Targeted comodulation supports flexible and accurate decoding in V1

Caroline Haimerl[1], Douglas A. Ruff[3], Marlene R. Cohen[3], Cristina Savin[1,2,†], and Eero P. Simoncelli[1,2 †]

[1] Center for Neural Science, New York University, New York, NY 10003, USA
[2] Center for Data Science, New York University, New York, NY 10011, USA
[3] Department of Neuroscience and Center for the Neural Basis of Cognition, University of Pittsburgh, Pittsburgh, Pennsylvania, USA.
[†] Co-senior authors

Abstract

Sensory-guided behavior requires reliable encoding of stimulus information in neural responses, and task-specific decoding through selective combination of these responses. The former has been the topic of intensive study, but the latter remains largely a mystery. We propose a framework in which shared stochastic modulation of task-informative neurons serves as a label to facilitate downstream decoding. Theoretical analysis and computational simulations demonstrate that a decoder that exploits such a signal can achieve flexible and accurate readout. Using this theoretical framework, we analyze behavioral and physiological data obtained from monkeys performing a visual orientation discrimination task. The responses of recorded V1 neurons exhibit strongly correlated modulation. This modulation is stronger in those neurons that are most informative for the behavioral task and it is substantially reduced in a control condition where recorded neurons are uninformative. We demonstrate that this modulator label can be used to improve downstream decoding within a small number of training trials, consistent with observed behavior. Finally, we find that the trial-by-trial modulatory signal estimated from V1 populations is also present in the activity of simultaneously recorded MT units, and preferentially so if they are task-informative, supporting the hypothesis that it serves as a label for the selection and decoding of relevant downstream neurons.

1 Introduction

Humans and animals are capable of flexibly switching between a multitude of tasks in different contexts, each requiring rapid, sensory-informed decision making. Incoming stimuli are processed by a hierarchy of circuits comprised of millions of neurons with diverse feature selectivity. How does the brain orchestrate these behaviors, which require selective gathering of the sensory information relevant for each task? Specifically, in the visual system, as neurons respond selectively to different features at different locations in the visual field (Fig. 1A), performance in a sensory-guided task relies on the ability to properly gather and combine their responses. This process – generally called "decoding" – needs to be both flexible to task changes and accurate.

Decoding of neural responses has been extensively studied, primarily for the purpose of building artificial prostheses, or with the objective of understanding information encoding. In particular, the "ideal observer" paradigm uses an optimal statistical decoder based on the response properties of the encoding population, which provides a performance bound against which behavioral data can be compared [1; 2; 3; 4; 5]. Two basic considerations suggest that such models do not provide a plausible framework for understanding how decoding can be achieved in later stages of cortical processing. First, computing optimal decoding weights requires full knowledge of each neuron's stimulus-response and noise properties, which seems implausible for a downstream decoding area to acquire and store. Second, decoding weights could perhaps be learned directly within the task, but this would require large numbers of repeated exposures, precluding the behavioral flexibility found in humans and animals.

As a potential alternative, task-specific information could be provided as a top-down signal that guides decoding. Top-down signals are thought to underlie attentional effects, which produce increases in response signal-to-noise ratio through selective increases in response gain [6; 7]. Although this can improve the accuracy of encoded information, it is not clear that it could serve to facilitate decoding. Here, we propose that task-specific information flow is achieved by altering the gain of the relevant neurons – not via an increase in mean, but rather through a shared stochastic modulatory signal. Evidence for shared, multiplicative noise has been reported across a variety of brain regions and tasks [8; 9; 10; 11] and can be described by stochastic modulators that affect neural population responses on different time scales [10] (Fig. 1A). The neural mechanisms underlying these modulatory effects remain unknown, and are likely to involve a combination of recurrent circuit dynamics and feedback signals (see Discussion). We remain agnostic to the mechanistic processes underlying the modulation and focus on their functional role. Analysis of visual responses suggests that modulatory effects are strongest in the most task-informative units [11; 12]. We propose that this

counterintuitive observation, in which noise is introduced in a way that is most deleterious to the encoded signal, provides a compensatory decoding benefit: the modulatory fluctuations serve as a "label" for the neurons that carry task-relevant information, enabling downstream stages to flexibly adapt their decoding strategies to the current task. We develop a modulator-guided decoder and show through simulations that modest levels of task-specific stochastic modulation of an encoding population can lead to a substantial overall benefit in decoding accuracy.

We examine the implications of this theory in the context of a visual experiment in non-human primates [13; 14]. Monkeys were trained to discriminate between two drifting windowed gratings, while recording a population of neurons in visual area V1, as well as a single units in downstream area MT. We show that the population activity in V1 exhibits fluctuations consistent with a shared modulator. Within the recorded population, this modulator preferentially targets task informative neurons. We demonstrate that this creates a functional label that can facilitate readout and reduce the amount of training required. We further show that the responses of simultaneously recorded units in area MT exhibit the same modulator labeling, supporting the hypothesis that the task-specific labeling is propagated through the visual hierarchy.

2 Results

Monkeys were trained to detect a small change in orientation of a Gaussian-windowed drifting sine grating (a "Gabor" function - Fig. 2A), while recording spiking responses of neurons in their primary visual cortex (area V1) and area MT, mostly in the form of multiunits (Fig. 2B) [14]. Three gratings were present simultaneously on the screen, but only one was relevant for the animal's decision. The location of the relevant stimulus was fixed for each block of trials, but switched randomly across blocks throughout an experimental session. The monkey was able to quickly adjust to the new stimulus location [14], reaching maximal performance levels after only a few trials (Fig. 2C-D). We aim to explain how the brain achieves such accuracy and flexibility in this type of task.

2.1 Encoding of local visual orientation in a V1 population

V1 neurons respond selectively to the orientation of local stimuli and the ensemble of neurons spans the visual field and all orientations. In the experiment by Ruff and Cohen [14], the stimulus size is roughly matched to V1 RF sizes at the eccentricity at which recordings are performed and the task requires detecting fine orientation changes (10-45°). Consequently, these stimuli selectively drive a small subset of neighboring V1 neurons, whose responses can be used to differentiate the stimulus orientation if combined in the right way. Since nearly all visual information passes through V1 [15], a downstream areas' sole source of information about fine orientation changes lies in the responses of those few V1 cells. Effectively, the downstream area needs to find those task-informative neurons, while ignoring the background chatter of activity from the remainder of the population.

Two of the three experimental stimuli are located so as to overlap the RFs of the recorded V1 population (Fig. 2A). If one of those stimuli is task-relevant, we expect a subset of the recorded neurons to provide information for the animal's decision ("relevant blocks/tasks") (see Fig. 2A from [14]). We quantified task-informativeness of each V1 unit with d' (the mean response change between stimulus 0 and 1 relative to response standard deviation) (Fig. 2E). Figure 2E shows the relationship between informativeness and responsiveness in three representative examples. First, we find units that are only weakly responsive to both stimuli and consequently cannot be informative about stimulus identity. Second, there are units that respond strongly but similarly to both stimuli which demonstrates that responsiveness is necessary but not sufficient. And third, we find units that respond strongly only to one of the two stimuli and hence have high informativeness. Overall, we find that a modest proportion of the recorded V1 units are significantly informative in the relevant task (monkey 1: 25.8%, monkey2: 18.4%). In contrast, only 4.3% of units are significantly informative in the control task (Fig. 2F). This is expected: the third stimulus is outside the RF of these neurons, and thus their activity should not be informative for discriminating stimuli in that region ("control blocks/task"). Finally, we observe that neurons that are informative in the relevant task have a particularly low $|d'|$ in the control task, reflecting their task-specificity (Fig. 2G).

From the perspective of a downstream circuit, it is necessary to read out selectively from a different subpopulation of V1 cells when the task changes from using the relevant to the control stimulus in order to make accurate decisions. This is difficult because the informative neurons do not differ sufficiently from the uninformative neurons with respect to their basic statistics (mean activity or overall variability, see Suppl. S2). This selection must happen quickly, in order to account for the observed behavioral flexibility of the monkey, whose performance reaches asymptotic levels roughly 5 trials after each task change within a session (Fig. 2D).

2.2 Shared noise corrupts encoding: the modulated Poisson population model

Neural responses to stimuli fluctuate from trial to trial, and these variations are usually interpreted as "noise" that corrupts the encoded stimulus information. While some of this noise is neuron-specific, some is correlated, and can be explained by shared sources of multiplicative trial-to-trial variability, suggesting that time-varying modulators influence the response of neurons [16; 10] (Fig. 1A). Experimental results suggest that some of these noise sources in visual areas preferentially target task-informative neurons [11; 12]. From an encoding perspective, it is counterintuitive that the system would deliberately corrupt the responses of task-informative neurons (Suppl. S1). However, we propose that this fluctuating modulator serves to "label" task-relevant neurons in primary visual area V1, and that this labeling is propagated to subsequent stages of processing, and eventually to decision areas in frontal cortex.

We illustrate this theory of decoding by constructing a model for V1 population encoding during a binary discrimination task. We assume modulated Poisson responses [10], with a rate that depends on the product of stimulus drive, $\lambda_n(s_t)$, and a stochastic modulation factor m_t (see Methods, and [17]):

$$k_{n,t} \sim \text{Poisson}\left(\lambda_n(s_t)\exp(w_n m_t)\right), \tag{1}$$

where $k_{n,t}$ is the spike count of neuron n at time t; the modulator m_t is a signal that varies randomly in time t with zero mean, and influences neuron n with weight w_n that is proportional to its task-informativeness.

