifically, phase resetting and frequency entrainment across many channels and frequencies, combined with an additive non-oscillatory response, produces the most realistic patterns. Other aspects of stimulus processing, and other experimental paradigms, can be investigated with g*enephys* —which is publicly available as a Python package[1].

Methods

A generative model of empirical electrophysiological signals: *genephys*

While the system is unperturbed, *genephys* is based on sampling spontaneously varying instantaneous frequency and amplitude (square root of power) time series, f^{rest} and a^{rest} respectively, that are analytically combined to form the sampled signal x. Amplitude and frequency are allowed to oscillate within ranges r^f and r^a. Instantaneous frequency here refers to angular frequency, from which we can obtain the ordinary frequency in Hertz as $\frac{F}{2\,\pi}\,f^{\text{rest}}$, where F is the sampling frequency of the signal. Both f^{rest} and a^{rest} are sampled separately from an autoregressive process of order one, endowing them with chaotic, non-oscillatory

[1] https://test.pypi.org/project/genephys, https://github.com/vidaurre/genephys

dynamics. Given autoregressive parameters $b^f, b^a < 1$, and Gaussian-noise variables e_{f_t}, e_{a_t}, I generate f^{rest} and a^{rest} for a given channel as

$$
\begin{aligned}
f_0^{\text{rest}} &= e_{f_0} \\
f_t^{\text{rest}} &= b_f\, f_{t-1}^{\text{rest}} + e_{f_t} \qquad \text{for t} > 0 \\
a_0^{\text{rest}} &= e_{a_0} \\
a_t^{\text{rest}} &= b_a\, a_{t-1}^{\text{rest}} + e_{a_t} \qquad \text{for t} > 0.
\end{aligned}
$$

Without stimulation, a phase time series φ^{rest} is then built as

$$
\begin{aligned}
\varphi_{0^{\text{rest}}_0} &= 0 \\
\varphi_t^{\text{rest}} &= \varphi_{t-1}^{\text{rest}} + f_t^{\text{rest}}
\end{aligned}
$$

Then, given some Gaussian-distributed measurement noise ϵ_t with standard deviation σ_ϵ, I build x in absence of stimulation as

$$x_t^{\text{rest}} = a_t^{\text{rest}} \sin \varphi_t^{\text{rest}} + \epsilon_t$$

This process is done separately per channel and per trial. Note that, under no stimulation, the channel time series are therefore asymptotically uncorrelated. We can think of them as dipoles in brain space. We can induce correlations by projecting these time series into a higher-dimensional space, which we can suppose to be at sensor space. Altogether, this generates chaotic oscillatory data similar in appearance to real data.

When a stimulus is presented, a perturbation is introduced into the system on the p channels that are stimulus-relevant, which can be a subset of the total number of channels. When a channel is not relevant, it does not respond; when it is relevant, it responds to the stimulus stochastically:

Prob (channel j responds | channel j is relevant) = θ_j
Prob (channel j responds | channel j is not relevant) = 0

Where θ_j is a hyperparameter representing a certain probability. In simpler words, relevant channels might respond in some trials, but not in others. In all the simulations, I set all channel probabilities to a single value, $\theta_j = \theta$, but other configurations are possible.

Given some stimulus presentation at time point τ, a channel can respond:

- By phase-resetting to a condition-specific target phase, and then, when the target phase is reached, by frequency entraining to a target frequency.
- By increasing the (absolute) amplitude of the ongoing oscillation following stimulus presentation (with no phase effect).
- By adding a transient, non-oscillatory response with a condition-specific amplitude.
- By adding a transient, oscillatory response with a condition-specific phase, amplitude and frequency.

In order to implement these effects, I use a piecewise response function: before the time of maximum effect t_{max} the response function is modelled as an exponential function; after the

time of maximum effect, the response function is modelled as a logarithmic function (which is slower to decay). This function is parametrised by: δ_1, reflecting the delay of the response in number of time points; and δ_2, reflecting how many time points it takes the exponential function to go from zero to its maximum before it changes to the logarithmic function. Therefore $t_{\max} = \delta_1 + \delta_2 + \tau$. I introduce some noise in δ_1 per trial to make the timing of the response to vary, as per $\delta_1 \sim \text{Uniform}(0, \sigma_{\delta_1})$, where σ_{δ_1} is a hyperparameters of the model. This delay noise could be fixed for all channels (absolute stochastic latency) or per channel (relative stochastic latency). See the **Supplemental Information** for a full mathematical specification of the response function.

With respect to the phase-reset and frequency entrainment effect, phase reset occurs before $t_{\max}$, when the target phase is reached. For each trial and channel,

$$\varphi_t = \varphi_{t-1} + (1 - g_t) f_t^{\text{rest}} + g_t \ \nabla_t \qquad \text{with}$$

$$\nabla_t = \begin{cases} \varphi^k - \varphi_t, & \text{if } \varphi^k - \varphi_t \in [-\pi, \pi] \\ \varphi^k - \varphi_t - 2\pi, & \text{if } \varphi^k - \varphi_t > \pi \\ \varphi^k - \varphi_t + 2\pi, & \text{if } \varphi^k - \varphi_t < \pi \end{cases}$$

Where, for each trial, φ^k is randomly sampled from a von Mises distribution with mean equal to the target phase $\overline{\varphi^k}$ and standard deviation σ_φ; ∇_t is the polar gradient between the target phase and the ongoing phase φ_{t-1}; and g_t is the value taken by the response function at time point t.

After $t_{\max}$, phase-resetting is over, and the phase entrainment period starts,

$$\varphi_t = \varphi_{t-1} + g_t \, f^k + (1 - g_t) \, f_t^{\text{rest}}$$

Now, the system entrains to a possibly stimulus-specific, possibly channel-specific, target frequency f^k, with a strength that slowly (logarithmically) decreases as t gets farther from $t_{\max}$.

With respect to the amplitude modulation of the ongoing oscillation, I use a similar response function (which may be parametrised differently). Here, I apply a multiplying factor to the amplitude time series as

$$a_t = (1 + g_t \, m^k) \, a_t^{\text{rest}}$$

where m^k is a proportional increment (for example, $m^k = 0.1$ would produce an increment of 10% in amplitude at $t_{\max}$). This effect applies to both the reset-phase and entrained frequency, if applicable.

