Flexible sensory information processing through targeted stochastic co-modulation

by

Caroline Haimerl

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Neural Science

New York University

May 2022

Cristina Savin

Eero P. Simoncelli

Dedication

Für meine Großmutter

Preface

Chapters 2, 3, 4 include results from analyzing experimental data collected by Douglas A. Ruff and Marlene R. Cohen and previously published in Ruff and Cohen (2016a). Chapter 2, 3, 4 include results available in a preprint on biorxiv Haimerl et al. (2021). Chapter 3 includes results published in Haimerl et al. (2019). Chapter 4 is mostly unpublished.

Abstract

Humans and animals can quickly adapt to new task demands while retaining capabilities developed previously. Such flexible sensory-guided behavior requires reliable encoding of stimulus information in neural populations, and task-specific readout through selective combination of these responses. The former has been the topic of intensive study, but the latter remains largely a mystery. Here we propose that targeted stochastic gain modulation could support flexible readout of task-information from an encoding population. In experiments, we find that responses of neurons in area V1 of monkeys performing a visual orientation discrimination task exhibit low-dimensional comodulation. This modulation fluctuates rapidly, and is stronger in those neurons that are most informative for the behavioral task. We propose a theoretical framework in which this modulation serves as a label to facilitate downstream readout. We demonstrate that the shared modulatory fluctuations found in V1 can be used to decode from the recorded neural activity within a small number of training trials, consistent with observed behavior. Simulations of visual information processing in a hierarchical neural network demonstrate that learned, modulator-induced labels can accompany task-information across several stages to guide readout at a decision stage and thereby fine-tune the network without reorganization of the feedforward weights. This allows the circuit to reach high levels of performance in novel tasks with minimal training, outperforming previously proposed attentional mechanisms based on gain increases, while also being able to instantly revert back to the initial operating regime once task demands change. The theory predicts that the modulator label should be maintained across processing stages and indeed we find that the trial-by-trial modulatory signal estimated from V1 populations is also present in the activity of simultaneously recorded MT units, preferentially so if they are task-informative. Overall, these results provide a

new framework for how intelligent systems can flexibly and robustly adapt to changes in task structure by adjusting information routing via a shared modulator label.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Humans and animals are able to flexibly adapt their behaviors according to dynamic objectives and changes in their environment. Imagine a walk through a tropical forest: a myriad of trees with busy birds searching for fruits; bushes hiding small animals in their quest for nuts and seeds; light reflecting off drops of water on the leaves and a small stream underfoot; in the distance, a few familiar faces. A scene like this contains a vast amount of details, and while it may be beautiful to observe in its entirety, we typically do not take in all of our environment simultaneously, but perceive it with an ever-changing set of intentions: find food high up in a tree, drink water from the stream, catch up with the group, is that a tiger in the bushes? While much is understood about how the brain extracts and represents sensory information, the means by which intentions guide selection and modulation of these representations to support flexible behavior remain unclear. This introduction aims to: First, define behavioral flexibility in sensory-guided decision-making tasks; Second, discuss sensory representation in the brain with a focus on vision; Third, lay out the challenges for reading out and using these representations flexibly when switch-

Figure 1.1 A busy jungle with rich sensory information. What matters depends on the internal intention.

ing between different tasks.

1.1 Behavioral flexibility

Behavioral flexibility describes the ability to switch from one task to another - to adjust to new rules or to find a new solution to a problem (Ionescu, 2012). It is observed across different species (Bronkhorst, 2015; Cheng and Frye, 2020; Cherry, 1953; Mante et al., 2013; Ruff and Cohen, 2016a; Tingley et al., 2014; Tolman et al., 1946) and is an important characteristic of biological intelligence. Static processing of sensory information, studied as "instrumental learning" (Drummond and Niv, 2020), model-free/habitual learning (Drummond and Niv, 2020) or "rigid stimulus-response memory systems" (Lee and Lee, 2013), is the direct mapping of a stimulus to a response, resulting in fast "automatic" behavior (such as the "knee-jerk" reflex), similar to what many machines perform. In contrast, behavioral flexibility requires merging sensory information about what is immedi-

ately present in the outside world with intention. Intention reflects objectives derived from internal state (e.g. hunger, thirst) or external changes in the environment (e.g. contextual background cues, see Lee and Lee, 2013). It does not need to be bound to the present sensory input, but may originate in previous experience or future expectations (Mante et al., 2013).

Flexible behavior in sensory-guided decision making has been studied in different fields, such as cognitive psychology, cognitive and systems neuroscience, and particularly in the fields of attention research and reinforcement learning. Consequently, it has been formalized in many experimental paradigms, of which I will highlight three representative groups here. First, visual attention: monkeys and humans are able to flexibly adjust visually-guided behavior based on changes in relevant spatial location (Ruff and Cohen, 2016a; Treue and Maunsell, 1996) or feature task relevance (Armbruster et al., 2012; Biró et al., 2019; Hayes and Allinson, 1998; Mante et al., 2013; Mohan et al., 2021; Ravizza and Carter, 2008). The change in task relevance can be signaled by an extrinsic sensory cue or be intrinsically motivated. In human behavioral studies, typical feature-based tasks that are being combined or alternated are object naming, categorization of digits as even/odd or larger/smaller, categorizing words or letters (Kiesel et al., 2010). The visual stimuli are typically bivalent, so that different tasks can be performed on them. Second, navigation: mice and rat navigating to a reward are able to adjust their paths when confronted with changes in their environment, such as paths in a maze being blocked (Tolman et al., 1946), starting position changes in the Morris water maze (Morris, 1984) or changes in the probability of a location being rewarded (Tingley et al., 2014). These behaviors are thought to rely on cognitive maps of the environment - complex internal representations that allow planning and flexibility. Third, behavior based on other sensory features: for example humans are able to concentrate on a particular source of sounds and suppress

other sources (the "Cocktail Party" problem, Bronkhorst, 2015; Cherry, 1953). Rats similarly show great auditory flexibility when switching between sound localization and pitch discrimination tasks while maintaining high performance (Rodgers and DeWeese, 2014). In the domain of touch, somatosensory sensitivity in humans can change depending on the probability that a task-relevant stimulus will appear on a particular part of the body (Johansen-Berg and Lloyd, 2000; Lindsay, 2020).

In most of these flexible behaviors, the sensory information or the statistics of the sensory environment stay the same, but their relative meaning for the subject changes depending on task details. These sensory-informed but rule-guided behaviors imply that the brain is able to capture and selectively use relevant information to guide a decision.

1.1.1 Engineering flexible behavior

Despite the immense advances in the field of Machine Learning and Artifical Intelligence through Deep Neural Networks and Recurrent Neural Networks integrating "attentional" mechanisms and meta-learning across different tasks (Caruana, 1997; Finn et al., 2017a; Vaswani et al., 2017), task switching or multitasking is still non-trivial for artificial agents (Lindsay, 2020). Arguably one of the biggest challenges is building representations that can represent the information needed for many tasks and then incorporating control processes that use these representations flexibly, within the constraints set by limited resources. Specifically, reorganizion of existing synapses to satisfy the demands of each task runs the risk of catastrophic loss of previous capabilities (Fusi et al., 2005; Kirkpatrick et al., 2017; Masse et al., 2018). Different approaches exist to improve task flexibility. For example, some meta-learning algorithms introduce hyperparameters that act across multiple tasks and aim to optimize performance, by improving starting parameters (Finn et al.,

2017a), orthogonalizing task representations (Yang et al., 2019), or modulating learning dynamics (Andrychowicz et al., 2016; Wang, 2017), among other methods. A different line of research focuses on episodic memory components that aim to increase the capacity for building associations across longer periods of time and allowing for flexibility through separate memory representations (Ritter et al., 2020, 2018). All of these approaches rely on significantly increasing the parameter space to avoid catastrophic forgetting. Resource constraints prevent the construction of *de novo* representations for each new task in the brain and set natural limits to the scalability of many of these algorithms. Therefore, while impressive performance can be achieved in a clearly specified set of tasks and given large sets of data and vast parameters spaces, the flexibility exhibited by biological agents given a wide range of tasks is, for now, out of reach.

1.1.2 Limitations to flexible behavior

In order to gain understanding of flexible biological behavior, it can be equally useful to consider its limitations. Some task switches are hard. There are sensory modalities that are prioritized and guide behavior preferentially, creating asymmetry when switching from one modality to another. For instance, vision tends to overrule other senses even if it is not informative for the task (Bertelson and Aschersleben, 1998; Lindsay, 2020; Spence, 2009). Similarly, within one modality there are sensory features that are especially salient and easily extractable, such as oriented edges (especially along the cardinal axis), spatial frequency and motion (Itti and Koch, 2001; Lindsay, 2020). Others require more effort and longer processing times, such as detecting combinations of features (Posner and Presti, 1987), combining information from different spatial locations (Itti and Koch, 2001; Lamme and Roelfsema, 2000; Lindsay, 2020), or reporting a less automated feature in the

presence of a more dominant one, such as color in the Stroop test (Scarpina and Tagini, 2017; Stroop, 1935). Another well documented phenomenon is visual search asymmetry where finding an object of type A among objects of type B can be a lot easier than finding an object of type B among objects of type A. For instance a curved line among straight lines is easy to find but not the other way around (Gupta et al., 2021). Furthermore, the frequency of task switches matters; block designs where a task stays the same for several trials tend to be easier to perform than alternations where tasks switch with every trial (Kiesel et al., 2010). A switching-cost effect remains even if task switches are strictly periodic and entirely predictable (Rogers and Monsell, 1995), are cue-instructed (Altmann, 2004; Meiran et al., 2000), or even voluntary (Liefooghe et al., 2009). However, which mechanisms underlie the different task switching behaviors and how they change as sequential task switching becomes parallel multitasking is unclear.

These "failures" of flexibility may in part be attributed to mechanistic limitations of the brain, such as the number of neurons and connections that can be sustained simultaneously or imperfect communication ("noise"). However, they may also teach us about when "hardwired" stability should be prioritized over flexibility (Ionescu, 2012), either to achieve the rapid behavior seen in stimulus-response behavior or because it provides a stable scaffold for learning, reducing the parameter space that needs to be considered when learning a new task. For instance, salient visual features likely emerge because of the way the brain represents information and makes certain features more easily available (e.g. by enabling linear readout and manifold separation (DiCarlo and Cox, 2007)). The next sections will review relevant properties of sensory feature representation (Sec. 1.2) and consider their impact on task-learning speed (Sec. 1.3.2).

1.2 Encoding of sensory information

In order to allow directed action in an environment, the brain has developed a neural architecture to both precisely and robustly represent sensory information in the world that has proven vital in the span of life and over evolution. Primary sensory areas can capture a wide range of stimulus information simultaneously and stably, while flexible changes due to task structure are thought to happen closer to the decision-making stages, such as prefrontal cortex in monkeys and humans (Mante et al., 2013; Tsotsos et al., 2019).

Given the importance of vision for humans and the dominance of vision research compared to the study of other sensory modalities (Hutmacher, 2019; Lindsay, 2020), we use visually-guided decision making as an example to discuss how sensory task information is *encoded* by the brain. Encoding properties of the visual processing hierarchy have been studied extensively since Hartline (1938) demonstrated the idea of receptive fields in frog retinal ganglion cells and Hubel and Wiesel (1959) comprehensively illustrated V1 neurons' tuning selectivity in cats. Here we will focus on visual encoding properties of primates. Neurons at the retina respond to highly localized light patterns (Carandini et al., 2005; Dayan and Abbott, 2005) made up of ON and OFF regions. Neurons downstream in the lateral geniculate nucleus (LGN) and primary visual cortex (V1) integrate over these local light responses, but maintain spatial receptive fields (RF) that make them sensitive to local patterns of light contrast (Fig. 1.2A) (Carandini et al., 2005). On a population level, neurons' RFs tile the visual field so that for any particular location a subset of neurons will be responsive. Within one location, different patterns of light drive neurons to varying degrees. For instance, V1 neurons tend to respond to small oriented edges or gratings and have preferred orientations and spatial frequencies.

As for physical space, neurons also tile the space of possible orientations and frequencies, illustrated in Fig. 1.2. Visual information is then processed by a hierarchy of brain regions, each spatially and functionally clustered. Neurons downstream of V1 create new response selectivity as they integrate information from previous representations. They progressively increase their RF size (Born and Bradley, 2005), their invariance to object rotation and scaling (Rust and DiCarlo, 2010), and form more complex feature preferences such as motion selectivity, form and face selectivity. These feature maps extract visual information hierarchically and are thought to build the basis for any visually guided behavior.

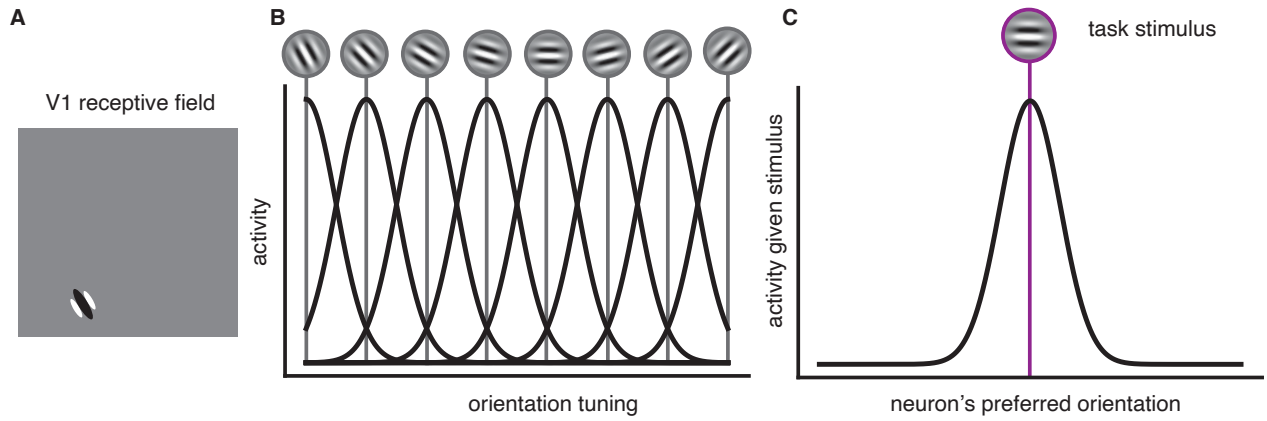

Figure 1.2 Illustrations of spatial and orientation tuning. A) Spatial RF of a theoretical example V1 neuron (see also Carandini et al., 2005). B) Cartoon tuning curves of neurons with same RF but different orientation preference. Orientation preference is indicated with vertical lines. C) Population response resulting from individual tuning curves in A, given the particularly oriented stimulus indicated in purple. Plotted is the activity level over neuron's preferred orientation.

1.2.1 Variability in sensory encoding

While neurons do respond to some feature more than others, their response rate (number of spikes) given one and the same stimulus shown in different experimental trials varies (Goris et al., 2014; Stein et al., 2005). These variations likely have multiple sources, for

instance, internal states (e.g. arousal or attention), recurrent dynamics, or an accumulation of molecular processes that are typically described as stochastic (e.g. fluctuations in ion concentration during synaptic communication) (Allen and Stevens, 1994). From a signal processing point of view, these processes are "noise" that corrupts the stimulus encoding and the ratio between the stimulus signal and the noise is a widely used measure for coding precision (signal-to-noise-ratio, SNR). This noise is considered a major limitation in sensory-guided decision making such as discrimination or detection, and there exists a large body of literature that relates the properties of neural noise to behavioral thresholds (Johnson, 1980; Kang et al., 2010; Renart and Machens, 2014; Shadlen et al., 1996). As there are different sources of noise, its structure can vary. Often noise is reported to be multiplicative, rather than additive, suggesting modulating factors that scale neuron's activity (Goris et al., 2014; Lin et al., 2015). Some of the noise is neuron-specific, but often times there are noise correlations between pairs of neurons, or even low-dimensional structure. Population coding allows to average over independent noise to get more robust estimates, but it can be limited by certain types of noise correlation structures (Zohary et al., 1994). Chapter 2 will explore the structure of such activity fluctuations in more detail.

1.2.2 Top-down gain modulation

In addition to the feedforward processing of information, neurons' activation is further influenced by feedback, that is, signals originating in downstream circuits (Felleman and Van Essen, 1991a; Gilbert and Li, 2013; Markov et al., 2014). Such top-down signals can affect sensory processing in a context, reward and task-dependent manner (Kuchibhotla et al., 2017; Niell and Stryker, 2010; Vinck et al., 2015), for instance, through gain modulation that multiplicatively scales neurons' activity up or down (Goris et al., 2014; Ra-

binowitz et al., 2015; Sherman and Guillery, 1998). Many functional roles have been ascribed to gain modulation (Reynolds and Heeger, 2009; Salinas and Abbott, 1997; Salinas and Thier, 2000) and the neuromodulatory systems involved are diverse (Ferguson and Cardin, 2020). A well-known example is top-down attentional modulation, in which relevant responses are selectively amplified (Carrasco, 2011; Reynolds and Chelazzi, 2004) and their covariability decreased (Cohen and Maunsell, 2009). Attentional modulation has been proposed as a means to selectively improve the accuracy of relevant encoded information (McAdams and Maunsell, 1999; Moran and Robert, 1985; Treue and Maunsell, 1996) since their amplification renders task-relevant information more salient. Chapter 2 will discuss the impact and limitations of gain modulation for task-specific information processing and propose a new role for gain fluctuations.

1.3 Task-flexible decoding

Visual areas extract general features from our environment. However, in order to support flexible behavior, this extensive representation needs to undergo selection and careful consolidation (Britten et al., 1996). Consider a discrimination task based on the basic sensory attribute of visual orientation. As outlined before, in primary visual cortex (area V1) of monkeys, neurons respond selectively to different orientations at different locations in the visual field. As a result, if the task involves a small stimulus or a small difference in visual orientation, we expect only a subset of neurons with particular tuning properties to carry the information relevant to that task (Fig. 1.3), while most neurons carry other signals that are uninformative for the task. It is worth emphasizing here that these "uninformative" neurons are not silent and consequently not by default ineffectual. Instead they need to be actively ignored somewhere downstream (Shadlen et al., 1996), similar

to the Cocktail Party problem, where a lot of voices have to be actively suppressed in order to follow a particular conversation. The performance in a sensory-guided task relies on the ability to properly identify and read out the task-informative responses while ignoring task-uninformative background activity.

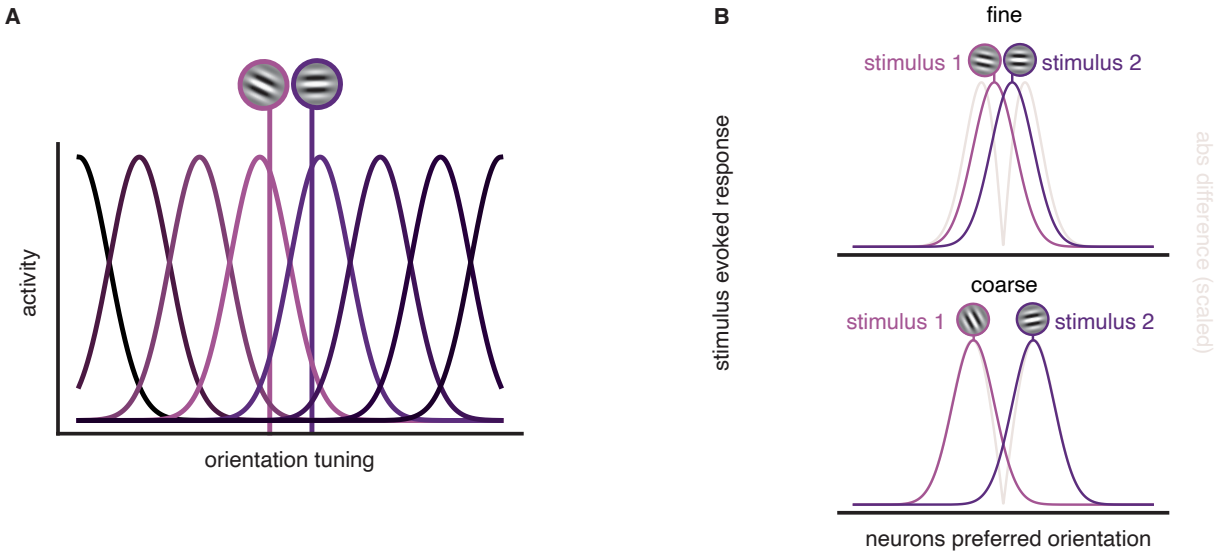

Figure 1.3 Illustrations of orientation decoding from a population with same RF. A) Tuning curves of neurons with different orientation preference relative to two stimulus orientations that need to be discriminated. B) Population responses in two example discrimination tasks, on top a fine discrimination between two similar orientations (10° difference) and below a coarse discrimination (70° difference) - plotted are the stimulus-evoked activity of neurons over their preferred orientation (shades of purple) and the absolute difference in response (grey) as a coarse proxy for information that can be extracted from each neuron.

The readout of stimulus information from any particular brain area is often conceptualized through the lens of *decoding* neural responses. Decoding provides a mathematical framework to study the process of appropriately combining the activity of single neurons or populations to select information required for tasks, such as discriminating between two stimuli or detecting the presence of a stimulus. Here we differentiate decoding approaches that aim to study how task-information is encoded in a population and those that aim to model biological decoding - the process of a brain area collecting incoming information to

guide behavior in a task. The first approach is applied to recorded activity from any brain area to study what information about an experimental stimulus is encoded, while the second approach encompasses the question of how and where the brain itself decodes information to solve a task. We will see that the methods that have been developed to study the first question are not necessarily suitable to answer the second.

1.3.1 The Ideal Observer Framework

Decoding has been extensively studied within the "Ideal Observer" paradigm which models the response properties of a population through an encoding model and then derives a statistically optimal decoder for this model (Berens et al., 2012; Britten et al., 1996; Dayan and Abbott, 2005; Geisler and Albrecht, 1997; Graf et al., 2011; Jazayeri and Movshon, 2007; Ma et al., 2006). As such, this approach relies heavily on Bayes' theorem which directly connects encoding and decoding in a probabilistic framework.

A probabilistic encoding model describes the response activity of a single neuron n in a particular visual area by translating an instantiation of a stimulus dimension s (for simplicity assumed to be 1-dimensional and called the stimulus) into predicted spiking response k_n through the conditional probability distribution $P(k_n|s)$. Fluctuations around the stimulus-evoked mean response are inherent to neural responses and captured by the stochasticity in $P(\cdot)$. Typically neural responses quantified as the number of spikes in a set temporal window are well described by Poisson variability since the variance tends to increase with the mean,

$$k_n \sim Poisson(f_n(s)), \tag{1.1}$$

where $f_n(s)$ is a function of the stimulus which expresses the mean response of neuron n

to the stimulus (as in the example of orientation tuning in Fig. 1.2C). Decoding is the reverse process and describes the probability of a particular stimulus given a specific neural activation $P(s|k_n)$. Bayes' theorem allows us to link the encoding $P(k_n|s)$ to the decoding probability of $P(s|k_n)$ for a single neuron n or a population with activity $\mathbf{k}$

$$P(s|\mathbf{k}) = \frac{P(\mathbf{k}|s) * P(s)}{P(\mathbf{k})}, \tag{1.2}$$

The total probability of any particular stimulus $P(s)$ in the world is generally undetermined and assumptions have to be made about the distribution of relevant stimuli. Once an encoding model is formalized and its parameters $\mathbf{f}(\cdot)$ estimated, we can decode the stimulus s from the spike responses $\mathbf{k}$ through maximum a posteriori (MAP) inference, which chooses the most likely stimulus given what we know and using Eq. 1.2. If we assume that all stimuli are equally likely (i.e. $P(s)$ is constant), the factor $\frac{P(s)}{P(\mathbf{k})}$ is independent of s and the MAP estimate simplifies to the Maximum Likelihood (ML) estimate (Dayan and Abbott, 2005). We can derive the ML estimate for a population of independent Poisson neurons by maximizing

$$P(\mathbf{k}|s) = \prod_n \frac{f_n(s)^{k_n} \exp(-f_n(s))}{k_n!}, \tag{1.3}$$

with respect to s. Following the standard approach of maximizing the mathematically simpler $\log P(\mathbf{k}|s)$ given that the position of an optimum does not change with a log-transformation, we set the derivative of the $\log P(\mathbf{k}|s)$ to 0 and solve for s, resulting in

$$\sum_n k_n \frac{f_n'(s_{ML})}{f_n(s_{ML})} = 0, \tag{1.4}$$

where $f_n'(s_{ML}) = \frac{\partial f_n(s_{ML})}{\partial s_{ML}}$ and assuming a population that uniformly spans the stimulus

range so that $\sum_n f_n(s)$ is approximately independent of s. The exact estimate depends on the form of $f_n(s_{ML})$ but for the classical Gaussian tuning curves around a preferred stimulus s_{pref} given by

$$f_n(s) = k_{max}\exp\left(-\frac{1}{2}\frac{(s - s_n^{(pref)})^2}{\sigma^2}\right), \tag{1.5}$$

we can derive the estimate s_{ML} as

$$s_{ML} = \frac{\sum_n k_n s_{pref}}{\sum_n k_n}. \tag{1.6}$$

The form of the tuning function varies with the type of stimulus feature. For example, circular variables such as direction require periodic tuning curves (e.g. the von Mises function $f_n(s) = k_{max}\exp\left(\frac{1}{\sigma^2}\cos(s - s_n^{(pref)})\right)$ (Fiscella et al., 2015; Jazayeri and Movshon, 2006)), however, the overall approach is the same.

There are several theoretical and practical applications of the Ideal Observer (IO) framework. It allows studying the accuracy and reliability of an encoding population (Dayan and Abbott, 2005) by providing an upper bound on how well information can be read out assuming a certain encoding model. This upper bound can than be applied to different brain areas to study changes in the type of information that is encoded. Rust and DiCarlo (2010) have found that the features encoded by V4 and IT encode natural images with similar discrimination precision but that IT is much better at preserving object identity over position, scale and context. Beck et al. (2012) used IO models to study decision making and compare the contribution of internal vs inference noise to explain behavioral performance. Jazayeri and Movshon (2007) used a Bayesian IO model to study biases in perceptual decision making and found that humans make biased estimates in discrimi-

nation tasks similar to what a Bayesian optimal decoder would suggest. Further on, the IO framework can identify the mathematical form of an optimal decoder. For instance, Eq. 1.6 suggests that the optimal decoding for Poisson stochasticity and Gaussian tuning curves is a linear weighting of each neuron's response by its preferred stimulus divided by the overall response. Other encoding models may lead to different decoder forms (for examples see Table 1.1). Given the decoding form, we can study the biological plausibility of different decoders resulting from varying encoding models and identify mechanistic limitations on which decoders could be approximately implemented by the brain given what we know about neural computations (Ma et al., 2006). Another application of the IO framework is the study of correlation structures among neurons (Berens et al., 2012; Franke et al., 2016; Kanitscheider et al., 2015a,b; Moreno-Bote et al., 2014; Pitkow et al., 2015). Specifically we can relax the unrealistic assumption of independent neurons and take into account biological interactions that cause dependencies which can be observed, for instance, as pairwise correlations. By estimating such correlations and integrating them in our encoding model we can test their effects on the encoding precision by deriving their respective optimal decoders (Kanitscheider et al., 2015a). Additionally, the IO framework has not just theoretical but also practical application. It is essential in the field of Brain Computer Interfaces, where it has proven successful in reading out and manipulating the activity of different brain areas with enough precision to guide artificial prostheses (Andersen et al., 2004) and even decode precise dynamics to form single written letters from activity of the human premotor area (Willett et al., 2021).

Despite its important application, the IO framework has several limitations. First, not all encoding models allow analytical derivation of a decoder (see Table 1.1) and there is a trade-off between biological accuracy of the encoding model and mathematical simplicity of the decoder. For instance, neuronal tuning is multidimensional but marginalizing for-

	Task	Encoding function	Noise assumption	ML decoder
(1)	Estimation	linear	Gaussian	linear
(2)	Estimation	log-Gaussian	Poisson	linear+normalization
(3)	Estimation	log-linear	Poisson	no closed form
(4)	Discrimination	mean SR	Poisson	linear

Table 1.1 Ideal observer ML optimal decoders resulting from different assumptions. SR refers to stimulus response.

mally over many stimulus dimensions can be nontrivial. Also, the above-mentioned study of correlation structure is limited by the interactions that we can capture with a tractable encoding model (typically low-dimensional correlation patterns). Second, the theoretical study of encoding and decoding of a population is ultimately limited by the neurons that can be recorded. The IO framework does not allow us to estimate the impact of neurons that do not respond much to the experimental stimuli for decoding, since they are typically discarded or mixed in with recording noise. In practice, we need to estimate the encoding model's parameters and this estimation is going to be limited by the number of trials that the recorded population was observed for. This has important implications for the comparison between simple and more complex encoding/decoding models. For instance, an encoding model assuming independent neurons may be less accurate than one that reflects the intricate interactions between neurons (such as Pillow et al., 2008), but it would require fewer parameters and consequently may be easier to fit with the data available (Berens et al., 2011; Kanitscheider et al., 2015a).

1.3.2 Learning task-specific decoders

From the perspective of a scientist who is trying to understand the recorded activity of a population, a lot of valuable knowledge can be gained from formulating an encoding model

and estimating its parameters given the maximum amount of data available. However, when considering how the brain solves a task, the biological and behavioral plausibility of learning the necessary parameters has to be taken into account. Specifically, identifying the IO optimal decoder assumes knowledge about the response properties of neurons, their tuning curves along a particular stimulus dimension, their stochasticity properties (details about their fluctuations around the mean response), and additional gain modulatory factors. It is unclear and arguably unrealistic that a decoding area in the brain could store all this information for many encoding neurons from different areas and many stimulus types. This framework therefore lacks a theory of how the optimal decoder could be learned online from neural activity.

An alternative approach is to learn decoding weights directly by minimizing a supervised loss function (Dayan and Abbott, 2005). The loss function depends on the specific task, for instance, in an estimation task we may assume that the decoder should minimize the mean squared error (MSE), but given a discrimination task where the aim is to differentiate two categories, a more appropriate loss function is the cross-entropy (CE) loss. There are cases where defining the loss function or assuming an encoding model to derive the MAP estimate can lead to the same decoding solution. The simplest is a linear dependency with Gaussian noise in the context of an estimation task. If the IO encoding model is linear Gaussian, $\mathbf{k}_t \sim N(\beta s_t)$, its optimal decoder coincides with the linear regression decoder aiming to minimize the MSE, $\mathbf{a} = (K^T K + \lambda I)^{-1} K^T \mathbf{s}$, where K is the spiking activity of all neurons to sufficiently many i.i.d. stimulus instantiations $\mathbf{s}$. These decoding weights can be learned online by using gradient descent optimization with a learning rate α

$$\mathbf{a}_{t+1} = \mathbf{a}_t - \alpha(\mathbf{a}_t \mathbf{k}_t - s_t)\mathbf{k}_t, \tag{1.7}$$

where the index t indicates the temporal progression of the learning samples.

Other IO encoding models also lead to simple linear decoders that can be approximated by linear regression, as in the case of a population of Poisson neurons in the context of a discrimination task (see Table 1.1). However, generally the IO decoding weights are not necessarily equivalent related to the ones derived through linear regression. Their similarity can, however, be assessed numerically. For instance, for the encoding model corresponding to Equ. 1.6, simulations show that the regression decoder weights for the log counts, $\log(\mathbf{k})$, are in fact proportional to the IO optimal weights, assuming that the encoding model is correct and stimulus samples are i.i.d.

More complex regression methods can improve decoding precision compared to linear regression, by taking into account higher order dependencies (e.g. polynomial regression or deep neural networks). If such regression models outperform a chosen encoding model's optimal decoder, this suggests that this encoding model may be insufficient. Conversely, a regression model that is arbitrarily complex may be unrealistic for what the brain can actually implement and use to solve a task. Even for a simple regression model, the biological plausibility of learning its parameters is questionable. Whether supervised feedback is available to the brain and can backpropagate to change readout of encoding neurons at a primary sensory area is highly debated (Crick, 1989). However, even at the normative level, regression is problematic - the mere number of samples that would be required to adjust decoding weights for any particular task from supervised feedback alone would prohibit flexible behavior as described in Sec. 1.2.

Specifically, the local RF and particular tuning properties discussed in Sec. 1.2 suggest that for any task involving a particular set of stimuli only a subset of the population carries information, while the rest of the population may be activated by other, task-

irrelevant sensory features (experimental evidence is further given in Chapter 2). This has two important implications for the performance of regression which can be illustrated even in the simple setting of a linearly dependent variable with Gaussian noise (simulating an observed variable $\mathbf{x} \sim N(\beta s, \sigma^2 I_N)$ and applying linear regression to these observations $\mathbf{x}$ to estimate s). Fig. 1.4A shows learning trajectories for different numbers of informative variables out of a total population of $N = 100$ variables quantified via the MSE; the MSE at convergence naturally decreases (Fig. 1.4B) and the initial learning slope increases as the number of informative variables in the population grows (Fig. 1.4C). As a consequence, the number of trials needed to reach a MSE criterion of 0.1 decreases with the number of informative variables (and may not even be achievable at all with very few informative variables). The performance is worse and learning is slower with fewer informative variables. While adding *"uninformative"* variables ($\beta_n = 0$) to a fixed informative set of variables in linear regression does not change the minimum MSE that can be achieved by that population (Fig. 1.4D-E), it can substantially impact the number of training trials that is required to reach a MSE criterion (Fig. 1.4F). Consequently, the many neurons with mainly task-irrelevant activation may be easy to discard if decoding weights are known (as their weights can be set to 0), but they pose a non-trivial challenge for learning as they obstruct finding and reading out from the informative neurons. Therefore supervised regression learning alone is likely insufficient to explain the flexible, few-trial learning, behavior shown by humans and animals.