The ideal observer optimal decoder for a binary discrimination task based on neural responses with modulated Poisson statistics compares a weighted sum of the neural responses with a decision threshold $c(m_t)$, that depends on the time-varying modulator (see Methods):

$$\sum_n a_n^{(\text{opt})} k_{n,t} > c(m_t), \qquad \text{with } a_n^{(\text{opt})} = \log(\lambda_n(1)) - \log(\lambda_n(0)). \tag{2}$$

where $a_n^{(\text{opt})}$ denotes the optimal decoding weights. It is useful to separately consider the amplitude and sign of the decoder weights, a_n. The weights with non-zero amplitude are associated with the small subpopulation of informative neurons (Fig. 1B, purple), while the zero weights eliminate the remaining active but uninformative (black) or inactive neurons (Fig. 1B, grey). The sign separates the subpopulation into two groups, according to their relative preference for the two stimulus alternatives. This decoder provides an upper bound on decoding performance given the encoding model, and motivates the use of a linear-threshold functional form for the solution.

2.3 A biologically plausible theory for flexible decoding with targeted modulation

The ideal observer chooses weights based on full knowledge of each neuron's mean responses to the stimuli of the current task (Eq. 2), but how would the brain achieve the optimal solution (or some approximation thereof)? A conventional approach to learning decoding weights without assuming such detailed knowledge is regression. This is feasible for a small set of neurons, but the number of training examples needed for accurate weight estimation grows with number of neurons, and a population of realistic size would require an enormous number of training examples [18]. The behavioral flexibility exhibited by the monkeys precludes such a solution. Instead, we seek a heuristic alternative that is flexible, accurate and does not assume unrealistic knowledge of the encoding process.

Consider first a decoder motivated by early work on neural binary discrimination/detection [19]. The idea is to average the response of two sub-populations ("preferred" and "anti-preferred") and then compare these averages. This solution corresponds to choosing decoding weights of unit amplitude, but differing in their sign (± 1), which indicates the sub-population assignments but ignores relative importance. This *sign-only* (SO) decoder can be learned from relatively few training trials (Suppl. S1), by comparing the average responses of each neuron to the two stimuli. The performance of the SO decoder falls as the ratio of informative to uninformative neurons decreases (Suppl. S1): Since all neurons must be included in one of the two sub-populations, the noise from the uninformative neurons corrupts the decision signal. For realistically small fractions of informative neurons [16; 1], the SO decoder cannot match the levels of performance seen in the monkeys.

We can improve on the SO decoder by adjusting the amplitude of the decoding weights to reflect the relative informativeness of each neuron. If we assume (as verified in section 2.6) that the modulation strength of each neuron reflects its informativeness, the decoder amplitudes may be estimated by comparing each neuron's response to the modulation signal. Specifically, if the decoder has access to the modulator (e.g., it is a globally broadcast signal), it can use its temporal correlation with each neuron's activity to estimate the amplitude of the decoding weight for that neuron:

$$|a_n^{(\text{MG})}| \propto \frac{1}{T}\sum_t m_t k_{nt}, \tag{3}$$

This *modulator-guided* (MG) decoder does not rely on knowledge of the stimuli or the mean responses of the encoding population, and its estimates of weights can converge rapidly (at the time scale of the modulator) [17].

To examine the accuracy of this decoding approach, we simulated 1000 model V1 neurons, 50 of them informative, in trials of a binary discrimination task and compared to the other decoders (Fig. 1D). Under these conditions, the SO decoder performs poorly, due to noise from the large number of uninformative neurons. The optimal (ideal observer) decoder provides an upper bound on the accuracy of a linear decoder, which deteriorates as modulation strength increases and more noise corrupts the encoded signal. The performance of the MG decoder reflects two opposing effects of modulatory noise (Fig. 1C). Increasing the modulator strength makes the shared variability more salient, which increases the precision with which decoding weights are estimated. On the other hand, increasing modulator strength also leads to more corruption of the stimulus signal, decreasing encoding precision. Within an optimal range of modulator strength these two effects can be balanced and the MG decoder nearly reaches the accuracy level of the ideal observer.

In practice, the performance of the MG decoder depends on how well-correlated the modulator couplings (w_n) are with the task-informativeness, and how quickly the modulator couplings can be estimated. We tested the robustness of the results with regard to the former by examining the effect of adding Gaussian noise to the modulation couplings. Overall, the performance decreases, but the nonmonotonic dependence of the MG decoder performance on modulator strength is preserved (Fig. 1E). Regarding the second point, the convergence rate of MG estimation of decoding weights is determined by the time scale of modulatory fluctuations, whereas convergence of weights estimated via regression would be determined by rate of trial-by-trial feedback. If the modulator timescale is short relative to that of the behavioral trials, the MG decoder offers a substantial advantage in flexibility over the optimal decoder.

2.4 Testable predictions of the theory

Our theory relies on and predicts several attributes of neural response. Since the experimental stimuli are chosen so as to drive a subpopulation of V1 neurons differentially we predict that specifically these cells should exhibit task-specific modulator labeling (Fig. 1D). Beyond V1, the task-relevant stimulus information needs to propagate through several stages before reaching decision-making areas. In particular, area MT receives primary afferent input from V1, but its receptive fields are substantially larger, integrating over their respective afferent inputs [20]. Task-specific stimulus information consequently diffuses. A decoder needs to focus on the subset of neurons whose responses are affected by changes in the responses of the relevant afferent V1 neurons (Fig. 1A). Thus, the decoding problem identified in V1 persists, perhaps worsening, in downstream areas. Our theory of modulator-labeling naturally provides a solution for this problem; if task-informative neurons can be labeled in V1 by shared fluctuations, this labeling is inherited by exactly those downstream neurons that receive their signal. This suggests that the functional V1-modulator should also label task-informative units in area MT.

2.5 Evidence for shared modulation in V1

Our theory posits the existence of shared fluctuations in the gains of task-informative V1 responses. We sought evidence for this by examining population recordings from macaque monkeys discriminating the orientation of drifting gratings [14]. We fit a modulated stimulus response model (*"modulated SR-model"*) to the recorded population of V1 neurons (see Methods). The model was fitted to each block separately, jointly estimating both the individual stimulus drive to each neuron, and the stimulus-independent within-trial shared variability (Fig. 3A). As a baseline, we consider a reduced model that only takes into account the stimulus response (*"SR-model"*). The SR-model accounts for response transients using separate parameters corresponding to different time bins during stimulus presentation and accounts for the changes in contrasts by two sets of time-specific parameters each for one contrast condition (see Methods). This specific model instantiation allows us to test the predictions of the theory with respect to the existence, dimensionality and time scale of shared modulation.

We found that 91% of blocks are better fit by the modulated SR-model than by the SR-model alone. Moreover, when comparing modulated SR-models with modulators of different dimensionality, we find that 72% of blocks are best described by a one-dimensional modulator (see Fig. 3C and Methods for details). We restricted subsequent analyses to these blocks. For a fraction of the population the SR-model does not improve prediction over a constant rate model, suggesting that those neurons do not respond to changes in stimulus contrast. Informative neurons tend to be those best fit by the SR-model, which is expected given that their selectivity to stimulus orientation also makes them more likely to be sensitive to changes in contrast.

This model comparison confirms the presence of one-dimensional modulation in our data (Fig. 3E), with an average estimated time constant of 75ms, smaller than the average trial duration (about 3s) as well as that of the stimulus presentation (200ms). In fact, it approaches the time resolution at which the data was binned (50ms). This suggest that the modulator indeed provides a much faster training signal to learn decoding weights relative to the trial-by-trial feedback. It is also worth noting that although the estimated modulatory signal is fluctuating at a fairly rapid timescale, we do not see any evidence that it is periodic or oscillatory.

The nature of the modulator is unclear, however, we verified in Fig. 3F that modulator values were similar for both stimulus contrast conditions which suggests that it does not depend on the stimulus (more details in Suppl. S4). We further tested whether the modulation could result from adaptation and found that the amount of adaptation was uncorrelated with the quality of the fit of the modulated SR-model (Suppl. S3). Moreover, there was no relationship between informativeness and strength of adaptation (Suppl. S3). We wondered whether the same results could be obtained using a simpler dimensionality reduction method, such as principal component analysis (PCA). Although the first component does provide a crude estimate of the modulator, the PCA analysis suggests high-dimensional structure, primarily because PCA is not well-suited to the Poisson-like variability of spiking neurons, nor to the modulatory (multiplicative) nature of their shared variability (Suppl. S6). The low-dimensional structure of the population activity is only revealed using a model that takes these properties into account.

2.6 V1 modulator targets informative units

A key assumption of the theory is that the modulator targeting is task-specific. In our experimental context, the strength of modulation should change based on whether task stimuli are within versus outside the RF of the neurons (see section 2.1). Indeed, we find that the overall strength of the estimated modulation significantly decreases in the control task, compared to the two relevant task conditions (Fig. 3G), which have indistinguishable modulation strengths.