With respect to the additive non-oscillatory response, we have:

$$z_t = g_t \, s^k$$

where s^k is sampled from a Gaussian distribution with stimulus-specific mean $\overline{z^k}$ and standard deviation σ_z.

With respect to the additive oscillatory response, we consider a sinusoidal oscillator, which is damped by the action of the response function g_t:

$$y_t = g_t \, \alpha^k \sin(\omega^k t + \gamma^k)$$

Given these elements, the signal is built as

$$x_t = a_t \sin \varphi_t + z_t + y_t + \epsilon_t$$

This model can be trivially extended to have correlated noise ϵ_t across channels, which would also make the channels non-independent.

An example of how the effect of stimulation looks for one trial is given in **Figure 1A**, where the effect of phase resetting and subsequent entrainment can be appreciated with some delay after stimulus presentation (indicated by the vertical dotted bar). Here, the sampled signal x is shown in blue on top, and the phase φ, frequency f, amplitude a, and additive response z are shown in red underneath. As shown in **Results**, phase resetting plus frequency entrainment and non-oscillatory signal addition are the elements that are necessary to generate decoding accuracy patterns as we observe in real data, while amplitude enhancement makes the existing effects to be stronger. **Figure 1B** shows the spectral properties of the signal, which are reasonably realistic. **Table 1** summarises all the hyperparameters that configure the model.

Parameter	**Math. notation**
Frequency autoregressive parameter	b_f
Amplitude autoregressive parameter	b_a
Frequency range	r_f
Amplitude range	r_a
Measurement noise standard deviation	σ_ϵ
Channel activation probability	θ
Number of relevant channels	p
Delay parameters	δ_1, δ_2
Target phase, mean	$\overline{\varphi^k}$
Target phase, standard deviation	σ_φ
Amplitude modulation	m^k
Additive non-oscillatory response, mean	$\overline{z^k}$
Additive non-oscillatory response, standard deviation	σ_z
Additive oscillatory response, phase	γ^k
Additive oscillatory response, amplitude	α^k
Additive oscillatory response, frequency	ω^k

Table 1: *genephys* configuration hyperparameters.

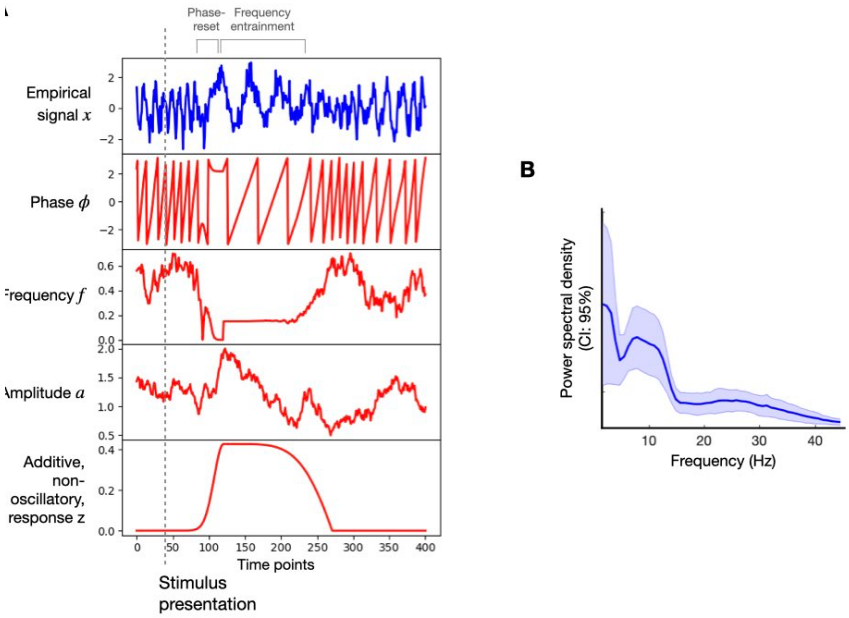

Figure 1. A: Single-trial example of the generated signal (in blue) and its constitutive com-ponents (in red): instantaneous phase, frequency and amplitude, as well as an additive non-oscillatory response component. **B**: Frequency spectrum of the data.

Decoding accuracy to characterise structured invariance

This paper focuses on exploring the patterns of differential activity between conditions across channels and throughout time. For each time point, a decoding method estimates an algebraic projection of the data onto one single dimension that maximally discriminate between conditions. Throughout the trial, this set of projections reflects information about stimulus processing and representation.

We use the temporal generalisation matrix (TGM), a $T \times T$ matrix of decoding accuracies (where T is the number of time points in the trial), such that one decoding model is trained per time point and tested on each one of the time points of the trial using cross-validation (King & Dehaene, 2014). This way, the diagonal of the TGM reflects how well can we can decode information time point by time point, while the off-diagonal of the TGM shows how well decoding models generalize to time points different from those where they were trained. As illustrated in **Figure SI-1**, where I show a real-data TGM from a subject performing a simple visual task (Cichy et al., 2016; Vidaurre et al, 2019), the TGM exhibits some characteristic features that we often see throughout the literature. First, there is a strong diagonal band that is relatively narrow early in the trial, surrounded by areas of worse-than-baseline accuracy. Then, the diagonal accuracy pattern becomes progressively wider and weaker, expanding until relatively late in the trial. Also, there are areas of higher-than-baseline accuracy stemming vertically and horizontally from the time points of maximum accuracy after a brief period of depression, typically alternated in an oscillatory manner with periods of lower-than-baseline accuracy.

Central to this paper is the observation that the variability of brain responses to stimuli are structured in a specific manner, and that the TGM can reflect various aspects of this structure —but not others. Below, we explore *genephys'* configuration space in relation to its ability to produce patterns of decoding accuracy similar to these from real data.