1.3.3 Impact of hierarchical sensory processing

Both the IO optimal decoder and supervised regression approaches described above assume decoding from one specific area, typically a recorded primary sensory area, but when

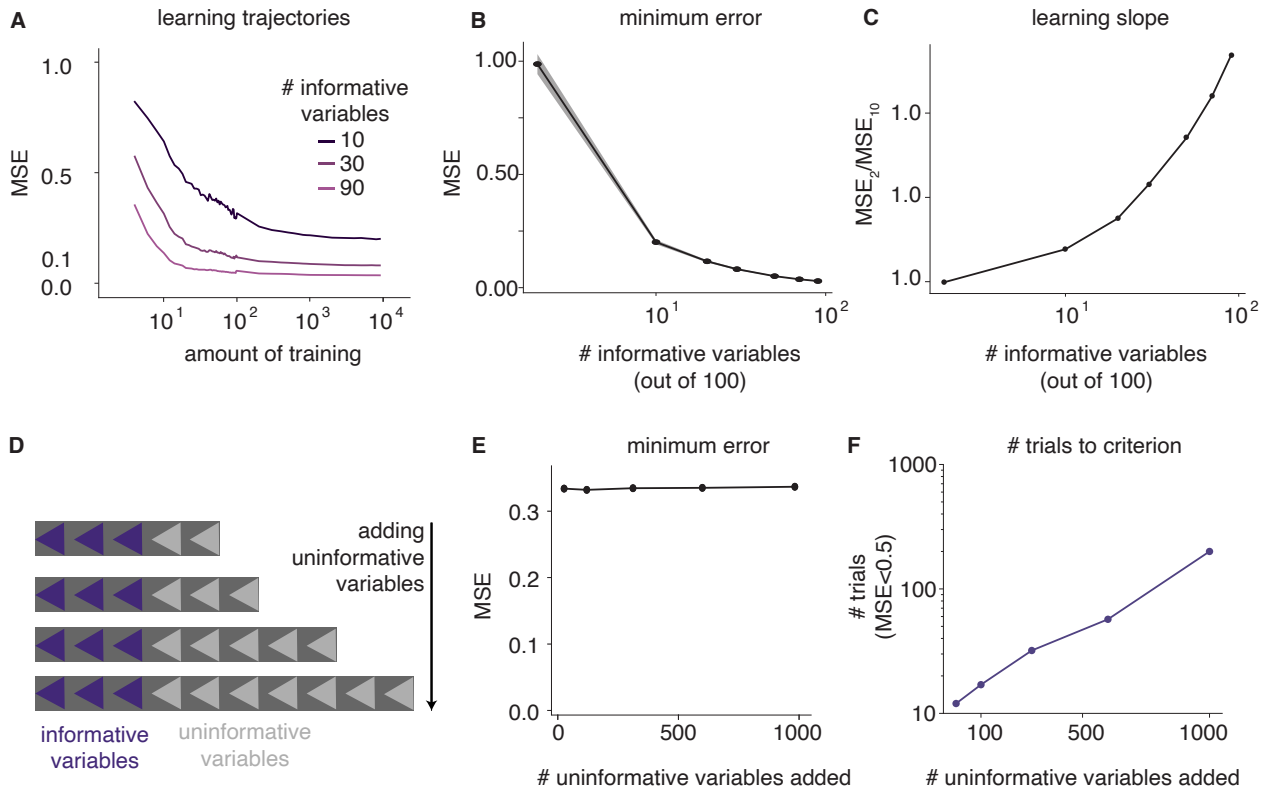

Figure 1.4 Ideal observer optimal decoder with sparsely informative population. Simulations are from a linear encoding model with a scalar hidden variable s and N observed stochastic variables $\mathbf{x} \sim N(\beta s, \sigma^2 I_N)$. A ridge regression model recovers s_t given $\mathbf{x_t}$ at trial t. Depending on the distribution of β a different number of trials is needed to learn how to read out s from $\mathbf{x}$. A) Shown are different learning trajectories if out of $N = 100$ variables, 10, 30 or 90 have a β_n value $\neq 0$ (termed *informative* variables) . B) The minimum error reached after the regression weights have converged, decreases with increasing number of informative variables. C) The decrease in MSE after the first 10 training samples ("learning slope") over the number of informative variables. D) Adding uninformative variables ($\beta_n = 0$) to a group of informative variables ($N_{inf} = 6$). E) The minimum error is constant over the number of added uninformative variables. F) The number of trials needed to reach a criterion performance increases with increasing number of uninformative variables.

studying how the brain could solve decoding, it is not just the "how" that is unclear, but also the "where/when". Given the modular and hierarchical structure of the brain, every brain area between sensory receptors and muscle activity takes the role of both encoder and decoder, depending on whether it is considered relative to the previous or the follow-

ing brain area. However, sensory representations are believed to be mostly stable while neurons in cognitive areas such as the prefrontal cortex, hippocampus or basal ganglia are strongly task-modulated, represent task rules and have widely varying tuning properties depending on the current demands (Kobak et al., 2016). Behavioral flexibility is therefore typically associated with higher order cognitive areas (Kang and Maunsell, 2020; Mante et al., 2013; Woolgar et al., 2015), suggesting that the problem of task-specific decoding happens at an advanced stage of processing. This introduces an additional challenge to the study of biological decoding: how do these higher order areas access the particular sensory information for a task (Kang and Maunsell, 2020)? Primary visual areas do not have sufficient connectivity to decision areas in the brain for direct readout, so information needs to propagate through the processing hierarchy before reaching decision areas. Three main issues emerge: first, once sensory information is lost due to a particular transformation at one stage of processing, downstream areas cannot recover this information (data-processing inequality) (Herzog and Clarke, 2014), which requires that general information is preserved before task information is selected later on. As a consequence, task-information is sparse and embedded in a large pool of non-informative encoded information. Second, another challenge pertains to the sparsity of a representation and can be illustrated with the previously described example of discriminating small patches of oriented gratings. Given the properties of visual encoding outlined in Sec. 1.2, small oriented edges optimally drive primary visual area V1, hence the information to discriminate these gratings is well contained and easily accessible in a few neurons of V1 (Froudarakis et al., 2014). Following V1, the encoded information undergoes a sequence of transformations that aims to build invariances and detect broader features (Hong et al., 2016; Yamins and DiCarlo, 2016). In this process, task-information mixes with task irrelevant signals before reaching decision areas. Specifically, downstream of V1, the increase in RF

size, feature complexity and invariance abstraction imply the dissipation of information about local gratings. So even if information is still fully available at the last stage, it is certainly much more spread across neurons. Such a diffuse representation can substantially slow down the learning of decoding weights. Again, this can be illustrated in a simple estimation task and linear Gaussian dependency (as in Fig. 1.4). Given a fixed amount of information in a population that can be either localized in a few neurons or spread across many (Fig. 1.5A), the estimation precision that can be achieved with unlimited data is unchanged (Fig. 1.5B), but given limited data localized information gives higher precision than spread information (Fig. 1.5C). Intuitively this makes sense because for localized information fewer variables have to "be detected". However, it leaves the question of how the brain may deal with decoding from such diffuse representation. Thirdly, as every processing stage will add noise as it propagates a signal forward, the information will be less reliable by the time it reaches decision making areas. Humans are able to react not only to single localized gratings, but even to as little as a single photon captured by the eye (Hecht et al., 1942; Tinsley et al., 2016), demonstrating extraordinary scaling ability given the neural noise and the miniscule fraction of elicited responses relative to overall background activity in the brain. It is unclear how such very particular, localized stimulation could inform and drive behavior.

1.3.4 How does the brain decode?

We illustrated how behavioral flexibility that may seem trivial and universal across species poses a fundamental challenge for artificial systems, as even small changes in a task (e.g. changing the location or scale) require very different sensory information and consequently rely on significant changes in the information-processing machinery. Both the IO optimal

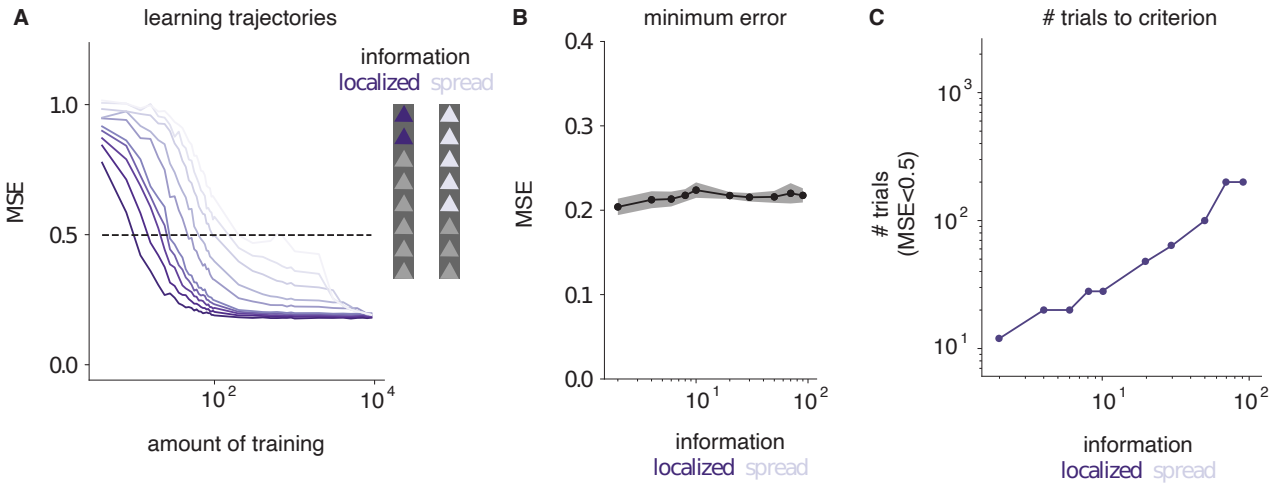

Figure 1.5 A) Learning trajectories for L1 linear regression models with $N = 1000$ variables of which either few variables are very informative (localized information) or many variables are a little bit informative (spread information). Here the total amount of information in the set of variables is approximately kept constant by decreasing the informativeness of any individual variable if more variables are informative $\left(\beta_{inf} = \frac{2}{\sqrt{D_{inf}}}\right)$. B) The minimum error over the number of variables over which informativeness is distributed. C) The number of trials needed to reach a criterion performance (MSE< 0.5) increases with increasing number of variables over which informativeness is distributed (within this range of information sparsity, L2 linear regression performs worse than L1).

decoder and regression-based approaches are fundamentally flawed as models for decoding in the brain (Dayan and Abbott, 2005). Storing either the multidimensional response properties of all encoding populations or the resulting decoding weights for many different tasks would allow quick switching between different task-specific decoders, but it is unrealistic for a decoding area to acquire and contain that much information, disqualifying the IO framework. On the other hand, learning decoding weights directly by minimizing a loss function would require large numbers of repeated exposures, a limitation that puts supervised regression at odds with the time scales of behavioral flexibility seen in humans and animals.

1.4 Outlook

This thesis explores the neural computations underlying flexible behavior in the context of a visual discrimination task. Chapter 2 presents data analysis of neural recordings in area V1 and medial temporal area (MT) provided by Douglas A. Ruff and Marlene R. Cohen and previously published in Ruff and Cohen (2016a). Chapter 3 introduces a novel theoretical framework for task-flexible decoding and demonstrates its ability to extract information from an encoding population. Predictions of the theory are then tested directly in the data and compared to both the IO framework and regression. Chapter 4 extends the theoretical framework to include hierarchical processing and to study learning, illustrating the plausibility of this novel theory in a hierarchical model of flexible visual processing.

Chapter 2

Structured variability during stimulus encoding in primary visual cortex

2.1 Introduction

In this chapter flexible behavior is studied in the context of a change detection experiment in non-human primates (Ruff and Cohen, 2016a,b), and together with its neural correlates in primary visual area V1. We quantify monkeys' ability to switch between different task variations presented in a block design and analyze how information about the task is encoded in primary visual neurons. We look at the encoding properties of the population specifically considering potential issues for decoding (described in Chapter 1) to better understand what challenges biological decoding in the brain needs to overcome in such a task and given these neural representations. Finally, we extract and study other sources of fluctuations in the neural activity that are not caused by the stimuli but impact the encoding precision (so called "noise").

2.2 Task design

Monkeys were trained to detect a small change in orientation/direction of a Gaussian-windowed drifting sine grating (Fig. 2.1A). Two to three gratings were present simultaneously, at high or low contrast levels, and spontaneously changed their orientation (coupled to an corresponding change in drift direction, Fig. 2.1A). However, the animals were rewarded only for responding to changes of one of these gratings, with the others acting as distractors. The location of the relevant stimulus was fixed within blocks of trials (Fig. 2.1B), switching randomly between blocks throughout an experimental session. The two task-orientations of a stimulus also changed between blocks. The task-relevant location is indicated by a few instructional stimulus presentations, and is selected randomly for each block within the session ($\sim 3-6$ blocks per session). We analyze each recording session by splitting it into the task-specific blocks of trials. In a trial, gratings flash on (200ms) and off (200-400ms) at the same orientation (repeated, 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 failed to detect 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). More details about the experiments can be found in Ruff and Cohen (2016a).

2.3 Behavioral performance

Once the overall reward structure of the task was learned, monkeys performed well; they were able to hold fixation through several presentations of one repeated stimulus, detect the change in orientation when presented, make the appropriate decision to react and execute an indicative saccade (Fig. 2.1C). Importantly, they performed this behavior flexibly and were able to quickly adjust to switches in task-relevant stimulus location (Ruff and Cohen, 2016a), reaching asymptotic performance levels roughly 5 trials after each task change (Fig. 2.1D). We aim to explain how the brain achieves this impressive combination of accuracy and flexibility. For simplicity we study the change detection through the lense of discrimination, in both cases the information required for accurate behavior is the difference between two stimuli.

2.4 Encoding of local visual orientation in a V1 population

Neurons in V1 respond selectively to the local orientation of visual stimuli, and the selectivities of the full population span all orientations and visual field locations. In the experiment, individual grating stimuli are roughly matched to V1 RF sizes at the eccentricity at which recordings are performed, and orientation changes are relatively small (10-45°, see Ruff and Cohen, 2016a), which restricts relevant stimulus information to a small subset of V1 neurons. Nearly all visual information passes through V1 (Felleman and Van Essen, 1991b), within which neurons of similar spatial and orientation selectivity are proximally located. The behavior of the monkey must rely on the responses of this relatively small localized subset of V1 neurons whose responses change with the stimulus orientation, while

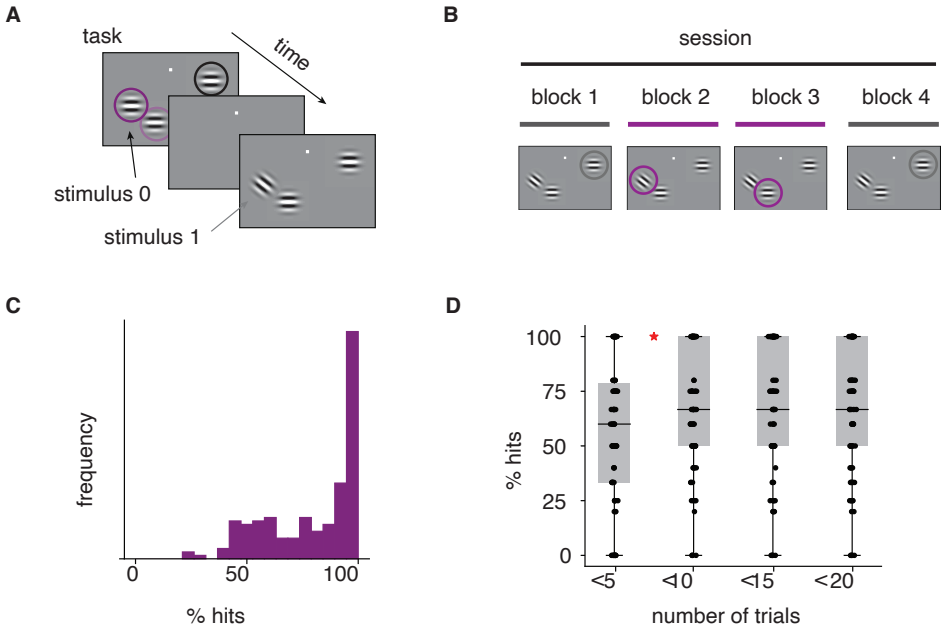

Figure 2.1 An orientation discrimination task with distractors. A) In each block of trials, two to three drifting gratings flash on and off on a screen and can change their orientation. One stimulus is selected as relevant for the task, and the monkey must report the change in its orientation with a saccadic eye movement (Ruff and Cohen, 2016a). B) The block design of the task. C) Distribution of behavioral performance across blocks, quantified by the % hits among hits and misses. 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; the boxes mark 25 and 75% quantiles, points indicate different blocks and the red star indicates a significant difference between the means of the two adjacent distributions (relative two-sided t-test, $p = 0.015$).

ignoring the background chatter of activity from the remainder of the population. Moreover, since the downstream decision-making area does not have access to V1 responses directly, the task relevant information must be traced as it progresses through various stages of visual processing.

2.4.1 Experimental methods

While monkeys discriminated the orientation of the local gratings presented to them on a screen, spiking responses of neurons in their primary visual cortex (V1) and middle temporal area (MT) were recorded simultaneously (Fig. 2.2A). For most of the recording sessions analyzed here 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 Ruff and Cohen, 2016a). The responses of "units" measured on each electrode can correspond to either clusters of multiple neurons or single neurons.

Two of the stimuli were positioned to drive the MT unit similarly and evoked responses in the recorded V1 population. The other stimulus, if present, was positioned outside of the MT and V1 receptive fields (RF) (see Fig. 2.2A, adapted from Ruff and Cohen, 2016a). In this chapter we focus on the V1 recordings as V1 constitutes a bottleneck for information transfer due to its ideal encoding properties for this task (see previous Sec. 2.4), but we will come back to MT in Chapter 4. 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 task blocks were interleaved within a session. V1 neural populations may overlap across sessions. 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 V1 units. We only include V1 units whose response to either one of the stimuli was at least 10% larger than their baseline value to avoid inclusion of noise channels into the analysis, since those could

bias the modulator targeting analysis. On average 88 V1 units ($\sim 90\%$) in a block showed stimulus modulation for one of the two stimuli placed within the MT RF (min 52, max 95). We exclude V1 units with a Fano factor > 5 standard deviations above the population average as this suggested especially many/diverse neurons in the unit. $0-3$ units per block population were excluded by this criteria with a mean of < 1 unit.

2.4.2 Quantification of neural informativeness

When one of two locations within the recorded V1 population's RFs was task-relevant, we expect a subset of the recorded V1 neurons to provide information for the animal's decision ("relevant tasks"). In contrast, the neurons should be uninformative when the third stimulus location is task-relevant as it lies in the opposite hemisphere ("control task"; Fig. 2.2A).

We quantified the task-informativeness of each V1 unit as the absolute difference in mean responses for the two orientations relative to response standard deviation, known as $|d^{'}| = \left|\frac{\mu_0-\mu_1}{\sqrt{0.5(\sigma_0^2+\sigma_1^2)}}\right|$ where μ_0 and σ_0^2, μ_1 and σ_1^2 are the means and variances of a unit's responses to the task-relevant stimuli 0 and 1, respectively. We compute informativeness across all stimulus presentations in behaviorally correct trials of the same block. 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 neural informativeness reliably. Varying this criterion does not qualitatively change the results. The first stimulus in a trial was always removed to allow for adaptation effects (Cohen and Maunsell, 2009). 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. We then take significantly informative units as those with a percentile >= 99.

Figure 2.2B shows the relationship between informativeness and responsiveness for three representative examples. First, an example unit that is weakly responsive to both stimulus orientations (for instance, because its RF does not overlap the stimulus location or because its preferred orientation was orthogonal to the stimulus) and consequently cannot be informative about stimulus identity (Fig. 2.2B, left). Cells like this one are the majority. Second, some units respond strongly but similarly to both stimuli (for instance, their orientation tuning curve may be centered between the stimuli; Fig. 2.2B, middle), showing that responsiveness is necessary but not sufficient for task-relevance. Third, some units respond strongly to only one of the two stimuli and hence have high informativeness (Fig. 2.2B, right). This separation of responsiveness and informativeness further illustrates the difference between a pair of neurons that is similarly tuned and a pair of neurons that is similarly informative; informativeness does not differentiate between which stimulus a neuron responds to more strongly and hence two neurons that are very differently tuned can still be equally informative.

Overall, for each relevant task block, a modest proportion of the recorded V1 units are significantly informative (monkey 1: 25.8%, monkey 2: 18.4%; non-parametric test, see Suppl. Sec. 2.4.2 for details), whereas only 2.4% and 6% of units are significantly informative in the control task (Fig. 2.2C). Neurons that are most informative in either of the relevant tasks have low $|d'|$ in the control task, reflecting their task-specificity (Fig. 2.2D). Across the two relevant tasks, unit informativeness is similar because of the close proximity of the two relevant stimulus locations. Specifically, on average only 9% and 11% of

units are informative in just one out of two relevant tasks while 14% are informative in both (and the remaining 66% are uninformative).

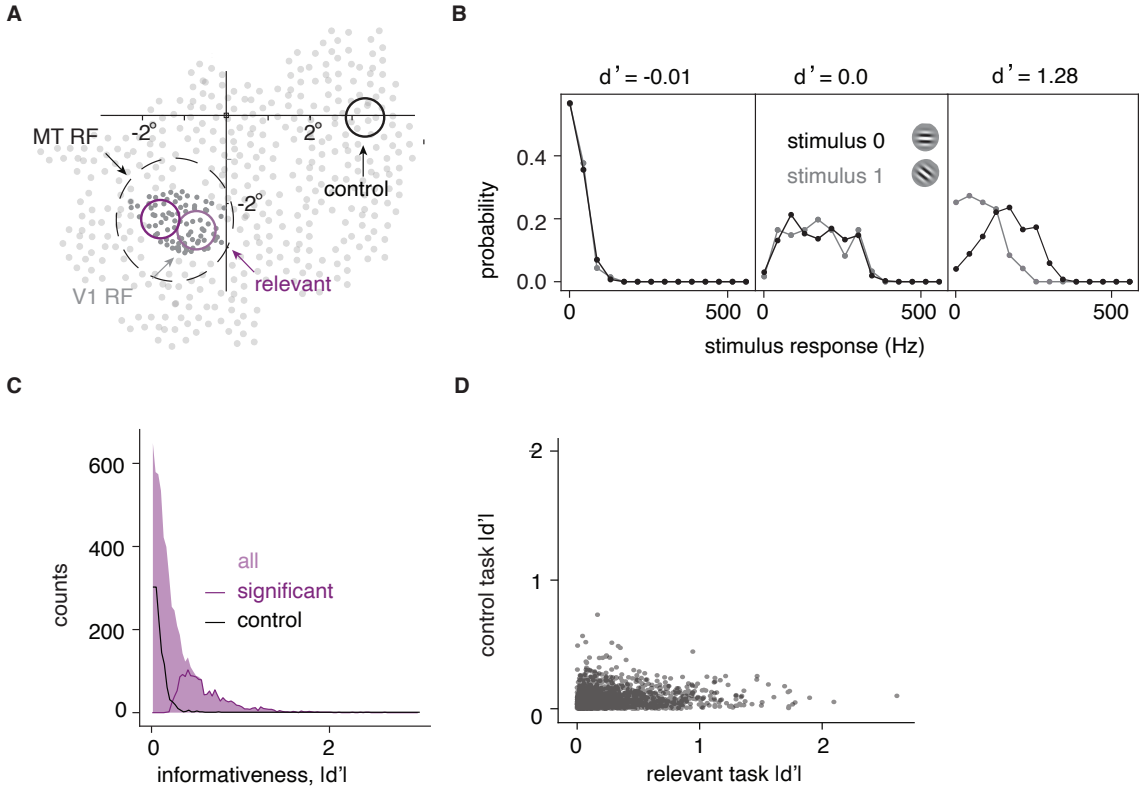

Figure 2.2 V1 neural informativeness in an orientation discrimination task. A) The recorded population of V1 neurons has RF centers (dark gray) close to one another and within the RF of a simultaneously recorded MT unit (Ruff and Cohen, 2016a). Two of the three stimuli locations are within the MT unit's RF ("relevant" - light and dark purple) and one is in the opposite hemisphere ("control" - black). Most V1 units have RF partially overlapping the relevant stimuli but not the control. Light grey dots illustrate the RF centers of other "imagined" unrecorded V1 neurons. B) The distribution of response rates over all stimulus presentations, to each of the two task stimuli for three example neurons with different d' values. C) The distribution of informativeness values, $|d'|$, over all blocks of relevant tasks and all V1 units (shaded purple). Lines indicate the subdistribution of neurons with significant informativeness (purple), and neurons in the control task (black). D) Relationship between the informativeness values in relevant and control tasks for units recorded in both tasks. Informativeness is always computed based on the changes in activity accompanying changes in the stimulus at the task-specific location. A and B adapted from Ruff and Cohen (2016a).

2.4.3 Statistics of V1 population responses for informative and uninformative units

We summarize single unit statistics across the V1 population and look for differences between task-informative neurons and task-uninformative neurons. We find that the distributions of mean firing rate in a trial is shifted slightly higher for the informative neurons (Fig. 2.3A). This is expected, since higher activity facilitates detecting systematic activity differences with stimulus orientation. However, the difference is small, informative units have an average rate in a trial of 42 spikes/second with a standard deviation of 22, while uninformative units have a mean of 37 spikes/second with a standard deviation of 20. We also find that the distributions of Fano factors for high contrast stimuli is shifted slightly higher for informative neurons (Fig. 2.3B, informative units have a mean Fano factor of $= 1.14 +/- 0.47$, uninformative units have a mean Fano factor $= 1.06 +/- 0.62$).

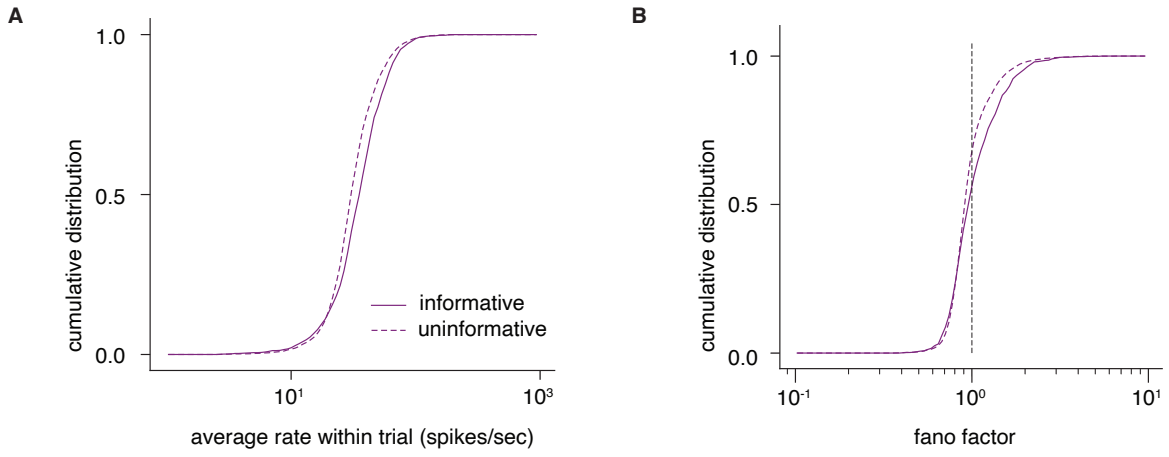

Figure 2.3 Basic statistics of V1 units. A) Average firing rate during a trial, separating the units into informative and uninformative subpopulations. B) Fano factor distributions for informative and uninformative units.

2.5 Implications for decoding

In this analysis we found that, within each task block, a different subset of V1 neurons carries task-relevant information. In order to make accurate decisions, a downstream circuit has to read out selectively from the currently relevant subpopulation. Moreover, the determination of this relevant subpopulation happens quickly: the monkey's performance reaches asymptotic levels after only a handful of trials (Fig. 2.1D). How can this flexible routing of information be achieved? Task feedback alone seems insufficient to robustly guide the selection of the informative V1 subpopulation. Moreover, as basic response statistics such as mean or variance of activity do not differ much between informative and uninformative neurons (see Fig. 2.3), they cannot guide this selection. Specifically, mean responses are modulated by many stimulus dependent factors, which do not necessarily have task-relevance, such as contrast (Reynolds and Chelazzi, 2004) and novelty (Jaegle et al., 2019). Modeling results suggest that tuning-specific gain increases in early stages of sensory processing may ultimately not have a strong impact on behavioral performance (Lindsay and Miller, 2018). Moreover, there are instances where behavioral performance benefits can be dissociated from increases in firing rates experimentally (Ni et al., 2018; Zénon and Krauzlis, 2012), suggesting that these are not the primary mechanisms that support task-specific processing and decoding in the brain. How the brain orchestrates the, mostly irrelevant, information from V1 to form a decision in a particular task is a mystery.

2.6 Functional structure in shared variability

Additionally to stimulus-evoked response, neurons activity fluctuates in time and varies from trial to trial, even if the same stimulus is presented. These fluctuations are often termed “noise" and have different sources, some of which is neuron-specific, but some is correlated across neurons (called “noise correlations”) (Averbeck et al., 2006; Cohen and Kohn, 2011; Huang et al., 2019). Here we use neural fluctuations or variability to refer to this stimulus-independent noise. These co-fluctuations have at times been attributed to shared stochastic gain modulation (Archer et al., 2014; Ecker et al., 2016; Goris et al., 2014; Lin et al., 2015; Rabinowitz et al., 2015) that introduces multiplicative temporal variability. Some empirical evidence suggests that this shared variability can have interesting task-specific structure (Bondy et al., 2018; Ni et al., 2018; Rabinowitz et al., 2015). Here we propose that such task-specific dynamic structure in the joint statistics of neuronal responses may be key to understanding flexible readout. The following sections discuss methods to extract different types of covariability from a population of neurons. I make use of these methods to test for shared modulation in the previously discussed V1 dataset and analyze its task-specific structure.

2.7 Methods for extracting shared modulation in neural responses

Shared modulation has been studied with different statistical tools; pairwise correlations give information about how specific neurons cofluctuate (Cohen and Maunsell, 2009; Goris et al., 2014; Ruff and Cohen, 2014; Rumyantsev et al., 2020), dimensionality-

reduction highlights axis of shared variability (Huang et al., 2019; Ni et al., 2018), and dynamical latent models specifically take into account temporal dependencies of observations (Macke et al., 2015; Yu et al., 2009). Following this is a short overview of the methods used to study modulation and their potential to increase scientific insights gained from data.

2.7.1 Pairwise correlations

Pairwise cofluctuations between neurons appear either as a result of stimulus variations ("signal correlations") or given the same stimulus ("noise correlations") (Cohen and Maunsell, 2009). Here we focus on the second. Noise cofluctuations can be quantified for example through the covariance, the Spearman correlation or the Pearson correlation coefficients. Usually one of these measures is applied to every possible neuron pairing and the mean is taken over all measurements. The mean of pairwise correlations has been shown to depend on the behavioral state of the animal (Ecker et al., 2014), differ between differently tuned neurons (Ruff and Cohen, 2014), change with task demands (Bondy et al., 2018; Cohen and Maunsell, 2009), and depend on whether neurons are from the same or different brain areas (Ruff and Cohen, 2016a). Pairwise noise correlations have been suggested to have varying functional implications for encoding depending on their structure (Abbott and Dayan, 1999; Lakshminarasimhan et al., 2018; Moreno-Bote et al., 2014; Zohary et al., 1994). Generally increasing the number of neurons that encode a stimulus increases the encoding precision, however, there exists extensive work that shows that specific types of correlations can impact encoding precision in the limit of a large number of neurons (Franke et al., 2016; Pitkow et al., 2015). Conclusions regarding the functional role, implications and origins of neural correlations tend to vary (Ecker et al., 2014;

Rumyantsev et al., 2020), which is in part because of the limitations of this measure. Taking the mean of all pairwise combinations implies the assumption that all pairs' cofluctuations share one origin and/or function and that differences are purely due to noise that can be averaged over. However, the distributions of correlation coefficients tend to be quite broad and the differences in mean relatively small in comparison, suggesting that there are many co-factors that influence correlations and cannot be well separated. For instance, it is likely that neurons share multiple sources of variability that differ in their source and function but contribute jointly to pairwise correlations.

2.7.2 Dimensionality-reduction

Dimensionality reduction methods such as Principal Component Analysis (PCA) or Factor Analysis (FA) are widely used tools to identify the main axes of variability in a neural population. They allow estimating the dimensionality of a neural population's joint activity and study changes in dimensionality due to extrinsic factors such as experimental stimuli (Stringer et al., 2019), or intrinsic factors like attention or learning (Huang et al., 2019; Ni et al., 2018). Recently they have also been used to study activation flow across different areas (Semedo et al., 2022). Importantly, they encompass the measure of noise correlations as the entire spectrum of principal components can exactly reconstruct the activation of each neuron. While pairwise noise correlations are difficult to visualize, dimensionality reduction methods allow plotting the principal axes of variability that drive a population in a simple 2-3 dimensional plot that facilitates understanding and can give insight into the computations that are being performed (assuming those computations are low dimensional) (Cunningham and Yu, 2014; Veuthey et al., 2020; Whiteway et al., 2020). However, there are several suboptimal assumptions that many of these

models make. PCA is deterministic, which means that it does not explicitly model neuron-individual fluctuations, which vary substantially within a population. This can be problematic as the estimated shared dimension of variation can be biased towards neurons with higher firing rate and consequently larger variance. Other methods such as FA do explicitly model individual variance but assume a Gaussian distribution which is typically problematic for non-negative, discrete neural data like spike counts (different if measurements reflect Calcium dynamics or other continuous signals). Therefore they require additional processing steps such as taking the logarithm or the square root of spiking activity to approximate a Gaussian distribution better. Further on, these measures assume i.i.d. data which is hardly satisfied in biological observations because of the natural time constants and dependencies of both neural mechanisms (refractory periods, bursting, etc.) and sensory stimuli. Finally, interpreting multiple components of shared variability in terms of their mechanistic source and function is non-trivial since these models merely differentiate them by their strength.

2.7.3 Dynamical probabilistic latent models

Traditionally information about neural computations is extracted through controlled repetitions of experimental trials, with the assumption that response properties are fixed across trials. For instance, when taking the mean and variance of the total spike count in a trial to characterize stimulus responses, we assume constant response properties within and across trials. Or when averaging across trials to apply static dimensionality reduction methods to within trial time points, we assume that across trial variations are irrelevant to the task. However, given the dynamic nature of neural activity and of many decision making tasks, important information may be missed when discarding trial-by-trial varia-

tions. Fig. 2.4A illustrates the potential impact of a slow population drift in a recording session on decision bounds computed by averaging across trials. Unless the slow drift is orthogonal to the decision axis, it will introduce a trial-specific bias. Fig. 2.4B shows varying dynamics across trials along the decision axis, which are lost when either considering only a particular time window, or when averaging across trials.

Probabilistic latent models attribute observed, often high-dimensional, data to a smaller number of hidden ("latent") sources with a probabilistic dependency between observed and latent (see Fig. 2.5) (Roweis and Ghahramani, 1999). They can be combined with stimulus response encoding models as the ones in Chapter 1, and jointly fitted to capture both extrinsic (stimulus-dependent) and intrinsic sources of variability (Archer et al., 2014; Macke et al., 2015). Dynamical probabilistic latent models explicitly take into account the temporal order of the observations within and/or across trials (see Fig. 2.5), by modeling and learning the structure of the temporal dependencies in the low-dimensional latent space (Roweis and Ghahramani, 1999). This allows extracting trial-specific dynamical trajectories from the observations. A common challenge in latent models is the interpretation of the latent space. The time constant of the latents' dynamics can help separate a multidimensional latent into putative separate sources of variability, which ultimately may aid interpretation of the latent and guide the study of underlying mechanism. For example, different types of neuromodulators may modulate activity at different spatial and temporal scales and feedforward versus feedback signals may have different dynamic profiles (Ferguson and Cardin, 2020; Semedo et al., 2022). In the last years a variety of latent dynamical models has been developed. Table 2.1 gives a summary of the modeling choices involved, together with their motivation and interpretation, and provides examples of usage in the literature. Overall this class of models has great potential to shed light on the dynamical modulation of neural responses (Duncker and Sahani, 2021).

question	modeling choice		interpretation	examples
What does the computation look like?	latent noise	continuous	continuous trajectories	SLG, LDS, GP (Macke et al., 2015; Yu et al., 2009)
		discrete	fixed number of states	SLG, HMM (Maboudi et al., 2018)
	latent dynamics	linear	limited dependencies	LDS (Macke et al., 2015)
		nonlinear	smoothness, e.g. periodicity	GP (Wu et al., 2017; Yu et al., 2009; Zhao and Park, 2017)
		Markovian	simple one-step dependency	LDS, HMM
How does each neuron/population reflect the latent?	mapping	linear	coupling strength	SLG, LDS, HMM, GCLDS (Archer et al., 2014; Gao et al., 2015; Macke et al., 2015; Zhao and Park, 2017)
		nonlinear	tuning curves	GP (Wu et al., 2017)
What does the "noise" in the data look like?	observations	Gaussian	continuous, e.g. calcium traces, transformed spiking data	SLG, LDS, HMM, GPFA (Yu et al., 2009)
		Poisson	discrete, e.g. spike counts	PLDS, PGPFA (Macke et al., 2015; Zhao and Park, 2017)
		Bernoulli	binary, e.g. finely binned spikes	GCLDS (Gao et al., 2015)
What are the stimulus response properties?	SR	in latent space	stimulus interacts with intrinsic dynamics	QLDS (Archer et al., 2014)
		in rate space	stimulus and intrinsic dynamics are separate	sPLDS (Macke et al., 2015)

Table 2.1 An overview of variants of latent dynamical models.

2.8 Modulation of V1 responses by a shared stochastic signal

To determine the structure of co-variability in our recorded V1 populations and its modulation across tasks, we fitted a modulated stimulus response model ("modulated-SR model") to the recorded population of V1 neurons in each block, using a Poisson latent dynamical system (PLDS, see Suppl. Sec.2.11.1.2 and Macke et al., 2015. This model jointly estimates the stimulus drive to each unit and the shared, within-trial variability across the population and across stimulus on-off periods (Fig. 2.6A, B). The stimulus response component ("SR model") accounts for stimulus-induced transients across multiple time bins of 50ms, with time-specific parameters for each contrast condition (see Suppl. Sec.2.11.1.1 for details) and independent Poisson noise. The shared, within-trial variability is modeled as a dynamic low dimensional stochastic signal, which multiplicatively modulates the stimulus responses of all simultaneously recorded units, with neuron-specific modulatory coupling strengths. This statistical framework allows us to probe the existence, dimensionality, and time scale of shared modulation in each block of our dataset, in a way that simpler dimensionality reductions methods cannot (Suppl. 2.11.4.4).

We found that 91% of blocks are better fit by the modulated-SR model than by the SR model alone (Fig. 2.6C), suggesting the existence of shared modulation in the V1 population. Moreover, varying the dimensionality of the modulator reveals that 72% of blocks are best described by a one-dimensional modulator (Fig. 2.6D; see Supplement for details). We restricted subsequent analyses to these blocks. The extracted modulator is unrelated to contrast variations in the stimulus (Suppl. 2.11.2.1) and fluctuates within and across trials at a fairly rapid timescale (Fig. 2.6B), with no evidence of oscillatory structure. The average estimated time scale of the fluctuations is 75ms (Fig. 2.6E and Suppl. 2.11.2.2) –

faster than the average trial duration ($\sim$ 3s) as well as the individual stimulus duration (200ms), and in fact approaching the time resolution of the time bins used to model the data (50ms). This fast time scale, together with the unimodal marginal statistics of the estimated modulator (Suppl. 2.11.4.1), differentiate it from previously reported on-off dynamics associated with selective attention (Engel et al., 2016).