More precisely, the theory predicts a close relationship between the degree of modulation of individual cells and their task informativeness (Fig. 1). We find that the inclusion of dynamical modulators in the SR-model improves performance preferentially for informative neurons, with positive correlations between informativeness and model fit (Fig. 4A), suggesting that they may be more strongly modulated than neurons that are uninformative for the task. To test this more directly, we examined the correlation of the estimated modulator couplings (w_n) with task informativeness ($|d'|$). A non-parametric comparison shows that informative neurons have a higher modulator coupling rank than uninformative neurons in their respective population (Fig. 4B). Although informativeness does depend on the overall firing rate of a unit (Fig. 2E), a partial correlation analysis (Fig. 4C) confirmed that firing rate differences cannot explain the inferred modulation targeting (see also [11]). We find significant correlations between firing-rate-corrected informativeness and modulator couplings in 84% of the blocks (Fig. 4D). Multiunits may partially corrupt our estimates of this relationship, but they do not produce artifactual targeting structure (Suppl. S5).

We also asked whether the relationship between informativeness and neural variability is specific to the fluctuations due to the shared modulator or arises because informative neurons are generally more noisy. Each unit has residual variability unexplained by the modulated SR-model, a proxy for its private noise. However, this private noise does not significantly correlate with informativeness (Fig. 4D); only 9% of correlations are significantly positive (19% significantly negative). Overall, this suggests that the increased variability in informative neurons is primarily coming from the shared modulator. The modulator coupling is also dissociable from traditional attentional effects on mean firing rate (see Suppl. S7).

Switching between the relevant and the control tasks leads to changes in estimated modulation coupling that correlate with the change in informativeness (r=0.12, p=0.002); the selection bias in the recorded sub-population does not allow to directly estimate the relationship between informativeness and coupling in the control task (as informativeness is invariably very close to zero, see Fig. 2G). When directly comparing modulation coupling between the relevant and control tasks for the same units, we find correlations between the two. This suggests that the fine targeting structure may not completely change on the time scale of the task switches. Instead, the modulation strength in the recorded subpopulation is dialed up and down when switching between the relevant to the control blocks, reflecting the overall change in informativeness of the population. The modulator coupling should determine how strongly neurons are weighted by upstream areas. Therefore, neurons that are strongly modulated should have a stronger influence on behavior. It has been proposed that the most informative neurons should have a stronger influence on behavior [21]. Despite V1's early position in the visual processing pipeline, we find this to be true already here (91% of blocks show significant correlations). More interestingly, even after controlling for both firing rate and informativeness, units that are more modulated are the ones that are more predictive of behavior (Fig. 4E). This relationship is not present for the neuron residual variance (Fig. 4). Furthermore, we do not find a relationship with behavioral correlation in other shared noise sources (Suppl. S6). Overall, the data supports the theoretical prediction that the modulator preferentially targets informative units and that the modulated units have a stronger impact on behavior.

2.7 Knowledge of the modulator allows rapid decoding

We set out to investigate the computational relevance of the 1-dimensional, targeted modulator extracted from V1. We decode stimulus identity from neural responses to each stimulus presentation using either the optimal decoder or our heuristic modulator-guided (MG) decoder to estimate weights for each unit. We find that the MG decoder performs nearly as well as the optimal decoder ($\sim$ 80% correct). This suggests that the strength and targeting precision of

modulation estimated in the data is sufficient to detect modulator-induced labels in V1.

The optimal decoder provides an upper bound on decodability assuming perfect knowledge of the response properties, but in practice those would need to be learned. Even in the small subpopulation of units recorded in our experiment, this takes many trials: the learned decoder performs at chance in the low-data regime (Fig. 5A). In contrast, the modulator-guided (MG) decoder finds informative units after only a few training examples, outperforming the optimal decoder. We quantify this effect across all data and find that the MG decoder reaches above-chance performance significantly faster than the optimal decoder and that the performance attained with minimal training is significantly higher relative to that of the corresponding optimal decoder. Our theory predicts that the advantage of the MG decoder lies in its ability to accurately estimate the decoding weights after a relatively small number of stimulus presentations. In fact, we find a strong correlation between the MG decoding weights obtained with minimal training and those estimated from almost all the available data, but this relationship does not hold for the learned optimal decoding weights (Fig. 5C).

Although significant, the difference in the number of trials required for above-chance performance may seem small. This should be seen as a lower bound on this effect due to two restrictions. First, the recorded subpopulation is biased towards informative neurons since the stimuli are placed as so to drive these neurons. A more realistic decoding method must operate over the entire V1 population. Second, the modulator may vary on a time scale faster than what our model can capture and consequently provide even faster estimation of the decoding weights. Thus, the fact that we find a significant benefit of the MG decoder provides strong support for our hypothesis that the brain could use such decoding to enable flexible task switching.

2.8 Modulatory fluctuations from V1 also present in informative MT units

Simulus information in V1 has to propagate through downstream areas in order to ultimately inform the monkey's decision. MT neurons selectively combine V1 afferents over spatial position and orientation, which constructs MT's receptive field properties, such as direction selectivity [20]. While the information for the current experimental task is highly localized in V1, due to the topographical layout of spatial position and orientation, it is likely more diffuse in MT. We conjecture that downstream decoding of this information is facilitated by propagating the labeling mechanism to MT.

We find that the individually recorded MT units vary in their task-informativeness (Fig. 6B) and noisiness (fano factor, Fig. 6A). The two measures are correlated across the MT units: the more informative units also have higher fano factors (correlation coefficient of 0.48, $p < 0.008$). However, the stronger prediction from the theory is that their responses should reflect the *same* modulator acting on V1. To test this, we first capture MT stimulus response using the same stimulus-dependent Poisson model (SR) as used for the V1 units, but including stimulus drift direction as an additional parameter (drift direction did not have predictive power for the V1 units, see also [14]). The SR-model provides a good fit for all units (see Supplement Fig. S11A), which is expected given that experimental stimuli were optimized to drive MT units. To test whether the V1 modulator has predictive power for MT responses we incorporate it as an additional input in the MT SR-model and find improvements in the fit of 73% of MT units (Fig. 6C). This confirms that the modulator found in V1 is not private, but indeed shared with downstream MT neurons.

If the modulator acted as a labeling mechanism in MT, it should preferentially target informative units. We find that the activity of task-informative units is better predicted by the modulator (Fig. 6E), suggesting that the same modulator that labeled informative V1 units also labels informative MT units. Further we find that only those V1 modulators with significant targeting structure are also predictive of MT activity (Suppl. S8). Altogether, these results support the idea that the modulation of task-relevant neurons in V1 is inherited by MT, allowing the propagation of labeling information in the visual processing hierarchy.

3 Discussion

Humans and animals are impressive in their ability to respond rapidly and precisely to a variety of sensory stimuli. Although substantial progress has been made in understanding how the attributes of these stimuli are encoded in the responses of neural populations, the means by which the brain achieves the flexible and accurate decoding necessary to use this encoded information in a task remains mysterious. The substantial literature on neural population decoding is primarily focused on comparisons of behavior to optimal decoders, or development of brain machine interfaces (BMI). Both assume detailed knowledge of encoding (stimulus responses, noise properties), and thus do not provide plausible explanations of how later stages of neural processing can themselves achieve decoding. Here we have proposed a framework for neural decoding in which targeted correlated noise provides a label for relevant cells. Using neural recordings obtained from primate areas V1 and MT while animals perform a perceptual discrimination task, we found evidence for this labeling scheme: fast, shared, modulation of V1 activity, preferentially affecting task-informative neurons. We demonstrated that this modulation can be used to estimate decoding weights using only a few trials.

Finally, we found that the V1 modulator explains variability of MT responses, especially for MT units whose responses are informative for the task. This suggests that this functional labeling of informative neurons in V1 extends at least to MT, and perhaps to later stages.

Shadlen and colleagues [19] explored the computational challenges faced by downstream circuits involved in decoding. They described three potential limitations on biological decoding that would reduce the behavioral performance of monkeys in a motion discrimination task, compared to predictions for an ideal decoder applied to a hypothetical population of independent neurons: "*correlated noise*" (which worsens performance since it cannot be averaged out by the decoder), "*suboptimally stimulated neurons*" (in which the decoder includes irrelevant neurons in computing its decision), and "*pooling noise*" (additional noise that arises in downstream circuits). Subsequent studies have suggested that the contribution of the last of these is quite small [22]. The first (correlated noise) does appear to limit encoding quality [19], but our theory proposes that it also plays an important role in reducing the effect of the second limitation (suboptimally stimulated neurons).

Suboptimally stimulated neurons are likely to pose a more serious problem than expected from analysis of most experimental data sets, since recorded subpopulations are generally not representative of the full population. Low-firing neurons are often overlooked or discarded, and experimental stimuli are often optimized to drive responses, introducing a strong selection bias for neurons informative for those specific stimuli. Under these circumstances, pooling over the wrong subset of neurons is less harmful to performance than it would be for the brain, which must select from a much larger set, many more of which are uninformative. In the data presented here, the stimuli driving the population are spatially localized, and the recorded population is sufficiently diverse to exhibit a range of different response levels and degrees of informativeness. This allowed us to examine the relationship between informativeness and modulator coupling, and to assess the decoding benefits derived from the modulatory labeling scheme. Decoding becomes more difficult if fewer neurons are informative (Suppl. S1), and thus, our conclusions regarding the benefits of targeted modulation for downstream readout are likely understated.