Results

In the experiments below, I confront *genephys* to classification-based decoding analysis to characterise the stochastic nature of the subspace of brain activity that carry stimulus-specific information. In each simulation, I generated 10 data sets per combination of parameters, each with N=250 trials worth of data and 400 time points per trial (i.e. one second for a sampling frequency of 400Hz). The total number of channels is set to 64.

Phase resetting and frequency entrainment underlie observed decoding patterns

In real data, we typically see oscillatory patterns in the TGM, indicating that the subspace of brain activity that carries information about the stimulus must have some oscillatory element. At least two distinct mechanisms may be behind this phenomenon: first, an ongoing oscillation might reset its phase and then entrain to a given frequency in a condition-specific fashion; second, a condition-specific oscillatory response might by added to the signal after stimulus presentation on top of the existing ongoing oscillation. I use *genephys* together with decoding analysis to compare between these two alternatives, showing that phase reset followed by frequency entrainment produce decoding patterns that are closer to what we observe empirically.

I ran simulations with a large number of relevant channels (p=32) and relatively low activation probability (θ=0.25). For frequency entrainment, I considered a diverse range of frequencies across channels; for the additive oscillatory response, I added an oscillation with a channel-specific frequency to each channel's ongoing signal. (We will see below that channel stochasticity and frequency diversity are both important to produce realistic decoding patterns). **Figure 2** shows TGMs (top) together with one channel average evoked responses (where each stimulus is represented by a different tonality of blue or red) for the sampled signal x and its various constitutive elements: ongoing phase φ, frequency f, amplitude a, and, when applicable, the additive response y. At the left, results are shown for phase reset plus frequency entrainment; here, the average evoked response shows a condition-specific modulation in the ongoing phase φ. At the right, results are shown for the additive oscillation; here, there is no effect in φ, and, instead, the additive oscillatory component y is shown at the bottom.

Altogether, although the TGM that corresponds to phase resetting plus frequency entrainment still lacks some aspects that we see in real data (such as the broadening of accuracy in later stages of the trial; see below) it shows more realistic patterns in the TGM than the additive oscillatory response. This is because the additive oscillation interferes with the phase of the ongoing oscillation if both have frequencies that are not too different. This, in turn, impedes cross-trial phase locking throughout the trial, cancelling any oscillatory patterns in the TGM. On the other hand, phase resetting of the ongoing oscillation and subsequent frequency entrainment allows for a more consistent cross-trial phase locking throughout the trial because there are not two coexisting oscillations interfering with each other. A hypothetical third possibility is that the ongoing oscillation gets supressed and then replaced by an additive oscillatory component; this scenario, however, is equivalent, from the generative process' point of view, to phase reset and frequency entrainment.

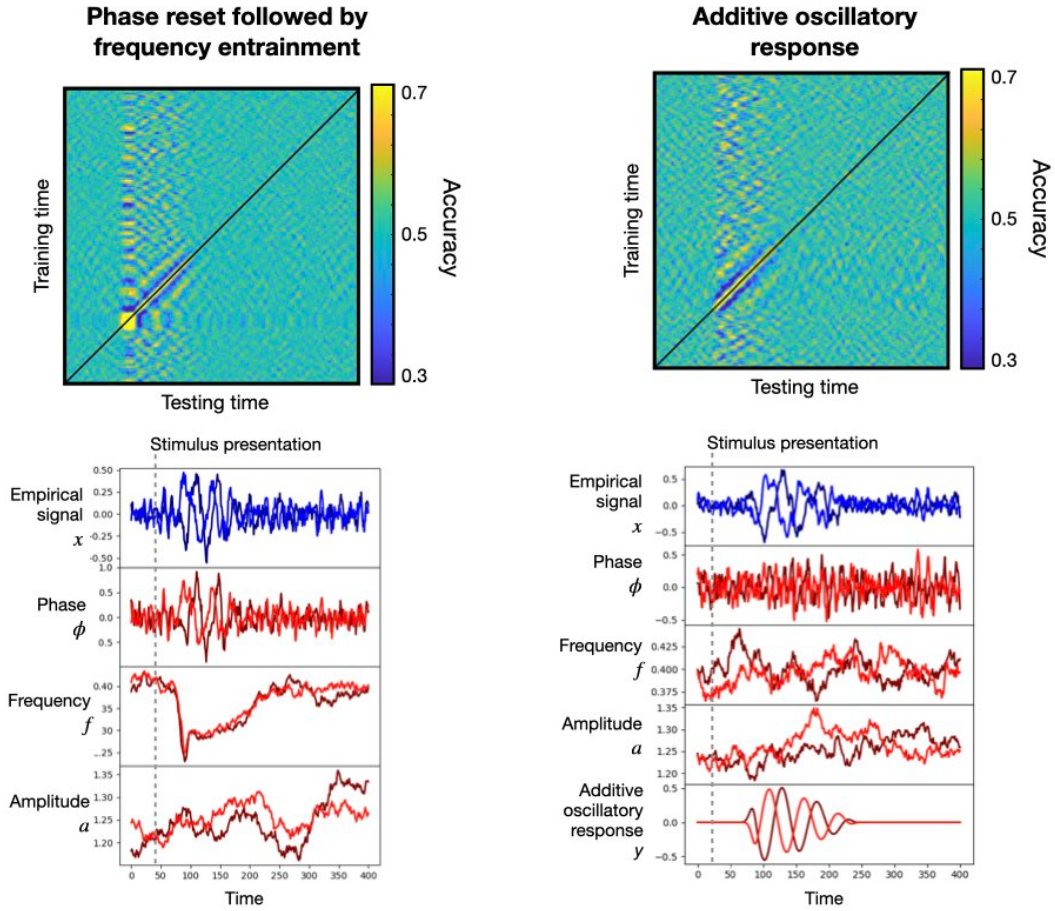

Figure 2. A configuration based on phase resetting and frequency entrainment (left) exhibits more realistic patterns than a configuration based on additive oscillatory responses (right). At the top, temporal generalisation matrix (TGM) from standard decoding analysis; on the bottom, average evoked responses for the sampled signal (blue) and its components (red), for the two conditions (represented as different tonalities of blue or red).