2.9 V1 modulator has functional targeting structure

The improvement in fit quality obtained by including the modulator varies across units (Fig. 2.6C), but is most prominent in task-informative neurons (Fig. 2.7A), suggesting that they may be more strongly affected. Within the model, the strength with which each neuron is modulated is governed by its associated modulator coupling weight. A non-parametric comparison revealed that informative neurons have larger coupling weights than uninformative neurons, and thus that the modulation is targeted toward task-informative neurons (Fig. 2.7B). Although informativeness correlates with the mean firing rate of a unit (Suppl. 2.11.3.1), a partial correlation analysis confirmed that firing rate differences cannot explain the inferred modulation targeting, as firing-rate-corrected informativeness and modulator couplings are still significantly correlated in 84% of blocks (Spearman r, $\alpha = 0.05$; Fig. 2.7C-E). The increased variability in the task-relevant neurons (Suppl. 2.4.3) is primarily due to the modulation; Residual variability unexplained by the modulated-SR model is generally not correlated with informativeness (Spearman r with $\alpha = 0.05$; Fig. 2.7E); only 9% of blocks have significantly positive correlations between residual variability and informativeness (19% significantly negative). While most of this residual variability is neuron-specific, we also find weak, structured correlations in pairs of units which suggest additional sources of shared noise not captured by the model (Suppl.

Fig. 2.10).

The modulator coupling is dissociable from traditional attentional effects on mean firing rate which have been suggested to improve encoding precision of particular attended stimuli (coupling and strength of attentional modulation are uncorrelated, Suppl. 2.11.4.2) and it cannot be explained by neural adaptation, as the degree of adaptation was uncorrelated with the quality of the fit of the modulated-SR model (Suppl. 2.11.4.3). Finally, the modulator structure does not arise from the fact that the response measurements are in the form of multi-unit spike counts (Suppl. 2.11.4.5).

2.10 Conclusion

Overall, the analysis reveals that V1 responses are modulated by a common fluctuating signal, and that the strength of this modulation in each unit reflects its task-informativeness. From an encoding perspective, this seems counter-intuitive since these fluctuations decrease the signal-to-noise ratio and on a population-level are most detrimental if directed to informative neurons (Suppl. 3.5.1.2). The next chapter discusses previous proposals for decoding strategies, how they may be affected by such a modulator and introduces a novel theory for how this targeted modulator may actually support flexible decoding. Subsequently, in Chapter 4.4 we explore how the modulator targeting exhibited here may be learned in a network.

2.11 Supplement

2.11.1 Computational model

2.11.1.1 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. The model is fitted exclusively to the stimuli within the RF of the population. 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 target (stimulus 1) response is only used to compute informativeness and for the decoding analysis. The stimulus is characterized by contrast (V1). Orientation is not one of the stimulus dimensions as it does not change during the repeated stimulus presentation (before the target appears). We find that the V1 units do not respond differentially to stimuli of different direction (compare also to Ruff and Cohen, 2016a). Stimuli are parametrized by a one-hot encoding vector at every point in 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 to capture potential delayed effects of the stimulus presentation, and an offset for base firing. The stimulus affects a unit's activity through linear coefficients $\boldsymbol{b_n}$, followed by an exponential nonlinearity that gives a rate of a Poisson process generating spike counts $\mathbf{k}_n$:

$$\boldsymbol{k}_{n,t} \sim \text{Poisson}\left(\exp\left(\boldsymbol{b_n}\mathbf{s}_t\right)\right) \tag{2.1}$$

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

$$L(\boldsymbol{b_n}) = \sum_t -(\boldsymbol{b_n}\mathbf{s}_t)^T\boldsymbol{k}_{n,t} + \exp\left(\mathbf{1}^T\boldsymbol{b_n}\mathbf{s}_t\right) + \alpha\boldsymbol{b_n}^T\boldsymbol{b_n} \tag{2.2}$$

The model was fitted to the trials of each block separately. The modulator statistics and targeting structure are therefore assumed to be constant across a block (data-limitations do not allows us to fit model statistics on a trial by trial basis). 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 stimuli.

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 (Benjamin et al., 2017) which gives "the fraction of the maximum potential log-likelihood gain (relative to the null model) achieved by the tested model" $\frac{\log L(\hat{y}) - \log L(\bar{y})}{\log L(y) - \log L(\bar{y})}$, where $\hat{y}$ is the estimation of the hypothesized model and $\bar{y}$ is the null model (constant-rate model that only fits the mean firing rate of each neuron).

2.11.1.2 Modulated SR model

Building up on the SR model, we look for population-wide low-dimensional modulator terms $\mathbf{m}$ that vary stimulus response both within and across trials. We use the frame-

work of Poisson Linear Dynamical Systems (PLDS, (Macke et al., 2015; Rabinowitz et al., 2015)) 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 $\mathbf{C}$ (equivalent in meaning to the coupling $\mathbf{c}$ in the theory). This joint model has the form:

$$\mathbf{k}_t \sim \text{Poisson}(\exp(\mathbf{C}\mathbf{m}_t + \mathbf{B}\mathbf{s}_t)) \tag{2.3}$$

$$\begin{aligned} \mathbf{m}_{t+1} &= \mathbf{A}\mathbf{m}_t + \epsilon \\ \epsilon &\sim \mathcal{N}(0, \mathbf{Q}) \\ \mathbf{m}_0 &\sim \mathcal{N}(0, \mathbf{Q}_0) \end{aligned} \tag{2.4}$$

where the modulator at time t, $\mathbf{m}_t$, is D-dimensional and the mapping $\mathbf{C}$ is N by D, with latent dimensionality $D \ll N$. t denotes time points within a trial across both stimulus presentation and inter-stimulus windows. Parameter $\mathbf{A}$ implicitly defines the modulator's time constant (see Suppl. 2.11.2.2 for conversion to seconds) 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 (Macke et al., 2015). 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 different data cross-folds.

Model Validation/Comparison: The same criteria are used to evaluate the modulated-SR model as described above for the SR model with one adjustment. Com-

puting 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 (Yu et al., 2009) 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.

2.11.2 V1 modulation

2.11.2.1 Modulator statistics

We tested whether the extracted V1 modulator could be exclusively caused by stimulus variations that are not sufficiently captured by our linear stimulus-response model. For this we look for variations in the statistics of the modulator with different stimuli. We computed the mean value of the modulator over all stimulus presentations in a block for low and high contrast stimuli and find that the distribution of modulator values is similar for different contrast conditions (Fig. 2.8A). We then compute modulator mean and variance respectively and separately for each of the four time windows of 50ms for which the stimulus was present. The mean of the modulator does not fluctuate much within stimulus presentations (Fig. 2.8B, top). Moreover, the differences between high and low contrast are small. The same applies to the estimated modulator variance which is similar within

different time bins of the stimulus presentations and across the two contrast conditions (Fig. 2.8B, bottom). Overall, this suggests that the modulator is unlikely to be merely accounting for residual stimulus responses that were not captured by the SR model, but is instead consistent with a stationary fluctuating modulatory source that is independent of the stimulus.

2.11.2.2 Estimating the time constant of the modulator

Given the maximum likelihood estimated parameters of the PLDS model, the latent dynamics of the modulator are given as $m_t = Am_{t-1} + \epsilon_t$. Taking the noise variance to zero gives: $m_t = m_0 A^t$, so that $\tau = -\frac{1}{\log(A)}$.

2.11.3 Modulator targeting

In Fig. 2.7B we compute the rank of each unit's modulator coupling in its own block-specific population and compare the distribution of significantly informative to uninformative units.

2.11.3.1 Partial correlation analysis for mean rate, coupling and informativeness

In Fig. 2.7C-E we used partial correlation to account for effects of mean firing rate on the relationship between coupling and informativeness. A linear regression lets us average out the linear effects of mean firing on informativeness (Fig. 2.9A) which explains most of the relationship between the variables (Fig. 2.9B), and gives us the unexplained, residual informativeness. We then compute the Spearman correlation coefficient between residual in-

formativeness (not explained by mean firing rate differences) and modulator coupling. We find that the relationship between informativeness and coupling is preserved (Fig. 2.9C-D). 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.

2.11.3.2 Excess correlations

The modulated SR model captures part of the shared variability between neurons. To test for additional structure, we look at the pairwise correlations expected from the modulated SR model versus those in the data. We find that the pairwise correlations are slightly larger in the real data than what would be expected due to differences in the response statistics (when simulated from the modulated SR model). More specifically, we find that those excess correlations are, on average, slightly larger between pairs of informative units than pairs of uninformative or informative-uninformative units. This suggests that there is additional structure, not captured by the modulator, which may be related to other sources of noise correlations (Kanitscheider et al., 2015b; Moreno-Bote et al., 2014).

2.11.4 Controls and comparisons

2.11.4.1 Comparison to On/Off states

The modulator extracted from our data has the form of normally distributed noise (see Fig. 2.8 and Fig. 2.11A-B). Other types of modulation have previously been described in the literature (Engel et al., 2016), with different statistics. Specifically On/Off states

modulate population activity in a binary manner, in the context of attention. We verify that our analysis is able to differentiate between these two different forms of modulation. Since the PLDS framework used for the modulated SR model assumes a Gaussian prior noise distribution (see details in Sec. 2.11.1.2), it is conceivable that the unimodality of the extracted modulator may be a consequence of this prior. In order to understand whether the influence of the prior could be strong enough to conceal true binary modulation, we fit the PLDS model to simulated data that has a binary ground truth modulator. We match the simulation statistics to our recordings, by using parameters estimated in an example session, but change the modulator statistics to reflect two discrete states. We simulate data from the resulting model and then repeat the same fitting procedure as that used on the real data. We find that the estimated modulation is strongly bimodal and reflects the binary modulator well, despite the model's prior (see Fig. 2.11C-D). We further find that the spiking patterns are qualitatively different in the simulations with the binary modulator, than in the real data, as they visually reflect the two different spiking regimes (see Fig. 2.11A-B).

2.11.4.2 Response modulation due to attention

It has previously been reported that neurons with receptive fields overlapping an attended stimulus increase their activity (Maunsell and Cook, 2002; Ruff and Cohen, 2014; Treue and Martínez Trujillo, 1999). However, here we do not find evidence that the increase in activity due to attention or task condition is specific to the task-informative units. Specifically, the correlation between attentional modulation and task-informativeness is close to 0 in the populations we were able to compare (sufficient trials and sufficient number of informative neurons, see main text for exclusion criteria) (Fig. 2.12A). The attentional

modulation in a single neuron is measured as the difference between response to high contrast stimuli in the relevant task condition (attend-in) minus in the control task condition (attend-out) divided by the sum of the two (Ruff and Cohen, 2016a). Similarly there is no significant correlation in the modulator coupling strength and the attention index in any of the blocks we were able to compare (3 pairs of blocks in the same session with good modulator fit; Fig. 2.12B).

2.11.4.3 Adaptation

Within a trial, stimuli are flashed on (for 200ms) and off (for 200-400ms) at low or high contrast several times before the target appears. We exclude the first stimulus presentation from our analysis, to avoid effects of initial transients or adaption (Cohen and Maunsell, 2009). Nonetheless, we wanted to test whether adaptation could interfere with our estimates of the informativeness of a unit or the modulated SR model fits. For this we define a summary statistic for adaptation in single units as follows: we group the last repeat stimulus responses from all trials by the number of repeats that preceded them. We compute the average of each group and the sign of the differences between them. We average over the signs to obtain a value between -1 (decreasing in response with increasing number of repeats) and 1 (increase in response with increasing number of repeats). Fig. 2.13A shows the distribution of adaptation indices over all blocks and units, compared to that of the subset of blocks well fitted by the modulated SR model. This distribution is broad overall, with no significant differences between populations that are well fitted by a model including modulation versus units that are badly fit by the model. Similarly, we see no systematic relationship between the adaptation index and the informativeness of a unit (Fig. 2.13B). Overall, these results suggest that classic adaptation cannot trivially explain

the effects of modulation or its targeting towards informative neurons.

2.11.4.4 PCA is insufficient to robustly find targeted modulation

We ask whether a simpler analysis based on dimensionality reduction (principal component analysis, PCA) of the fitted SR model residuals would suffice for robust modulator estimation in the V1 data. We find that neither the eigenspectrum of the data (Fig. 2.14A, C), nor that of the residuals (Fig. 2.14B, D) reveal low-dimensional structure. Nonetheless, the first principal component roughly aligns with the projection from latent space, $\mathbf{C}$ estimated by the modulated SR model (Fig. 2.14E) providing a noisy version of the estimated modulator coupling (Fig. 2.14F). We further find that PCA results are highly dependent on the stimulus condition. Depending on whether PCA residuals are computed on only high contrast stimulus residuals or on any contrast stimulus residuals, the resulting PC axis can vary substantially (Fig. 2.15A, residuals computed from the SR model). Interestingly we also see that the variance explained by the first PC axis is smaller in the control task vs the relevant task if any stimuli are included, but larger if only high contrast stimuli are included (Fig. 2.15B-C). This suggests that standard dimensionality reduction of residuals might not be sensitive enough to detect low dimensional modulator structure. Taking the square-root of the activity before applying PCA (a common preprocessing step for homogenizing the data variability (Yu et al., 2009)), does not change the results qualitatively. Instead explicit latent dynamical models jointly fitted with a SR model are required to detect the modulator. This suggests that such low-dimensional targeted modulation may be more ubiquitous than one would expect from previously reported analyses.

2.11.4.5 Multiunits

Many of the responses in the analyzed dataset are likely a composite of multiple neurons, and we wanted to examine the effects this could have on our results. We use mathematical analyses and numerical simulations, both based on a simple encoding model (2.16A), to ask how estimates of informativeness and targeted modulation are affected by pooling together the activity of several (potentially similarly tuned) neurons.

Estimating informativeness from multiunits. We model multiunit activity as the sum of activities of pairs of neurons, i,j, each described as a Poisson processes with spike count distribution $\sim$ Poisson $\left(\lambda_{i,j|s} \exp(c_{i,j} m_t)\right)$, as in the main text theory (Eq. 2.1). The informativeness of the multiunit may be written as:

$$d' = \frac{\left(\mu_{i|s_1} + \mu_{j|s_1}\right) - \left(\mu_{i|s_2} + \mu_{j|s_2}\right)}{\sqrt{\frac{1}{2}\left(\left(\sigma^2_{i|s_1} + \sigma^2_{j|s_1} + 2\mathrm{Cov}_{i,j|s_1}\right) + \left(\sigma^2_{i|s_2} + \sigma^2_{j|s_2} + 2\mathrm{Cov}_{i,j|s_2}\right)\right)}}, \tag{2.5}$$

where we have used the standard relationship for the variance of a sum of correlated variables.

Given the doubly stochastic nature of the single neuron responses (formally, a log-Gaussian Cox process, see Snyder and Miller 2012), the average firing rate of neuron i in response to a stimulus s is:

$$\mu_{i|s} = \mathbb{E}\left[\lambda_{i|s} \exp(c_i m_t)\right] = \lambda_{i|s} \exp\left(\frac{c_i^2 \sigma_m^2}{2}\right), \tag{2.6}$$

where σ_m^2 is the variance of the modulator. The corresponding variance is

$$\begin{aligned}\sigma^2_{i|s} &= \mathbb{E}\left[\lambda_{i|s}\exp(c_i m_t) + \lambda^2_{i|s}\exp(2c_i m_t)\right] - \mu^2_{i|s} \\ &= \lambda_{i|s}\exp\left(\frac{c_i^2\sigma_m^2}{2}\right) + \lambda^2_{i|2}\exp\left(2c_n^2\sigma_m^2\right) - \lambda^2_{i|s}\exp\left(c_i^2\sigma_m^2\right).\end{aligned} \tag{2.7}$$

Finally, the covariance of responses for neurons i and j given a stimulus s is

$$\begin{aligned}\mathrm{Cov}_{i,j|s} &= \mathbb{E}\left[\lambda_{i|s}\lambda_{j|s}\exp\left((c_i+c_j)m_t\right)\right] - \mu_{i|s}\mu_{j|s} \\ &= \lambda_i\lambda_j\exp\left(\frac{(c_i^2+c_j^2)\sigma^2}{2}\right)\left(\exp\left(c_i c_j\sigma^2\right) - 1\right).\end{aligned} \tag{2.8}$$

In the limit when the modulator variance is very small ($\sigma_m^2 \to 0$), we have $\mu_{i|s} \to \lambda_{i|s}$, $\sigma^2_{i|s} \to \lambda_{i|s}$ and $\mathrm{Cov}_{i,j|s} \to 0$, so the informativeness becomes:

$$d' = \frac{\lambda_{i|s_1} + \lambda_{j|s_1} - \lambda_{i|s_2} - \lambda_{j|s_2}}{\sqrt{\frac{1}{2}\left(\lambda_{i|s_1} + \lambda_{j|s_1} + \lambda_{i|s_2} + \lambda_{j|s_2}\right)}}. \tag{2.9}$$

We will now show that, in this limit, and assuming the neurons in the multiunit have the same stimulus preference, the absolute value of the multiunit informativeness is bounded from above by the sum of that of the two component neurons, i.e.

$$\frac{|\lambda_{i|s_1} + \lambda_{j|s_1} - \lambda_{i|s_2} - \lambda_{j|s_2}|}{\sqrt{\frac{1}{2}\left(\lambda_{i|s_1} + \lambda_{j|s_1} + \lambda_{i|s_2} + \lambda_{j|s_2}\right)}} \leq \frac{|\lambda_{i|s_1} - \lambda_{i|s_2}|}{\sqrt{\frac{1}{2}\left(\lambda_{i|s_1} + \lambda_{i|s_2}\right)}} + \frac{|\lambda_{j|s_1} - \lambda_{j|s_2}|}{\sqrt{\frac{1}{2}\left(\lambda_{j|s_1} + \lambda_{j|s_2}\right)}} \tag{2.10}$$

To simplify notation in the proof, we first introduce variables for the sum and difference responses, $\gamma_{i/j} = \lambda_{i/j|s_1} + \lambda_{i/j|s_2}$ and $\beta_{i/j} = \lambda_{i/j|s_1} - \lambda_{i/j|s_2}$, so that the inequality above

becomes:

$$\frac{|\beta_i + \beta_j|}{\sqrt{\gamma_i + \gamma_j}} \leq \frac{|\beta_i|}{\sqrt{\gamma_i}} + \frac{|\beta_j|}{\sqrt{\gamma_j}}.$$

Multiplying by all denominators yields

$$|\beta_i + \beta_j|\sqrt{\gamma_i\gamma_j} \leq |\beta_i|\sqrt{\gamma_j(\gamma_i + \gamma_j)} + |\beta_j|\sqrt{\gamma_i(\gamma_i + \gamma_j)}.$$

Under the assumption that the two neurons in the multiunit have the same stimulus preference, $|\beta_i + \beta_j| = |\beta_i| + |\beta_j|$, and one can rearrange the terms as

$$|\beta_i|\left(\sqrt{\gamma_i\gamma_j + \gamma_j^2} - \sqrt{\gamma_i\gamma_j}\right) + |\beta_j|\left(\sqrt{\gamma_i\gamma_j + \gamma_i^2} - \sqrt{\gamma_i\gamma_j}\right) \geq 0.$$

Hence, we conclude that in the limit when the modulation is weak and neurons share the same stimulus preference (and thus same sign for optimal decoding weights) the informativeness of a multiunit is upper bounded by that of its component units.

Using a similar derivation, one can also show that the informativeness of a multiunit is lower bounded by the average of the informativeness of the two neurons that compose it:

$$\frac{|\lambda_{i|s_1} + \lambda_{j|s_1} - \lambda_{i|s_2} - \lambda_{j|s_2}|}{\sqrt{\frac{1}{2}\left(\lambda_{i|s_1} + \lambda_{j|s_1} + \lambda_{i|s_2} + \lambda_{j|s_2}\right)}} \geq \frac{1}{2}\left(\frac{|\lambda_{i|s_1} - \lambda_{i|s_2}|}{\sqrt{\frac{1}{2}\left(\lambda_{i|s_1} + \lambda_{i|s_2}\right)}} + \frac{|\lambda_{j|s_1} - \lambda_{j|s_2}|}{\sqrt{\frac{1}{2}\left(\lambda_{j|s_1} + \lambda_{j|s_2}\right)}}\right) \tag{2.11}$$

While these constraints are derived under extreme assumptions, we can show numerically that the same intuition holds in the more relevant scenario when the modulation is not

negligible (Fig. 2.16). For moderate modulation strength, $\sigma_m = 1$ and $w = [0, 1]$, and random targeting (w coupling assigned uniformly randomly to neurons with different tuning properties), the summed informativeness, $|d'_i + d'_j|$, upper bounds the multiunit informativeness for both pairs of neurons with aligned stimulus preference and pairs of neurons with dissimilar tuning (Fig. 2.16B). Note that tuning similarity, as measured by the inner product of the individual units' tuning functions, does not strongly affect multiunit informativeness (Fig. 2.16C).

Finally, we find the process of pooling neural responses alone cannot induce positive correlations between modulation and informativeness. The effect of modulation on a multiunit, as estimated via the modulator-guided decoder, takes the form $\mathbb{E}\left[(k_i + k_j)m\right] = \sigma_m^2 \left(\lambda_i c_i \exp\left\{\frac{\sigma_m^2 c_i^2}{2}\right\} + \lambda_j c_j \exp\left\{\frac{\sigma_m^2 c_j^2}{2}\right\}\right)$. This estimated modulator coupling does not show targeting towards informative multiunits if the single unit coupling is unrelated to the single unit informativeness. When the modulator is targeted, some of this structure is preserved on the multiunit level (Fig. 2.16D). This is why using our modulator-guided decoder on the V1 data still performs well. Overall, these results suggest that analyzing multiunits underestimates the true informativeness of the underlying neurons. In itself it does not induce dependencies between informativeness and modulation, as used by the decoder proposed in the theory.

Impact of multiunits on the estimation of modulator targeting. Last, we wanted to understand how the presence of multiunits impacted the fitting and interpretation of our model. The following analysis was performed for simulations of either single or multiunits and the results regarding targeting structure were equivalent. Only the multiunit scenario is reported as its results include the single unit case.

To investigate the impact of multiunits on the model fitting procedure, we use the same modulated SR model that pools pairs of neurons, with various degrees of informativeness. We simulate a population of multi-units with data-matched statistics (in terms of number of units and firing rate distribution) including pairs of neurons that are either coupled to the global modulator or not (Fig. 2.16A). We consider two targeting scenarios: 1) preferential targeting towards task informative neurons, as hypothesized in the theory (Fig. 2.17A-D) and 2) random targeting, with coupling strengths that are independent of neuron informativeness (Fig. 2.17E-H). We apply the data analysis pipeline used on the experimental data to the resulting artificial datasets to fit a modulator and assess the properties of the corresponding PLDS estimated coupling strengths.

In the first scenario, we have a diversity of degrees of informativeness for the single units (Fig. 2.17A) and strong correlations between single unit modulation coupling and informativeness (Fig. 2.17B). The corresponding estimated multiunit informativeness remains upper bounded by the sum of the informativeness of the component neurons, as before. The correlation between modulator coupling and informativeness is weaker than for single units, but follows the same trend (Fig. 2.17D). This suggests that if targeting is presented in the single neurons, then analysis of multiunit measurements is likely to underestimate the degree of targeting.

The picture is quite different when no targeting is enforced (specifically, modulator couplings are random and independent). Since higher modulator coupling introduces noise and decreases a neuron's informativeness, random targeting leads to an anti-correlation between informativeness and coupling strength (Fig. 2.17F). The sum of single unit d' values remains an upper bound for multiunit informativeness, with less variability than in the targeted scenario (Fig. 2.17G). Importantly, there are no spurious positive correlations be-

tween the estimated multiunit coupling and informativeness when structured targeting is not present in the single units (Fig. 2.17H). Hence, the presence of multiunits alone cannot explain the modulator targeting we find in the data.

In the V1 data, different multiunits are likely to have different numbers of neurons driving their responses. This variability will further bias estimates of modulator strength towards larger multiunits. The opposite effect occurs with informativeness: larger multiunits will in general have more diverse responses, reducing their estimated informativeness. Thus, variability in multiunit size induces *negative* correlations between modulation strength and informativeness. In summary, it is likely that our empirical estimates based on the data underestimate the degree of targeting V1 neurons.

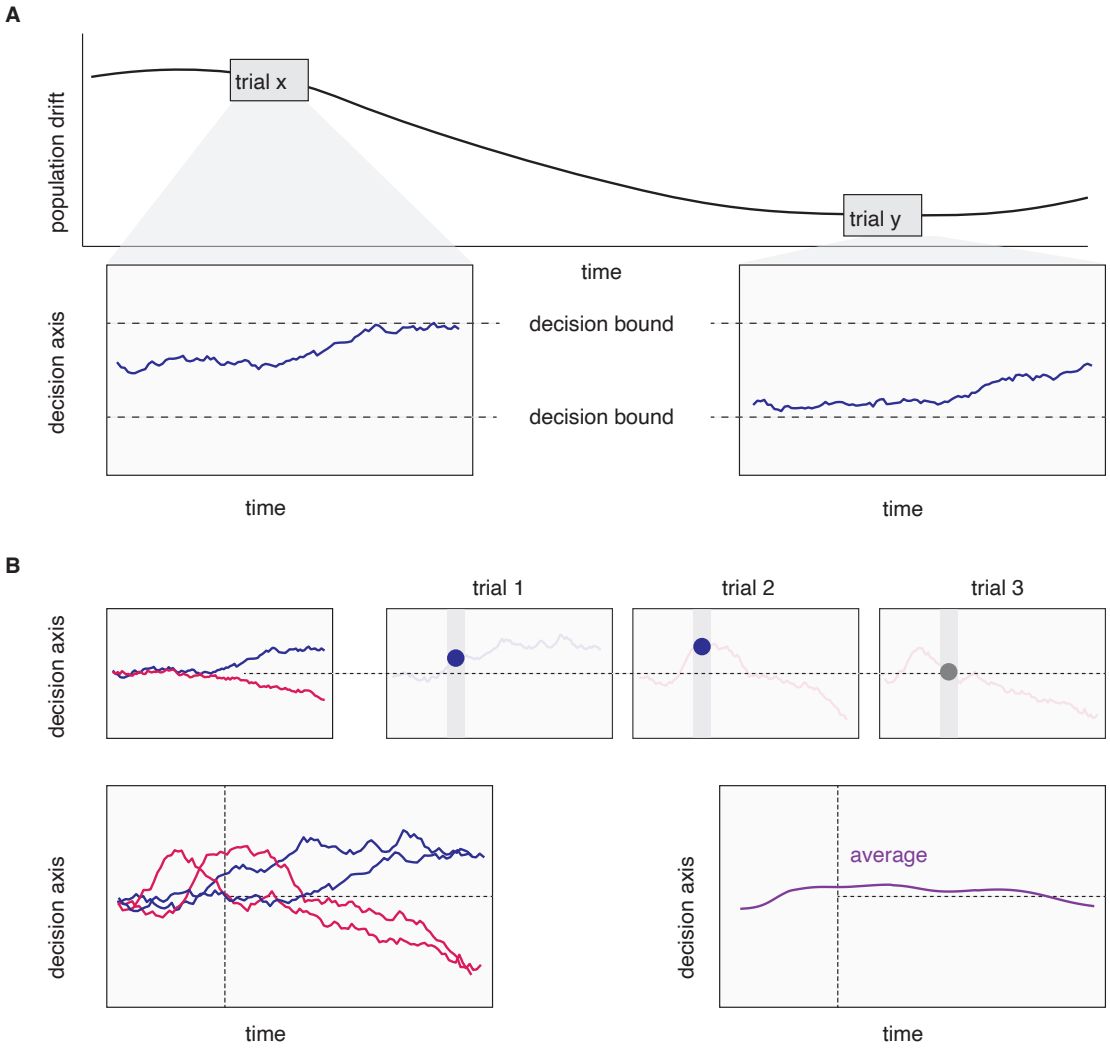

Figure 2.4 Effects of across and within trial averaging. A) Slow drift across trials (top) influences the population activity and can bias a read out decision axis if drift and readout are not orthogonal. Bottom plots show two example trials x and y where the population activity is projected on a task-specific decision axis and propagates to one of two decision bounds. B) Differences in within-trial trajectories of neural population activity projected on a decision axis. Taking a snapshot of the evolving activity at a particular time bin (shaded region and dot) in the trial and inferring the subject's decision in a binary discrimination task based on a threshold (dotted line), may obscure important dynamics that differ throughout the trial (indicated by colored lines, where color indicates final decision). On the other hand, considering many time bins within a trial (bottom left) but averaging across trial-specific trajectories (bottom right) leads to a flat mean estimate which falsely suggests the absence of decision making dynamics.

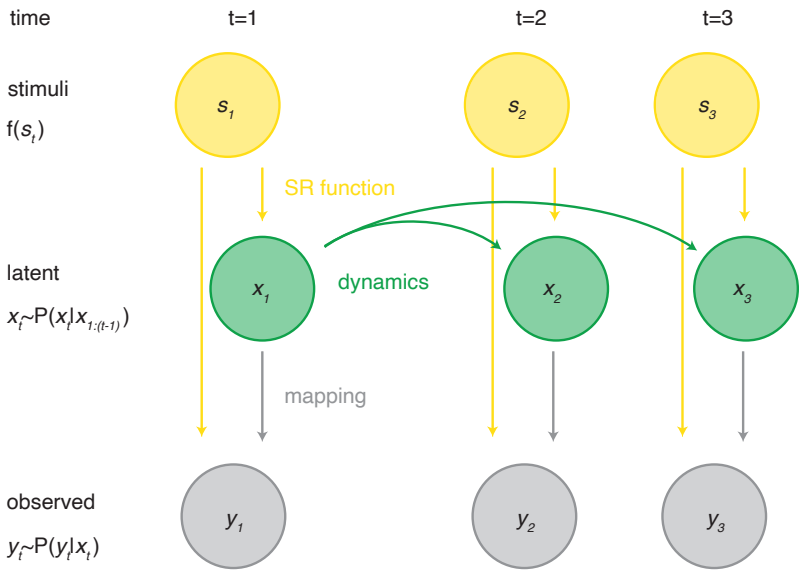

Figure 2.5 An illustration of the main components of latent dynamical models. A low-dimensional latent variable $\mathbf{x}_t$ varies in time t with a certain probability distribution $P(x_t)$ and temporal dependencies $x_t|x_{1:(t-1)}$ that express the latent dynamics (green arrows). A mapping function defines how the latent influences the higher dimensional observed variable $\mathbf{y}_t$ (grey arrows). The stimulus may influence the latent and/or the observed variable directly via a stimulus response (SR) function (yellow arrows).

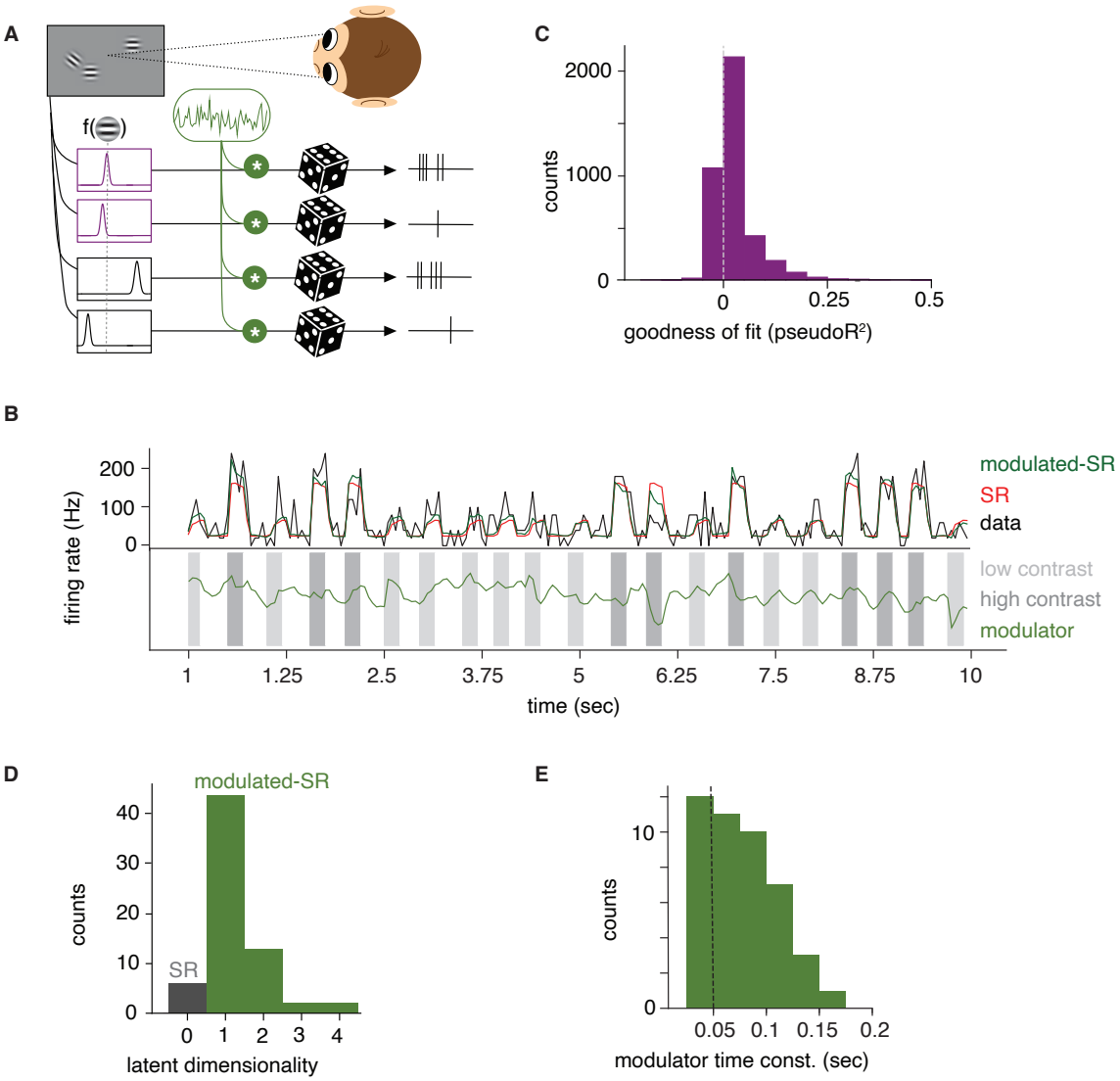

Figure 2.6 Estimating the modulator in the recorded V1 population. A) An illustration of the modulated stimulus response model: Each neuron's tuning function specifies its base response to a stimulus; this rate is modulated by a time-varying shared source of multiplicative noise (green), with spiking 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) The distribution of pseudo-R^2 values over all neurons in blocks that were best fitted by a 1-dimensional modulated-SR model. D) Summary for the dimensionality of best fitted models across relevant tasks (see Methods for Details). E) The distribution of estimated time constants over all blocks that were best fitted by a 1-dimensional modulated-SR model.