Shared response oscillations have been proposed to underlie the representation of common features in subpopulations of neurons [23]. The "communication through coherence" (CTC) theory [24; 25] refines this idea in an encoding-decoding framework, in which a top-down oscillatory modulator projects to both encoding neurons with the same feature selectivity, and to the decoding network that needs to read out from them. Three important distinctions are; First, the oscillations in [24] target feature-selective rather than task-informative neurons. These could be the same for a detection task, but differ for discrimination, as used in our experimental data. Second, the CTC decoder in [24] was assumed to use a fixed, instead of a modulator-dependent, threshold, which is suboptimal 2.2. Third, although our theory would apply to oscillatory modulation, we do not rely on this additional restriction. The V1 modulation estimated from our data seem to favor dynamics that are very fast (close to the bounds of what we can estimate given our time binning resolution), but stochastic, with no evidence of periodic structure. Finally, at the conceptual level, the communication through coherence framework describes a fixed labeling strategy based on tuning properties alone, while our framework proposes modulatory labeling adapted to task structure.

Shared low-dimensional noise is well documented in the cortex [16; 10; 9; 11]. The effects on neural activity are primarily multiplicative [10], although additive components have also been reported [9]. The time scales of these shared noise sources differ, suggesting a variety of underlying mechanisms. Mostly, however, they appear to occur at slower time scales, reflecting for instance global fluctuations in attention across trials [11]. In contrast, our modulated SR-model is build to capture fast, within-trial, shared covariability, with estimated time constants on the order of tens of milliseconds. Slow systematic changes may occur in this dataset analyzed here, but they operate at a time scale that makes them orthogonal to our theory. In particular, slow multiplicative, low-dimensional noise may serve other functional roles, such as encoding uncertainty in visual areas [26; 27], but it cannot serve as a labeling mechanism of the type proposed here. Such variability would convey information about informativeness on a time scale slower than that needed for single trial feedback and decoder learning.

A large body of research describes the effects of top-down "attention" on perception, and measurements of cortical responses of primates provide evidence that these effects arise because of changes in the gain of neural responses [28; 29; 30]. Generally, these have been described as increases in mean response of spatially localized neural sub-populations, although recent studies also document decreases in response variability and correlation [16; 31]. These changes are consistent with an increase in the signal-to-noise ratio (SNR) of the local sensory representation, and are hypothesized to underlie the observed improvements in perceptual discrimination performance. Our theory is not aimed at explaining attentional effects, and we find that across units, modulator coupling is unrelated to the strength of attentional modulation, suggesting that it may arise from separate mechanisms (Suppl. S7). The functional role proposed here is different in that fluctuations of gain provide a *label* for informativeness, rather than increasing the gain to boost encoded SNR. As a consequence the targeting not only follows retinotopic location, but more specifically task-informativeness. While attentional modulation has been found to be tuning-specific [32; 33; 34], we do not find evidence that it is further specific to task-informative neurons (Suppl. S7). Interestingly, Zénon&Krauzlis find that inactivating the SC yields attentional deficits in behavior despite the fact that the attentional modulation of mean activity is preserved [35]. Our theory offers an interpretation of this observation: SC inactivation disrupts the covariablity labeling that allows

downstream circuits to decode the relevant information, but not the increases in firing rate that improves encoding SNR. If so, we would predict that targeted manipulation of the SC should directly affect the strength and targeting of shared modulatory fluctuations. In conclusion, our theory does not replace but may coexist with the known attentional effects.

Our theory is agnostic to the source of the modulator and the circuit mechanisms required for flexibly targeting neurons in a task-specific manner. Our findings that the estimated modulator in V1 is predictive of the activity of task-relevant simultaneously recorded MT units suggest that the modulator structure is shared across sensory regions, through either top-down or bottom up signals. Dynamic changes in noise correlations across task conditions (i.e. differences in modulator coupling) could themselves arise through either local circuit dynamics [36] or top-down mechanisms [12]. Past analyses of V1 activity suggest that a top-down source of the modulation changes pairwise correlations in V1 in a task-dependent manner [12]. Nonetheless, anatomical considerations and the relative sparsity of top-down connections seem to argue that the reorganization of noise correlations needs to also involve local recurrent dynamics. Moreover, theoretical models suggest that complex noise correlation structure can be produced locally [36], triggered for instance via recruitment of inhibitory neurons [37]. Given the currently available data, we cannot determine whether the modulator originates in V1 locally and propagates to MT in a feedfoward manner, or whether the signal is fed back from MT to V1. In order to target task-relevant neurons, the modulator could take advantage of the topographic organization of sensory codes present in some areas (for instance, orientation-specific columns in V1), without explicit knowledge of individual tuning functions. Once the label is established it may propagate to and be used in downstream areas without topographic organization. If this kind of spatially localized modulation was indeed an organizing principle of neural activity, it would predict that only features that are spatially localized at some stage in the brain can be flexibly decoded. In particular, a comparison of performance in tasks that rely on such features against those that rely on features with spatially diffuse encoding would be expected to expose fundamentally different processing and learning strategies. Interestingly, Nienborg and Cumming (2014) found that V1 neurons choice probability was significantly larger for orientation discrimination than for disparity discrimination, suggesting that V1 shows decision-related activity only if the task features are laid out in the columnar organization [21]. Future experiments, that simultaneously record from populations of neurons across multiple stages of sensory processing and across learning of different tasks, may be able to 1) estimate region specific modulators and the causal link between them and 2) study the targeting structure over behavioral learning.

The lack of a biologically plausible theory of neural decoding strongly limits our understanding of neural computation. Resolving the puzzle of how information is routed through brain regions has direct impact on the study of sensory and cognitive dysfunction, including clinical applications such as brain-machine interfaces (BMI) [38]. Moreover, flexible task-dependent information routing in hierarchical networks also remains a key open question in the field of machine learning, and biological brains may thus provide crucial computational insights for the improvement of engineered computational systems.

4 Figures

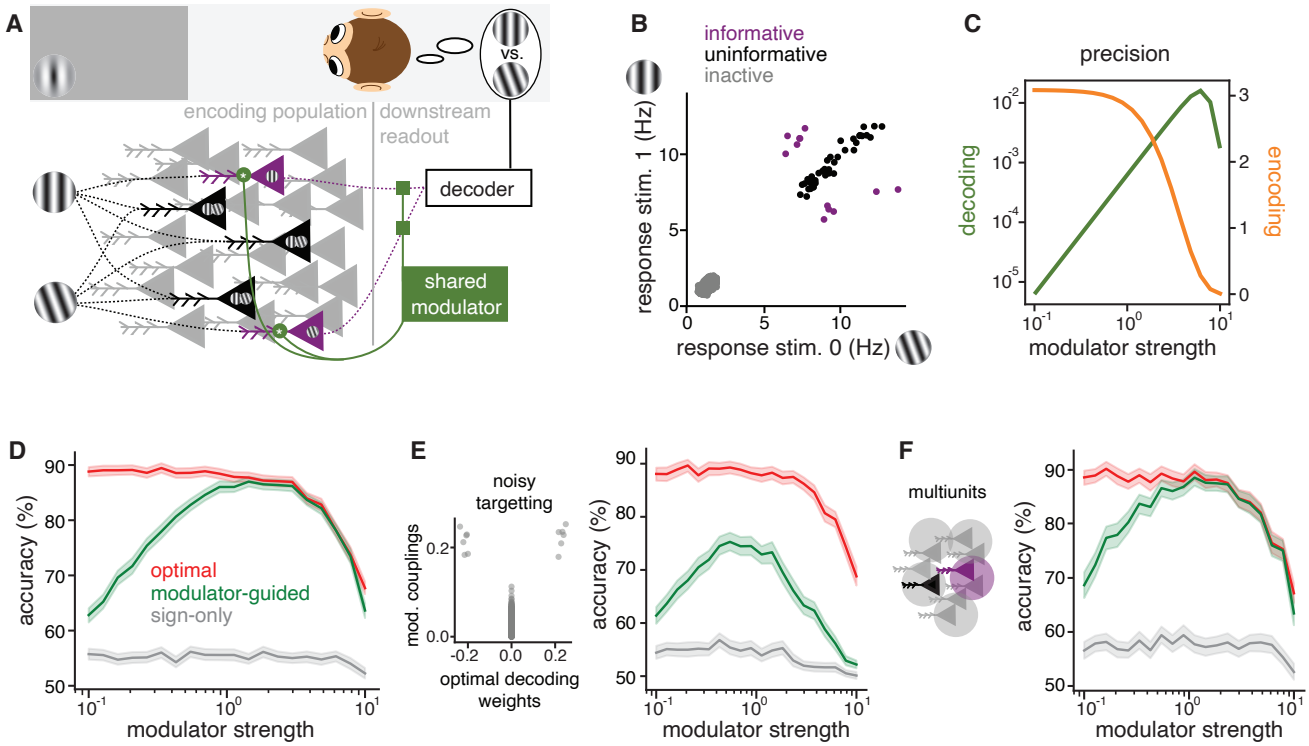

Figure 1: Theory of modulator guided decoding A) A population of neurons with diverse tuning properties respond to one of two task-specific grating stimuli; non-responsive neurons (grey), responsive but uninformative neurons (black), informative neurons (purple). A shared modulator (green) injects multiplicative noise to the informative neurons and sets the linear weights of a decoder to inform the decision of the monkey. B) The average response of neurons of the three subpopulations to two task stimuli. There are 12 informative, 38 uninformative and 4950 inactive neurons. C) Effects of increasing modulator strength on encoding and decoding, respectively. Encoding is measured by the SNR, while decoding precision is quantified as the variance of the decoding weights of the modulator-guided decoder. D) Performance of three different decoders in simulations of a discrimination task with increasing modulator strength. E) Same comparison for an imperfect relationship between informativeness and modulator coupling (independent gaussian noise added to the optimal coupling shown in left panel). F) Decoder performance comparison for simulated multiunits, obtained by summing the activity of random pairs of neurons.