The distribution of stimulus information across channels is high-dimensional and stochastic

Next, I use *genephys* to show that the stimulus-specific information spans a large number of channels, and do so stochastically. As in the previous section, I apply phase resetting and subsequent frequency entrainment, with no amplitude enhancement or additive responses.

In this simulation, I varied the number of relevant channels p as well as the probability of activation θ for the relevant channels, so that the subset of channels that activate is different for every trial. **Figure 3** shows the TGMs for various combinations of p and θ, from a configuration where there are few relevant channels that always respond ($p = 2$, $\theta = 1.0$) to another where there are many relevant channels that only respond sparsely ($p = 32$, $\theta = 0.125$). The top row represents one run of the sampling/decoding procedure, and the bottom row represents the cross-run average.

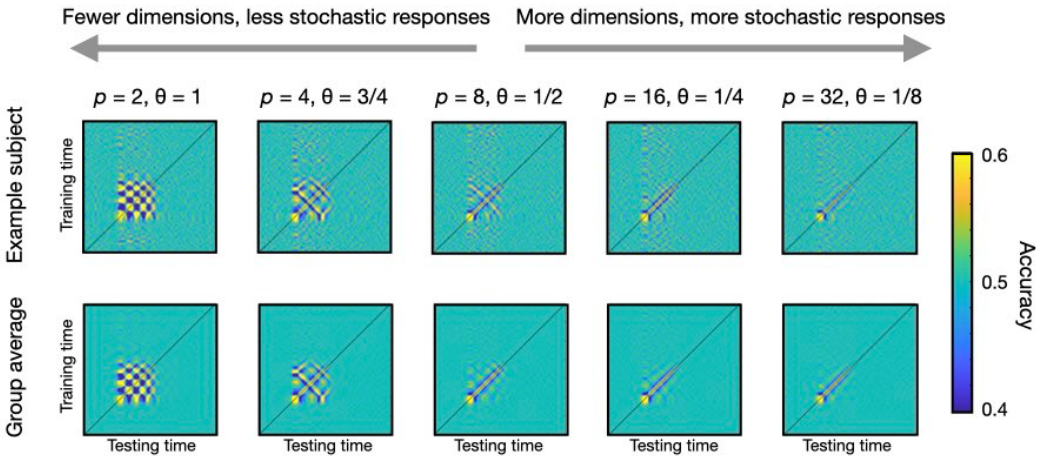

Figure 3. Having more dimensions (channels) carrying stimulus-specific information, but with a larger degree of stochasticity, produces more realistic decoding patterns than having fewer dimensions with a lower degree of stochasticity. Here stochasticity referred to the channels' probability of activation.

This experiment shows that having only a few relevant channels produces a chequer pattern not observed in real data, where the TGM has a relatively stylised diagonal extending from the point of maximum accuracy (here, around t = 250ms). This indicates that in real data, the subspace of the data that contains information about the stimulus needs to be high-dimensional, i.e. that the amount of relevant channels must be high. However, this must occur in combination with an unstable informational contribution of the channels —here in the form of probabilistic participation—, or else the decoding accuracy becomes unrealistically perfect.

Channel participation must therefore be unstable, but are there other types of stochasticity besides probabilistic channel participation that can generate results as we observe in real data? A different alternative to model instability is that channels always respond, but they do so by resetting to a noisier phase. To explore this possibility, I ran a new simulation where I manipulated the standard deviation σ_φ of the phase resetting distribution, while fixing $\theta = 1.0$ and $p = 32$ (i.e. relevant channels always respond).

Figure SI-2 shows that by having a noisier distribution of the target phase, we can also generate realistic TGMs with unperfect accuracies and features as we typically observe in real data. For ease of comparison between the two sources of variability (channel probabilistic activation and noisy phase resetting), **Figure SI-3** shows TGMs over a 2D grid of choices for θ and σ_φ (averaged across 10 simulations), and **Figure SI-4** shows the correlations between each pair of TGMs across configurations. As observed, there are high similarity of results across various configurations of the space of hyperparameters, showing that the TGM is indeterminate in disambiguating whether the true variability is caused by one or the other cause.

In summary, we observe that the subspace of brain activity that carries out information about the condition is highly stochastic at the single trial level, and that this can owe to at least two different causes. While these two types of variability are distinct and potentially point at different neurobiological mechanisms, the standard use of decoding analysis —based on cross-trial averages in both training (i.e. when estimating the regression coefficients) and testing (i.e. when calculating the accuracy)— cannot fully disambiguate between them.

Frequency entrainment span multiple frequencies

In the previous sections, I showed that an unstable phase reset across multiple channels, followed by frequency entrainment, can generate the decoding patterns we see in real data. Next, I demonstrate that frequency entrainment must span a diversity of frequencies across channels —i.e. different channels must entrain at slightly different frequencies that are not multiples of each other. In real data, this frequency diversity could manifest, for example, as a gradient of frequencies from primary towards more associative areas.

As mentioned, real data normally yield a relatively tight band of high decoding accuracy along the diagonal, often accompanied from contiguous parallel bands of below-baseline. Critically, the fact that we do not usually observe a chequer pattern means that the trajectory of phase across channels does not repeat itself periodically. I show this here empirically using *genephys*. **Figure 4** shows different regimes of frequency entrainment, for one run of the simulation (top) and across runs (bottom). In all cases, frequencies are the same for the two stimuli, so nothing can be decoded from the frequency of entrainment. In **Figure 4a**, all channels entrain to the same frequency (0.30 in angular frequency). This produces an unrealistically regular chequer pattern due to the channels becoming coherent to each other after phase reset, such that a decoder trained at a certain time point will become equally accurate again after one cycle at the entrained frequency. We can observe this effect for a single run and on average. Similarly, in **Figure 4b**, the channels entrain to one of three frequencies: 0.15, 0.30 or 0.45. In this scenario, we observe a frequency aliasing effect that makes the decoders to reoccur at the rate of the slowest frequency. This is also not what we observe in real data. In **Figure 4c**, every channel entrains at a different frequency, ranging from 0.15 to 0.45 and no frequency is an exact multiple of another. Now, the decoding results are a much better approximation of what we observe in real data.