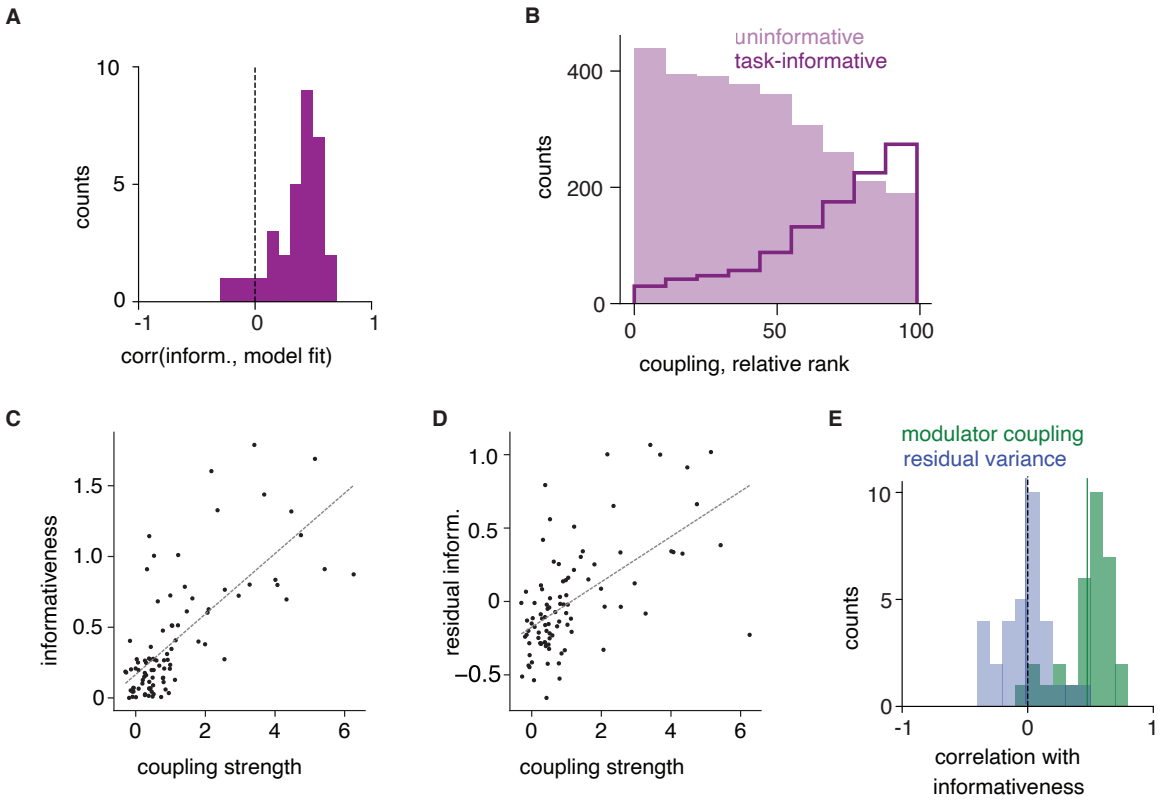

Figure 2.7 The targeting structure of the modulator reflects the current task. A) Distribution of the correlations between the individual unit's model fit (pseudo-R^2) and their informativeness. (78% of blocks have significant positive correlations between informativeness and model fit, Spearman r, $p < 0.05$) B) Relative population rank of modulator coupling strength for significantly informative (dark purple line) and uninformative (light purple shading) neurons. The rank is computed for each block-specific model, then rank values are pooled across blocks for each population respectively (see Suppl. Sec. 2.11.3). C) Informativeness over coupling strength in an example block's model fit. D) Residual informativeness (unexplained by linear effects of mean firing) over coupling strength in same example as H. E) Partial correlation analysis assessing the dependence between informativeness and modulation strength, after controlling for differences in firing rates. Distribution of correlation coefficients obtained by partial correlation analysis across blocks (green, 84% of blocks significant Spearman r) and a similarly obtained distribution that uses the modulated-SR model residual response variance as a proxy for neuron individual variance and instead of modulator coupling (blue).

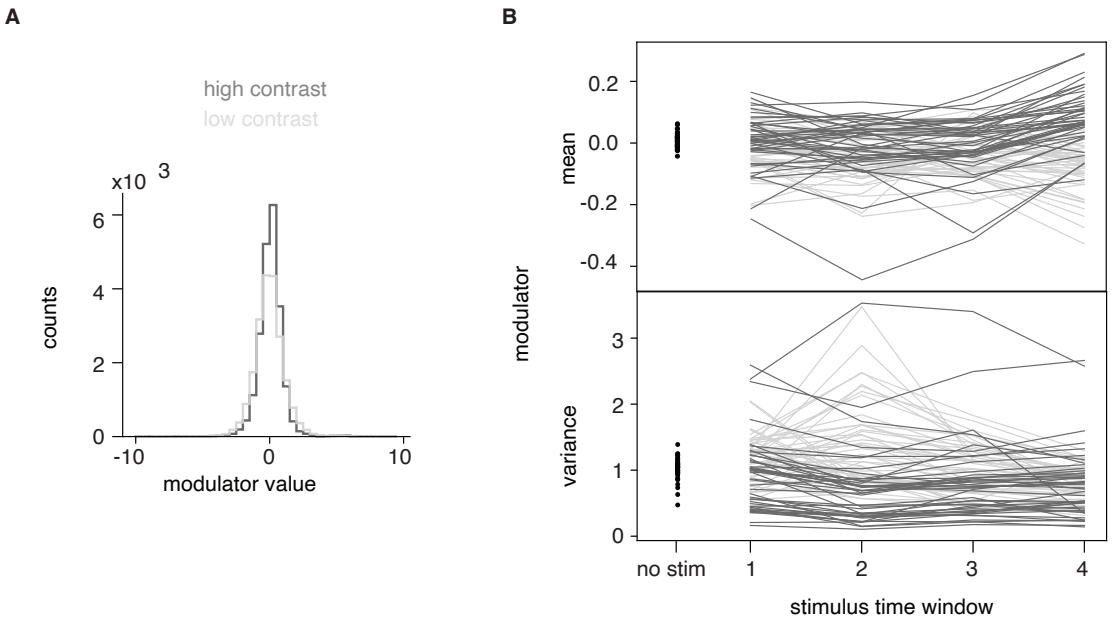

Figure 2.8 Inferred modulator statistics. A) The distribution of modulator values during high/low contrast stimulus presentations. B) Every line represents the mean (top) and variance (bottom) of the modulator in a block, estimated from the different time bins of a stimulus presentation and for low and high contrast. Dark grey lines represent high contrast and light grey low contrast presentations.

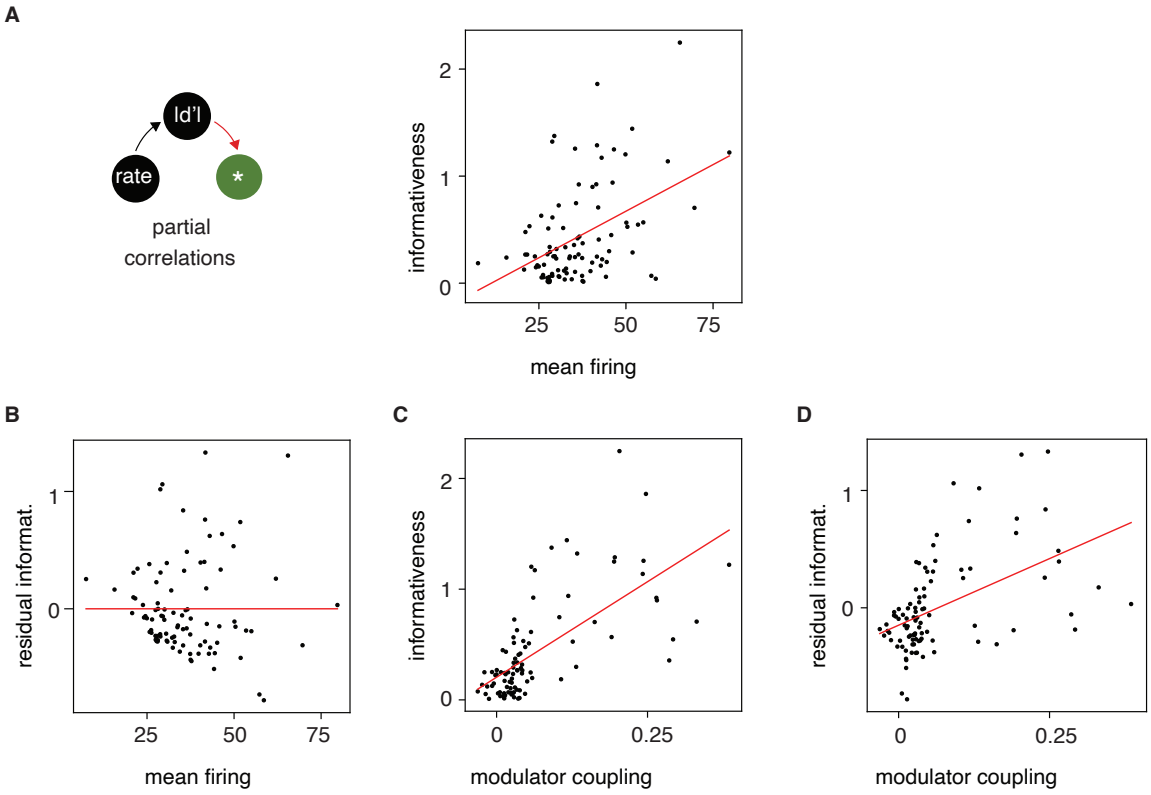

Figure 2.9 Partial correlation analysis for mean rate, coupling and informativeness. A) Dependence between |d'| and mean firing. B) Residuals of linear fit as a function of firing rate are unstructured. C) The relationship between informativeness and coupling. D) The same for residual informativeness (unexplained by differences in mean firing).

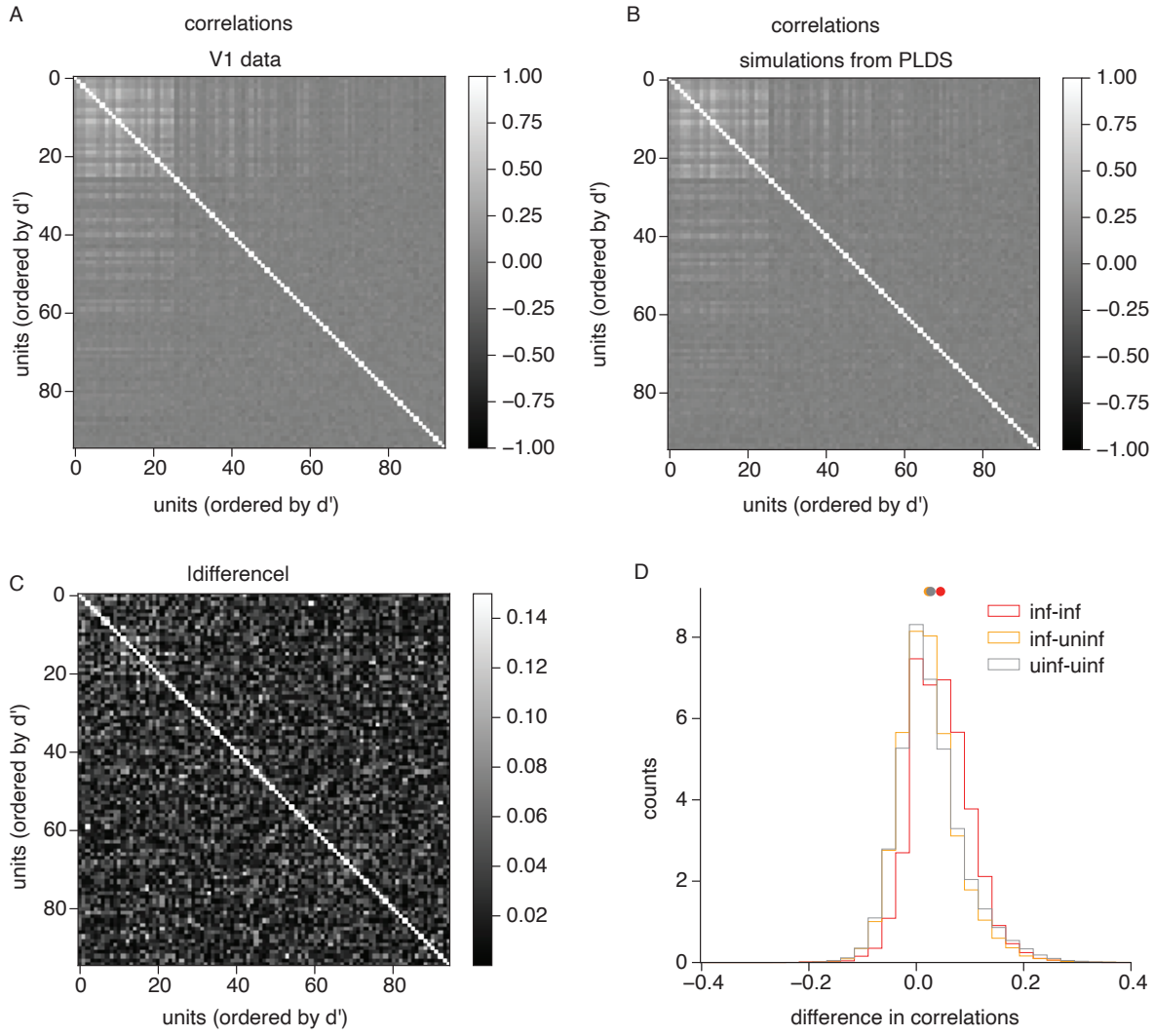

Figure 2.10 Excess noise correlations. A) Pairwise noise correlations of a population in an example block computed on high contrast stimulus presentations. The color bar indicates Pearson correlation coefficient. B) Pairwise noise correlations in simulations from the same example modulated SR model. C) Difference between pairwise noise correlations in data (A) and in simulations (B). Colors indicate difference in Pearson correlation coefficient. D) Distribution of differences in pairwise correlation coefficients over all blocks. Colors indicate the type of pairs; Red for two informative units, yellow for one informative and one uninformative units, grey for two uninformative units. Points indicate the mean of the respective distribution.

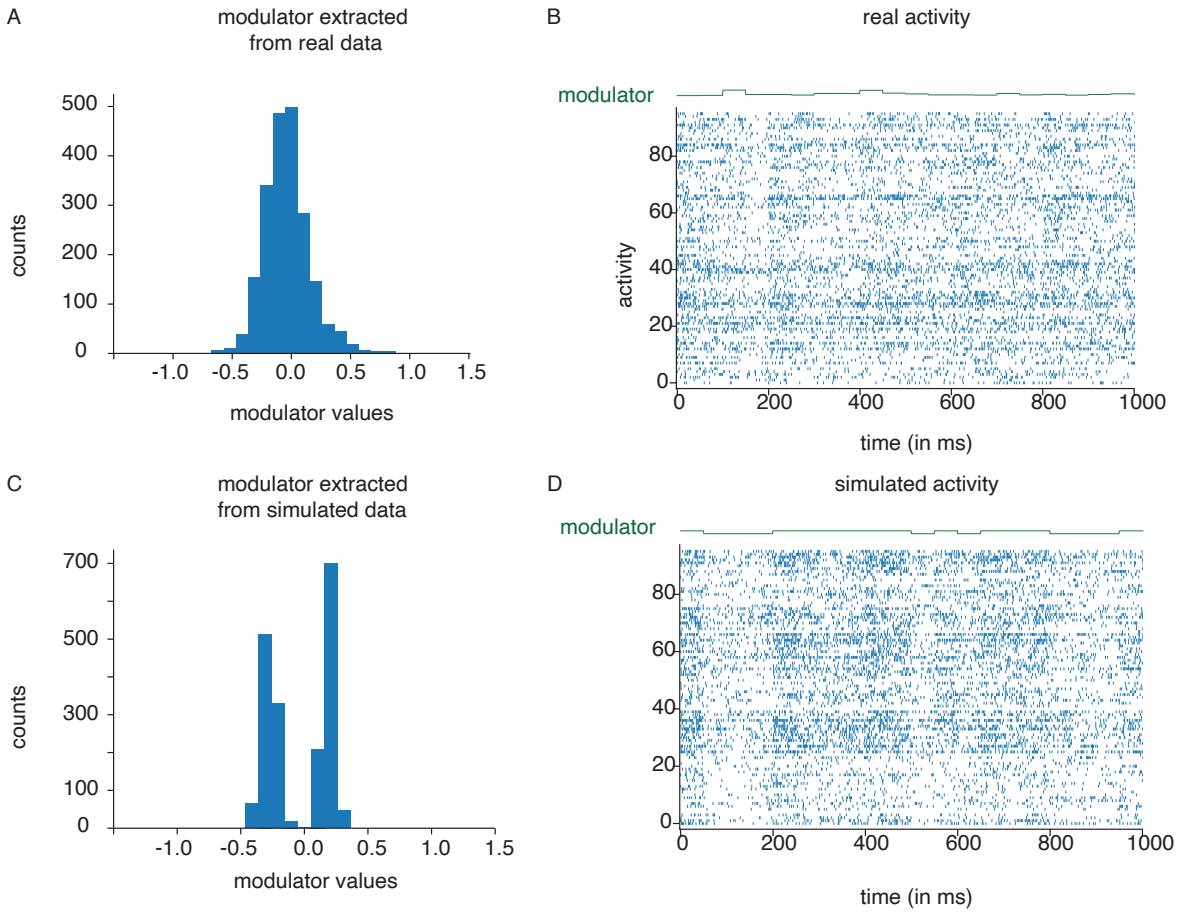

Figure 2.11 On/Off states A) Modulator distribution extracted from an example block population. B) Population spiking activity for for one second of that same example block. C) Modulator distribution extracted from simulations from the same model fit but using an artificial bimodal modulator instead. D) Spiking activity from an example second simulated from that model.

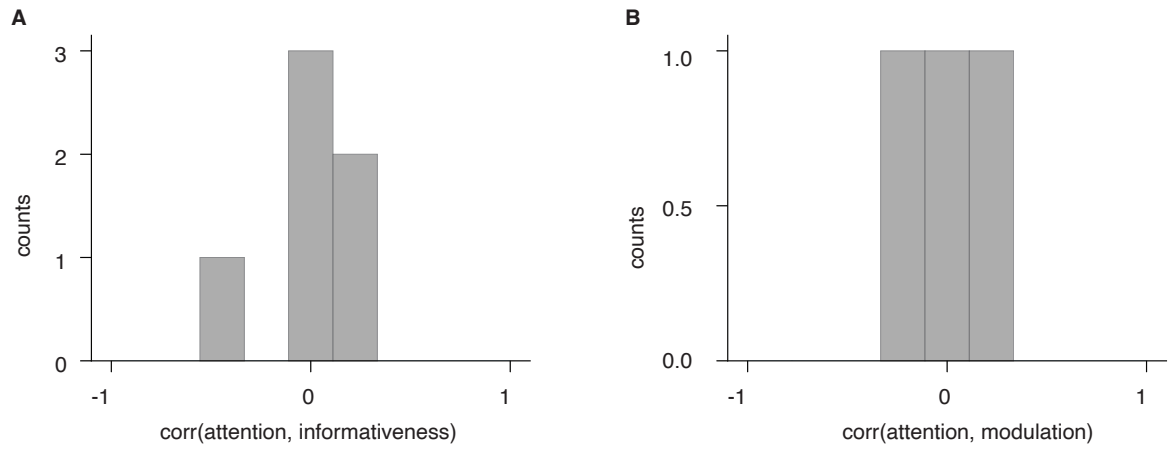

Figure 2.12 Distribution of Spearman correlation coefficients over blocks between A) attentional modulation index and informativeness and B) between attentional modulation index and modulator coupling.

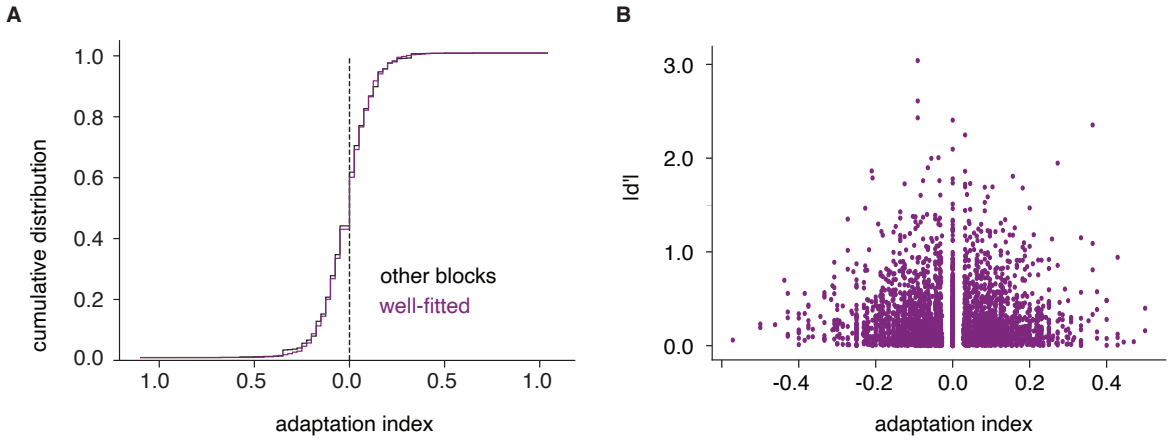

Figure 2.13 Effects of adaptation. A) The distribution of adaptation indices for blocks well fit by mod-SR (purple) and all blocks (black). B) Informativeness of each unit, measured by $|d'|$ as a function of the unit's adaptation index; no linear dependency, Spearman $p = .36$

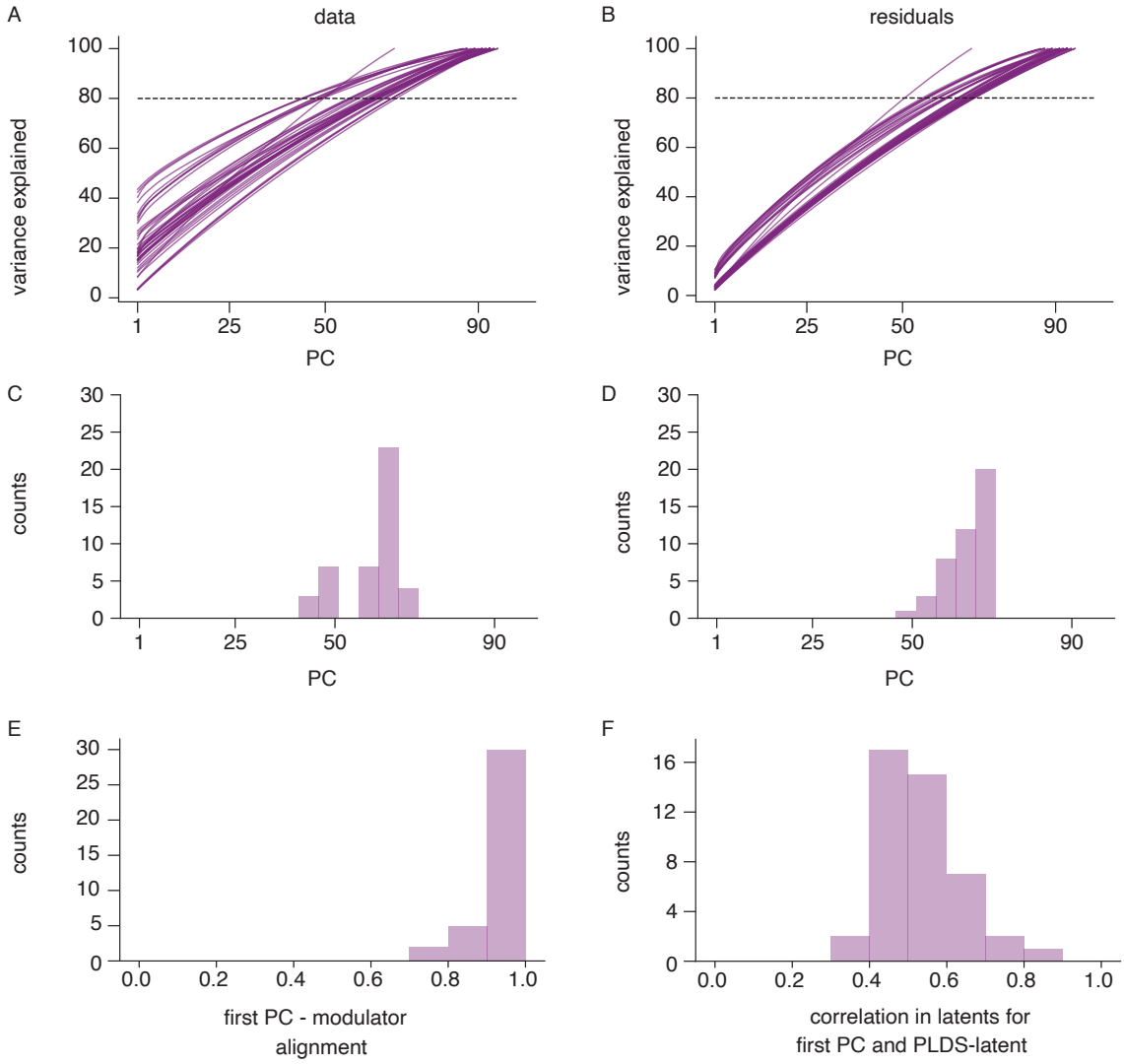

Figure 2.14 PCA analysis of V1 population activity. Variance explained by principle components of A) population responses and B) SR residuals; we apply PCA to the concatenated trials of a block, each datapoint is the activity of a unit in a 50ms time bin. C) A histogram of the minimum number of PCs required to explain 80% of the variance in the data and D) SR residuals. E) We take the cosine similarity between the eigenvector corresponding to the first PC of the residuals and the modulator coupling for every block. Here we plot the distribution over all blocks. F) We compute the correlation between the trial-concatenated modulator found by the PLDS and by PCA on the residuals. We plot the distribution of correlation coefficients over all blocks.

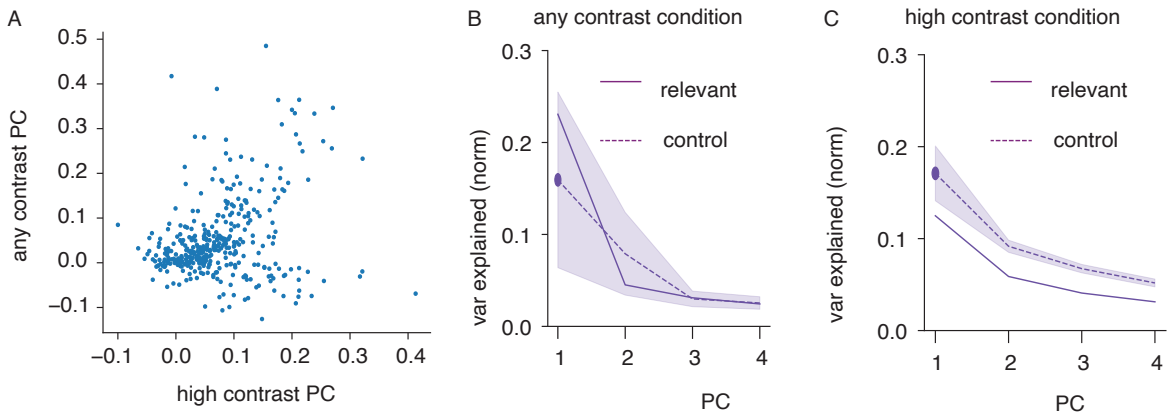

Figure 2.15 Dependency of PCA analysis on stimulus contrast. A) First PC computed on all stimulus presentations in the control task, over first PC computed on high contrast stimulus presentations. Each point represents a unit's loading on the PC axis. The graph pools over all control task blocks. If a population's first PC had a negative mean (mean of first eigenvector $<$ 0) the entire vector is rotated by -1 to increase visual comparability. B) Normalized variance explained for first 4 PC axis extracted from residuals of any stimulus presentations using the SR model fitted, for either the relevant or the control tasks. Lines show averages over all blocks and shaded region shows the sd. C) As B but for high contrast stimulus presentations.

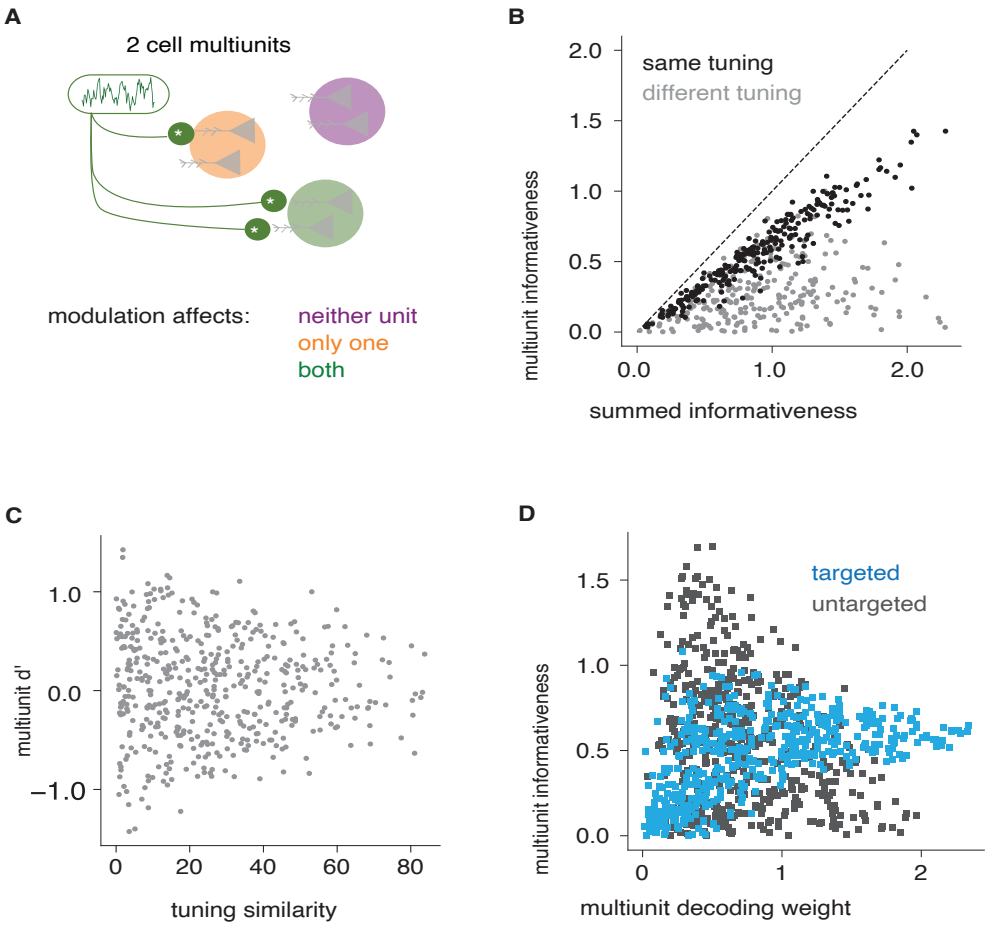

Figure 2.16 Effect of multiunits on key analysis measurements. A) We model multiunits by summing over the activity of two model Poisson neurons with modulated rates. B) Informativeness of multiunit $|d'_{(i,j)}|$ versus the sum of informativeness of the component units, $|d'_i| + |d'_j|$. C) Multiunit d' as a function of the cosine similarity between the tuning of the individual neurons. D) Multiunit informativeness as a function of estimated multiunit modulator-guided decoding weights for a simulated targeted and untargeted population.

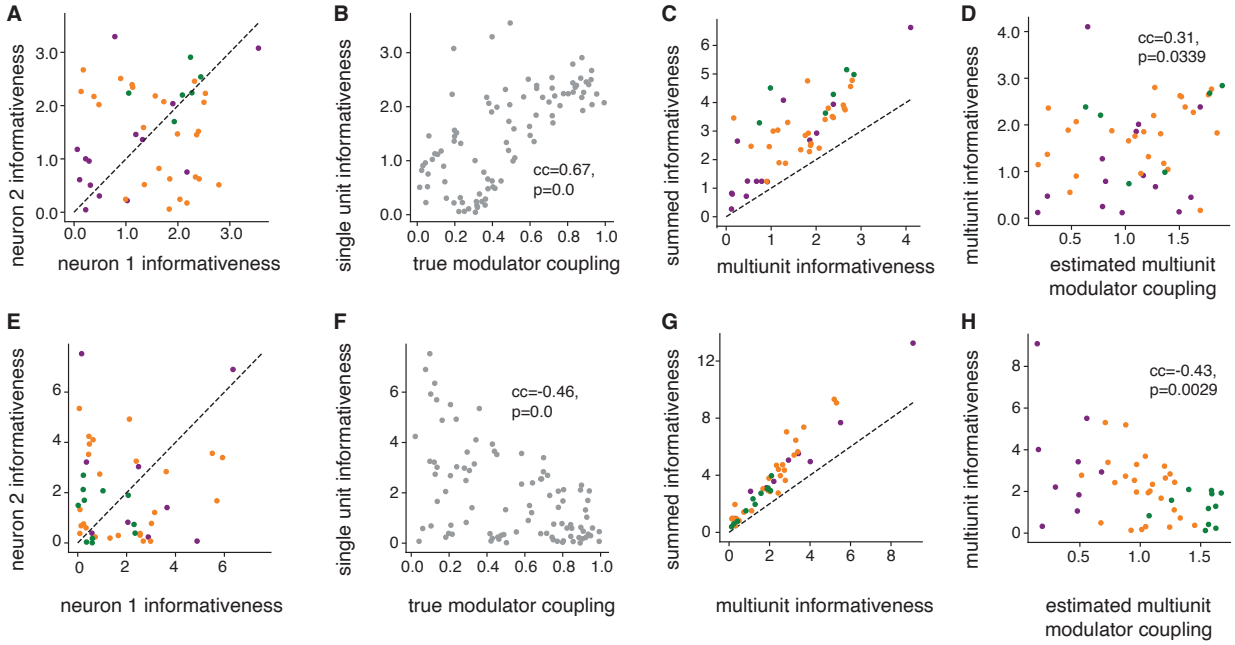

Figure 2.17 Effect of multiunits on model fitting. A) The informativeness of individual units comprising the set of simulated multiunits; colors mark type of modulation for each pair, as in Fig. 2.16. B) Corresponding modulator couplings are correlated with single unit informativeness (the 'targeted modulation' scenario). C) Multiunit informativeness versus sum of single neurons $|d'|$. D) Modulator coupling estimated using the PLDS model and its correlation to multiunit informativeness. E-H) as above, but for the 'untargeted' scenario.

Chapter 3

A functional role for shared targeted modulation in decoding

3.1 Introduction

The computational challenges faced by downstream circuits involved in decoding have been explored in seminal work by Shadlen and colleagues (Shadlen et al., 1996), who enumerated three potential factors that could reduce an animal's behavioral performance compared to predictions of an ideal observer optimal decoder operating on a hypothetical population of independent neurons: "*suboptimally stimulated neurons*" (in which the decoder includes irrelevant neurons in computing its decision), "*correlated noise*" (which worsens performance since it cannot be averaged out by the decoder), and "*pooling noise*" (additional noise in downstream circuits, whose contribution , however, was later shown to be small (Osborne et al., 2005)).

Here we propose a theory that uses stochastic gain modulation, a source for correlated noise (see Chapter 2), to alleviate the problem of suboptimally stimulated neurons. Specif-

ically, the modulator described in Chapter 2 fluctuates rapidly and is targeted to task-informative neurons. The modulator itself is not related to the task but its targeting structure provides information about which neurons are informative and its fast time constant allows any information in the modulator to be quickly accessible, potentially on the time scale of single trials. We propose that the modulatory fluctuations serve as a "label" for the task-relevant neurons, so that downstream circuits can easily identify and use these signals. We posit that the decoder makes use of the modulator itself (or the modulator-induced covariability) when assigning appropriate decoding weights to each neuron. We construct such a *modulator-guided decoder*, and show through simulations that moderate levels of task-specific stochastic modulation of an encoding population can lead to a substantial overall benefit in decoding accuracy, while keeping the assumed knowledge about the encoding population at a biologically plausible level. Thus, structured noise may be an essential feature of brain computation.

3.2 Targeted modulation can facilitate decoding

To study and test our hypothesis, we simulate encoding in a population of stimulus-selective, noise-modulated Poisson neurons and compare statistically optimal *ideal observer* decoders, that have full knowledge of the stimulus-selectivity and modulatory structure of the encoding population, with *biologically plausible* decoders, that must operate with limited knowledge of the encoding population. For this, we use a variant of the doubly stochastic modulated-SR model introduced in Chapter 2 with static stimulus-dependent firing rates, and a one-dimensional, shared, temporally-independent stochastic modulator m_t (see Suppl. 3.5.1.1):

$$k_{n,t}(s) \sim \text{Poisson}\left(\lambda_n(s)\exp(c_n m_t)\right), \tag{3.1}$$

where $k_{n,t}(s)$ is the spike count of neuron n at time t in response to stimulus s; the modulator m_t is drawn independently from a Gaussian distribution with zero mean and variance σ_m^2 (reflecting the fast fluctuations in the data), and influences neuron n with coupling weight c_n, which is set to be proportional to the neuron's task-informativeness. Overall modulation strength in the population is determined by the modulator variance ($\text{Var}(m_t c_n) = \sigma_m^2 c_n^2$ - see also Churchland et al., 2011). Since the degree of modulation affects not only variability but also mean responses, we explicitly correct for the mean increase, given by $\frac{\sigma_m^2 c_n^2}{2}$, to isolate the effects of modulator-induced co-variability.