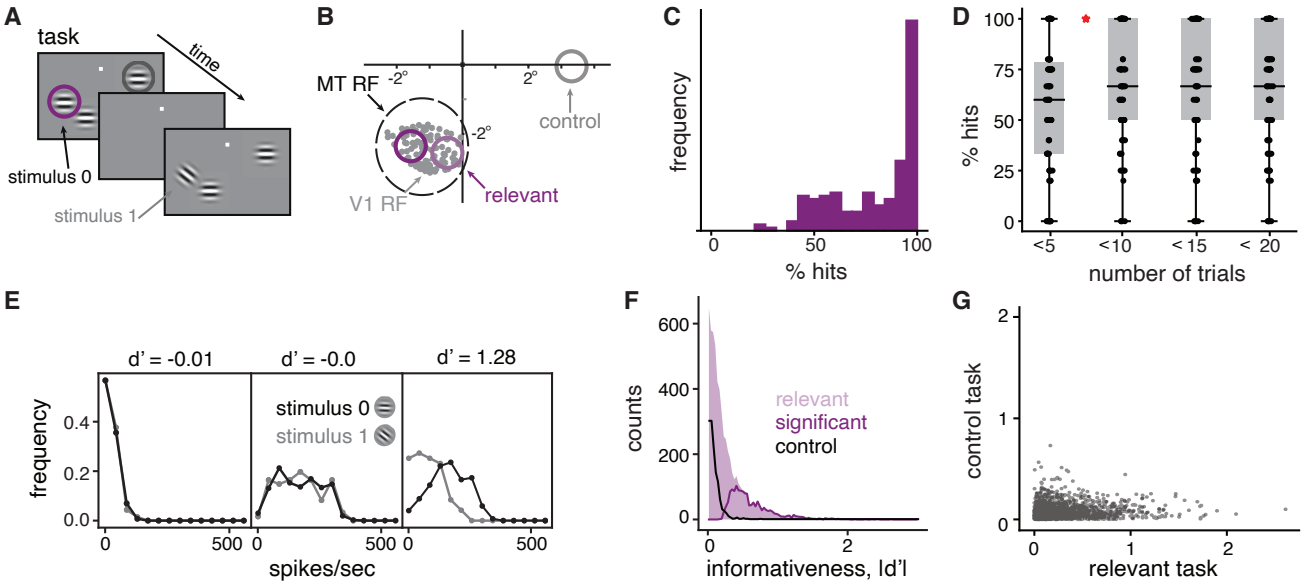

Figure 2: Monkeys perform an orientation discrimination task on one (purple) of three grating stimuli, while activity of neurons in V1 is recorded. A) Three drifting gratings flash on and off on a screen and can change their orientation. One stimulus (purple) is relevant for the task and its changes in orientation need to be reported by the monkey [14]. B) The recorded population of V1 neurons has receptive field centers (gray) close to one another and within the receptive field of a simultaneously recorded MT unit [14]. Two of the three stimuli lie within the MT units receptive field (light and dark purple) and one well outside (grey). C) Distribution of behavioral performance across block, quantified by the % of hits. D) Changes in behavioral performance as a function of time within a block. Each block is split in sets of 5 consecutive trials and the performance measure is computed within each set; points indicate different blocks and the red star indicates a significant difference between the means of the two adjacent distributions (relative t-test, $p = 0.015$). E) The distribution of response rates to each of the two task stimuli for three example neurons with different d' values. F) The distribution of informativeness values, $|d'|$, over all blocks of relevant tasks and all neurons (purple); purple line represents the distribution of the subset of neurons with significant informativeness. The black line represents the informativeness of neurons in the control task. G) Relationship between the informativeness values in relevant and control tasks for units recorded in both tasks.

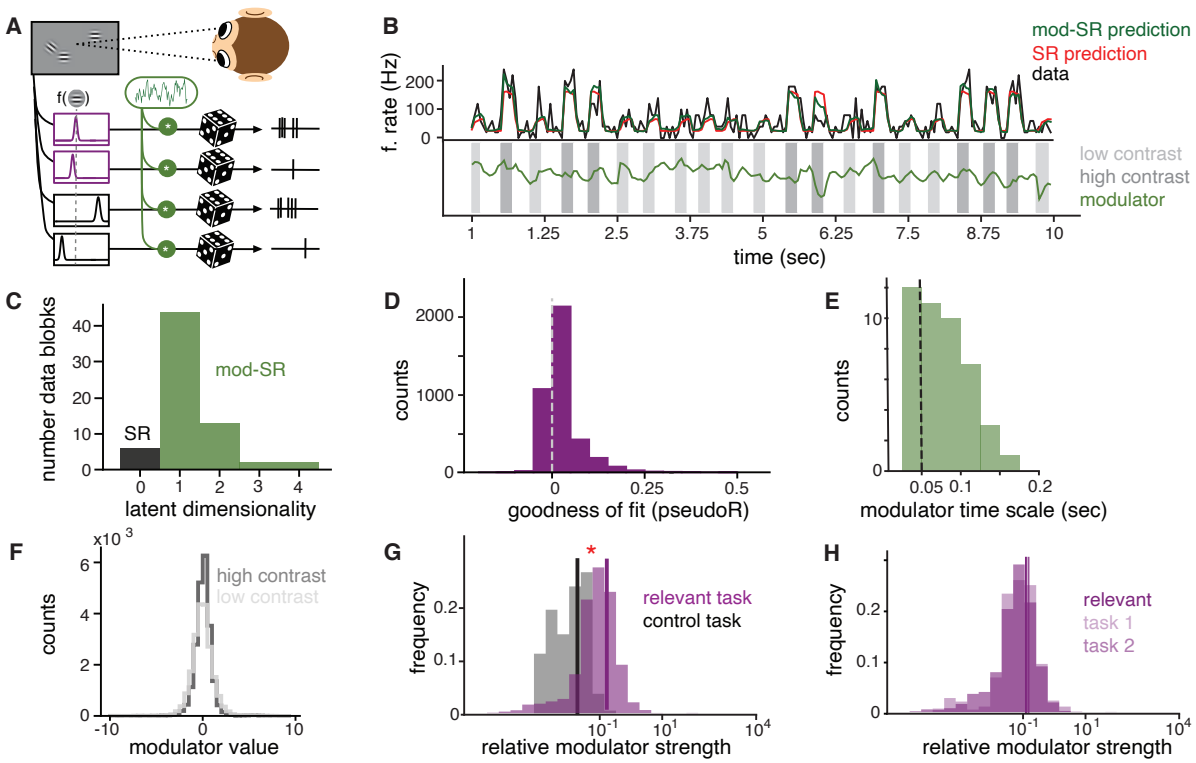

Figure 3: Estimating the modulator in the recorded V1 population. A) An illustration of the modulated stimulus response model: Each neuron's tuning function specifies its base response to a stimulus; this rate is modulated by a time-varying shared source of multiplicative noise (green), with spiking modeled by a Poisson process. B) An example unit's activity over concatenated test trials of a block and the corresponding prediction of the SR-model and the modulated SR-model. Bottom row shows the estimated trajectory of the modulator. C) Summary for the dimensionality of best fitted models across relevant tasks. D) The distribution of pseudoR values over all neurons in blocks that were best fitted by a 1-dimensional modulated SR model. E) The distribution of estimated time constants over all blocks that were best fitted by a 1-dimensional modulated SR model. F) The distribution of modulator values during high/low contrast stimulus presentations. G) The distribution of relative modulator strength across all relevant task blocks (purple) and all control task blocks (black); we quantify relative modulator strength as the variance in the modulator relative to that of the stimulus. The star indicates significant difference between the two distributions (U-test, $p < .001$). H) Same as in G, but comparing the two relevant tasks against each other.