In summary, these results show that it is not only important for the stimulus-relevant subspace of activity to be high-dimensional in space, but also in frequency.

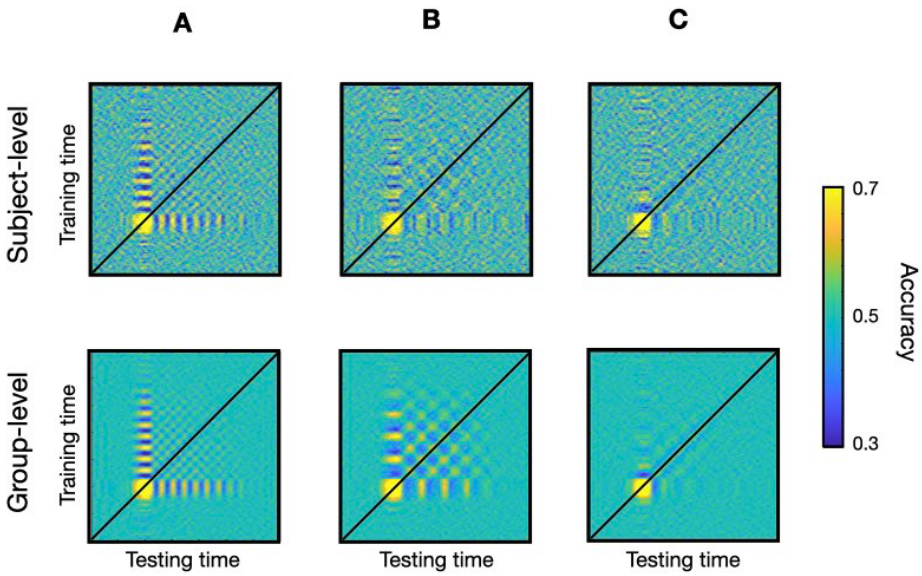

Figure 4. Channels must entrain to a wide range of frequencies. **A**: All channels entrain at the same frequency. **B**: Channels entrain at three different frequencies that are multiple of each other. **C**: Each channel entrain at a different frequency, and no frequency is a multiple of another frequency. Top, results from a single simulation (subject); bottom, group average.

Frequency entrainment is accompanied by condition-specific additive activity

So far, I have generated data that exhibits some of the periodic features we observe in decoding analysis results. In real data, however, we normally also see a general broadening of decoding accuracy (see **Figure SI-1**). This is often interpreted as neural representations becoming more stable at latter stages of the trial, which is putatively linked to memory encoding and higher cognitive processes. I here show that this effect can be reproduced in synthetic data through an additive, slowly-progressing response that is not oscillatory.

Specifically, I set up a response function such that, after stimulus presentation, the additive response ramps up to a condition-dependent target value ($-\pi/4$ for one condition and $+\pi/4$ for the other), and then slowly decays for a period of around half a second (200 time points). The rest of the simulation was set up as in **Figure 2A**. **Figure 5** shows the effect of adding this response, for both the TGM and the average response function (the average of the additive non-oscillatory response is depicted as the last curve of the right panel). The spread of accuracy in the TGM closely resembles what we observe in real data.

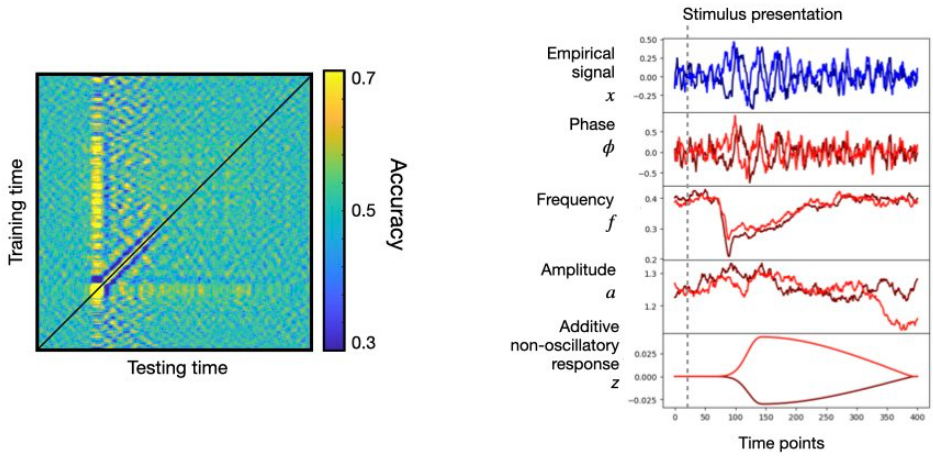

Figure 5. An additive non-oscillatory response gives rise to a slow expansion of decoding accuracy in the TGM toward the end of the trial. Left, TGM; right, average evoked response.

Amplitude increases modulate the size of the effect

Are modulations in the amplitude of the stimulus-specific oscillation necessary for the effects we observe in real data? Next, I show that they are not, but they can still occur and enhance the already existing patterns.

I generated various data sets, all with the same phase modulation based on phase resetting and subsequent frequency entrainment. In each of them, I applied a different amount of amplitude enhancement, from no increase to a ten-fold increase —which, critically, did not entail any effect on the signals' phase. Note that an amplitude effect without a phase effect would not result in any phase locking across trials, and therefore could not lead to any decoding accuracy. Having applied decoding analysis, I looked at two characteristic features of the TGM: (i) the diagonal, and (ii) a vertical cut stemming from the time point of maximum accuracy, which contains information about recurrence. **Figure 6** shows the results. As observed, the main features are present in all runs regardless of the size of the amplitude effect, although they were more prominent for higher amplitude modulations; for example, frequency entrainment could not be clearly observed beyond two cycles when no amplitude modulation was applied, even though it was present in the generative model.