For a binary discrimination task, $s \in \{0, 1\}$, the ideal observer's optimal decoder for the modeled population compares a weighted sum of the neural responses with a decision threshold, $z(m_t)$, that depends on the time-varying modulator:

$$\begin{aligned} \sum_n a_n^{(\text{opt})} k_{n,t}(s) &> z(m_t) \qquad \text{with,} \\ a_n^{(\text{opt})} &= \log(\lambda_n(1)) - \log(\lambda_n(0)) \qquad \text{and,} \\ z(m_t) &= -\sum_n \exp(m_t c_n)(\lambda_n(1) - \lambda_n(0)), \end{aligned} \tag{3.2}$$

where $a_n^{(\text{opt})}$ denotes the optimal decoding weights[I]. These are independent of the modulator and equivalent to those derived from an independent Poisson model (for derivation see Suppl. Sec. 3.5.1.3). The optimal decoder relies on perfect knowledge of the modulator m_t,

[I] For brevity, 'decoder' refers to both the stimulus readout, and its corresponding optimal discriminator.

the stimulus selectivity of the neurons, $\lambda_n(s)$, and the coupling weights c_n. We can relax these requirements, by assuming that the modulator is unknown, and only the modulator-marginalized stimulus selectivity of the cells is available (i.e., the stimulus response averaged over possible modulators - see Suppl. Sec. 3.5.1.3). This would be, for instance, the view of an experimentalist recording stimulus responses without access to ongoing modulation. We refer to this solution as the *modulator-marginalized optimal* (mm-optimal) decoder. Due to the particularities of the Poisson noise model, this second decoder also computes a weighted sum over responses:

$$a_n^{(\mathrm{mm})} = \log(\lambda_n^*(1)) - \log(\lambda_n^*(0)). \tag{3.3}$$

But it compares this weighted sum to a *fixed* threshold, that does not depend on the time-varying modulator:

$$z^{(\mathrm{mm})} = -\sum_n \left[\lambda_n^*(1) - \lambda_n^*(0)\right], \tag{3.4}$$

where $\lambda_n^*(s)$ is the mean response of the nth neuron averaged (marginalized) over possible modulator values. For the encoding model in Eq. (3.1), $\lambda_n^*(s) = \lambda_n(s)$, which means that the mm-optimal decoding weights are the same as those used in the optimal decoder (i.e., $a_n^{(\mathrm{mm})} = a_n^{(opt)}$). Hence, in the case of a binary discrimination task, the mm-optimal decoder is able to achieve an unbiased estimate of the decoding weights from the stimulus responses, without knowing the modulator. However, it does lead to systematic time-dependent biases in the decoder threshold and therefore to biased decisions.

The decoding weights are non-zero only for the small subpopulation of informative neurons (Fig. 3.1A, purple), with their signs indicating preference between the two stimulus alternatives. Zero weights eliminate the activity of active but uninformative (Fig. 3.1A, black)

or inactive (Fig. 3.1A, grey) neurons.

These decoders provide upper bounds on decoding performance given the encoding model if modulation is either fully known or marginalized. They motivate the use of a linear-threshold functional form for the readout, but use weights that rely on full knowledge of each neuron's mean responses to the stimuli of the current task. Given that the modulator is one-dimensional, it is much easier to estimate or to relay a copy of it than the high-dimensional neural response properties. We only consider the optimal, not the mm-optimal, decoder in the following comparisons, since the knowledge required is very similar (see Table 3.1). The challenge for a downstream circuit is to find a way to approximate the optimal weights, when provided only with incoming spikes, the task feedback, and potentially the one-dimensional modulator, but without explicit knowledge of the stimulus encoding model. How can the brain achieve this? The conventional means of learning decoding weights is regression. Although this is feasible for a small set of mostly informative neurons, the number of training examples needed for accurate weight estimation grows significantly with population size (see Table 3.1 and problem formulation in Chapter 1) (Hair et al., 2014; Kanitscheider et al., 2015b). The behavioral flexibility exhibited by the monkeys precludes such a solution. Instead, we seek a heuristic alternative that learns faster.

3.2.1 Heuristic decoders

Consider first a decoder motivated by early work on neural binary discrimination (Shadlen et al., 1996), where the idea is to split all neurons into two sub-populations ("preferred" and "anti-preferred") and then compare their average responses. This solution only assigns decoding signs ($a_n^{SO} \in \{-1, 1\}$), which indicate relative stimulus preference,

but ignores the relative importance of different neurons by weighting all their responses equally. We refer to this approach as the *sign-only* (SO) decoder (see Suppl. 3.5.1.5 for details). It can be learned relatively quickly (Suppl. 3.5.1.6), and if all neurons in a population were informative, learning the signs would provide an accurate readout of task information. However, its performance falls rapidly as the fraction of informative neurons decreases (Suppl. 3.5.1.7) and for realistically small fractions of informative neurons (Britten et al., 1996; Cohen and Maunsell, 2009), the SO decoder cannot match the levels of performance seen in the monkeys (Suppl. 3.5.1.7). The explanation for this becomes obvious in an illustration of an encoding population with diverse tuning properties; Fig. 3.1A shows average responses of simulated neurons with diverse stimulus tuning features to two task-specific stimuli. Only a small fraction of neurons are responsive, while the large majority of neurons respond weakly ("inactive"). Even though the individual noise of each inactive neuron is small by definition (Fig. 3.1A, grey points), together their task-irrelevant responses eventually dominate the relevant stimulus signal (see Fig. 3.1B), and since all neurons must be included in one of the two sub-populations, the noise from the inactive and uninformative neurons corrupts the decision signal. In order to discount the inactive neurons, they should be assigned decoding weights with smaller amplitudes. However, these weights cannot be assumed to be known, but must be learned/adapted based on information readily available to upstream circuits.

Since informative neurons necessarily have to show activity during a task, one simple heuristic rule is to set decoding weights proportional to the mean spike count of their associated neurons:

$$|a_n^{(\mathrm{RG})}| \propto \frac{1}{T} \sum_t k_{nt}(s). \tag{3.5}$$

For this decoder, the sign of the weights must again be learned (as for the SO decoder).

The time-invariant threshold is set optimally. This *rate-guided* (RG) decoder improves decoding accuracy over the SO decoder by excluding the inactive neurons that do not respond to the stimuli (Fig. 3.1A, grey points). Fig. 3.1B shows that while the SO decoder's performance drops to chance level with increasing numbers of inactive neurons, the RG decoder is much less affected. However, the RG decoder is still far from optimal. In particular, it cannot exclude neurons that are active, but respond similarly to both stimuli (and are thus uninformative - Fig. 3.1A, black points).

The modulator could deliver this missing differentiation through its task-specific targeting structure. Here we propose a simple local rule for estimating the amplitude of the decoding weights as a function of the relative strength in modulation of each neuron, which in turn reflects its relative informativeness. We define a *modulator-guided* (MG) decoder that uses temporal correlations with each neuron's activity to estimate its decoding weight amplitude as:

$$|a_n^{(\mathrm{MG})}| \propto \frac{1}{T} \sum_t m_t k_{n,t}(s). \tag{3.6}$$

Our heuristic learning rule results in estimates of the form (see Suppl. 3.5.1.4):

$$\mathbb{E}\left[|a_n^{(\mathrm{MG})}|\right] = \overline{\lambda}_n \sigma_m^2 c_n, \tag{3.7}$$

which scale with the average response of neuron n across stimuli, $\overline{\lambda}_n$, and the modulator variance, σ_m^2. For this to be an unbiased estimate of the optimal decoding weights, we need the modulation strength to scale as $c_n = \overline{\lambda}_n^{-1} |a_n^{(\mathrm{MC})}|$. This additional assumption for the encoding model will not affect the optimal decoding weights, but will change

Decoder	SR knowledge	Modulation knowledge	Degrees of freedom
optimal	$\lambda_n(s)$ (optimal)	m_t, c_n	2N+N+T
mm-optimal	$\lambda_n^*(s)$ (mm)	σ_m, c_n	2N+N+1
MG	none	m_t	T
RG	none	none	0
SO	none	none	0

Table 3.1 Knowledge assumed by each of the five decoders (ideal observer optimal decoder with modulator knowledge: optimal; modulator marginalized: mm-optimal; modulator-guided: MG; rate guided: RG; sign only: SO - see text for details). Last column gives the dimensionality of variables that are assumed known or need to be estimated from neural responses, with N the number of neurons in the population, and T the number of time points that is used for training. SR stands for stimulus response.

the expression for the optimal threshold (see Eq.3.2). We use this bias-corrected encoder here. Empirically, we have found that the positive effects of modulation on decoding remain, even in the absence of de-biasing. For simplicity we assume that the MG threshold has the optimal functional form, as defined by the optimal decoder (Eq.3.2, for details see Suppl. Sec. 3.5.1.4).

The MG decoder does not rely on knowledge of the response properties of the encoding population, but it assumes access to the modulator (e.g., it is a broadcast signal). This has important implications for learning the decoder; The MG weight estimates converge rapidly, on the time scale of the modulator fluctuations which have been shown to be much faster than a trial (Sec. 2.7). Once the informative neurons have been identified, their decoding sign is determined based on explicit trial feedback, which only requires a handful of trials for small populations (Suppl. 3.5.1.6). For simplicity, the amplitude and sign were estimated separately here. Nonetheless, they can also be learned jointly using a form of local online learning based on eligibility traces (Suppl. 3.5.1.10) (Gerstner et al., 2018; Izhikevich, 2007).

3.2.2 Decoder accuracy

We compared the performance of different decoders listed in Table 3.1 in a binary discrimination task, based on simulated responses of a large population of V1 neurons with a small fraction of informative neurons (5%, Fig. 3.1A; see also Suppl. 3.5.1.7 for variations in percentage of informative neurons). The statistically optimal decoder provides an upper bound on the accuracy of a linear decoder with full knowledge of the encoding population and the modulator, while the SO decoder provides a lower bound on achievable performance. The optimal decoder's accuracy deteriorates as the modulator increases in amplitude, corrupting the encoded signal (Fig. 3.1D). This reinforces the point that, unlike other forms of noise correlations (Kanitscheider et al., 2015b; Moreno-Bote et al., 2014), the modulator-induced covariability is strictly detrimental for encoding (Suppl. 3.5.1.2). While the performance of the MG decoder is limited by this corruption as well, it also benefits from a stronger label in the informative neurons (Fig. 3.1C). Its performance follows an inverted U-shape as a function of modulation amplitude, reflecting the tradeoff of these two opposing effects (Fig. 3.1D). MG decoding performance is maximized at an intermediate modulation amplitude, where it attains an accuracy close to that of the ideal observer, a result which remains robust to variations in population size (Suppl. 3.5.1.8).

We study this optimum with respect to modulator strength by looking more closely at the encoding and decoding processes separately. For encoding, we predict the signal-to-noise ratio using Fisher's Linear Discriminant (FLD):

$$\mathrm{SNR} = \frac{\left(\mathbf{a}^T(\mu_\mathbf{1} - \mu_\mathbf{0})\right)^2}{\mathbf{a}^T\Sigma_1\mathbf{a} + \mathbf{a}^T\Sigma_0\mathbf{a}} \tag{3.8}$$

for the optimal decoding weights $\mathbf{a} = \mathbf{a}^{(optimal)}$ (Fig. 3.1C), where $\mu_{\mathbf{s}}$ and Σ_s are the mean and covariance of the neural responses to stimulus s. Decoding accuracy can be estimated by the MSE of the MG-estimated decoding weights relative to the theoretical optimum. Given that the MG-decoding weights are unbiased, the MSE is given by the variance of the estimator, which decreases in inverse proportion to T (see Suppl. Info. S2):

$$\mathrm{Var}\left[|a_n^{(\mathrm{MG})}|\right] = \frac{\sigma_m^2}{T}\left(\overline{\lambda_n}(1+\sigma_m^2 c_n^2) + \overline{\lambda_n^2}\mathrm{e}^{\sigma_m^2 c_n^2}(1+4\sigma_m^2 c_n^2) - \overline{\lambda}_n^2 c_n^2 \sigma_m^4\right), \tag{3.9}$$

where $\overline{\lambda}_n$ and $\overline{\lambda_n^2}$ are the mean and second moment of the neural response across stimuli.

In practice, the performance of the MG decoder could depend on how strongly correlated the modulator couplings, c_n, are with task-informativeness. We tested the robustness of our decoding results by weakening this correlation by adding noise to c_n. We find that although performance decreases overall, the nonmonotonic dependence of the MG decoder performance on modulator strength is preserved (Fig. 3.2A). Interestingly, the optimal modulation amplitude shifts towards the range estimated from the data (see next Sec. 3.3.1), suggesting that physiologically, the degree of modulation may be well-matched to the precision of the modulator targeting. Other non-modulatory noise sources deteriorate performance of all decoders similarly, but do not change the relative performance of the MG decoder to the optimal decoder (see Suppl. Sec. 3.5.1.9). Given that our measurements mostly include multiunits, we tested the impact on decoding and found that the results are qualitatively robust to such measurement noise (Fig. 3.2B).

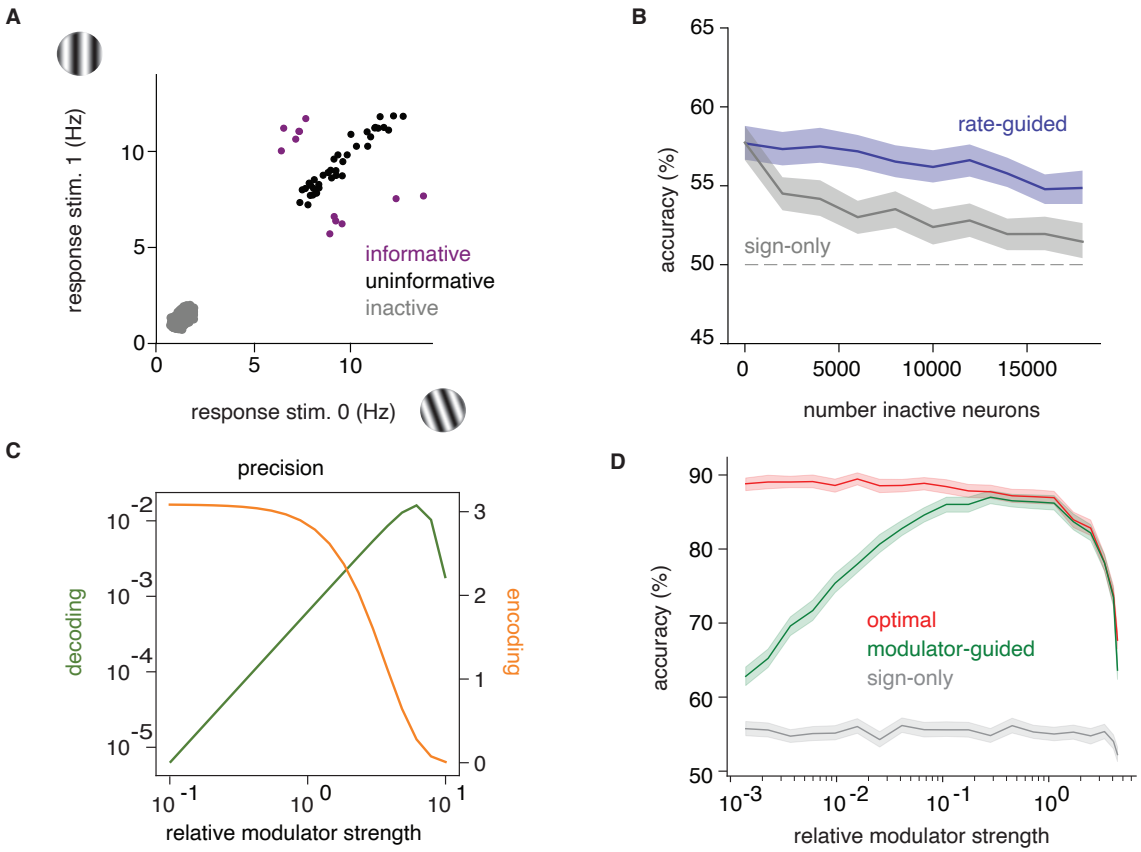

Figure 3.1 Theory of modulator-guided decoding. A) The average response of neurons of the three subpopulations to two task stimuli. There are 12 informative, 38 uninformative and 4950 inactive neurons. B) Mean performance of RG and SO decoders as the number of inactive neurons is increased. The RG decoder downweights inactive neurons, thus allowing it to maintain better performance than the SO decoder. C) Effects of increasing modulator strength on encoding and decoding, respectively, with modulator coupling weights equal to informativeness. 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 1000 model V1 neurons, 50 informative, with increasing relative modulator strength (mean and 95% confidence interval).

3.3 Testing theoretical predictions in V1 data

The modulator-guided decoder theory makes several predictions which can be examined in an experimental context that includes a dynamically changing task, like the one described

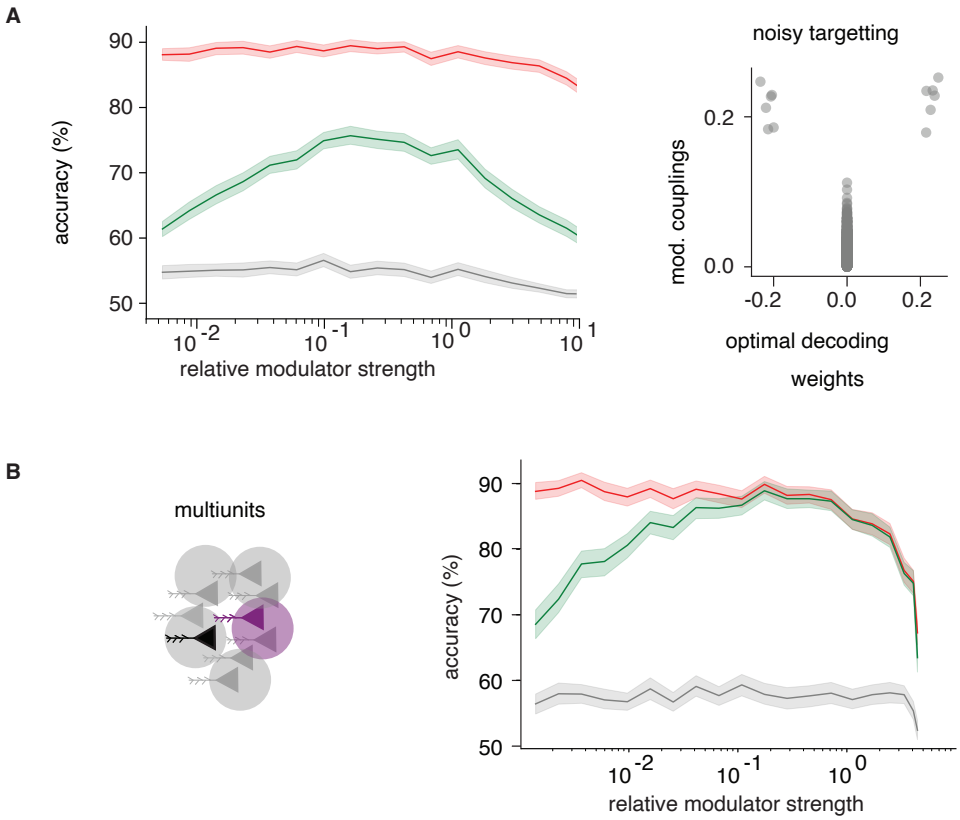

Figure 3.2 Controls for the theory of modulator-guided decoding. A) Same comparison as in 3.1D but with modulator coupling weights equal to informativeness corrupted by Gaussian noise. Right panel shows noisy coupling compared to optimal decoding weights. B) Decoder performance comparison for simulated multiunits, obtained by summing the activity of random pairs of neurons.

in Chapter 2. In particular, the influence of low-dimensional (shared) noise should shift with the task, so as to continue to preferentially target task-informative neurons. Moreover, a modulator-guided decoder should perform better in the low-data regime than either an optimal decoder with learned parameters or regression. Here we test both predictions in the monkey V1 data described in Chapter 2.

3.3.1 V1 modulator is task specific

For the recorded V1 population, the theory predicts that the co-variability of neural responses should change based on whether they are task-informative. Given that the recorded V1 population is informative in the relevant tasks but not the control task (Fig. 2.2D) we expect differences in overall modulator strength across tasks and in neuron-individual modulation strengths.

3.3.1.1 Overall modulator strength

Given the modulated stimulus response model in Eq. 2.3, the modulator and stimulus affect the neural response through the mapping functions **c** and **b** respectively (see details in Suppl. 2.11.1.2). When assessing the overall modulation strength in the population, both the mapping **c** and the modulator variance need to be considered jointly (as scaling up the mapping and decreasing the variance leaves results unchanged).The overall modulator strength is therefore quantified as the variance of the modulator multiplied by the coupling norm $\sqrt{\sum_n c_n^2}$ where n indicates the neuron. The overall stimulus drive is quantified as $\sum_n \sum_i Var(s_i b_{n,i})$, where i indicates the stimulus dimension.

The overall strength of the estimated modulation significantly decreases in the control task relative to stimulus induced variations (Fig. 3.4A). This difference is driven by a change in the modulator strength, which decreases in the control (non-parametric Wilcoxon U-test $p \ll 0.0001$). The stimulus (contrast) induced variance is relatively constant across the task-conditions ($p > 0.05$) (Fig. 3.3). In comparison, the two relevant task conditions have indistinguishable statistics of overall modulator strength (Fig. 3.4B). This difference in overall modulator strength is explained by the theory as a change in labeling, from the

recorded subpopulation that is informative and hence labeled for the relevant tasks, to an unrecorded subpopulation that is informative in the control task.

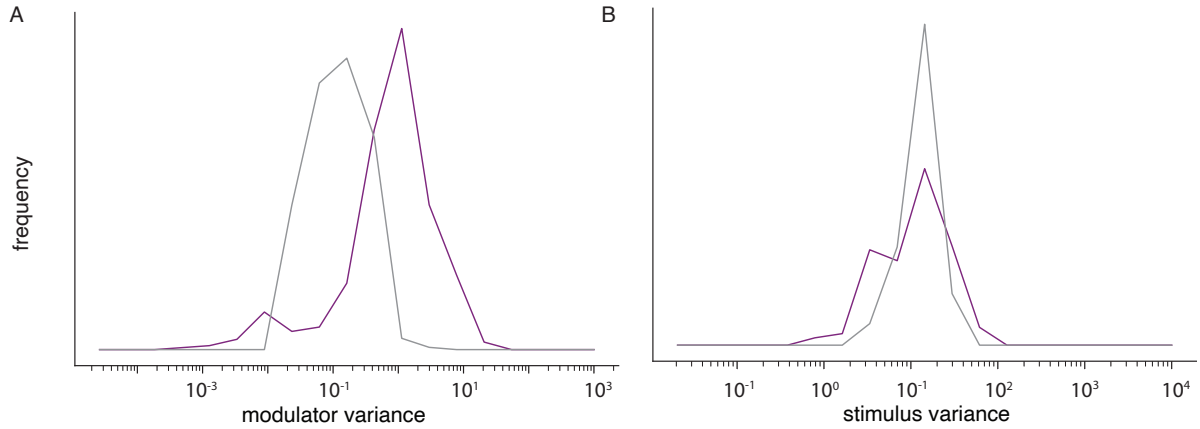

Figure 3.3 Modulator strength and stimulus strength analyzed separated. A) Modulator strength (variance of modulator with unit vector coupling) in relevant (purple) versus control (grey) task over all blocks. B) Stimulus variations in relevant and control task.

3.3.1.2 Task specificity of neuron-individual modulation

The fine coupling itself is correlated in relevant and control task blocks (Spearman $r = 0.5$ with $p < 0.001$), but since the modulator variance itself is dialed down in the control task, the modulation that neurons receive is overall substantially less, reflecting the change in informativeness of the entire subpopulation (see Fig. 2.2). The comparison between the two relevant tasks is limited by the proximity of the two relevant stimulus locations, as only few units are exclusively informative in one task (see Chapter 2). However, despite the reduced sample size, we find a significant correlation between the difference in informativeness in the two relevant tasks and the difference in coupling rank (Spearman correlation, $r = 0.16$ with $p < 0.05$), so that units that are informative in only one of the two tend to also have higher coupling in that task.

In our framework, decoding weights are approximated by estimating coupling strengths, and thus neurons with large coupling (and thus strongly modulated) should have a stronger influence on behavior. Despite V1's early position in the visual processing stream, we find this to be true in our data; 91% of blocks show significant correlations (Spearman r, $\alpha = 0.05$) between modulator coupling and a unit's correlation with the monkey's behavior computed as a $|d'|$ of neural responses, with categories defined by the animal's choices rather than stimulus identity. Specifically, 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, over the squared mean variance $|\frac{\mu_1 - \mu_2}{\sqrt{.5 * (\sigma(a_1^2 + \sigma_2^2}}|$ where $\mu_{1,2}$ and $\sigma_{1,2}^2$ are the means and variances of activity corresponding to the two choices. This gives us an estimate of how involved a unit may have been in the choice of the animal. Potential confounds in this analysis are not only overall firing rates, but also the informativeness of a unit, as the most informative neurons would be expected to have a stronger influence on behavior (Haefner et al., 2013; Nienborg and Cumming, 2014). We therefore use a partial correlation with two covariates, firing rate and informativeness (using multivariate linear regression). Even after controlling for these confounds, it remains the case that units that are more modulated are the ones that are also more predictive of behavior (Fig. 3.4C). This relationship is not present for the residual response variance (Fig. 3.4C). Furthermore, we do not find a relationship with behavioral correlation in other shared noise sources (Suppl. 3.5.2.1), which suggests that the shared modulator-induced fluctuations are particularly relevant for downstream processing.

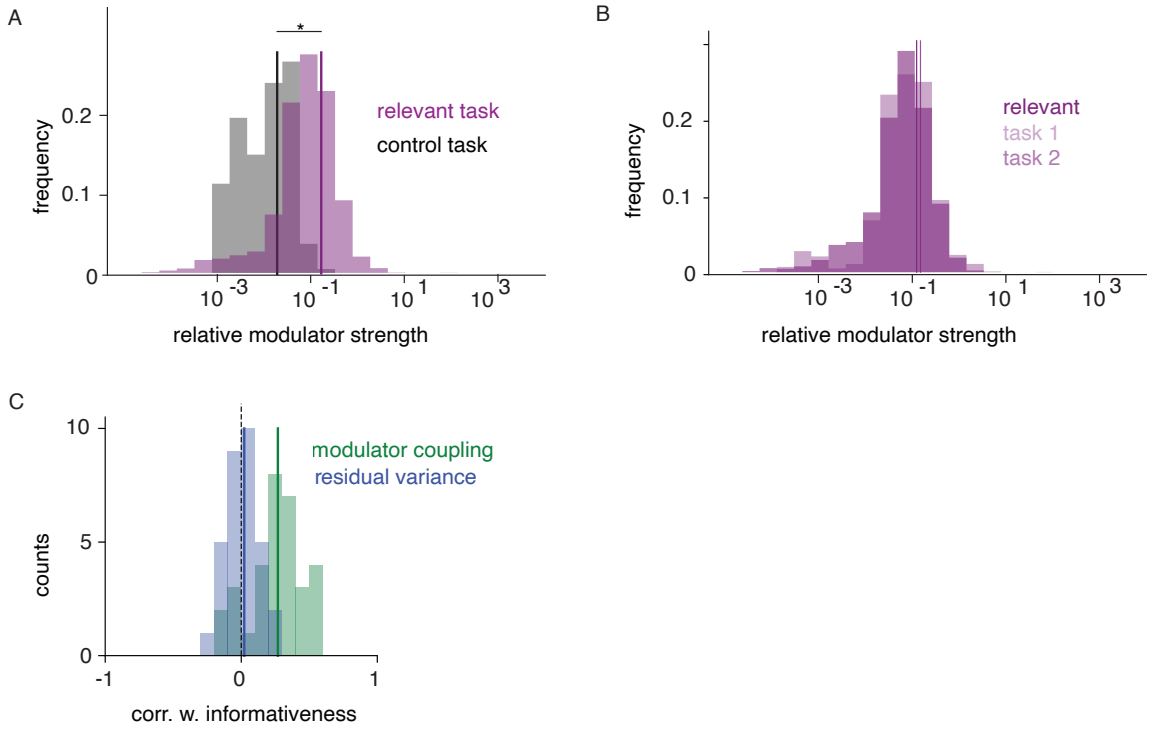

Figure 3.4 V1 modulator is task-specific. A) The distribution of relative modulator strength across all relevant task blocks (purple) and all control task blocks (black); we quantify relative modulator strength as the variance in the modulator (with coupling being unit vectors) relative to that of the stimulus. The star indicates significant difference between the two distributions (U-test, $p < 0.001$). B) Same as in A, but comparing the two relevant tasks against each other ($p = 0.45$). C) The distribution of correlation coefficients between modulator coupling (green) or residual response variance (blue) and the residual behavioral relevance of a unit's activity (correlation with behavior), obtained by regressing out informativeness and mean firing rate.

3.3.2 Knowledge of the modulator allows rapid decoding

The most direct prediction of the theory is the ability of the MG decoder to set appropriate decoding weights for the recorded V1 responses, and to do so rapidly, with limited data. To test these predictions, we decoded the stimulus identity from V1 responses using our heuristic MG decoder and compared its performance with that of the ideal observer

for the estimated (modulated-SR) encoding model (see details in Suppl. Sec. 3.5.2.2). We found that the MG decoder performance is close to that of the optimal decoder ($\sim$ 80% correct) when all the available data is used to estimate the decoding weights (using Eq. 3.6 for the MG decoder and Eq. 3.2 for the optimal decoder). This suggests that the strength and targeting precision of the estimated modulator is sufficient to guide decoding.

The optimal decoder provides an upper bound on decodability assuming perfect knowledge of the response properties, but a downstream decoding area would presumably need to learn those parameters (see $\lambda_n(s)$ in Eq.3.2) through trial-by-trial feedback, or otherwise know and store them in advance, which is unrealistic for many tasks. Specifically setting the optimal decoder weights for a modulated, as for an independent Poisson process, requires information about the mean rate to the stimuli. Even in the relatively small subpopulation of units recorded in our experiment, this requires many trials: the learned "optimal" decoder performs at chance in the low-data regime (Fig. 3.5A). Similarly, learning decoding weights directly through logistic regression requires several training trials before performing above chance (Fig. 3.5A). In contrast, the modulator-guided (MG) decoder finds informative units after only a few training examples, as it estimates the modulator coupling on the time scale of the modulator itself instead of learning from task feedback. It outperforms the learned optimal decoder and logistic regression in the small training sample regime (comparing MG against either learned optimal or regression-based decoder significant; t-test $p < 0.0001$, see Fig. 3.5A). We quantify this effect across all data and find that the MG decoder reaches above-chance performance significantly faster than the learned optimal decoder (t-test, $p < 0.0001$, Fig. 3.5B) and that the performance attained with minimal training is significantly higher relative to that of the learned optimal decoder (t-test,$p = 0.01$). The MG decoder also reaches above-chance performance significantly faster than a regression-based decoder (t-test $p < 0.001$) and learned optimal and

regression-based decoder do not differ significantly (t-test for minimal training and performance $p > 0.05$). Our theory predicts that the advantage of the MG decoder lies in its ability to accurately estimate the decoding weights quickly. Indeed, we find a strong correlation between the MG decoding weights obtained with minimal training and those estimated from all available data, but this relationship does not hold for the learned optimal decoding weights or the regression weights (Fig. 3.5C).

Although significant, the difference in the number of trials required for above-chance performance may seem small. It is likely that the benefits of modulation are substantially underestimated, due to two experimental limitations. First, the recorded subpopulation is biased towards informative neurons since the stimuli are placed so as to drive these neurons. The animal must decode the information present in the entire V1 population, with a much lower percentage of informative neurons. Under such conditions, finding the few informative neurons from task feedback becomes even harder (see Fig. 1.4 and Suppl. 3.5.1.7). Second, the modulator may vary on a time scale faster than the stimulus-presentations of the experiment or the model, which would allow an even faster estimation of the decoding weights (Eq. 3.6 could also be applied to single spikes). Finally, additional sources of co-variability not considered in the theory but found in the V1 data (Sec. 2.7 and Suppl. 2.11.3.2), do not seem to interfere with modulator-based estimates of the decoding weights as suggested by the performance of the MG decoder. Thus, the benefits of the MG decoder for the V1 data provide strong support for the hypothesis that the brain could use such decoding to enable flexible task switching.

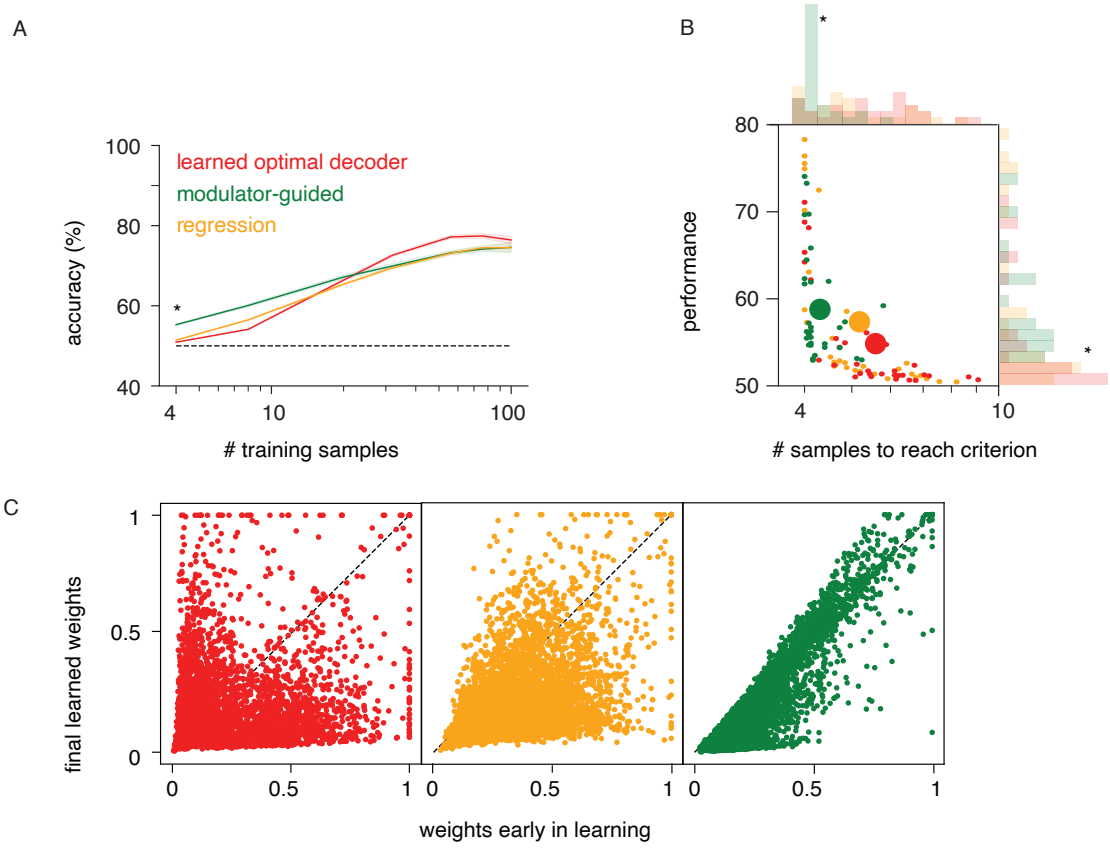

Figure 3.5 V1 modulator facilitates decoding. A) Decoding from the recorded V1 population; Performance of the modulator-guided decoder, the learned optimal decoder or logistic regression for an example block population with increasing number of training samples (shown are mean and its standard error). Black star indicates significant differences between the optimal and the MG decoder. B) Performance with minimal training against minimal number of training samples (stimulus presentations) needed to reach above chance (50%) performance, for each block. Black stars indicates significant differences between the learned optimal and the MG decoder. C) Decoding weights estimated with maximum training (90% of all stimulus presentations) versus with minimal training (1%) for the optimal (red), the logistic regression (orange) and modulator-guided (green) decoders.

3.4 Discussion

Here we proposed that a functionally targeted stochastic modulator could dynamically label informative neurons, facilitating their flexible and accurate task-specific readout. We

showed that a modulator-guided linear decoder, in which weights are estimated through correlation of responses with the modulator, can achieve near-optimal performance. We investigated how parameters of the encoder (proportion of inactive neurons, and active but uninformative neurons) impact performance and found that these dictate a choice of modulator strength that best balances the disruptive effects of correlated noise on encoding against its positive effects for decoding. Importantly, performance is invariant to other parameter changes, such as size of the population and baseline firing rate, demonstrating the robustness of the modulation labeling scheme to circuit details. We then tested the predictions of the theory for flexible information readout from V1 using the targeted modulation extracted in Chapter 2. We found that the modulation strength is functionally different between the relevant and control task and that the modulation itself can be exploited by a decoder that is aware of it, to accurately read out task information from the V1 population after observing only a few trials.