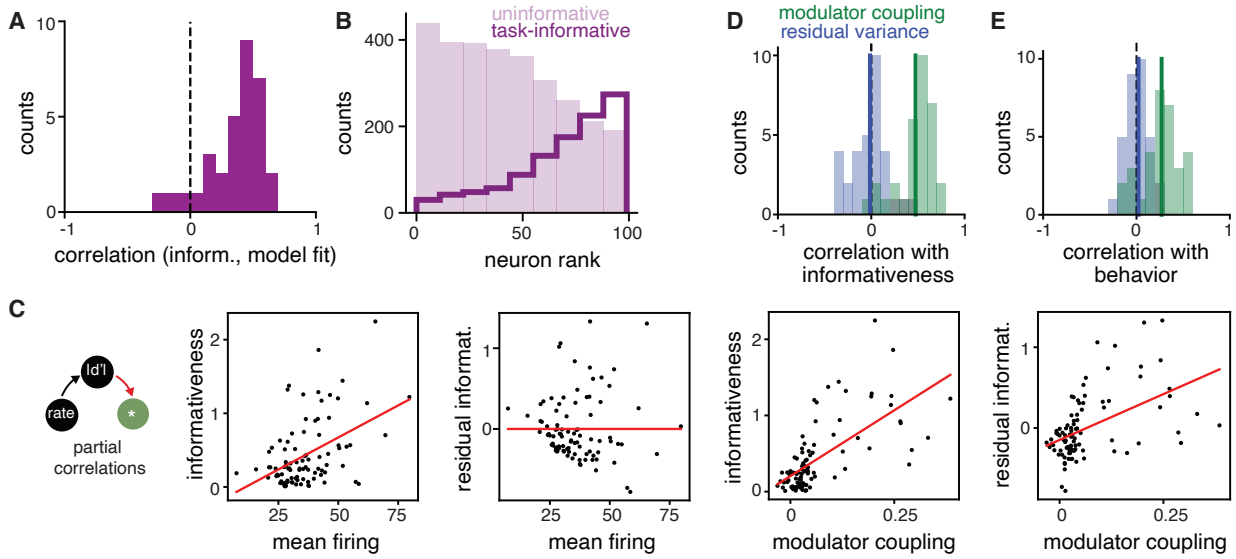

Figure 4: The targeting structure of the modulator reflects the current task. A) Distribution of the correlations between the individual unit's model fit (pseudoR) measure against their informativeness and plot the distribution of correlation coefficients. B) relative population rank of modulator coupling for significantly informative (dark purple line) and uninformative (light purple shading) neurons. C) Partial correlation analysis assessing the dependence between informativeness and modulation strength, after controlling for differences in firing rates. In order: dependence between |d'| and mean firing; residuals of linear fit as a function of firing rate are unstructured; the relationship between informativeness and coupling; same for residual informativeness (unexplained by differences in mean firing) D) Distribution of correlation coefficients obtained by partial correlation analysis across blocks (green) and a similarly obtained distribution that uses the modulated SR model residual variance as a proxy for neuron individual variance and instead of modulator coupling (blue). E) The distribution of correlation coefficients between modulator coupling (green) or residual variance (blue) and the residual behavioral correlation of a unit's activity, obtained by regressing out informativeness and mean firing rate.

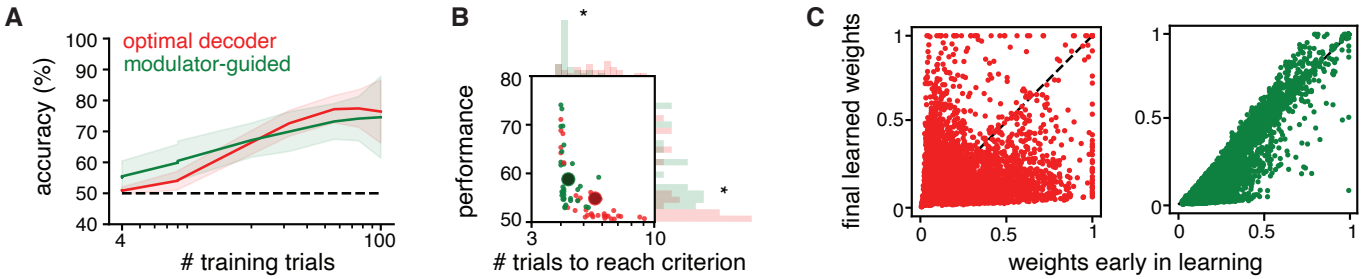

Figure 5: Decoding from the recorded V1 population. A) Performance of the modulator-guided decoder or the learned optimal decoder for an example block population with increasing number of training trials. B) Performance with minimal training against minimal number of training trials needed to reach above chance (50%) performance, for each block. Black stars indicate significant differences between the means of the distributions (t-test, $p < 0.0001$ for minimal training, $p = 0.0116$ for performance). C) Decoding weights estimated with maximum (90%) training versus with minimal (1%) training for the optimal (red) and modulator-guided (green) decoders.

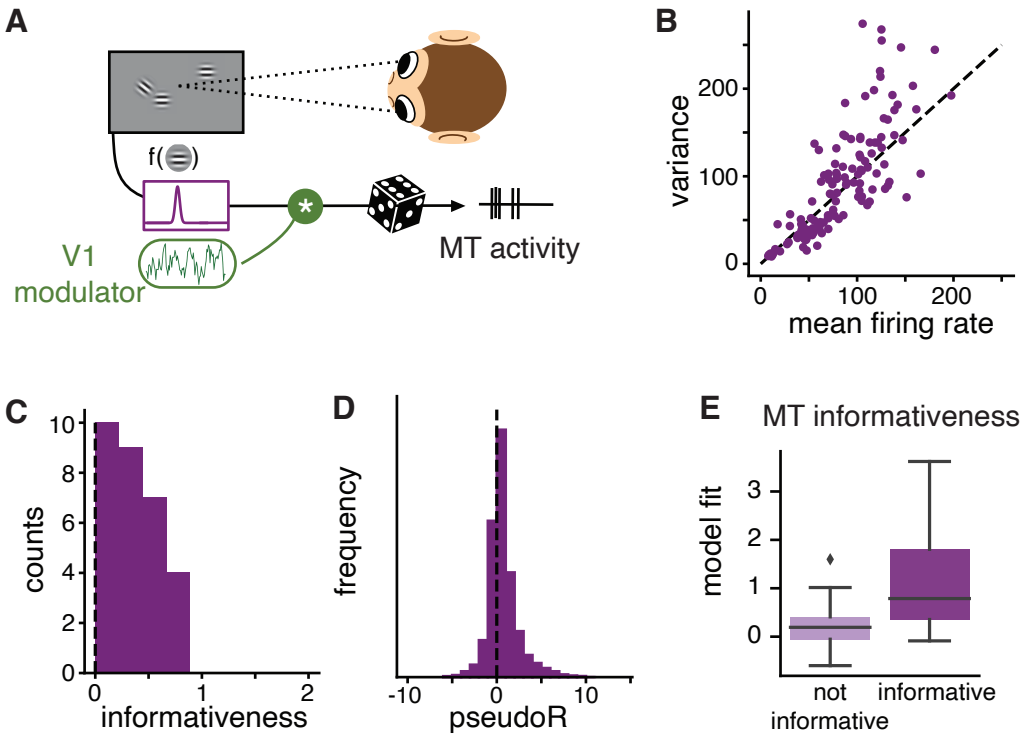

Figure 6: Effects of V1 modulator on simultaneously recorded MT units. A) Schematic of the model; the spiking of each MT unit is specified by a tuning function potentially multiplicatively gated by the modulator estimated from V1 activity, with Poisson noise. B) Stimulus response variance as a function of mean firing for all MT units, and stimulus presentations. C) Distribution of informativeness values for all MT units in relevant tasks. D) Distribution of pseudoR values obtained by comparing the log-likelihood of the SR model that includes the V1 modulator as an additional dimension against the SR model. E) Improvement in fit quality for the SR+V1 modulation model, grouping MT units into informative and uninformative.

5 Methods

Theoretical framework

We briefly describe the framework for modulator-guided decoding here, with an extended analysis available in [17]. In analogy to the experimental data, we simulated a binary discrimination task, which requires discriminating $s = 0$ from $s = 1$ on the basis of the activity of a population of N neurons. Neural responses are modeled as Poisson draws with a stimulus-dependent firing rate, which is itself modulated by a time-varying noisy signal, m_t, shared across neurons. Specifically, the firing rate of neuron n is given as:

$$k_{nt}(s, m_t) \sim \text{Poiss}\left(\lambda_n(s) g_n(m_t)\right), \tag{4}$$

where $\lambda_n(s)$ is the stimulus response function of the neuron, and t indexes time within a trial. For simplicity, the modulator m_t is 1-dimensional, modeled as i.i.d. Gaussian noise with zero mean and variance σ_m^2; the nonlinearity $g_n(\cdot)$, here an exponential, ensures that the final firing rate is positive.

The degree of modulation is neuron specific, parametrized by modulation weights w_n, which we take to be proportional to the n-th neuron's ability to discriminate the two stimuli, $w = |\log(\lambda_n(1)) - \log(\lambda_n(0))|$. This leads to a final expression for our encoding model of the form:

$$k_{nt}(s, m_t) \sim \text{Poiss}\left(\lambda_n(s) \exp\left(w_n m_t\right)\right), \tag{5}$$

We divide by the expected increase in mean rate due to the modulator given by $\exp\left(\frac{\sigma_m^2 w_n^2}{2}\right)$ to compensate for systematic differences in mean firing rate due to neuron-specific modulation strength. This parametrization ensures that any benefits of targeted modulation cannot be trivially explained by an increase in firing rates. Overall modulation strength in the population is determined by the modulator variance ($\text{var}(m_t) = \sigma_m^2$ - see also [8]).

Given this modulated Poisson encoding model, an ideal observer with complete knowledge of both stimulus response properties $\{\lambda_n(s)\}$ and modulation $\{w_n, m_t\}$ provides an upper-bound on task-decision accuracy. This optimal decision is made based on the sign of the log odds ratio, which compares the probability of the two stimuli under the full model. For our specific encoding model, this reduces to comparing a weighted linear combination of the observed neural spike counts against a modulator-dependent time-varying threshold (see [17] for details):

$$\sum_n a_n^{(\text{opt})} k_{nt} > c^{(\text{opt})}(m_t), \tag{6}$$

with weights:

$$a_n^{(\text{opt})} = \log(\lambda_n(1)) - \log(\lambda_n(0)), \tag{7}$$

and time-varying threshold:

$$c^{(\text{opt})}(m_t) = -\sum_n \exp(m_t w_n)\left[\lambda_n(1) - \lambda_n(0)\right], \tag{8}$$

Our modulator-guided heuristic decoder has access to the modulator m_t and the neural responses k_{nt}, but no detailed knowledge of the encoding model. Instead, it learns approximate decoding weights based on co-fluctuations of the two within a trial, by a simple learning rule:

$$|a_n^{(\text{MG})}| = \frac{1}{T}\sum_t m_t k_{nt} \tag{9}$$

The above expression only provides the magnitude of the decoding weight, with the signs separately estimated by comparing responses to the two stimuli. Estimation of the sign requires few trials for informative, strongly responding, neurons but will be noisy for uninformative neurons which are, however, excluded by the decoding weight magnitude (see also [17] and Suppl. S1).