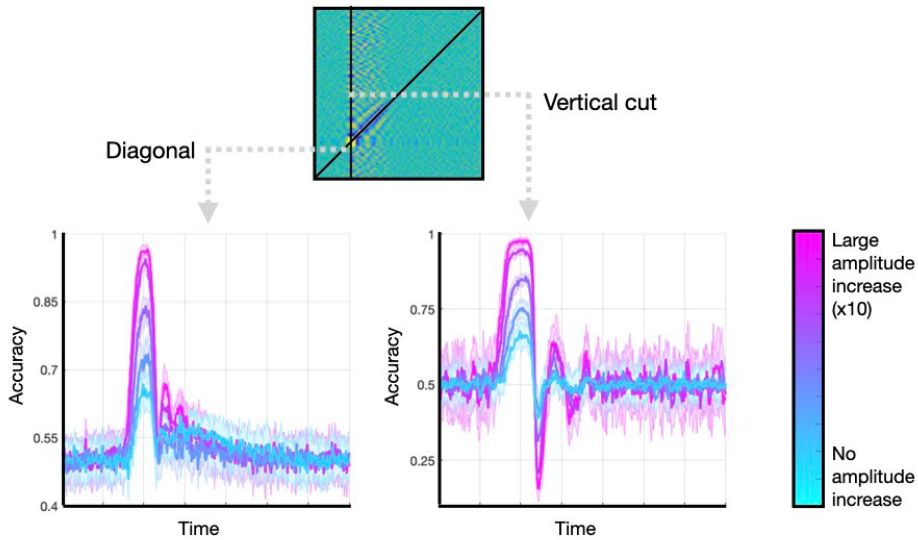

Figure 6. Amplitude enhancement of the stimulus-relevant oscillation is not strictly necessary to produce realistic results, but it can enlarge the effects. Two features of the TGM are highlighted: the diagonal, and a vertical cut at the time point of maximum accuracy.

Discussion

The brain's responses to even the simplest stimuli are characteristically variable. This is not surprising, given the brain's plasticity, and that it never stops generating ever-changing endogenous activity. It also speaks to the brain's degeneracy (Edelman & Gally, 2001) —i.e. the fact that there are many possible neural trajectories that can achieve a single purpose. How this variability translates into behaviour and experience is however an open question. Anyhow, it seems reasonable that, at some level, brain responses must keep some invariant aspects so that our perceptual experiences remain stable. Here, I investigated, using a generative model, the most stable aspects of brain responses to stimuli as seen through the lens of decoding accuracy, which is by definition based on averaging.

Previous work has analysed the nature of brain responses to perceptual stimulation using average evoked responses (Sauseng, Klimesch, Gruber, Hanslmayr, & Doppelmayr, 2007), arguing either for a predominant role of additive responses (Shah, et al., 2004; Mazaheri & Jensen, 2006) or phase resetting (Makeig, et al., 2002). These studies looked at the average response to a given stimulus, but did not investigate what aspects of the data carry information about the identity of the stimulus; this is the focus of this paper.

It is important to note that the list of effects that I have explored is non-exhaustive. For example, I have considered additive oscillatory responses in the form of sinusoidal waves. Another possibility could be to have additive oscillatory responses with a tendency to spend more time in certain phases (for example, having wider peaks than throughs); in this case, even in the absence of phase locking between trials, we could potentially have a significant evoked response due to phase asymmetry (Nikulin, et al., 2007). For simplicity, I have considered independent sources of activity that exhibit correlations only through the induced effect. In practice, brain areas are in constant communication even in the absence of stimulation. Alternatives where sources are modelled as coupled oscillators (Breakspear, Heitmann, & Daffertshofer, 2010; Cabral, et al., 2014), or at least where there is correlated noise, are also

possible. Also for simplicity, I have considered that, at a given time point, every channel has only one fundamental frequency (without counting the additive responses, which have their own temporal dynamics); in practice, multiple frequencies coexist and interact, potentially modulating the TGM. Finally, the inherent stochasticity of the stimulus-specific space of activity can take various forms. Here I have explored two —noise in the phase distribution and a probabilistic activation of the channels—, but others are also possible: for example through condition-specific delays, either absolute across channels or relative (Vidaurre, Cichy, & Woolrich, 2021). Importantly, I have shown that standard decoding analysis can differentiate between these explanations only to some extent. In future work, alternatives to standard decoding analysis might be used to disentangle these aspects (Vidaurre, Myers, Stokes, Nobre, & Woolrich, 2019). In order to explore the space of configurations more effectively, an automatic exploration of the hyperparameter space using, for instance, Bayesian optimisation (Lorenz, Hampshire, & Leech, 2017) is an alternative for future consideration.

Overall, these results suggest that the stable aspects of brain activity regarding stimulus processing are underpinned by phase reorganisation and frequency entrainment, in a space that is high-dimensional in both the spatial domain (it must span many channels) and the frequency domain (it must involve a great diversity of frequencies). This effect is accompanied by a non-oscillatory additive response, and, optionally, by an amplitude modulation. The results also show that, above and beyond these average patterns, the stimulus-specific subspace of brain responses is highly stochastic at the trial level. Moving forward, the evidenced role of ongoing oscillations in stimulus processing, together with its high dimensionality in space and frequency, point at a relevant role for largely distributed neural ensembles besides, or instead of, more localised circuits.

Acknowledgements: DV is supported by a Novo Nordisk Foundation Emerging Investigator Fellowship (NNF19OC-0054895) and an ERC Starting Grant (ERC-StG-2019-850404).

Bibliography

Dowling, J. E. (1987). *The retina: an approachable part of the brain.* Harvard University Press.

McIntosh, A. R., Kovacevic, N., & Itier, R. J. (2008). Increased Brain Signal Variability Accompanies Lower Behavioral Variability in Development . *PLOS Computational Biology*, e1000106.

Steven, L. J., & Kappenman, S. E. (2011). *he Oxford handbook of event-related potential components.* Oxford university press.

Axel, R. (1995). The molecular logic of smell. *Scientific American*, 154-159.

Haxby, J. V., Connolly, A. C., & Guntupalli, S. (2014). Decoding Neural Representational Spaces Using Multivariate Pattern Analysis. *Annual review of neuroscience*, 435-456.

King, J.-R., & Dehaene, S. (2014). Characterizing the dynamics of mental representations: the temporal generalization method. *Trends in cognitive sciences*, 203-210.