3.4.1 The decoding challenge

In the beginning of this chapter we highlighted two factors that can reduce an animal's behavioral performance: *suboptimally stimulated neurons* and *correlated noise* (Shadlen et al., 1996). The first factor has likely been underestimated in experimental data, since the neurons recorded are not necessarily representative of the full population. For any particular task, most neurons do not carry relevant information, and thus only contribute noise if included in decoding. Moreover, experimental procedures often undersample or deliberately discard low-firing neurons, and experimental stimuli are often optimized to drive responses, introducing a strong selection bias for neurons informative for those specific stimuli. Thus, decoding from a recorded population is less harmed by inclusion of subop-

timal neurons than it would be for the brain, which must operate on the full population. In such circumstances, finding the few informative neurons and setting appropriate decoding weights becomes more difficult, and more essential (Suppl. 3.5.1.7). As such, our conclusions regarding the benefits of targeted modulation for downstream readout are likely understated.

The study of the effects of the second factor, correlated noise, on neural coding is extensive. Correlations can either facilitate or impede the encoding of stimuli, with consequences for the downstream readout (Cohen and Kohn, 2011). In particular, differential correlations, which reflect the stimulus sensitivity of a pair of neurons, are information limiting; they restrict the encoding benefits that would otherwise arise from increasing population size (Moreno-Bote et al., 2014), but also support coding robustness (Pitkow et al., 2015). Experimentally, such correlation structure has been recently detected in mouse V1 (Rumyantsev et al., 2020). The form of covariability that we identify is also information limiting, in that the modulator-induced fluctuations are strictly detrimental for encoding (Sec. 3.2). However, previous results on pairwise correlations are not directly comparable to our analysis here; while shared modulation does introduce pairwise interactions, inferring modulation from pairwise correlations is not as straightforward - it is not clear that pairwise interactions would necessarily lead to fast, low dimensional shared co-variability at the level of the population, of the kind we find in our data. Moreover, the effects of noise correlations have been studied mainly from an encoding perspective, or in the ideal observer framework, whereas we have focused on potential implications for biological decoding. Irrespective of correlation structure, identifying appropriate decoding weights using regression requires many trials (Kanitscheider et al., 2015a), so flexible decoding remains a problem.

Historically, ideal observer models have ignored the presence of modulation, yet have provided good approximations of behavioral performance. Our mm-optimal decoder provides a possible explanation for this incongruity: an experimenter that measures tuning functions by averaging neural responses in the presence of unaccounted-for modulation is effectively marginalizing over it. Optimal decoding weights derived from these estimates are in fact correct, but the use of a fixed decision threshold is suboptimal. This suboptimality is relatively minor in the context of our simulations, but could prove more substantial when fit to physiological data, depending on the structure and strength of the modulator.

3.4.2 The modulator

Our encoding model assumes multiplicative noise since, to our knowledge, there is no evidence that additive noise is functionally targeted. Moreover, experimental reports are conflicting as to whether additive noise is a common phenomenon (e.g. Goris et al., 2014 argue that an additive noise model is inconsistent with their data). Should it be there, task-invariant additive noise would decrease the performance of all decoders, but would not qualitatively change our results (see Fig. 3.8). Consistent with the results of the data analysis in Chapter 2, we here assumed a single task-specific signal that underlies the correlated noise within the population. This is further consistent with previous results from other areas such as (Huang et al., 2019; Rabinowitz et al., 2015) which showed that V4 noise correlations were largely captured using a one dimensional modulator per hemisphere. Alternatively, one could introduce several Gaussian modulators, that combine linearly to jointly gate neural responses. This model would be harder to parameterize, but the net effect would be similar. Additional modulators that are not targeted would reduce the SNR of all neurons and negatively affect all decoders, but again, should not qualita-

tively change the results. The fast time scale of the modulator identified in the data is essential to the theory as it introduces sufficient variability to enable quickly reading out which neurons are strongly modulated. It also suggests, that the modulation does not reflect slow diffuse neuromodulators, but rather the effect of low-dimensional top down signals recruiting local circuitry.

3.4.3 Relationship to attention and other shared variability

Modulation due to top-down attention can facilitate sensory encoding, and has been shown to selectively affect neural responses, including increases in mean response (McAdams and Maunsell, 1999; Moran and Robert, 1985; Treue and Maunsell, 1996), decreases in response variability (Mitchell et al., 2009), and decreases in noise correlations (Cohen and Maunsell, 2009; Mitchell et al., 2009; Rabinowitz et al., 2015; Ruff and Cohen, 2014; Rust and Cohen, 2022), all of which increase the signal-to-noise ratio (SNR) of the local sensory representation. These benefits for encoding are distinct from the modulatory effects we have explored here. First, they tend to operate on a time scale of a task condition (minutes) or stimulus presentations (seconds), whereas the estimated modulation fluctuates on a time scale of tens of milliseconds. Second, we have shown that the targeting of modulation primarily follows task-informativeness instead of tuning properties. While attentional gain boosts have been shown to be tuning-specific (Maunsell and Cook, 2002; Ruff and Cohen, 2014; Treue and Martínez Trujillo, 1999), we do not find evidence that they are specifically shared between task-informative neurons (Suppl. 2.11.4.2). In our data, the estimated modulator coupling is unrelated to the strength of attentional modulation of the mean, suggesting that it may arise from separate mechanisms. This observation seems consistent with the observation that inactivating the superior colliculus

(SC) disrupts the behavioral benefits of attention, but not the attentional modulation of mean responses (Zénon and Krauzlis, 2012), and related results documenting a similar dissociation between increases in mean and improvements in behavior over learning in V4 (Ni et al., 2018). In the context of our theory, we hypothesize that SC inactivation may selectively disrupt the strength or targeting of modulation, affecting the propagation of task-relevant information to decision areas, a prediction that can be validated experimentally.

While pairwise correlations at the level of the full population have been reported to decrease in the relevant task condition (Ruff and Cohen, 2016a), the neuron-specific modulation reported here increases when those neurons become relevant for the task. This suggests that neural covariability likely reflects different sources of modulation, potentially subserving different roles (see also Chapter 2). Importantly, we here model fluctuations across the entire trial, including stimulus on/off periods and take into account the temporal dependencies, which allows us to study modulation at a fine time scale. Previous analysis focused exclusively on fluctuations across repeated presentations of a single stimulus (Huang et al., 2019; Rabinowitz et al., 2015; Ruff and Cohen, 2016a), which may reflect different modulatory signals (Suppl. 2.11.4.4).

In fact, while our modulator accounts for a significant portion of the variance of measured responses, there exist additional sources of variability in the data, reflected in the residual pairwise correlation structure, which are explained by neither the stimulus nor the modulator (Suppl. 2.11.3.2). These residual correlations may arise from other sources (e.g. common feedforward inputs) and subserve distinct functional roles, such as improving encoding precision (Pitkow et al., 2015). Critically, the additional noise correlations do not hinder the ability of the modulator-induced fluctuations to serve as a label for downstream

decoding (Sec. 3.3.1).

The identified modulator is distinct from slow multiplicative, low-dimensional noise reported in other contexts, which may serve other functional roles such as encoding uncertainty in visual areas (Festa et al., 2020; Hénaff et al., 2020). It is also distinct from gain changes due to fluctuations in attention which happen on the time scale of seconds (Denfield et al., 2018). Such signals cannot serve as a labeling mechanism of the type proposed here, since the fluctuations would convey information on a time scale slower than that needed for single trial feedback and decoder learning. Choice-related feedback signals have also been shown to modulate neural activity on a trial-by-trial basis, but they again occur on a slower time scale of several hundreds of milliseconds or seconds (Bondy et al., 2018; Engel et al., 2015). The modulatory process of our theory does not replace but coexists with these additional forms of gain modulation.

3.4.4 Other labeling theories

In many brain areas modulation with a periodic structure has been found (Buzsaki and Draguhn, 2004). Such shared oscillatory structure induces low-dimensional covariability and one of its proposed functions is to bind the representation of common features in subpopulations of neurons (Singer, 1999). The "communication through coherence" (CTC) theory (Akam and Kullmann, 2014, 2012) 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 them out. This theory differs from our own in three important ways. First, the oscillations in Akam and Kullmann (2012) target feature-selective rather than task-informative neurons. These could be the same for a detection task, but differ for a discrimination such as that used

in our experiment. Second, the CTC decoder in Akam and Kullmann (2012) was assumed to use a fixed (as opposed to a modulator-dependent) threshold, which is suboptimal (Sec. 3.2). Third, although our theory would apply to oscillatory modulation, it does not rely on this additional restriction. The V1 modulation estimated from our data seem to favor dynamics that are fast (close to the bounds of what we can estimate given resolution of temporal binning), 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 theory proposes modulatory labeling adapted to task structure, transmitting information about informativeness of neurons for a particular task.

3.4.5 Conclusion

Here we proposed a new role for shared but targeted gain fluctuations specifically in task-specific decoding. We demonstrated through theory, data analysis and computational modeling that the shared variability in task-informative encoding neurons described in Chapter 2 can be used as a label to guide decoding. This novel way of thinking about modulatory "noise" shifts the focus from coding through mean response with variability being a nuisance, to actually transmitting key information through neural variability. It opens up new experimental questions to explore regarding the mechanistic details of this modulation and the generality and limitations of its potential use.

3.5 Supplement

3.5.1 Theory

3.5.1.1 Theoretical framework for decoding from a neural population

We briefly described the framework for modulator-guided decoding above but will extend on it here. We simulated a binary discrimination task analogous to that used in the experiment, 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:

$$k_{n,t}(s, m_t) \sim \text{Poisson}\left(\lambda_n(s) \exp\left(c_n m_t\right)\right), \tag{3.10}$$

where $\lambda_n(s)$ is the stimulus response function of the neuron, and t indexes time within a stimulus presentation. Given the data results, we model the modulator m_t as 1-dimensional i.i.d. Gaussian noise with zero mean and variance σ_m^2; the nonlinearity $\exp(\cdot)$ ensures that the final firing rate is positive. The degree of modulation is neuron specific, parametrized by modulation weights c_n, which we take to be proportional to the n-th neuron's ability to discriminate the two stimuli, $c_n = |\log(\lambda_n(1)) - \log(\lambda_n(0))|$. We divide the firing rate by the expected increase in mean rate due to the modulator given by $\exp\left(\frac{\sigma_m^2 c_n^2}{2}\right)$ 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 simply explained by an increase in firing rates. Overall modulation

strength in the population is determined by the modulator variance (see also (Churchland et al., 2011)). The relative modulator strength in Fig. 3.1 is quantified as a ratio of modulator-induced variance to stimulus-induced variance.

3.5.1.2 Encoding analysis precision

If the modulator coupling is unstructured (e.g., c_n are distributed randomly), then the modulator can be viewed as a global noise source that decreases the discriminatory power of V1 responses proportional to its variance. If the modulator coupling is targeted towards informative neurons (e.g., c_n proportional to difference in mean responses to stimuli), then the harmful effect becomes stronger, as noise is introduced specifically where it most impacts the encoded signal. We quantify the signal-to-noise ratio, using a Fisher Linear Discriminant:

$$\mathrm{SNR} = \frac{\left(\mathbf{a}^\top(\mu_\mathbf{1} - \mu_\mathbf{0})\right)^2}{\mathbf{a}^\top\Sigma_1\mathbf{a} + \mathbf{a}^\top\Sigma_0\mathbf{a}}, \tag{3.11}$$

where $\mathbf{a}$ denotes the decoding weights, and $\{\mu_s, \Sigma_s; s \in [0,1]\}$ are the population mean and covariance for the two stimuli. We evaluate this measure for the optimal decoding weights $\mathbf{a} = \mathbf{a}^{(\mathrm{MC})}$(see main text, optimal decoding weights, Eq. 3.2) and find that discriminability (SNR) decreases faster if modulation is targeted (Fig. 3.6A).

3.5.1.3 Derivation of optimal decoders

Given the modulated Poisson model, and assuming that the modulator m_t and the modulator coupling c_n are known, the log probability of the stimulus s at time point t given

spike counts k_{nt} of the whole population, $n = \{1, 2 \dots N\}$, becomes:

$$\mathrm{L}(s) = \sum_{n}^{N} k_{nt} \left(\log \lambda_n(s) + c_n m_t - \frac{\sigma_m^2 c_n^2}{2} \right) - \sum_{n}^{N} \lambda_n(s) \exp \left(c_n m_t - \frac{\sigma_m^2 c_n^2}{2} \right). \tag{3.12}$$

When discriminating between two stimuli $s = \{0, 1\}$ under this model, the optimal decision is given by the sign of log-odds ratio, $\mathrm{L}(s = 1) - \mathrm{L}(s = 0) \gtrless 0$, which translates into the following expression:

$$\mathrm{L}(0) - \mathrm{L}(1) = \sum_{n}^{N} k_{nt} \log \left(\frac{\lambda_n(0)}{\lambda_n(1)} \right) - \sum_{n}^{N} \exp \left(c_n m_t - \frac{\sigma_m^2 c_n^2}{2} \right) (\lambda_n(0) - \lambda_n(1)). \tag{3.13}$$

This corresponds to thresholding a weighted combination of individual neural responses, with optimal decoding weights:

$$a_n^{(\mathrm{MC})} = \log \left(\frac{\lambda_n(0)}{\lambda_n(1)} \right). \tag{3.14}$$

Since m_t is a known constant, it does not influence the decoding weights themselves, but merely changes the threshold. For the same reason, the optimal decoding weights remain unchanged whether the modulator is known (IO optimal decoder), or whether its effects are marginalized over (mm-optimal decoder).

3.5.1.4 The modulator-guided decoder

Our modulator-guided heuristic decoder assumes 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, using a simple learning rule:

$$|a_n^{(\mathrm{MG})}| = \frac{1}{T}\sum_t m_t k_{n,t}, \tag{3.15}$$

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 their decoding weight magnitudes (see Suppl. 3.5.1.6).

Here we analyse the properties of the MG decoders estimate of modulation strength in a neuron n. First, its mean is:

$$\begin{aligned}
\mathbb{E}\left[|a_n^{(\mathrm{MG})}|\right]_{\mathrm{P}(k,s,m)} &= \mathbb{E}\left[mk_n\right]_{\mathrm{P}(k,s,m)} && (3.16)\\
&= \mathbb{E}\left[m\lambda_n(s)\mathrm{e}^{mc_n-\frac{\sigma_m^2c_n^2}{2}}\right]_{\mathrm{P}(s,m)} && (3.17)\\
&= \mathbb{E}\left[\lambda(s)\right]_{\mathrm{P}(s)}\mathbb{E}\left[m\mathrm{e}^{mw-\frac{\sigma_m^2c_n^2}{2}}\right]_{\mathrm{P}(m)}, && (3.18)\\
&= \overline{\lambda}_n\int m\mathrm{e}^{mc_n-\frac{\sigma_m^2c_n^2}{2}}\mathrm{e}^{-\frac{m^2}{2\sigma_m^2}}dm && (3.19)\\
&= \overline{\lambda}_n\sigma_m^2c_n, && (3.20)
\end{aligned}$$

where $\bar{\lambda}_n$ denotes the average activation of the neuron, $\bar{\lambda}_n = \sum_s \mathrm{P}(s)\lambda_n(s)$; we have used the encoding model and the fact that s and m are independent (Eq. 3.18) and m_t is i.i.d. Gaussian with zero mean and variance σ_m^2 (Eq. 3.19). Under the assumption that $c_n = \log\frac{\lambda_n(1)}{\lambda_n(0)}$, the MG estimates of the decoding weights are biased. While the scaling with σ_m^2 could be easily corrected for by appropriately rescaling the threshold, the neuron-specific

$\overline{\lambda}_n$ bias is problematic. One could correct this bias by a slight adjustment of the encoding model, i.e. assuming $c_n = \frac{1}{\overline{\lambda}_n} \log \frac{\lambda_n(1)}{\lambda_n(0)}$. This will not change the optimal decoding weights $a^{(\mathrm{MC})}$, but will affect the expression of the optimal threshold.

The variance of the estimator can be computed in a similar way:

$$\begin{aligned} \mathrm{Var}\left[|a_n^{(\mathrm{MG})}|\right] &= \frac{1}{T}\left(\mathbb{E}\left[m^2 k_n^2\right] - \mathbb{E}\left[m k_n\right]^2\right) && (3.21) \\ &= \frac{1}{T}\left(\mathbb{E}\left[m^2 k_n^2\right] - \left(\overline{\lambda}_n \sigma_m^2 c_n\right)^2\right) && (3.22) \end{aligned}$$

The second moment term can be computed as:

$$\begin{aligned} \mathbb{E}\left[m^2 k_n^2\right]_{\mathrm{P}(k,s,m)} &= \mathbb{E}\left[m^2\left(\lambda_n(s)\mathrm{e}^{mc_n - \frac{\sigma_m^2 c_n^2}{2}} + \left(\lambda_n(s)\mathrm{e}^{mc_n - \frac{\sigma_m^2 c_n^2}{2}}\right)^2\right)\right]_{\mathrm{P}(s,m)} \\ &= \overline{\lambda}_n \int m^2 \mathrm{e}^{mc_n - \frac{\sigma_m^2 c_n^2}{2}} \mathrm{e}^{-\frac{m^2}{2\sigma_m^2}} dm + \overline{\lambda_n^2} \int m^2 \mathrm{e}^{2mc_n - \sigma_m^2 c_n^2 - \frac{m^2}{2\sigma_m^2}} dm \\ &= \overline{\lambda}_n \int m^2 \mathrm{e}^{-\frac{(m-\sigma_m^2 c_n)^2}{2\sigma_m^2}} dm + \overline{\lambda_n^2} \mathrm{e}^{\sigma_m^2 c_n^2} \int m^2 \mathrm{e}^{-\frac{(m-2\sigma_m^2 c_n)^2}{2\sigma_m^2}} dm \\ &= \overline{\lambda_n}(\sigma_m^2 + \sigma_m^4 c_n^2) + \overline{\lambda_n^2} \mathrm{e}^{\sigma_m^2 c_n^2}(\sigma_m^2 + 4\sigma_m^4 c_n^2) \end{aligned}$$

where $\overline{\lambda_n^2} = \sum_s \lambda_n^2(s)\mathrm{P}(s)$ denotes the second moment of $\lambda_n(s)$ and we have used the fact that the second moment of a Poisson distribution with mean λ is $\lambda + \lambda^2$, the fact that each of the two integrals is the second moment of a gaussian. This holds for any setting of c_n (with or without unbiasing).

Lastly, the covariance for the decoding weights of pairs of neurons n, l takes the form:

$$\begin{aligned}
\mathrm{Cov}\left[|a_n^{(\mathrm{MG})}|, |a_l^{(\mathrm{MG})}|\right] &= \frac{1}{T}\left(\mathbb{E}\left[m^2 k_n k_l\right] - \mathbb{E}\left[mk_n\right]\mathbb{E}\left[mk_l\right]\right) && (3.23)\\
&= \frac{1}{T}\left(\overline{\lambda}_{nl}\mathbb{E}\left[m^2 \mathrm{e}^{m(c_n+c_l)-\frac{\sigma_m^2(c_n^2+c_l^2)}{2}}\right] - \overline{\lambda}_n\overline{\lambda}_l\sigma_m^4 c_n c_l\right) && (3.24)\\
&= \frac{1}{T}\left(\overline{\lambda}_{nl}\mathrm{e}^{\sigma_m^2 c_n c_l}\left(\sigma_m^2 + \sigma_m^4(c_n+c_l)^2\right) - \overline{\lambda}_n\overline{\lambda}_l\sigma_m^4 c_n c_l\right) && (3.25)
\end{aligned}$$

where $\overline{\lambda}_{nl} = \sum_s \lambda_n(s)\lambda_l(s)\mathrm{P}(s)$ is related to the signal correlations of the two neurons.

We assume that the MG threshold has the optimal functional form, as defined by the optimal decoder (Eq.3.2. To maintain biological plausibility, we replace the true c_n (which requires precise knowledge of the encoding model) in the threshold with estimates $|\tilde{a}_n^{(\mathrm{MG})}|$. Furthermore, the difference in firing rates $[\lambda_n(1) - \lambda_n(0)]$ is replaced by an empirical estimate $\Delta\lambda$; this is determined as a function of the estimated decoding weights, the learned signs and one free parameter per informative subpopulation (two parameters in total). It measures the population average change in activity as a function of the stimulus and can easily be learned within a few trials.

3.5.1.5 Sign-only decoder

As a lower bound of performance, we use a weightless "sign-only" decoder that subtracts the summed responses of two subpopulations (i.e., a linear decoder with weights ± 1):

$$a_n^{(\mathrm{SO})} = \mathrm{sign}(\lambda_n(1) - \lambda_n(0)), \tag{3.26}$$

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

3.5.1.6 Learning the signs of decoding weights

We have separated the problem of approximating the optimal decoding weights into two sub-problems: estimating the magnitudes of the weights, $|a_n|$, and estimating their corresponding signs (i.e. their preferred stimulus). For the modulator-guided decoder, the first estimation happens within trials, based on correlations between individual neural responses and the modulator, whereas the signs are learned from explicit feedback given at the end of each trial. Here we simulate informative neurons with different strengths of modulation and examine the correctness of sign estimation as a function of number of training examples (Fig. 3.6D). We find that, in general, the number of trials needed to estimate the signs is small, about 10 trials for the moderate modulator strengths in our simulations. Hence, if the decoding mechanism can identify the few informative neurons and attribute negligible weights to the rest, finding the signs of the informative neurons is fast.

3.5.1.7 Fraction of informative neurons

We tested the influence of percentage of informative neurons in the encoding population on these results. The decoding problem of identifying task-informative neurons is particularly difficult when only very few of the active neurons are task-informative. In experiments, the percentage of informative neurons varies depending on the intrinsic tuning properties of the cells (e.g. width of tuning curves), and extrinsic task properties (e.g. coarse vs. fine discrimination). In our simulations, varying the percentage of informative neurons serves as a proxy for both. We simulated a population of neurons (Fig. 3.6B), varying the percentage of neurons that are task-informative. Mean firing rates of all neurons in the population were the same, but the firing rates of informative neurons were

stimulus-modulated by ±5%. We found that the modulator-guided decoder matched the performance upper-bound for the entire range (Fig. 3.6C). In contrast, the performance of the sign-only decoder is suboptimal, and suffers as the number of neurons that are uninformative increases (Fig. 3.6C, grey). This is expected given that the sign-only decoder groups all neurons into two subpopulations based on stimulus preference, and simply compares their total firing rates. When all neurons are similarly informative, the decoder performs optimally, but otherwise performance is well below the ideal observer bound.

3.5.1.8 Size of population

We varied the size of the simulated population while keeping the % of informative neurons fixed at 5%. We find that MG decoding qualitatively performs similar when population size is increased to $N = 2000$, $N = 4000$ and $N = 10000$ neurons (Fig. 3.7).

3.5.1.9 Effect of additive noise

We test the robustness of our results to additive non-targeted noise in (Fig. 3.8). We reproduce Fig. 3.1D but add Gaussian noise to the firing rates. This decreases the performance of all decoders as the SNR decreases. The qualitative results stay unchanged; the MG decoder reaches optimal performance within a range of modulator strengths.

3.5.1.10 Learning MG weights and signs jointly via eligibility traces

We illustrate a potential learning rule for estimating the modulator-guided decoding weights in a biologically plausible way, using eligibility traces online learning. The key idea

is to use eligibility traces, updated online on the time scale of modulator fluctuations, to estimate the degree of modulation of individual neurons, then combine these correlations with the explicit task feedback received at the end of each trial.

First, one eligibility trace integrates evidence of modulation in neuron n over time t which is independent of the stimulus presentation.

$$e_{n,t+1} = \alpha e_{n,t} + (1-\alpha) k_{n,t} m_t. \tag{3.27}$$

On the same time scale, we use another eligibility trace for the overall firing rate of neurons to correct for the bias (see Sec. 3.5.1.4)

$$r_{n,t+1} = \alpha r_{n,t} + (1-\alpha) k_{n,t} \tag{3.28}$$

Second, information about the signs of the decoding weights comes from trial feedback and is tracked down in the form of a rescaled error:

$$b_{n,i} = (\hat{s}_i - s_i) \cdot \hat{k}_{n,i}, \tag{3.29}$$

where $\hat{k}_{n,i}$ denotes the total spike count during the ith stimulus presentation, and $\hat{s}_i = \sum_n \hat{c}_n \hat{k}_{n,i}$ is the estimated stimulus category.

Whenever an error occurs in trial i (decision made about a stimulus was incorrect) we make an error-dependent update. If no negative task-feedback is provided, only the amplitude of the MG weight is updated, while the sign is preserved.

$$\Delta c_n = b_{i,n}\frac{e_{n,t}}{r_{n,t}} + \delta(b_{n,i})\text{sign}(\hat{c}_n)\frac{e_{n,t}}{r_{n,t}} \tag{3.30}$$

where δ is the Kronecker delta. $\frac{e_{n,t}}{r_{n,t}}$ corresponds to the estimation of the absolute MG weights and is combined with $b_{i,n}$ which corresponds to the gradient of a regression loss. Finally, these changes are integrated to provide the estimated decoding weight:

$$\hat{c}_n = \gamma\hat{c}_n + (1-\gamma)\Delta\hat{c}_{n,i,t} \tag{3.31}$$

In numerical simulations, we find that this learning rules allows a reasonably robust online estimation of decoding weights (Fig. 3.9A). Specifically, we simulated a population of 20 neurons for 200 stimulus presentation, each lasting 200ms, at a time resolution of 50ms. Neural responses during a stimulus presentation are generated according to our encoding model:

$$k_{n,t} \sim Poisson(\exp(\lambda_n(s)\exp(c_n m_t))) \tag{3.32}$$

with the modulator drawn independently from a zero mean, unit variance Gaussian distribution. The stimulus-response rate $\lambda_n(s)$ is set so that neurons differentiate between the two task categories to varying degrees, with about half the population preferring one stimulus and the other half the other stimulus (Fig. 3.9B). The eligibility traces that contribute to the estimation of the absolute weight of $\hat{c}_n$, Eq. 3.27 and Eq. 3.28, integrate information with a time constant given by $\alpha = 0.9$, while the weight updates happen at the time scale of stimulus presentations, with a learning rate $\gamma = 0.999$. Estimated weights $\hat{c}_n$ converge to values that preserve the ground truth after 40 trials (Fig. 3.9B), with decoding

performance showing around 90% accuracy.

3.5.2 Data analysis

3.5.2.1 Relationship of other noise sources to behavior

We have found that the modulator coupling is higher in neurons that also show a strong correlation with the behavioral choice of the monkey, indicating that they are preferentially recruited for the decision. As a control, we want to know whether this relationship extends to other shared sources of noise in the population. We take advantage of the fact that the first PC is correlated with the PLDS extracted modulator, while the second PC is uncorrelated with the modulator and accounts for a very similar fraction of variance explained. We repeat the original analysis with PC1 as a proxy for the modulator and PC2 as another shared noise sources. As already reported for the modulator, PC1 is predictive of a unit's correlations to behavior, but this effect does not extend to the second PC (Fig. 2.14). This dissociation supports the idea that the relationship between single cell fluctuations and behavior is specific to the modulator and not to other sources of noise in the neural responses.

3.5.2.2 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. 3.2). 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. This threshold is suboptimal as shown by the theoretical results in Eq. 3.2 but it is more robust to the noise in the data and therefore performs better in the limited data regime. 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 (see Eq. 3.15). 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.

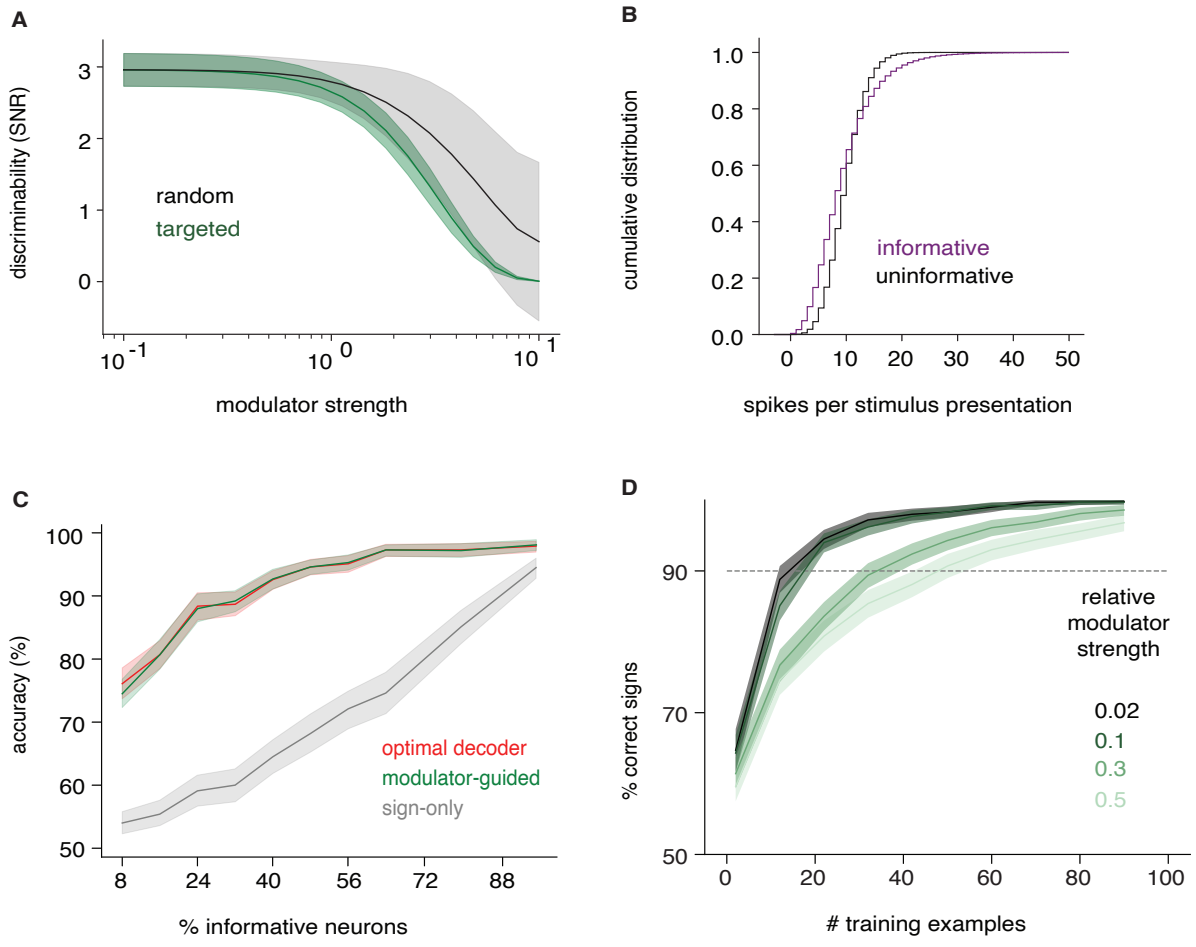

Figure 3.6 Simulations of decoding from V1. A) The effect of modulation on stimulus SNR, as measured by the Fisher Linear Discriminant, for unstructured and targeted modulator coupling. N=100 neurons, 50 inactive, 12 informative, 38 uninformative. B-C) Effect of fraction of informative neurons on decoding performance. B) Firing rate distributions in a simulated population; all neurons are similarly active, but uninformative neurons do not change their responses as a function of the task relevant stimuli while informative neurons are modulated by $\pm 5\%$; N=50 neurons. C) Decoder performance as a function of the fraction of informative neurons (constant total population of 50 neurons, for details see text). D) The percentage of correctly estimated decoding signs as a function of the number of training examples. Different colors correspond to varying relative modulator strengths (see Sec. 3.5.1.1 for details).

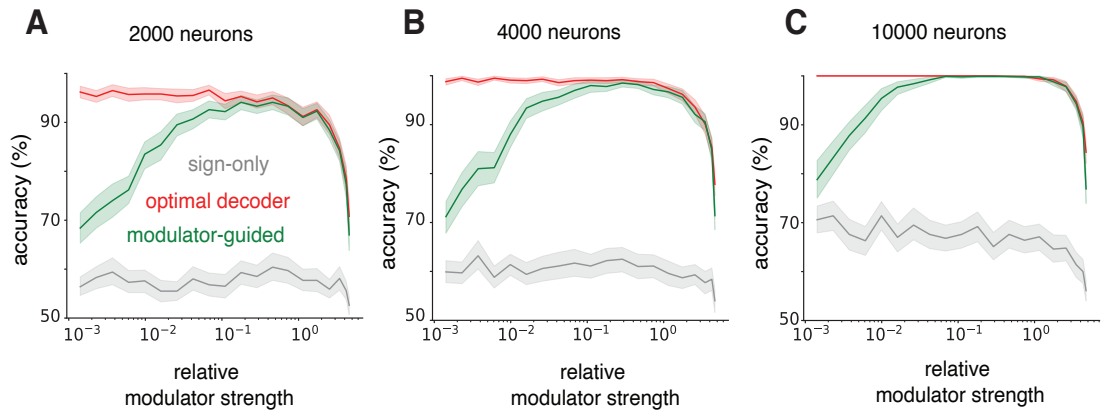

Figure 3.7 Performance with varying size of the population. A) As main Fig. 3D but using 2000 instead of 1000 neurons. B) and C) as A but with increasing population size.

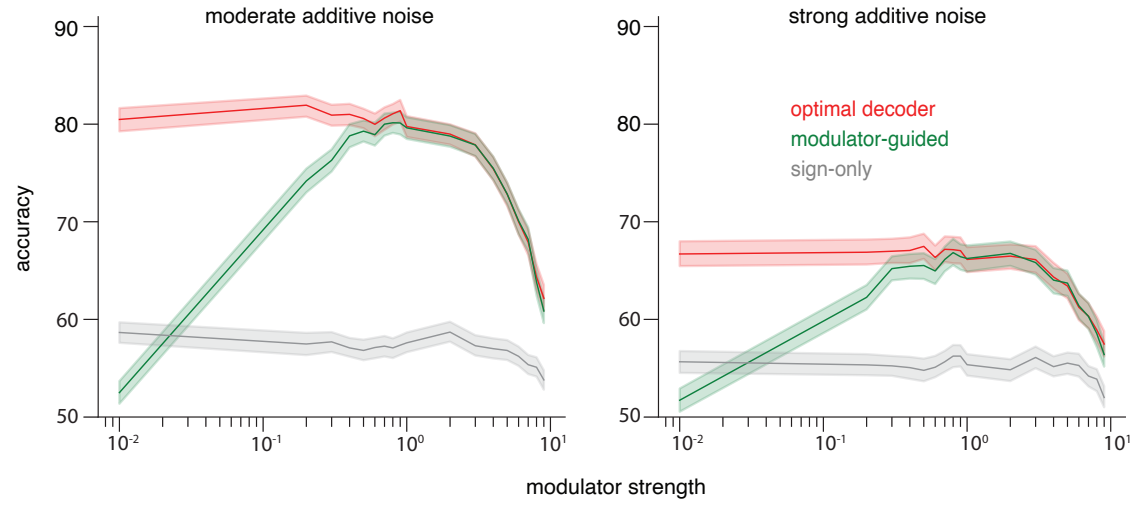

Figure 3.8 Robustness of model to perturbations in firing rates. Decoder performance with moderate and high levels of Gaussian noise added to the firing rates defined by Eq.3.1.