As a lower bound of performance, we define a weightless decoder that subtracts the summed responses of two subpopulations, which corresponds to a linear decoder with weights ± 1.

$$a_n^{(\text{SO})} = \text{sgn}(\lambda_n(1) - \lambda_n(0)), \tag{10}$$

where $\text{sgn}(\cdot)$ is the signum (sign) function.

5.1 Experiments

In experiments by Ruff and Cohen [14], two adult male rhesus monkeys performed a motion direction change detection task on one out of 3 oriented drifting gratings at high or low contrast on a screen. Which grating is task-relevant

is indicated by a few instruction stimulus presentations and varies in blocks within the session ($\sim 3-6$ blocks per session). The experimenters implanted a 10 by 10 microelectrode array (Blackrock Microsystems) in area V1 and a recording chamber with access to area MT, allowing simultaneous recordings in the two areas (see details in [14]). Units can be either multiunits or single unit clusters. Two stimuli were positioned to drive the MT unit similarly and one stimulus was positioned outside of the MT RF (see Fig. 2B from [14]). Within a recording session changes in one out of the three stimuli had to be reported. We analyze each recordings session by splitting it into the task-specific blocks of trials. We analyzed 67 blocks of 20 recording sessions across two monkeys where the task-relevant stimulus was positioned in the RF of the population (relevant tasks) and 20 blocks of 20 sessions where the stimulus outside of the RF was task-relevant (control task). Control and relevant tasks where interleaved in blocks within a session. In a trial, gratings flash on (200ms) and off (200-400ms) at the same orientation (repeats, stimulus 0) until a change occurs at an unknown time (target, stimulus 1).

In each block we analyze $21-109$ trials where the monkey either detected the target (hit) or missed it (miss). We drop any trials where the monkey did not finish the task in a hit or miss. Trials where one of the distractor stimuli changed orientation were also excluded from the analysis here. A block then provides an average of 54 trials, each with several stimulus repeats ($s=0$, each 200ms) interrupted by breaks (200-400ms) and completed by a target presentation ($s=1$, orientation-change). We only include blocks that show a minimum of 20 valid trials (77 out of 90 blocks), as simulations suggest that about 20 trials are the minimum necessary to estimate informativeness reliably. Varying this criterion does not qualitatively change the results. The first stimulus in a trial was always removed to accord for adaptation effects [16]. Stimuli vary in contrast and orientation and are randomly interleaved. In the control task condition the two stimuli within the RF were presented either together or one by one. The individual stimulus presentation allows us to assess responsiveness of units. We only include units whose response to either one of the stimuli increased by a minimum of 10% compared to their baseline value. On average 88 units ($\sim 90\%$) in a block showed stimulus modulation for one of the two stimuli placed within the MT RF (min 52, max 95). Other units are excluded since they cannot be distinguished from non-neural noise. We exclude units with a rate < 0.1 or a Fano factor > 5 standard deviations above the population average.

5.2 Informativeness of a unit

The informativeness of a unit is quantified by $\left|d^{'}\right| = \left|\frac{\mu_0-\mu_1}{\sqrt{(\sigma_0^2+\sigma_1^2)/2}}\right|$ where μ_0 and σ_0^2, μ_1 and σ_1^2 are the means and variances of a unit's responses to stimulus 0 and stimulus 1, respectively. We compute informativeness across all stimulus presentations in behaviorally correct trials of the same block. We do not average across blocks to avoid behavioral changes or slow drift effects. To determine whether a unit is significantly informative, we simulate a null-distribution of d' values by comparing mean and variance of random subsets of stimulus 0 responses. This allows us to compute a percentile for the true d' value that gives us an estimate of whether such a value could be the result purely of sampling noise or whether there is a systematic difference in the neural response between the two stimuli.

5.3 Stimulus-Response (SR) Model:

The stimulus response model is a Linear-Nonlinear Poisson (LNP) single neuron model which is fit to the data by maximizing the log-likelihood of neural activity under the model. We analyze activity binned within 50ms time windows. Due to the asymmetry in the data available before and after stimulus orientation change we only model responses to the repeated stimulus orientation (stimulus 0) at varying contrasts. The model is fitted exclusively to the stimuli within the RF of the population, the target (stimulus 1) response is only used to compute informativeness and for the decoding analysis. The stimulus is characterized by contrast (V1) or contrast+direction (MT). Orientation is not one of the stimulus dimensions as it will not change before the target comes on. We find that the V1 units do not respond differentially to stimuli of different direction (compare also to [14]). The stimulus is parametrized by a one-hot encoding vector at every point in time $\mathbf{s}_t$ with 4 time-windows for each 200ms stimulus presentation, resulting in 8 stimulus dimensions for the contrast-specific V1 model. We add one after-stimulus dimension and an offset for base firing. The MT model includes the same stimulus dimensions, plus one dimension indicating the stimulus drift direction. The stimulus affects a unit's activity through a linear coefficient $\boldsymbol{\beta_n}$, followed by an exponential nonlinearity that gives a rate of a Poisson process generating spike counts $\mathbf{k}_n$:

$$\boldsymbol{k}_{n,t} \sim \text{Poisson}\left(\exp\left(\boldsymbol{\beta_n}\mathbf{s}_t\right)\right) \tag{11}$$

We can optimize for $\boldsymbol{\beta}_n$ by maximizing the log-likelihood of the data:

$$L(\boldsymbol{\beta_n}) = -(\boldsymbol{\beta_n}\mathbf{s}_t)^T\boldsymbol{k}_{n,t} + \exp(\mathbf{1}^T\boldsymbol{\beta_n}\mathbf{s}_t) + \alpha\boldsymbol{\beta_n}^T\boldsymbol{\beta_n} \tag{12}$$

Out of 67 blocks 6 were excluded from subsequent analyses due to bad population-average fits, meaning that in those populations only very few neurons responded to contrast changes in stimulus.

MT SR-model with V1 modulator: We fit the same SR-model to MT as previously fitted to V1 but include direction of stimulus drift. To test whether the V1 modulator has predictive power for MT activity, we add the (normalized) modulator as an additional predictive variable and refit the model.

Model Validation/Comparison: The models are cross-validated using 90% of trials to train and 10% to test. Three criteria where used to validate the SR-model; log-likelihood of test data under the model, variance explained by the model and the pseudo R^2 [39] which gives"the fraction of the maximum potential log-likelihood gain (relative to the null model) achieved by the tested model" $\frac{\log L(\hat{y}) - \log L(\bar{y})}{\log L(y) - \log L(\bar{y})}$. The null model here had the same form but no stimulus-related dimensions so that the only explanatory variable was the average activity.

5.4 Modulated Stimulus-Response Model

Building up on the SR-model, we look for population-wide low-dimensional modulator terms $\vec{m}$ that vary stimulus response both within and across trials. We use the framework of Poisson Linear Dynamical Systems (PLDS, [40; 11]) as our modulated SR model, which allows us to make use of the temporal dependencies within a trial while treating different trials as independent. While the SR-model is fit independently for each neuron, the modulator terms of the PLDS are shared across the population and influence each unit's activity through a linear mapping function C. This joint model has the form:

$$\mathbf{k}_t \sim \text{Poisson}(\exp(\mathbf{C}\mathbf{m}_t + \mathbf{B}\mathbf{s}_t)) \tag{13}$$

$$\begin{aligned} \mathbf{m}_{t+1} &= \mathbf{A}\mathbf{m}_t + \epsilon \\ \epsilon &\sim \mathcal{N}(0, \mathbf{Q}) \\ \vec{m}_0 &\sim \mathcal{N}(0, \mathbf{Q}_0) \end{aligned} \tag{14}$$

where the modulator at time t, $\mathbf{m}_t$, is D-dimensional and the mapping $\mathbf{C}$ is N by D, with latent dimensionality $D \ll N$. Parameter $\mathbf{A}$ implicitly defines the modulator's time constant and $\mathbf{Q}, \mathbf{Q}_0$ are the noise covariances for the modulator. The full model is fitted using the EM algorithm with a Laplace approximation (see [40]). All model fitting is cross-validated (see below). We test for up to 4 modulatory dimensions; we cannot exclude the possibility of higher dimensionality due to restrictions imposed by noise and sample size. We found that the fitted parameters were very similar across the different cross-folds.

Model Validation/Comparison: The same criteria are used to evaluate the modulated-SR as described above for the SR-model with one adjustment. Computing the likelihood of the test data requires an estimation of the modulator trajectories from the population activity at each trial. For this we use a leave-one-out approach [41] where first the model parameters are fitted for all neurons on training trials and then the modulator trajectory is predicted for the testing trials using all but one unit. This allows a double cross-validation, in time and neural space. This estimated modulator trajectory can then be used to compute the log-likelihood of the left-out unit's activity under the model. To accord for the uncertainty in the estimation of the modulator trajectory we sample from the estimated distribution of modulator trajectories, computing the respective log-likelihoods for the left-out units and then take the mean of these log-likelihoods as an approximate measure of true fit quality.