Dawson, G. D. (1954). A summation technique for the detection of small evoked potentials. *Electroencephalography & clinical neurophysiology*.

Stokes, M., Wolff, M., & Spaak, E. (2015). Decoding Rich Spatial Information with High Temporal Resolution . *Trends in Cognitive Sciences*, 636-638.

Croner, L., Purpura, K., & Kaplan, E. (1993). Response variability in retinal ganglion cells of primates. *Proceedings of the National Academy of Sciences*, 8128-8130.

Freeman, W. J. (1978). Spatial properties of an EEG event in the olfactory bulb and cortex. *Electroencephalography and Clinical Neurophysiology*, 586-605.

Vidaurre, D., Cichy, R. M., & Woolrich, M. W. (2021). Dissociable Components of Information Encoding in Human Perception. *Cerebral Cortex*, 5664-5675.

Stein, R. B., Gossen, E., & Jones , K. E. (2005). Neuronal variability: noise or part of the signal? *Nature Reviews Neuroscience*, 389–397.

Garrett, D. D., Samanez-Larkin, G. R., MacDonald, S. W., Lindenberger, U., McIntosh, A. R., & Grady, C. L. (2013). Moment-to-moment brain signal variability: A next frontier in human brain mapping? *Neuroscience & Biobehavioral Reviews*, 610-624.

Freeman, W. J. (2000). *How brains make up their minds.* New York: Columbia University Press.

Vidaurre, D., Myers, N., Stokes, M., Nobre, A. C., & Woolrich, M. W. (2019). Temporally unconstrained decoding reveals consistent but time-varying stages of stimulus processing. *Cerebral Cortex*, 863-874.

Nikulin, V., Linkenkaer-Hansen, K., Nolte, G., Lemm, S., Müller, K., Ilmoniemi, R., & Curio, G. (2007). A novel mechanism for evoked responses in the human brain. *European Journal of Neuroscience*, 3146-3154.

Edelman, G., & Gally, J. (2001). Degeneracy and complexity in biological systems. *Proceedings of the National Academy of Sciences* , 13763-13768.

Breakspear, M., Heitmann, S., & Daffertshofer, A. (2010). Generative models of cortical oscillations: neurobiological implications of the Kuramoto model. *Frontiers in human neuroscience*, 190.

Cabral, J., Luckhoo, H., Woolrich, M., Morten, M., Mohseni, H., Baker, A., . . . Deco, G. (2014). Exploring mechanisms of spontaneous functional connectivity in MEG: How delayed network interactions lead to structured amplitude envelopes of band-pass filtered oscillations. *NeuroImage*, 423-435.

Lorenz, R., Hampshire, A., & Leech, R. (2017). Neuroadaptive Bayesian optimization and hypothesis testing. *Trends in Cognitive Sciences*, 155-167.

Pfurtscheller, G., & Lopes da Silva, F. (1999). Event-related EEG/MEG synchronization and desynchronization: basic principles. *Clinical Neurophysiology*, 1842-1857.

Kragel, P. A., Koban, L., Feldman Barrett, L., & Wager, T. D. (2018). Representation, pattern information, and brain signatures: from neurons to neuroimaging. *Neuron*, 257-273.

Supplemental methods

Mathematical specification of the response function

Assuming the stimulus was presented at time point τ within a given trial, and that both t and τ are expressed in number of time points, the response function is piecewise, with a left exponential part and a right logarithmic part, being $t_{\max}$ the time point when exponential changes to logarithmic (and where the function takes the value 1.0 at either side). Mathematically:

$$g_t = \mathrm{G}(t;\ \tau) = \ 0.0,\ \text{if}\ t\ <\ \delta_1 + \tau$$

$$g_t = \mathrm{G}(t;\ \tau) = \frac{e^{-\frac{0.5\,(\delta_1+\delta_2+\tau-t)^2}{C_1}}}{C_2}\ ,\ \text{if}\ t\ >\ \delta_1 + \tau\ \text{and}\ \ t \leq \delta_1 + \delta_2 + \tau$$

$$g_t = \mathrm{G}(t;\ \tau) = \ -\log\left(1 + \left(\frac{\mathrm{t}-\tau-\delta_1-\delta_2-1}{T_2}\right)^{\varsigma}\right) + C_3,\ \text{if}\ t > \delta_1 + \delta_2 + \tau$$

$$g_t = \mathrm{G}(t;\ \tau) = \ 0.0,\ \text{if}\ t > \tau + \delta_1 + \delta_2 + \tau + 1 + \ T_2,$$

where

- δ_1 reflects the delay of the response in number of time points,
- δ_2 is how many time points it takes the exponential function to go from zero to its maximum before it changes to the logarithmic function,
- $t_{\max} = \delta_1 + \delta_2 + \tau$ is the time point of maximum response (i.e. the changing point between the two functions),
- C_1 is a constant such that $\mathrm{G}(\delta_1 + \tau\ + 1;\ \tau) \leq\ 0.001$,
- C_2 is a normalisation constant such that $\mathrm{G}(t_{\max};\ \tau) = 1.0$ from the left side.
- ς determines the shape of the logarithmic function,
- T_2 represents the duration in time points of the logarithmic response function, and
- C_3 is a normalisation constant such that the logarithmic function is also bounded between 0.0 and 1.0, and $\mathrm{G}(t_{\max};\ \tau) = 1.0$ from the right side as well.

Note that, thanks to the normalisation constants, both the left and the right side of the response function take values between 0.0 and 1.0.