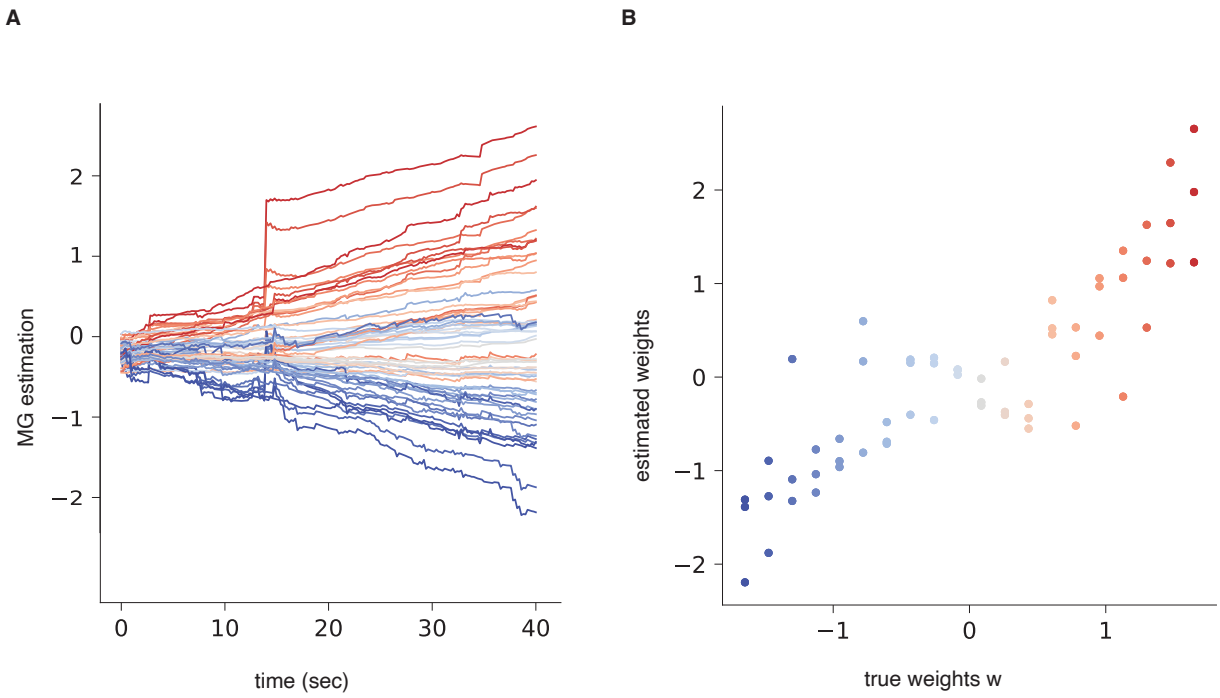

Figure 3.9 Eligibility Trace A) Estimation of weights $\hat{c}_n$ over learning; each stimulus presentation lasts 200ms. Individual lines correspond to decoding weights (combining results from 3 simulations). Color gradient indicates the rank of the corresponding ground truth decoding weight in the population, with red and blue representing opposite tuning preferences. B) Final estimates $\hat{c}_n$ after learning compared to the optimal decoding weights. Colors as in A. Both axes are z-scored.

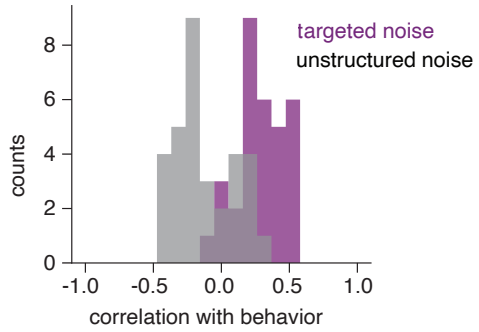

Figure 3.10 Relationship of other noise sources with behavioral correlation. We plot the distribution of correlation coefficients across blocks between behavior and the first PC (purple) or the second PC (grey).

Chapter 4

Hierarchical visual processing with learned targeted modulation

4.1 Introduction

Visual information processing is hierarchical, and task-relevant information needs to propagate through several stages before reaching decision-making areas. Empirical evidence suggests that irrelevant sensory representations are not filtered out until an integration stage in higher cortical areas such as PFC (Mante et al., 2013). Since receptive field sizes increase across stages of processing (Born and Bradley, 2005), task-specific information that is localized in a primary visual area like V1 will diffuse in the subsequent visual layers, making the task of identifying the subpopulation of relevant readout neurons even harder the further downstream (see Introduction). Studying decoding from primary sensory areas directly means sidestepping the substantial challenges that arise due to the dissipation and corruption of information as it flows through stages of sensory processing before reaching decision areas. The decoding problem studied in the previous sections there-

fore needs to be reassessed taking into account hierarchical processing.

As a separate issue, while so far the modulator-guided decoding theory has assumed the correct modulator targeting to be already present in the circuit, the right degree of modulation for each neuron in a task needs to also be learned from experience. It is not clear whether the modulator-guided readout can still facilitate flexible and accurate task performance if coupling needs to be learned, or how its modulatory fluctuations can be used in a hierarchical architecture.

Here, we develop a model that is both hierarchical and learns task-specific coupling. We augment a feedforward hierarchical network, a general model of the visual processing hierarchy (Kriegeskorte, 2015; Yamins and DiCarlo, 2016; Zhuang et al., 2021), to include gain modulation circuits in an early stage (*"encoder gain"*), and a modulator-gated readout mechanism in the last stage (*"decoder gain"*). The unmodulated base network is initially trained to solve a general location-invariant digit classification task (standard MNIST 10-digit classification LeCun and Cortes, 2010). In a subsequent phase, the modulator becomes active and targeting of the modulator is trained to optimize performance on a more specialized task , thus fine-tuning the network for the new task without any reorganization of the feedforward weights. This labeling signal is task-specific and ephemeral, allowing the network to instantly revert to the initially-trained state once task demands are removed. We find that the modulated network learns substantially faster than retraining the base network's parameters or using classic attentional mean boosts, both for single tasks and in a continual learning scenario where the task switches repeatedly. Empirical exploration of the effects of injecting the modulator at different stages of the network reveals that its labeling is most effective when applied to layers in which task-specific information is concentrated in a subpopulation - an "informativeness bottleneck". We use

this hierarchical modeling framework to illustrate how the V1 modulator label detected in Chapter 2 may affect downstream areas, and test the predictions in data of MT activity recorded simultaneously with the V1 population.

4.2 Stochastic modulation labeling in a hierarchical network

Our approach builds on the model of stochastic co-modulation introduced in Chapter 3, which provides a theoretical framework for decoding information from large neural populations, of which only a small fraction carry task-relevant information. It postulates that fast low-dimensional co-fluctuations targeting the task-informative subset serve to label the information for use by a decoder. A "modulator-guided" decoder can then use these fluctuations to estimate the correct decoding weights for the task, achieving high levels of performance within a handful of trials, with minimal explicit task feedback.

We generalize this framework to a hierarchical feedforward neural network, in which neurons linearly combine their inputs, together with a bias term, and pass the result through a nonlinear activation function ($\exp(\cdot)$ for the first, "encoding", layer in accordance with the original formulation in Chapter 3, and ReLU(.) thereafter). We incorporate a stochastic modulator which fluctuates on a significantly faster timescale (indexed $t = 1, \ldots, T$) than the stimulus presentation trials (indexed $k = 1, \ldots, K$), with two distinct effects on the network. First, the modulator, $m_{kt} \sim N(\mu_m, \sigma_m^2)$, controls the gains of all neurons in the encoding layer, via learnable coupling strengths $\mathbf{c}$ (Fig. 4.1, "task-specific encoder gain"):

$$\mathbf{h}_{kt}^{(1)} = \exp\left(\mathbf{W}^{(1)}\mathbf{s}_k + m_{kt}\mathbf{c} + \mathbf{b}^{(1)}\right), \tag{4.1}$$

where $\mathbf{h}_{kt}^{(1)}$ is a vector of activities of the encoding layer for trial k and time t, $\mathbf{s}_k$ the multi-

dimensional stimulus vector and $\mathbf{W}^{(1)}$ and $\mathbf{b}^{(1)}$ are the weight matrix and bias terms of the encoding layer. Unlike the model of Chapter 3, this modulator affects both the mean and the variance of the neural responses, combining traditional deterministic gain boosting with stochastic labeling. We have chosen this formulation in order to facilitate fast learning of the coupling strengths, but it is also more realistic biologically (since modulation of mean and co-variability coexist in the cortex), and provides an opportunity to directly compare to models of attention that rely on deterministic gain boosts (where $\mu_m > 0$ and $\sigma_m = 0$).

As with the model of Chapter 3, we assume the modulator is also available at the output stage of the network, and can be used to guide decoding. This is implemented as an adaptable decoder gain, $\mathbf{g}$, on the neurons in the final (Jth) processing layer, which directly map into the network output (Fig. 4.1, "modulator-gated decoder gain"):

$$\mathbf{h}_{kt}^{(J)} = \mathbf{g} F\left(\mathbf{W}^{(J)}\mathbf{h}_{kt}^{(J-1)} + \mathbf{b}^{(J)}\right), \tag{4.2}$$

with F a rectifying nonlinearity (here a ReLU).

The strength of both encoder/decoder gain mechanisms adapts over time to fine-tune the network's operation on a new task. First, the coupling strengths $\mathbf{c}$ in the "labeled" encoding layer are optimized based on explicit feedback so as to maximize network performance on the task (using backpropagation). Second, in the final layer the decoder gains g are adjusted based on the correlation of neural activity with the modulator, following the modulator-guided estimation rules proposed in Chapter 3:

$$\mathbf{g} = \frac{1}{KT}\sum_{kt} \bar{m}_{kt}\bar{\mathbf{h}}_{kt}^{(J)}, \tag{4.3}$$

where $\bar{m}_{kt}$ and $\bar{\mathbf{h}}_{kt}^{(J)}$ denote the mean-subtracted modulator and neural activity, respectively.[I] Importantly, this rule is independent of stimulus or reward, and only requires the modulator as a 'key' to identify responses of task-relevant neurons in the last layer. All feedforward weights remain unchanged throughout this task-specific learning – their values are assumed to reflect a slower optimization process on a general set of tasks. The task-specific adaptation is only applied to the modulation coupling strengths, c_n, in the encoding layer – a parameter set of size N_1, compared to retraining of the full set of $\sum_{j=0}^{J-1} N_j N_{j+1}$ network weights (N_0 = denotes the input dimension). By concentrating learning on the coupling strengths, fine tuning based on stochastic modulation can rely on fewer training examples to robustly improve performance on the specific task at hand.

4.3 Fine-tuning MNIST digit recognition in the presence of distractors.

To validate the idea of a stochastic modulator guiding task-specific information flow in hierarchical networks, we used task variations built around MNIST digit recognition. We first defined a location-invariant version of digit recognition, in which downscaled MNIST images are embedded in a noisy background, at different spatial locations (Fig. 4.2A; full image size 28×28) and must be identified regardless of their position. We pretrained the circuit on this 'general task', then used stochastic modulation to fine-tune the resulting network to perform 'specific' tasks which involve binary classification of two specific digits confined to one specific image quadrant, in the presence of randomly chosen distractor

[I] Correlations are computed from the fluctuations of the modulator and the neural responses at the fast time scale t, integrated over the time scale of single trials, during which the stimulus is constant.

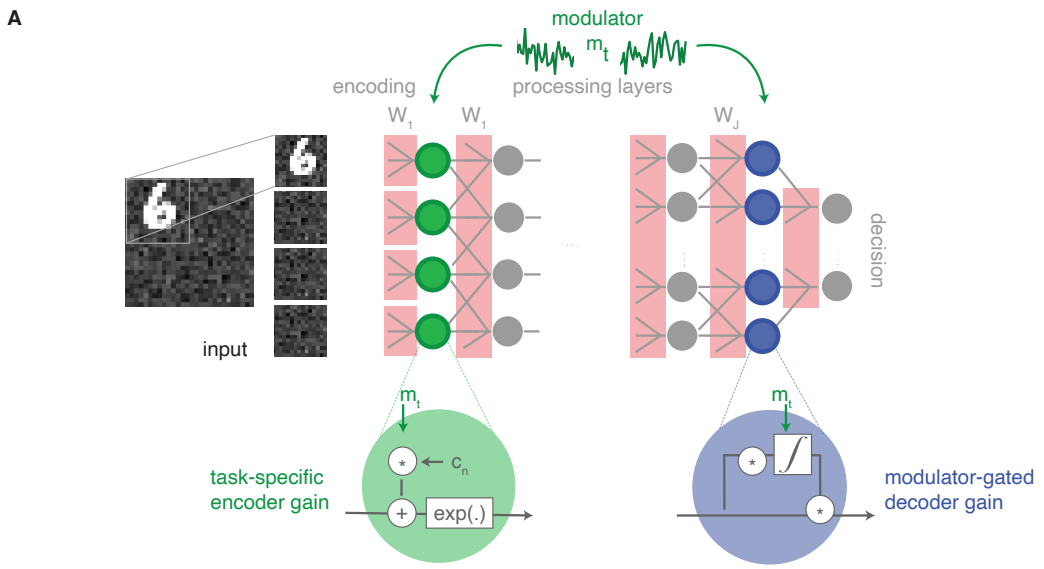

Figure 4.1 Network with stochastic modulation. A feedforward network with J layers maps input images into categorical outputs. Neurons in the encoding layer have localized receptive fields (within one of 4 image quadrants), while all other layers are all-to-all connected. A stochastic modulator induces correlated gain fluctuations in the encoding layer, with neuron-specific coupling strengths c_n ("encoder gain", green circles). Activities of neurons in the last layer are adaptively gated based on within-trial correlations between the modulator and their stimulus-driven responses ("decoder gain" blue circles).

digits appearing in the other three quadrants (Fig. 4.2B). Different instances of the specific task vary in the choice of the relevant digit pair and quadrant, but all are subtasks of the general digit classification problem, and thus the information needed to solve them should be present within the network after pre-training. However, since the network only experiences digits in isolation during training, output neurons in the base network will respond to the combination of task relevant and distractor digits. The objective of the stochastic modulation refinement is to focus the readout on those neurons that carry the task-relevant information.

We use a 3-layer feedforward network with stochastic modulation of activity in the encoding layer, followed by an all-to-all connected "processing" layer, and a final "decision" layer

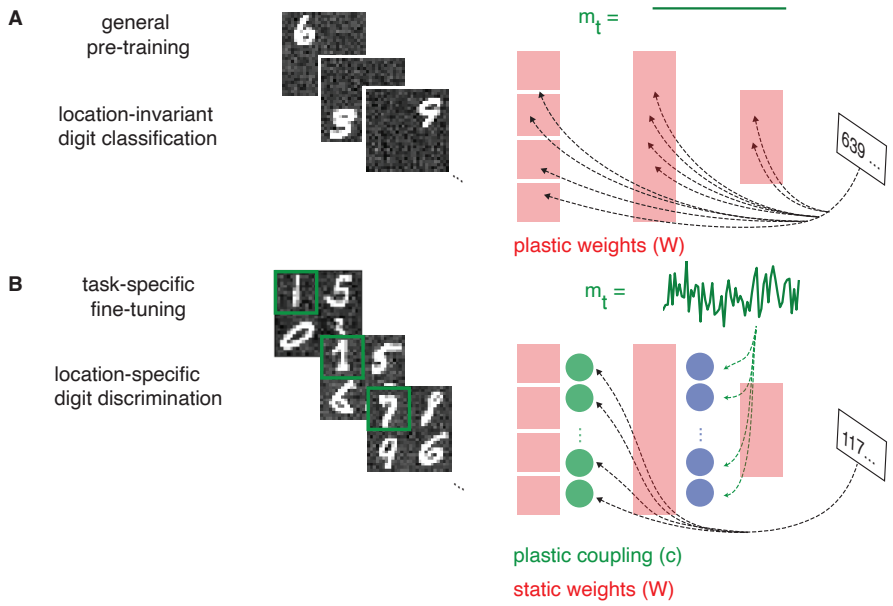

Figure 4.2 Pretraining and fine-tuning. A) During pretraining, feedforward weights $\mathbf{W}^{(j)}$ are optimized (via backpropagation) on a general categorization task (here, location-invariant MNIST digit classification), with the modulator disabled (i.e., set to zero). B) The network is fine-tuned for binary classification of a specific pair of digits, localized within a specific spatial quadrant (here, '1' vs. '7' in the upper left quadrant), in the presence of distractors. The feedfoward weights $\mathbf{W}^{(j)}$ are held fixed, and the modulator coupling strengths, c_n are trained (via backpropagation). Output gains (blue) are automatically adjusted based on correlation with the modulator (Eq. 4.3), without task feedback.

that maps into a categorical output (softmax). The encoding layer receives local information about the input, with 'receptive field'-like weights from one of four image quadrants. The modulator has to isolate the subset of neurons that are target-relevant, which in this simple architecture means neurons that encode both the task-relevant input quadrant and whose responses differentiate the task-specified digit pair. The hypothesis is that learning will target the modulation specifically towards these neurons, and that the labeling of their responses will propagate through the densely connected layer so that the modulator-gated readout can adjust the decoder gain to help perform the task.

The unmodulated network was pretrained on a 'generic' recognition problem: identify

a single digit at an arbitrary location within the image Fig. 4.2A. Weights were optimized using conventional backpropagation with Adam (Kingma and Ba, 2015), using the MNIST training set (modulator $m_{kt} = 0$ and the gain terms $\mathbf{g} = \mathbf{1}$). For details on hyperparameters and their optimization, please see Suppl. Sec. 4.7.1.1. The location and background noise of each image were drawn independently for each image, uniform for location and using additive i.i.d. Gaussian pixel noise for the background (std=0.1, for image pixels in the range $[0, 1]$, training dataset includes 4000 images). The extent of pretraining ensures that the network reaches good performance on the 10-class digit classification; the trained network also exhibits good performance on two-digit categorization at any location, in the absence of distractors (Fig.4.3A), but falls to near-chance levels when distractor digits are introduced. During task-refinement, learning alternates between updating the modulator coupling $\mathbf{c}$ by backpropagation, and updating the decoder gains $\mathbf{g}$ using correlations estimated within a single trial ($T = 100 - 500$) according to Eq. 4.3. To simplify the comparison to other forms of feedback-based learning and avoid any interactions between intrinsic network noise and feedback-based learning, the network dynamics are deterministic (modulator is held constant) during the backpropagation steps, and modulator stochasticity is only introduced in the second step (m_{kt} drawn i.i.d. from $\mathcal{N}(0, 0.1)$). Hence, the effects of modulation on the coupling gradients are indirect, via its effects on the decoder gain.

To assess the effectiveness of this combined learning, labeling, and decoding procedure on specialized task performance, we compared it to three alternatives. The first uses backpropagation to relearn all feedforward weights (initializing from the pretrained weights), which we term 'retraining' ($m = 0$, $\sigma_m^2 = 0.0$). The second uses an attention-like deterministic gain boost ("attentional modulator"), in which the feedforward weights are fixed, but the responses in the encoding layer are amplified by scale factors learned via backpropaga-

tion ($m = 1$, $\sigma_m^2 = 0.0$); since our procedure also includes a boost in mean responses, this control provides a natural lower bound on the benefits of stochastic gain modulation.

4.3.1 Modulator label allows for efficient and effective fine-tuning

Evaluating the performance of different learning algorithms on example digit pair tasks reveals systematic differences in the speed of learning, with the stochastic modulator outperforming its competitors by a substantial margin (Fig. 4.3B). We quantified the systematicity of these observations across all tasks by measuring the initial slope of learning, estimated with linear regression over the first 50 measurements (Fig. 4.3C), and by the number of training examples required to reach an accuracy criterion of 70% (Fig. 4.3D). These measures confirm that using a learned stochastic modulator to fine-tune the network to the requirements of the new task is faster / requires less data than the other methods. The complete retraining is generally much slower, presumably because it needs to tune a much larger number of parameters. Importantly, the stochastic modulator generally improves over the attentional modulator, despite the fact that the feedback-based part of learning (the encoder gain $\mathbf{c}$) is identical in the two conditions.

4.3.1.1 Modulator learns task-specific targeting structure

To better understand the nature of the adapted modulation solution, we linearly projected the modulation strengths (after training on 400 examples) back into the pixel space, as a means of visualizing which features in the input are enhanced via modulation (Fig. 4.4A). The modulation is seen to preferentially affect localized patterns that reflect both common and distinct features of the two task-relevant digits, within the task-relevant quad-

rant. This spatial specificity is expected, given that task structure requires that quadrants containing distractors should be ignored, and given that the quadrant structure is explicitly mirrored in the encoding layer. However, we see additional structure that shows that the learned modulator coupling is also targeting neurons with task-specific feature selectivity. In fact, further analysis shows that within the subgroup of spatially relevant neurons there is a strong positive relationship between task-specific coupling strength and informativeness measured by $|d'|$ (Fig. 4.4B).[II] The informativeness distribution are skewed, with task informativeness concentrated in relatively small subpopulations of neurons that are distinct across tasks (Fig. 4.4C). The learned coupling correspondingly changes across tasks to reflect these differences in task informativeness (Fig. 4.4C: neurons that are more informative in one compared to another task tend to also have higher coupling strength in that task). This confirms that the task-specific targeting of modulation posited in the original theory can be directly learned from experience.

4.3.1.2 Results generalize to deeper architecture

It seems likely that the initial modulator label might lose its specificity as it propagates through many layers of distributed nonlinear processing. To test how intermediate levels of processing affect learned stochastic modulation we extended the network introduced above by an additional all-to-all connected processing layer. Repeating the experiments, we confirmed that our learned stochastic modulation still functions, even when the label needs to propagate further (Fig.4.5A). The experiments on the new architecture qualitatively reproduced the speed and sample efficiency improvements of the stochastic modulator over al-

[II] Note that although $|d'|$ is easy to compute across layers, it only provides a coarse measure of informativeness by ignoring the effects of network nonlinearities. See Chapter 5 for an extended discussion.

ternative learning procedures (full characterization in Suppl. 4.7.1.2), but a direct comparison between stochastic modulation in the 3- vs. 4-layer architecture does reveal a modest slow-down of learning, despite the fact that baseline performance was statistically matched between the two (Fig.4.5B, C).

4.3.1.3 Modulator targeting best if task information is concentrated in a subpopulation

Thus far, we have assumed that the primary effects of modulation are directed towards the encoding layer of the network, but does that need to be the case? In principle, the stochastic label should be most efficient when only a small fraction of the modulated population carries task-relevant information (see Chapter 1 Sec. 1.3.3 for intuition). In contrast, if this information were uniformly distributed across the entire layer, stochastic modulation would not help at all. Hence, the presence of task-specific information bottlenecks seems to be a critical consideration for deciding where to direct stochastic modulation for maximum effect.

We vary the network architecture across two dimensions, sparsity and modulator placement, to study the impact of localizing information about features in the input in different ways (Fig. 4.6). Intuitively, we expect that sparse and localized connectivity will result in features (e.g. locations) being represented in subsets of neurons, whereas broad or all-to-all connectivity will make feature information less localized across neurons. Indeed, this is the case in our simple networks: the informativeness distribution is substantially broader in the processing (all-to-all connected) layer, compared to the encoding (Fig. 4.7A; $|d'|$ estimated using the task digit pair, for the pretrained network). In contrast, when both layers have local connectivity, the information distribution in the processing layer is similarly sparse as that of the encoding layer (Fig. 4.7D). Among only the neurons in the task-

relevant quadrant, the processing layer has more highly informative neurons, due to its additional nonlinearities that allow for more specific feature selectivity (inset in Fig. 4.7D).

Putting these observations together with the idea that modulation should be directed to the task informativeness bottleneck, we hypothesized that applying the modulation to the processing layer should negatively affect the ability of the stochastic modulator to fine-tune the network when the processing layer is all-to-all connected. The opposite should hold when the processing layer weights are localized. We test these predictions in a four-layer network. To avoid potential confounds caused by across-layer differences in the neural nonlinearities, we use ReLU$(\cdot)$ as the activation function in all layers, and modify first-stage modulation in Eq. 4.1 as follows: $\mathbf{h}_{kt}^{(1)} = \text{ReLU}\left(\mathbf{W}^{(1)}\mathbf{s}_k\right)\exp\left(m_{kt}\mathbf{c} - \mathbf{b}^{(1)}\right)$. When the processing layer is all-to-all connected, we find that directing the modulator towards the encoding layer yields faster learning and better end performance (Fig.4.7B). Quantifying the number of training examples required to reach criterion performance across tasks reveals a systematic shift in the distribution across the two scenarios, confirming that early modulation is preferable for this architecture (Fig.4.7C). Repeating the same analysis in the architecture where both layers are spatially localized leads to the opposite conclusion (Fig.4.7E-F). Here we find that performance is better if the modulator is injected in the processing layer (Fig.4.7E), with results robustly reproduced across task instances (Fig.4.7F). Overall, these results confirm our expectation that stochastic modulation is most effective when directed towards bottlenecks of task-relevant information.

4.3.1.4 Seamless switching back to the general task and continual learning

One of the immediate appeals of gain modulation (either stochastic or deterministic) as a mechanism for task-specific information routing is that, since the feedforward weights

are unchanged, returning to original performance in the general task is instantaneous (Fig. 4.8A). The combination of this capability and the increase in speed of learning makes stochastic modulation a remarkably effective mechanism for adapting (and unadapting) to specific tasks. In contrast, weight retraining alters the entire network in a way that cannot be easily undone. The extent of these changes during task retraining depends on the task itself and the training duration. In extreme cases, parameter retraining to restore initial capabilities may take just as long as the original pretraining.

The stochastic modulator's ability to quickly adapt to changing task circumstances may also prove beneficial in continual learning situations, especially when switching between tasks that share some task-relevant features. To test this idea, we continuously trained the same network on a sequence of digit-pair categorization tasks that share a common location as well as the identity of one of the two digits classes. In this case, we again find that the stochastic modulation model generally outperforms both weight retraining and deterministic gain boosts (Fig. 4.8B). Moreover, we see learning savings across episodes, with the later tasks requiring fewer examples to reach plateau performance. We quantify this effect by measuring the total number of trials required for criterion performance for tasks 2 and 3 in a 3-task sequence, for both continual learning and a control scenario where the same tasks are learned in isolation directly after pretraining (Fig.4.8C). The distribution of this measure of learning speed across different instantiations of the tasks is systematically lower for continuous learning than isolated learning (Fig.4.8C). Hence, task fine-tuning by stochastic modulation is even more effective during continual learning with across-task overlap.

4.3.2 Intermediate conclusion

In the numerical experiments presented here, we have used spatial locality as a convenient knob for directly controlling the informativeness bottlenecks in the network. Nonetheless, the mechanism also exploited informativeness in the shape of the relevant digits, whose feature locality was inherited from pretraining on the general task. This confirms that the idea of targeting task-relevant features is general, and applies to any aspect of the stimulus that the network encodes. Two important principles guide placement of the modulator. First, targeted layers should be task-specific representational bottlenecks (i.e. informativeness should be sparsely distributed in the population). Second, if multiple layers exhibit such structure (as in the example of a locally connected processing layer), then placing the modulator closer to the decision yields better performance, likely because the stochastic modulator has to propagate through fewer layers, and/or because the learning signals are also stronger when backpropagated through fewer layers (Srivastava et al., 2015). The sparsity of feature representations is determined through complex interactions between network architecture, statistics of the training data, and details of pretraining including the algorithm and choices of regularization. Networks whose feature representations are inherently localized in space, and across distinct channels may particularly benefit from stochastic modulation. Biologically, such feature maps are ubiquitous in cortical sensory processing. Spatially localized receptive fields with selectivity to different image features have been discovered and studied throughout the visual hierarchy, and the sparsity of the associated neural responses appears to be conserved in different areas (Rust and DiCarlo, 2012). Interestingly, Nienborg and Cumming 2014 find that V1 choice probabilities were significantly larger for an orientation discrimination task than a disparity discrimination task, suggesting that task-relevant feature maps are important for neurons to drive behav-

ior.

To conclude, there is growing interest in the machine learning community in developing more flexible, adaptive neural models. Attention mechanisms inspired by the brain have already been shown to improve performance of deep learning models (Lindsay, 2020), but both few-shot learning after distribution shifts and continual learning remain key open problems. Current machine learning algorithms approach these problems from many different angles, from optimizing the network's initial conditions for subsequent training (as in MAML, see Finn et al., 2017b), to probabilistically detecting changes in the input distribution (Wang et al., 2021), or learning segregate representations across tasks to begin with (Duncker et al., 2020; Kirkpatrick et al., 2017; Masse et al., 2018). Our model adds the paradigm of stochastic modulation to this list, opening the door for new biologically-inspired advances in machine learning.

4.4 Orientation discrimination of small localized gratings

We can use the modulated hierarchical network model to test the efficiency of the modulator in guiding information across processing stages in the experiments described in Chapter 2 and generate new experimental predictions. To actualize this, we put special emphasize on modeling known features of the visual processing hierarchy in the brain and slightly adapt the previously introduced network architecture. The previously trainable encoding layer of the network is replaced by a fixed V1-like set of localized oriented filters, whose responses are then propagated through two processing layers of neurons with increasing RF size, and finally read out by a decision stage (Fig. 4.9A; details in Suppl. Sec. 4.7.2). To reflect previous experience, connections between stages following the

encoding layer are again pre-trained (via backpropagation), to solve the general location-invariant digit classification task (Fig. 4.9A), in the absence of the modulator. As a result of this optimization the model is capable of discriminating complex visual features. Analogous to the V1 experiment, we use stochastic modulation to fine-tune this network to the task of discriminating the orientation of local gratings (Fig. 4.1A). We first adjust the decision circuit to the new data categories (10 possible orientations, see Suppl. Sec. 4.7.2 for details). Then the network needs to perform a binary discrimination task involving two orientations at a fixed location (Fig. 4.9C). As in the actual experiment, distractors are placed at other locations in the image, something which the network has not encountered during the previous episodes of learning.

The shared, stochastic gain modulation affects the encoding layer via its neuron-specific coupling parameters (the encoder gain) and the responses of neurons in the last layer are combined with the modulator-gated decoder gains g_n, which tune the readout of the decision circuit to the specific task (as in Fig. 4.1A and Fig. 4.2B). Again the rationale of this model is that if task-informative neurons can be modulator-labeled in the V1 stage, then this labeling will be inherited downstream by exactly those neurons that receive their signal. Thus their co-variability can guide decoding at the decision layer.

4.4.1 Performance in simulations

We assess the efficiency of the modulator-based solution by comparing it again to two alternative models, both of which adapt based on experience within the task, but differ in their parameter complexity. At one extreme, we consider the system that relearns the connection strengths between all layers de novo ("retraining"). At the other extreme, we consider a fixed network that only relearns the final decision layer weights ("readout only").

Retraining all network weights requires many training examples to reach good performance (defined as $> 80\%$ accuracy; Fig. 4.10A), likely due to the high dimensionality of the parameter space. Retraining only the readout results in poor performance, possibly because the presence of distractors renders the pre-trained representation insufficient for effective category discrimination. Compared to alternative models, fine-tuning the network via the modulator substantially reduces the amount of task-training required to reach criterion performance (Fig. 4.10A).

To disambiguate the effects of modulation on neural variability versus mean responses, we compare to the attentional gain model which deterministically boosts the encoder gain (Lindsay and Miller, 2018). We find that targeting of attentional gain modulation can be learned faster than retraining all the connections, but it does not reach the same performance as the stochastic modulator given limited training. This suggests that the separation of stimulus information and task relevance into two information channels, carried by the mean and variance of neural activity, respectively, further enhances the separability of the stimuli at the decision stage.

When investigating the properties of the learned solution, we find that in the encoding layer the learned couplings are highest for the most task-informative neurons (5% highest $|d'|$, Fig. 4.10B, see Suppl. 4.7.2.1 for details), similar to what we see in the data (Fig. 2.7). Although the modulator only affects the responses of these neurons directly, we find that informative neurons in the downstream processing layer are still preferentially correlated with the modulator (Fig. 4.10C). This suggests that task relevance can indeed propagate along the hierarchy in parallel to the stimulus information.

4.5 V1 modulator label is preserved in downstream MT

The hierarchical model predicts that task-specific modulation introduced in V1 should label task-informative neurons in downstream areas. We look for signatures of such labeling in simultaneously recorded MT activity included in some of the sessions. MT neurons are known to receive direct input from V1 (Maunsell and Van Essen, 1983) and selectively combine these afferents to construct their receptive field properties, such as motion selectivity (Born and Bradley, 2005; Movshon et al., 1986). Their receptive fields are larger and more complex, responding to localized gratings with different combinations of position, speed and orientation (Movshon et al., 1986; Simoncelli and Heeger, 1998). Given anatomical considerations, we expect correlated activity in V1 to drive MT to some extent. What is specific to our theory is the prediction that the degree of inherited modulation should reflect the task informativeness of individual MT units.

We find that responses of individually recorded MT units that cover the two relevant stimulus locations (Fig. 2.2A) vary in their task-informativeness (Fig. 4.11A) and show different degrees of supra-Poisson variability (Fig. 4.12A), suggesting different levels of modulation (Goris et al., 2014). The two measures are correlated across the MT units, with informative units having higher Fano factors (correlation coefficient of 0.48, $p < 0.008$). To test whether the excess variability arises due to V1 modulation, we compared two models of MT activity. The first is based on the visual stimuli alone ("SR"); it resembles the V1 SR model, but includes stimulus drift direction (consistent with previous literature Movshon et al., 1986, drift direction did not have predictive power for the V1 units, see also Ruff and Cohen, 2016a, but it did have a strong effect on MT activity). The second model additionally conditions on the (normalized) modulator extracted from the

model fitted to the simultaneously recorded V1 units ("SR+V1 modulation"; Fig. 4.12B). We verify the model fit over multiple cross-folds and find almost exclusively good model fits quantified by a comparison to a constant rate model through the pseudo-R^2 measure (Fig. 4.11B). This is expected given that experimental stimuli were optimized to drive the particular MT unit in a session. The inclusion of the V1-estimated modulator improved the fit for 73% of the MT units (measured as difference in pseudo-R^2, see Methods; Fig. 4.12C). Importantly, this effect is preferentially observed in task relevant units, which show a significantly larger model fit improvement relative to the uninformative units (t-test, $p = 0.01$; Fig. 4.12D).

While most V1 modulators extracted from the data show a strong targeting towards informative neurons (significant Spearman correlations between coupling and informativeness), a few outliers do not. We look for differences between targeted and untargeted modulators with respect to their predictability for MT. We find that only those V1 modulators that are well targeted to informative neurons have predictive power for their respective MT units (Fig. 4.11C); the few outlier blocks without structured targeting could not explain MT variance. This could be because of differences in fit qualities where for some blocks the estimated V1 modulator coupling is too noisy and hence does not reveal targeting and also prevents modulator estimates precise enough to be predictive of MT. Alternatively, the untargeted V1 modulators may reflect a different kind of shared noise in the population that is private and not propagated to MT.

The fact that both V1 and MT units are co-modulated as a function of their task informativeness is consistent with our theory, but does not exclude alternative patterns of information flow, such as top-down influences of MT on V1, or independent modulation of both areas from an external signal. To more directly address the nature of the modulation

in MT we take advantage of a smaller set of MT population recordings (partly published in Ruff and Cohen, 2016a). Despite the technical differences in recording procedure, this data recapitulates the same overall statistics, with 60% of the MT units having a significant part of their variability explained by the V1-estimated modulator. When independently extracting a modulator from the joint MT population responses ("SR+MT modulation"), we find that this population model better explains individual unit responses than the SR model (in 72 out of 73 blocks the modulated SR average fit is better, Suppl. 4.7.3). The extracted modulator has mostly consistent statistics across stimulus contrast variations (see Suppl. 4.7.3) and has similar time constants as those separately extracted in V1 (mean 61ms, s.d. 20ms). Lastly, there is a significantly positive correlation between modulator coupling and informativeness across blocks (Pearson r= 0.24, $p < 0.0001$, Fig. 4.13A), suggesting that the same structure seen in V1 is qualitatively replicated in MT responses. Are these properties inherited from V1? We find that the cross-correlogram of the V1 and MT-extracted modulators is maximal at a time lag that is consistent with feedfoward propagation from V1 to MT (Fig. 4.13B), however, additional data and finer temporal precision will be required to test the statistical significance of this relationship. Altogether, our analysis of MT responses supports the idea that the modulation of task-relevant neurons in V1 is shared with task-informative neurons in MT, allowing the propagation of labeling information towards decision areas.

4.6 Discussion

Hierarchical processing of sensory information allows sequential building of complex sensory representations, but also creates simultaneous sensory maps at different stages. As a consequence different dimensions of sensory information are explicitely represented at one

point and afterwards diffuse as new computations create representations along different dimensions. This has behavioral implications. Tasks that are based on information that is localized at one stage, instead of spread across different representations, can be solved easier and faster (Posner and Presti, 1987). Schneider and Logan (2009) define task switching as the dynamic selection of a task (and its representation) among many available ones, suggesting the switching between different available representations. However, given the hierarchical form of processing, an area where a decision needs to be formed does not physically connect to all previous representations of information. So how can information still be accessed flexibly and with great precision?

We have proposed a framework enabling flexible behavior, in which stochastic modulation adaptively and transiently fine-tunes the hierarchical processing of task-relevant features, while retaining a stable network 'backbone' for general computation. The key idea is that rapidly varying modulatory noise injected into the task-relevant subset of neurons in an early processing layer propagates through the feedforward synapses together with the primary stimulus information, and serves to guide the readout at the final decision stage. The selection of targeted neurons is learned from trial feedback on the current task. We explored the properties of this mechanism in multilayer feedforward networks trained on variants of MNIST digit classification. We found that task specific targeting of the modulator can be learned from small numbers of examples, yielding substantially more efficient task adaptation than attention-like deterministic gain modulation, or retraining of the feedforward network as a whole. Moreover, we found that modulation is most effective when injected into the layers in which task-specific information is concentrated in a small fraction of the neurons.