5.5 Modulator targeting

In Fig. 4B we compute the rank of each modulator coupling in its own block-specific population and compare the distribution of significantly informative to uninformative units. In Fig. 4C we used partial correlation to test for a relationship between unit's modulator coupling and task-informativeness in each block not explained by differences in overall rate. First we compute the linear effect of firing rate on informativeness to define the unexplained, residual informativeness. We then compute the Spearman correlation coefficient between residual informativeness and modulator coupling. For the correlation analysis we exclude blocks with less than 15 informative neurons since a linear correlation is not sensible in this case. Varying this criterion does not change the results qualitatively.

5.6 Correlation with behavior: relationship between behavioral choice and neural activity

We compute the difference in target-response for trials where the target was correctly detected by the monkeys to those where the monkey missed the target. This gives us a rough estimate of how involved a unit was in the choice of

the animal. To asses the relationship with modulator strength we use a partial correlation with two covariates, firing rate and informativeness (multivariate linear regression).

5.7 Decoding

We train each decoder on a training data set that includes a balanced number of stimulus 0/1 presentations at high and low contrast. Decoder performance is tested on held out data. To asses how training-efficient decoders are, we vary the number of data points used for training between 4 samples (stimulus 0 and 1 at low and high contrast) and all but 4. For the optimal decoder we use the same Maximum Likelihood approach as described above in the theory. This requires estimating the mean response to each stimulus to then use the log-ratio as a decoding weight for a unit (see Eq. (7)). For a closer comparison to the theory, the number of spikes is summed over the 200ms stimulus presentation. For simplicity we here compare against a constant threshold which is optimized on the training data. The modulator-guided (MG) decoder estimates the modulation strength of each unit by taking the inner product between the unit's activity (in 50ms bins) and the modulator values. It uses these estimates as absolute decoding weights, with signs determined from trial-level feedback, comparing the response to one stimulus versus the other. Importantly, the decoding weights are estimated using the finer resolution of 50ms bins since that is the time scale at which the modulator varies. Again the linear weighted sum taken using the MG decoding weights is compared against a constant threshold.

References

[1] Britten, K. H., Newsome, W. T., Shadlen, M. N., Celebrini, S. & Movshon, J. A. A relationship between behavoiral choice and the visual responses of neurons in macaque MT. *Visual Neuroscience* **13**, 87–100 (1996).

[2] Geisler, W. S. & Albrecht, D. G. Visual cortex neurons in monkeys and cats: Detection, discrimination, and identification. *Visual Neuroscience* **14**, 897–919 (1997).

[3] Dayan, P. & Abbott, L. F. *Theoretical neuroscience* (MIT Press, Cambridge, MA, 2005). 0-262-04199-5.

[4] Ma, W. J., Beck, J. M., Latham, P. E. & Pouget, A. Bayesian inference with probabilistic population codes. *Nature Neuroscience* **9**, 1432–8 (2006).

[5] Jazayeri, M. & Movshon, J. A. A new perceptual illusion reveals mechanisms of sensory decoding. *Nature* **446**, 912–915 (2007).

[6] Carrasco, M. Visual attention: the past 25 years. *Vision Res* 1484–1525 (2011).

[7] Reynolds, J. H. & Chelazzi, L. Attentional modulation of visual processing. *Annual Review of Neuroscience* 611–647 (2004).

[8] Churchland, A. K. *et al.* Variance as a signature of neural computations during decision making. *Neuron* **69**, 818–831 (2011).

[9] Lin, I. C., Okun, M., Carandini, M. & Harris, K. D. The nature of shared cortical variability. *Neuron* **87**, 644–656 (2015).

[10] Goris, R. L., Movshon, J. A. & Simoncelli, E. P. Partitioning neuronal variability. *Nature Neuroscience* **17**, 858–865 (2014).

[11] Rabinowitz, N. C., Goris, R. L., Cohen, M. R. & Simoncelli, E. P. Attention stabilizes the shared gain of V4 populations. *eLife* 1–24 (2015).

[12] Bondy, A. G., Haefner, R. M. & Cumming, B. G. Feedback determines the structure of correlated variability in primary visual cortex. *Nature Neuroscience* **21**, 598–606 (2018).

[13] Ruff, D. A. & Cohen, M. R. Stimulus dependence of correlated variability across cortical areas. *Journal of Neuroscience* **36**, 7546–7556 (2016).

[14] Ruff, D. A. & Cohen, M. R. Attention increases spike count correlations between visual cortical areas. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **36**, 7523–34 (2016).

[15] Felleman, D. J. & Van Essen, D. C. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex* **1**, 1–47 (1991).

[16] Cohen, M. R. & Maunsell, J. H. Attention improves performance primarily by reducing interneuronal correlations. *Nature Neuroscience* **12**, 1594 (2009).

[17] Haimerl, C., Savin, C. & Simoncelli, E. Flexible information routing in neural populations through stochastic comodulation. *Advances in Neural Information Processing Systems* **32**, 14402–14411 (2019).

[18] Hair, J. F., Black, W. C., Babin, B. J. & Anderson, R. E. *Multivariate Data Analysis* (Pearson Education Limited, Essex, 2014), 7th edn.

[19] Shadlen, M. N., Britten, K. H., Newsome, W. T. & Movshon, J. A. A computational analysis of the relationship between neuronal and behavioral responses to visual motion. *Journal of Neuroscience* **76**, 1486–1510 (1996).

[20] Born, R. T. & Bradley, D. C. Structure and function of visual area MT. *Annu Rev Neurosci.* 157–89 (2005).

[21] Nienborg, H. & Cumming, B. G. Decision-related activity in sensory neurons may depend on the columnar architecture of cerebral cortex. *Journal of Neuroscience* **34**, 3579–3585 (2014).

[22] Osborne, L. C., Lisberger, S. G. & Bialek, W. A sensory source for motor variation. *Nature* **7057** (2005).

[23] Singer, W. Neuronal synchrony: A versatile code review for the definition of relations? *Neuron* **24**, 49–65 (1999).

[24] Akam, T. E. & Kullmann, D. M. Efficient "communication through coherence" requires oscillations structured to minimize interference between signals. *PLoS Computational Biology* **8** (2012).

[25] Akam, T. & Kullmann, D. M. Oscillatory multiplexing of population codes for selective communication in the mammalian brain. *Nature Reviews Neuroscience* **15**, 111–122 (2014). 1309.2848v1.

[26] Hénaff, O. J., Boundy-Singer, Z. M., Meding, K., Ziemba, C. M. & Goris, R. L. Representation of visual uncertainty through neural gain variability. *Nature Communications* **11** (2020).

[27] Festa, D., Aschner, A., Davila, A., Kohn, A. & Coen-Cagli, R. Neuronal variability reflects probabilistic inference tuned to natural image statistics. *bioRxiv* (2020).

[28] Moran, J. & Robert, D. Selective attention gates visual processing in the extrastriate cortex. *Science* **229**, 782–784 (1985).

[29] Mcadams, C. J. & Maunsell, J. H. R. Effects of Attention on the Reliability of Individual Neurons in Monkey Visual Cortex proportionally and does not improve the selectivity of single neurons, as measured by the width of their tuning curve (Vogels and Orban. *Neuron* **23**, 765–773 (1999).

[30] Treue, S. & Maunsell, J. H. Attentional modulation of visual motion processing in cortical areas MT and MST. *Nature* **382**, 539–541 (1996).

[31] Mitchell, J. F., Sundberg, K. A. & Reynolds, J. H. Spatial Attention Decorrelates Intrinsic Activity Fluctuations in Macaque Area V4. *Neuron* **63**, 879–888 (2009).

[32] Treue, S. & Martínez Trujillo, J. C. Feature-based attention influences motion processing gain in macaque visual cortex. *Nature* **399**, 575–579 (1999).

[33] Maunsell, J. H. & Cook, E. P. The role of attention in visual processing. *Philos Trans R Soc Lond B Biol Sci.* **357**, 1063–72 (2002).

[34] Ruff, D. A. & Cohen, M. R. Attention can either increase or decrease spike count correlations in visual cortex. *Nature neuroscience* **17**, 1591–7 (2014).

[35] Zénon, A. & Krauzlis, R. J. Attention deficits without cortical neuronal deficits. *Nature* **489**, 434–437 (2012).

[36] Huang, C. *et al.* Circuit models of low-dimensional shared variability in cortical networks highlights. *Neuron* **101**, 1–12 (2019).

[37] Middleton, J. W., Omar, C., Doiron, B. & Simons, D. J. Neural Correlation Is Stimulus Modulated by Feedforward Inhibitory Circuitry. *Journal of Neuroscience* **32**, 506–518 (2012).

[38] Andersen, R. A., Musallam, S. & Pesaran, B. Selecting the signals for a brain–machine interface. *Current opinion in neurobiology* **14**, 720–726 (2004).

[39] Benjamin, A. S. *et al.* Modern machine learning outperforms GLMs at predicting spikes. *bioRxiv* (2017).

[40] Macke, J. H., Buesing, L. & Sahani, M. Estimating State and Parameters in State Space Models of Spike Trains. In *Advanced State Space Methods for Neural and Clinical Data*, 137–159 (Cambridge University Press, 2015).

[41] Yu, B. M. *et al.* Gaussian-Process Factor Analysis for Low-Dimensional Single-Trial Analysis of Neural Population Activity. *Journal of Neurophysiology* **102**, 614–635 (2009).