Supplemental figures

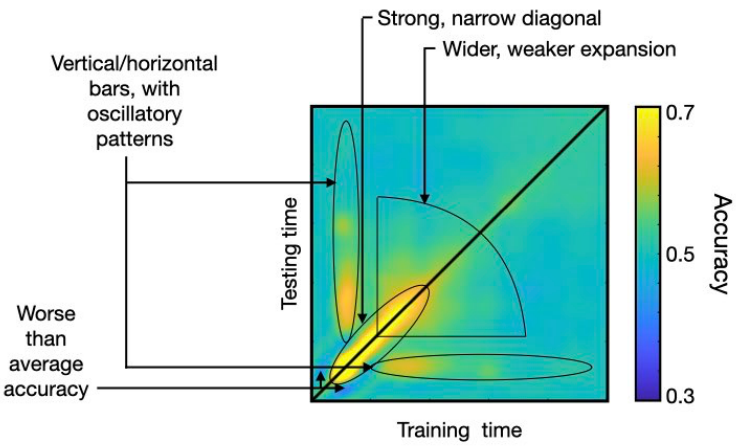

Figure SI-1: Example of real TGM, where the different characteristic features have been highlighted.

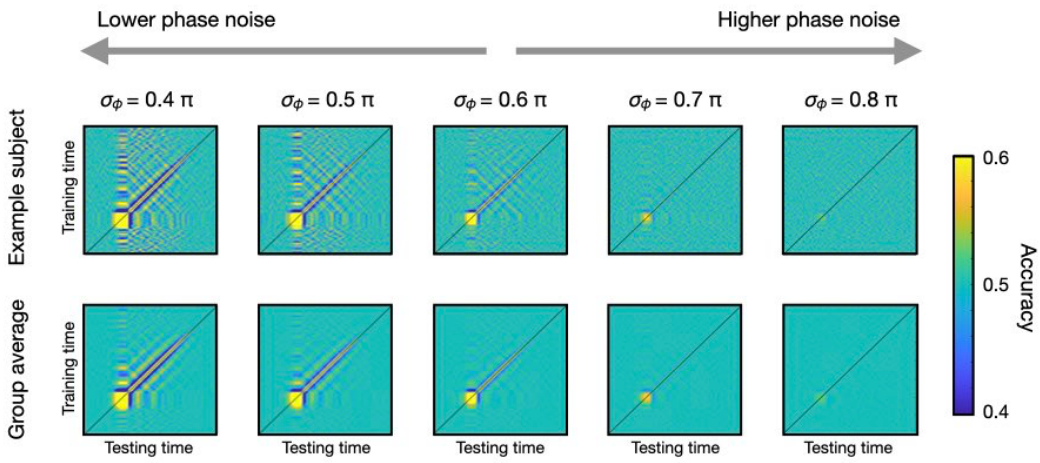

Figure SI-2: Increasing phase noise produces more realistic decoding patterns, akin to having a probabilistic channel activation.

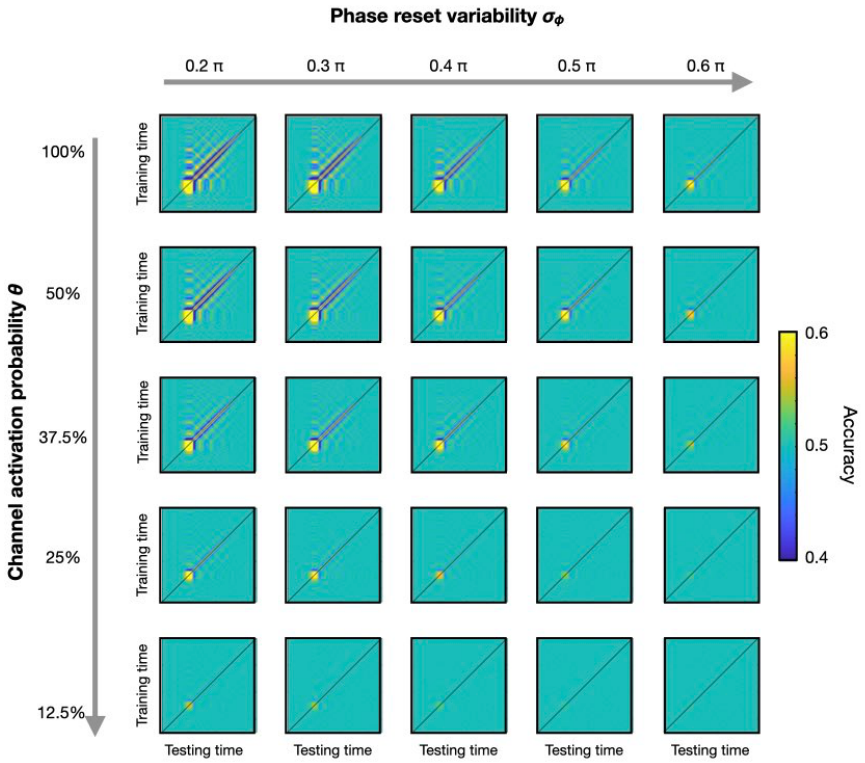

Figure SI-3: TGMs across different choices of phase noise and channel activation probability.

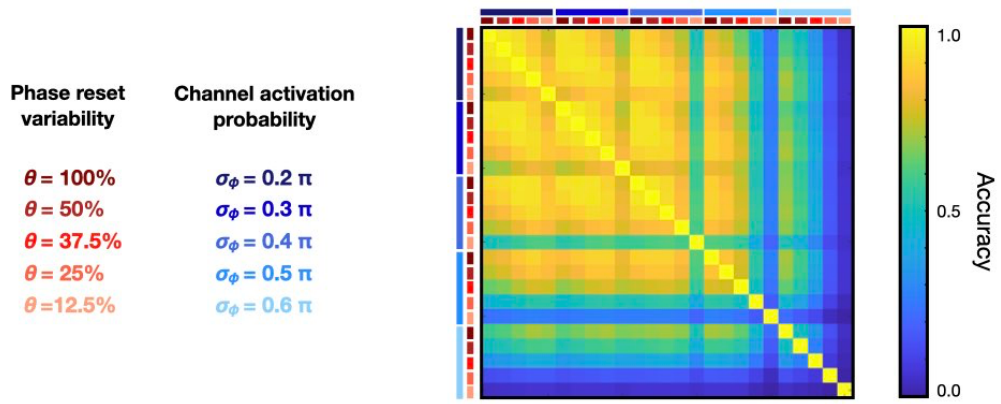

Figure SI-4: Correlation of the vectorised off-diagonal elements of the TGMs, across the different configurations shown in Figure SI-3.