We tested this fine-tuning mechanism in an orientation discrimination task that seems

straightforward from the perspective of an ideal observer operating on V1 activity but becomes difficult for a downstream decision circuit that first, does not have ideal observer knowledge about previous stages, and second only has access to task information after it dissipates and combines with other irrelevant information across multiple stages of visual processing (see also Chapter 2). We show that targeted modulation in V1 allows both stimulus identity and task relevance to be carried in parallel and guide readout at the decision stage. In contrast, attentional gain boosts that combine both types of information into mean responses are less effective, consistent with previous modeling observations (Lindsay and Miller, 2018). Importantly, the changes introduced via the modulation are task-specific and ephemeral, allowing the network to instantly disengage from the task, and revert to the pre-task state by reducing the strength of the modulator. Finally, in the experimental data we found evidence for the propagation of the modulator extracted from V1 to informative neurons in downstream area MT, preserving the modulator label, as suggested by the theory.

4.6.1 Source of the modulator and its targeting structure

Our theory is agnostic regarding the source of the modulator, the means by which it is available to downstream circuits and the circuit mechanisms underlying its flexible task-specific targeting. Dynamic changes in shared noise structure across tasks could arise through either local circuit dynamics (Huang et al., 2019) or top-down mechanisms (Bondy et al., 2018; Haefner et al., 2016), and later propagate to downstream regions in parallel with the stimulus information. Given the sparsity of top-down connections relative to the full population size (at least, in V1), the reorganization of modulation likely needs to involve local recurrent dynamics. The initial targeting could exploit the

topographic organization of sensory codes present in some areas (in our experimental context, orientation-specific columns in V1), modulating spatially-localized clusters of neurons in V1. This has not been explored in the learning of the coupling here but it may reduce the amount of experience-driven fine-tuning required as it could decrease the parameter space.

4.6.2 Labeling for information processing in a hierarchy

We found that stochastic modulation signals injected early in a hierarchical network remains reasonably effective in labeling and guiding decoding several stages later. This was not a foregone conclusion: models of attention in deep convolutional networks for object categorization have documented instances when attentionally-induced increases in activity of early layers fail to propagate to decision circuits (Lindsay and Miller, 2018). Intuitively, deterministic gain modulation intermingles information about the stimulus (the responses) with information about task relevance (the gain), making it difficult or even impossible to disentangle the two at later stages of processing (Liu et al., 2009). In contrast, the variability signal of the stochastic modulator is essentially orthogonal to the stimulus information, and thus can serve as an accessible label for the relevant stimulus information, analogous to the role of the carrier signal in FM radio transmission. We showed here that such a label could propagate to and be used in downstream areas without further reliance on topographic localization later on. If this kind of spatially localized modulation was indeed an organizing principle of neural activity, it would predict that flexible decoding is most effective for tasks relying on sensory features that are localized in some brain area where they can then be labeled. 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. In support of this, 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 localized in the columnar organization. Moreover, in a task involving higher order features, Koren et al. (2020) found neural variability was high in V4, but not V1 suggesting that the modulator could target later stages of processing depending on the task. Future work is needed to determine the neural mechanisms that determine the location and form of targeting across tasks.

4.6.3 Limitations and future work

In the examples presented here, learning the targeting relies on backpropagation of error signals, which is not only biologically-unrealistic but also has the practical disadvantage that the learning signals get weaker as the depth of the network increases (Srivastava et al., 2015). This is a common problem in training of deep and convolutional neural networks, where clever architectural additions such as skip connections provide a way of speeding up learning (He et al., 2016; Huang et al., 2017). In the specific context of stochastic co-modulation we have the advantage that we only need to update the modulator couplings, and the intermediate backpropagation signals do not need to be represented explicitly. As such, it should be possible to train a separate network to directly generate the required signals in parallel to pretraining (since the backpropagation operations are architecture-specific, but not task-specific), similar to synthetic gradients (Jaderberg et al., 2017; Marschall et al., 2020). Once the learning signal is available in the modulated circuit, the update of individual modulator strengths is local and Hebbian in form, so it

could be implemented with synaptic plasticity.

The simple MNIST task performed by a 3-4 layer network here illustrated well that the modulator label can pass more than one or two layers but is insufficient to test propagation across many stages. It also limits the number of representations that may form in the network and consequently does not allow to study the optimal placement of the modulator beyond the second layer (first processing stage).

4.6.4 Outlook

The theoretical considerations and experimental evidence presented here suggest that the primary mechanism for task-specific information routing in the brain could be structured covariability, rather than increases in response amplitudes. It is likely that both processes contribute to the behavioral improvements we typically attribute to attention, with boosts in mean responses serving to improve the initial encoding of the stimulus, and targeted covariability facilitating task-specific signal transmission and decoding.

4.7 Supplement

4.7.1 Hierarchical information propagation with learned stochastic modulation

4.7.1.1 Training

The loss for pretraining and task fine-tuning is defined by the crossentropy function with 10-categories and parameters were optimized using backpropagation with Adam (Kingma and Ba, 2015). The learning rate for pretraining is 1e−4 with a batch size of 200 images and 20 batches used for training (resulting in a total of 4000 images).

For the task-training, the batch size is reduced to 2 images, to allow testing performance in the low-sample regime. The total number of batches may vary and is specified in each main text figure. The modulator-coupling learning (stochastic or attentional modulator) is stable for a learning rate of 1e−3 to 1e−4 (Fig.4.14). We use the slower learning rate of 1e−4. Given the small batch size and the many parameters that need to be adjusted, similar learning rates lead to unstable learning trajectories for retraining (Fig.4.14). We optimized the retraining learning rate hyperparameter so as to achieve stable learning, measured by the variance of the across-runs final performance. We used a grid search with log-spacing and found that a learning rate of 1e−6 provided a low-variance learning performance similar to that of pretraining and modulator based learning at their respective learning rates (measured in % correct, see Fig.4.14). For the task-learning there is an L_1-norm penalty term applied to weights and coupling ($\lambda = 0.1$). For the modulator learning, the coupling parameters where initialized i.i.d. from the uniform distribution $[0.9, 1.1]$.

4.7.1.2 Architectural variations

Here we provide an extended comparison of the 3-layer vs. a 4-layer networks (extensions of Fig. 4.5 in the main text). We trained both networks on the same 10 tasks, differing by digit pair and ran 10 experiments for each, differing by random seed. A direct comparison between the two architectures shows that the initial learning slopes and baseline performance for the stochastic modulator are very similar across the two architectures (Fig.4.15). Nonetheless, the 4-layer network tends to require more training to reach the criterion of 70%. The two networks' learning slopes, baseline performance and training to criterion correlate across tasks and simulations , e.g. those tasks that require more training in the 3-layer network also require more training in the 4-layer network (see Fig.4.16).

4.7.2 Fine-tuning to orientation discrimination

We use a 4 layer artificial neural network that maps an image stimulus with 3136 pixels to one out of 10 categories, corresponding to digits '0' to '9', or different orientations. The first encoding layer includes 2560 neurons, whose receptive fields are fixed and modeled as Gabor filters, with varying location and orientation uniformly distributed across space (16x16 spatial grid) and phases (10 orientations). The responses of these neurons propagate through two processing layers with 9000 and 7840 neurons, respectively, with spatially-local connections between layers. Responses in the second processing layer then map into the 10 task categories (fully connected). Neurons in the initial encoding layer have an exponential nonlinearity, to match the single layer decoding theory. Processing and decision layers use a more traditional rectified linear nonlinearity (ReLU).

The modulator affects encoding neurons through coupling terms c_n, which modulate the

neuron's stimulus-dependent responses, as described above for single layer decoding (but without the Poisson noise).

$$h_{n,t}^{(0)} = \exp\left(\mathbf{w}_n^{(0)}\mathbf{s} + m_t c_n\right), \tag{4.4}$$

where $h_{n,t}^{(0)}$ is the activity of neuron n in the encoding layer, $\mathbf{w}_n^{(0)}$ are the weights from the input to this neuron. Neurons in the last processing layer include a gain term g_n that multiplicatively scales their activity up or down, before it is read out by the decision layer:

$$h_{n,t}^{(2)} = g_n \mathrm{ReLU}\left(\mathbf{w}_n^{(2)}\mathbf{h}_t^{(1)} + b_n^{(2)}\right), \tag{4.5}$$

where $b_n^{(2)}$ is a neuron-specific bias, optimized together with the weights $\mathbf{w}_n^{(2)}$ during pre-training. The gain g_n is strictly positive and it is learned using the same MG correlation rule (Eq. 4.3).

There are three stages of learning. First, we pre-train the network weights (i.e. the connection weights of processing layers, $\mathbf{w}_n^{(1)}$ and $\mathbf{w}_n^{(2)}$ and the decision layer readout $\mathbf{w}_n^{(3)}$) to solve a digit classification task (locally placed MNIST digits (LeCun and Cortes, 2010) with image presentation and pixel-specific i.i.d. additive Gaussian noise on the background). In this stage the modulator is disabled (set to zero) and all gain terms are set to their default value, $g_n = 1$. Second, we keep network processing fixed, and train a new decision layer for orientation discrimination (10 orientation categories). During this stage, the input consists of images that include a single, local oriented grating at various positions on the screen (14x14 possible positions on the 16x16 grid, outer-most positions excluded) with background noise. Third, during the task condition we refine the resulting circuit to perform orientation discrimination in the presence of distractors. Specifically, we

set a fixed task location where one of two oriented gratings are shown and need to be differentiated, while varying distractor gratings are shown at other locations. The modulator is active; At the time scale of stimulus presentations, it is set to a constant value of $m = 1$ and the coupling strengths c_n are optimized by backpropagation to solve the task. However, at its own fast time scale t, the modulator varies with 100 time points per stimulus presentation and is modeled as independent Gaussian draws $m_t \sim \mathcal{N}(0, 0.1)$ ("stochastic modulator"). These fluctuations are then used to adapt the gains in the last processing layer in response to the modulator.

We compare the performance of our model ("stochastic modulator", 2560 parameters for backpropagation, 7840 parameters using MG gain adjustment as in Eq. 4.3) to three controls. The first allows full retraining of all connections in the network ("retraining", 256690 parameters). Second, only retrains the decision layer weights ("decision layer retraining", 78410 parameters). In both of these approaches the modulator is still zero and the gain terms unused. Third, all network weights are fixed, but the modulator is active $m = 1$ and the modulator coupling c_n are optimized for the task ("attentional modulator", 2560 parameters).

4.7.2.1 Informativeness analysis:

We assess the informativeness of neurons in the pretrained network using the task stimuli (oriented grating at task-location with distractor gratings at other locations and background noise). We quanitfy informativeness again using $|d'| = |\frac{\mu_1 - \mu_2}{\sqrt{0.5(\sigma_1^2 + \sigma_2^2}}|$ where $\mu_{1/2}$ are the mean responses to each task category respectively, and $\sigma_{1/2}^2$ is the variance in responses due to the distractors and the background noise.

4.7.3 Extension on MT population analysis

The modulated SR model is fit to a population of 24 units. The SR model includes direction and contrast. The fit of the SR model is good across all blocks compared to a constant rate model (4.17A). The modulator further improves this fit in all but one block (4.17B). The modulator is not significantly dependent on the stimulus contrast in 72% of blocks (4.17C). For the other 28% of blocks the modulated SR model does not manage to separate stimulus response and the modulator. We exclude those blocks from the analysis to avoid confounds.

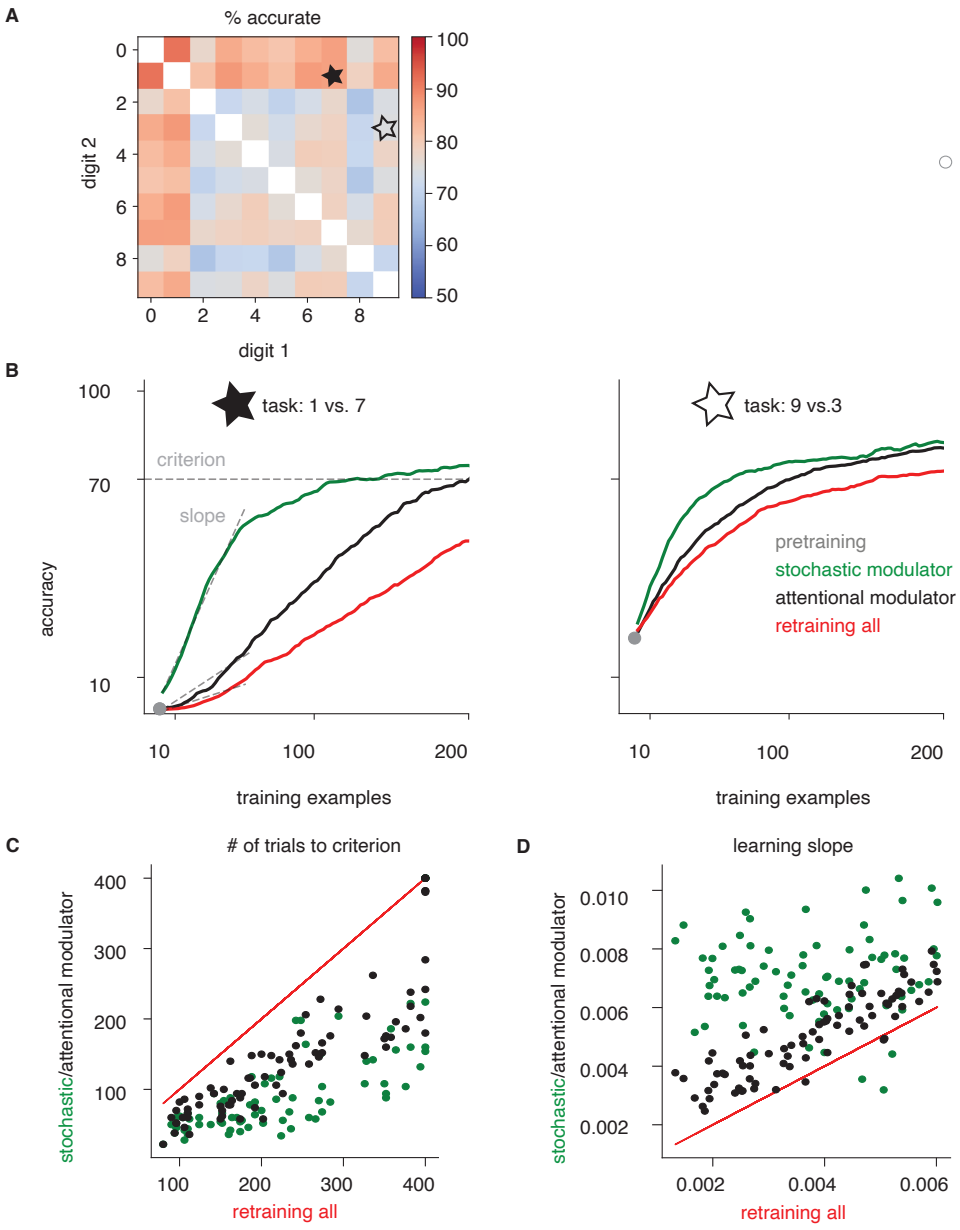

Figure 4.3 Performance comparison. A) Average performance (% correct) after pretraining, for discriminating digit pairs at any location without distractors. B) Two-digit classification accuracy for two example pairs. Grey dot indicates the baseline performance of the pretrained network. Lines represent averages over 10 simulations for each learning procedure. C) Number of training examples required to reach a criterion performance of 70% accuracy for the modulator-dependent methods compared to training needed when retraining all weights. D) Initial slope of performance improvement during learning over different two-digit classification tasks, relative to that of retraining. Slopes are estimated by linear regression on performance over the initial 50 training samples (indicated in B).

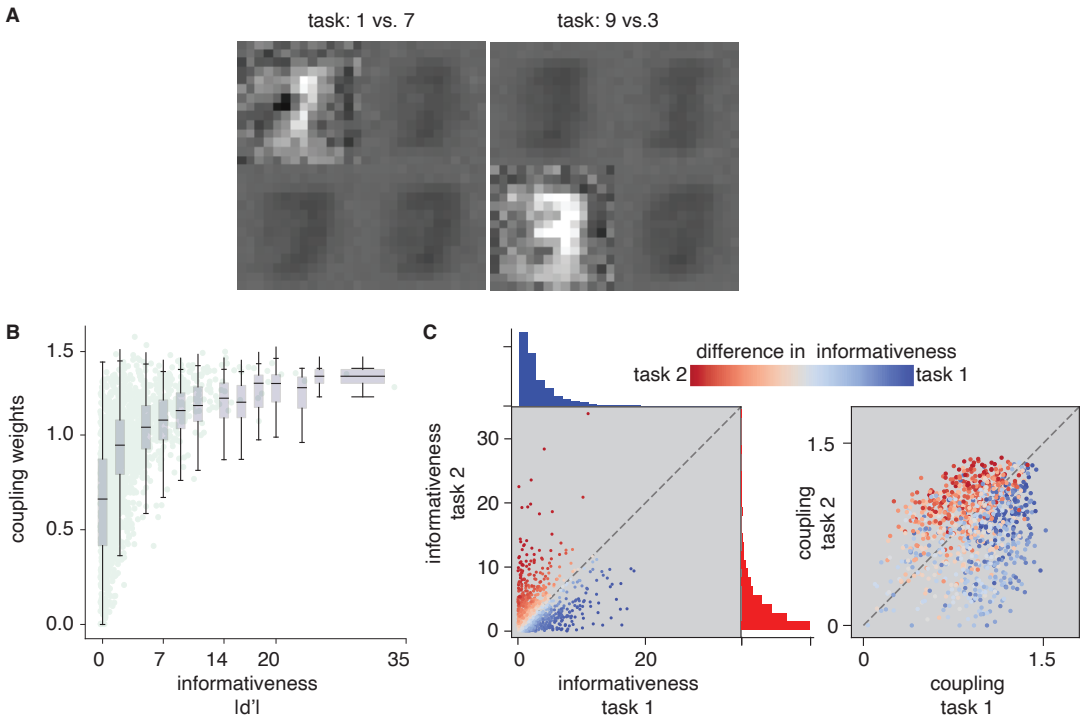

Figure 4.4 Learned coupling structure. A) Learned coupling strengths mapped back to the input space for two tasks involving different digits and locations; coupling strengths are standardized (z-scored) before averaging. B) Comparison of modulator coupling strength and informativeness ($|d'|$) for all first-stage neurons with receptive fields in the task-relevant input quadrant. C) Comparison of task informativeness of first-stage neurons in the task-relevant input quadrant for two tasks that involve different digit pairs within the same quadrant (left). Comparison of coupling strengths (right, same neurons, tasks, and colormap as left).

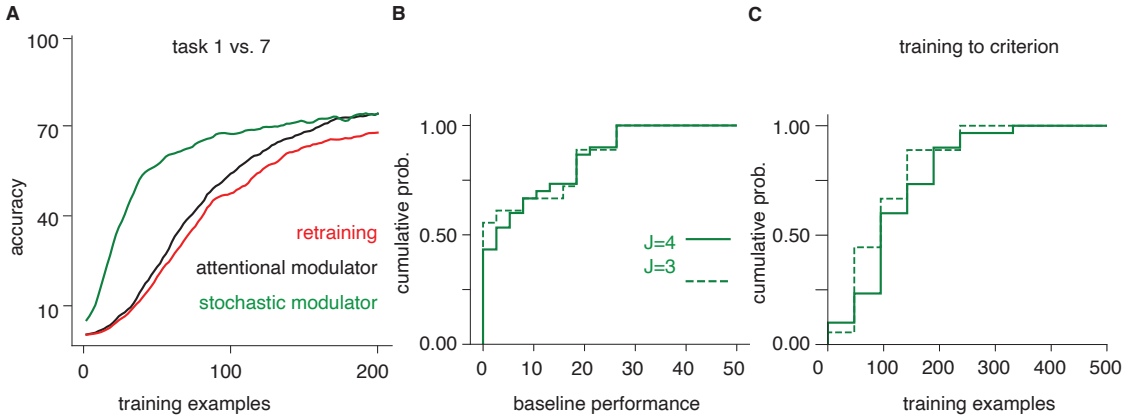

Figure 4.5 Stochastic modulation robust to changes in architecture. A) Performance comparison for architecture with two all-to-all intermediate layer. B) Distribution of baseline performance of the pretrained network for J=3 vs. J=4 layers. C) Corresponding distribution for the number of training examples needed to reach the 70% performance criterion.

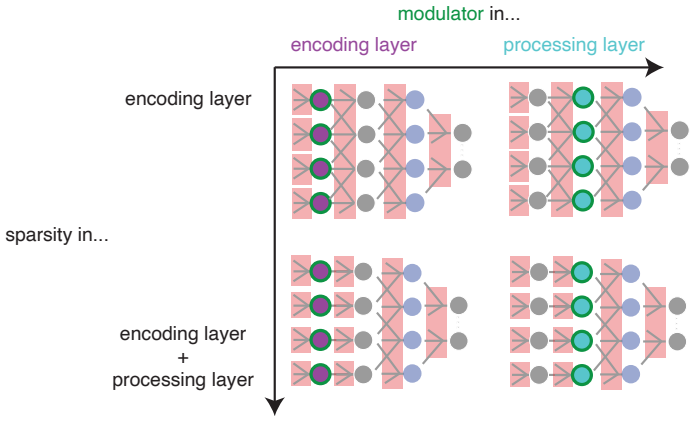

Figure 4.6 We vary the architecture along two dimension, sparse connectivity in either only the encoding layer or the encoding and processing layer, and modulation in either the encoding or the processing layer.

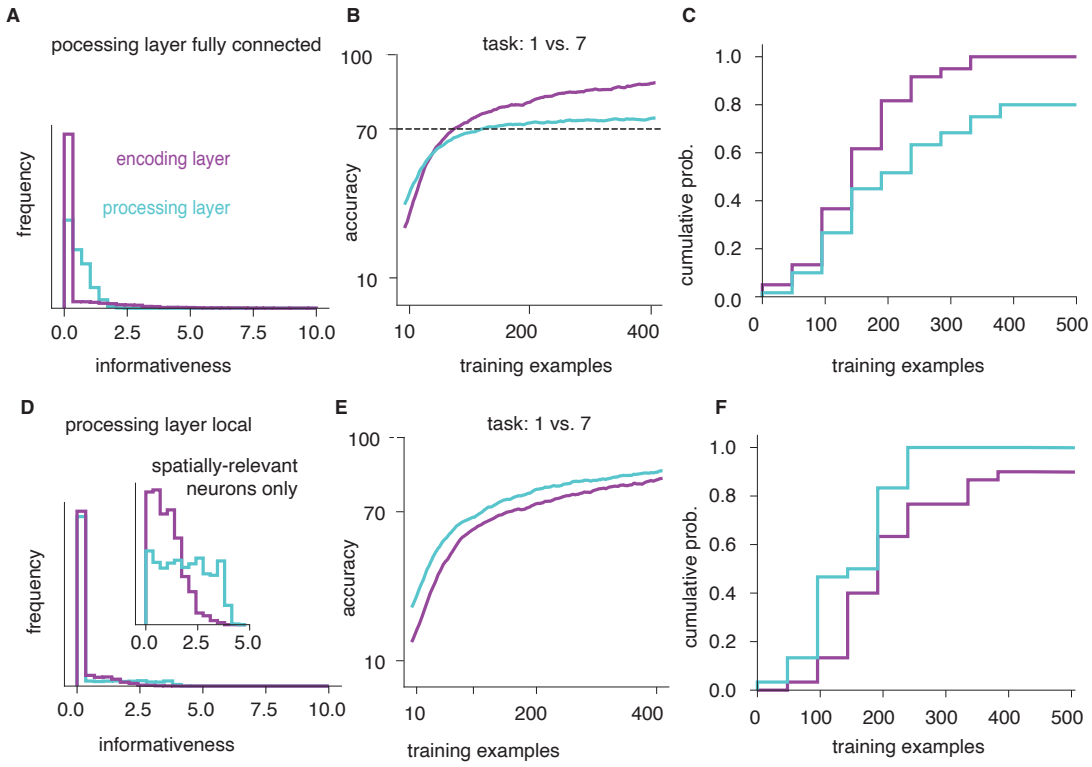

Figure 4.7 Stochastic modulation robust to changes in architecture. A) Informativeness ($|d'|$) distributions for encoding neurons (purple) and processing neurons (cyan) after pretraining when only the encoding layer is sparse and the processing layer is all-to-all connected; all neurons included. B) Comparing effects of directing modulation in either the encoding (purple) or the processing (cyan) layer with respect to accuracy on an example task. C) Number of training examples needed to reach criterion for many tasks. Network as in B&C. D) Same as A, but here the processing layer is locally connected. Additionally, the inset shows the distributions of informativeness for only those neurons that have their RF in the task-relevant location (spatially-relevant neurons only) E-F) Same as B-C, but with the network as in D.

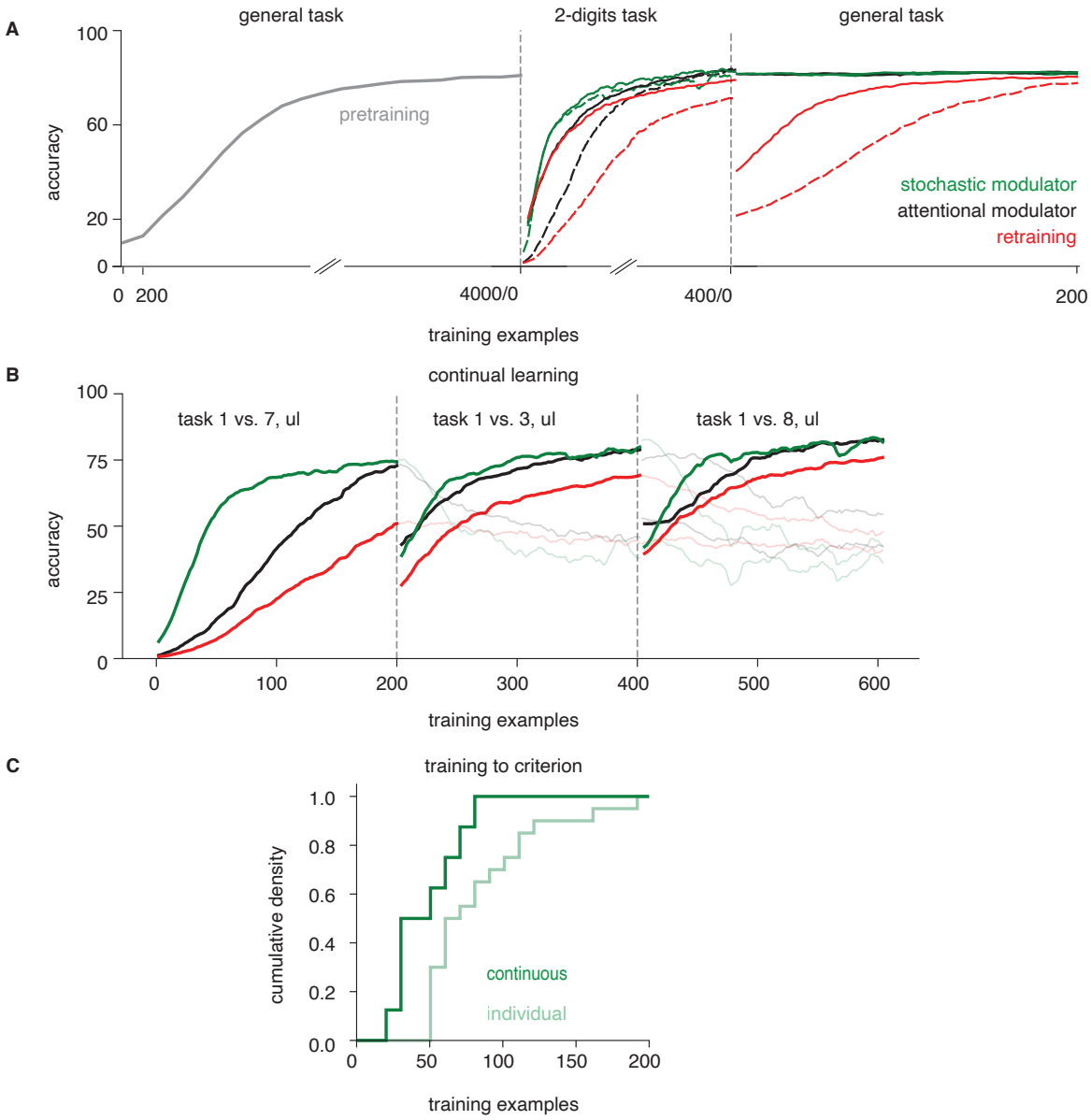

Figure 4.8 Online task switching. A) Evolution of the network's categorization accuracy over learning, from pretraining, to specific task, and returning to the general task; solid and dashed lines show results for two example tasks, respectively. B) A continual learning experiment with several task switches; all tasks include digit '1' as one of the categories and the same up-left location. Thick lines show performance for currently active task, thin likes track performance in the old tasks. C) Cumulative distributions of the number of training examples required to reach 70% performance if tasks are learned in sequence (dark green) or in isolation (light green); the first task was excluded from analysis.

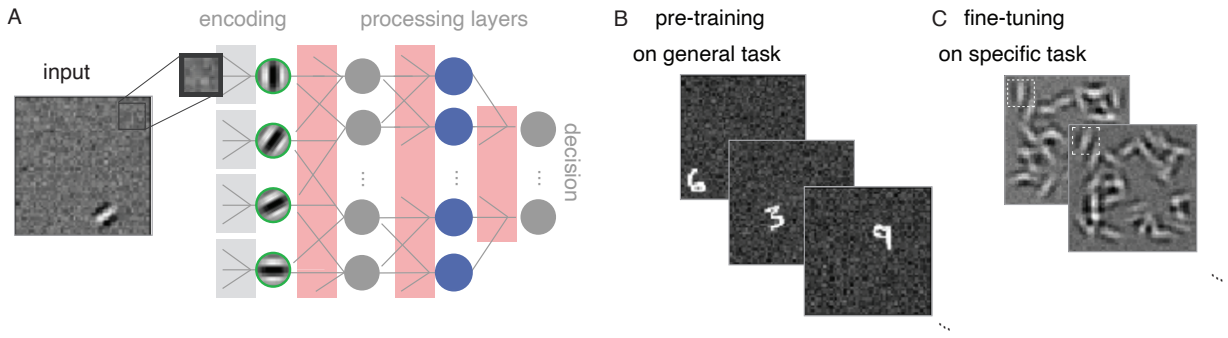

Figure 4.9 Fine tuning for an orientation discrimination task. A) Network with an encoding layer consisting of 2560 neurons with fixed Gabor filters with varying orientation and RF location, two locally connected processing layers and a fully connected decision layer. B) Pre-training on a spatially invariant version of the classic MNIST classification. C) Task training involves binary discrimination of grating orientation at a particular location in the presence of distractors.

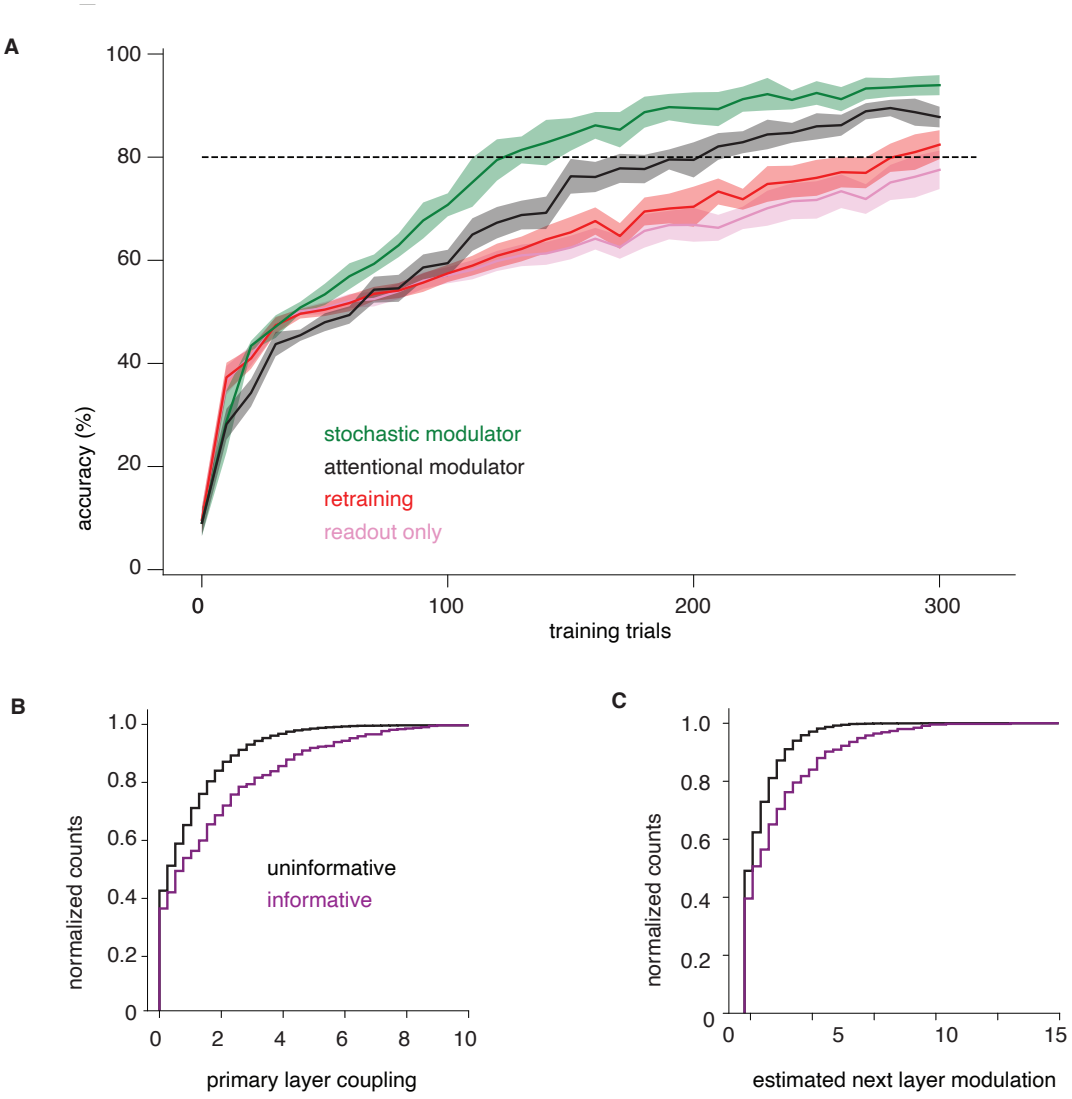

Figure 4.10 Performance in orientation discrimination task. A) Performance of different decoding strategies, as a function of the amount of data used for task training. B) Distribution of task-optimized modulator coupling for most informative neurons (5% highest $|d'|$ values) vs. all other neurons at the encoding layer. C) Estimated neuron-specific modulation at the first processing layer for most informative neurons vs. the rest.

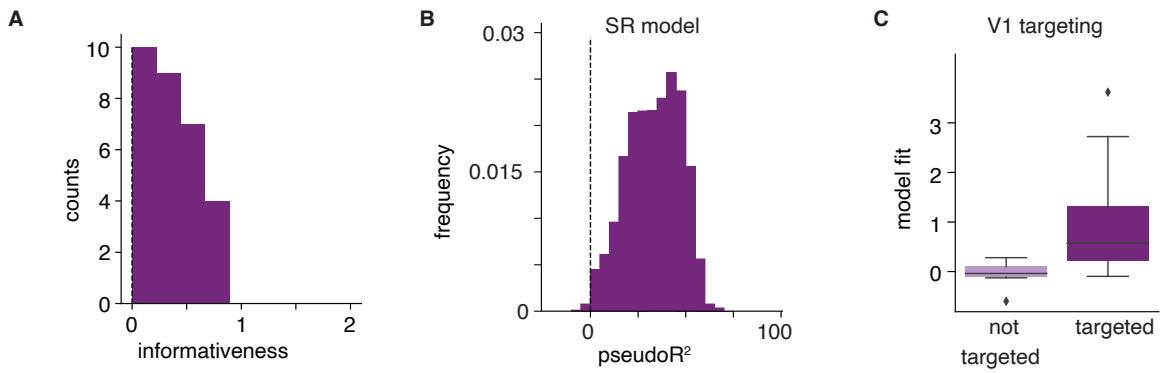

Figure 4.11 A) Distribution of informativeness ($|d'|$) over all MT units from the single unit recordings. B) The fit quality for the MT SR model is quantified by pseudoR which compares the log-likelihood of the SR model against a simpler constant rate Poisson model. We plot the distribution of pseudoR values for all units and different cross-folds. C) Blocks are split into subsets for which the estimated V1 modulator targets preferentially informative neurons (as measured by significant correlations between modulator coupling and informativeness) and blocks without significant targeting. We plot the respective distributions of model fit quality (pseudoR).

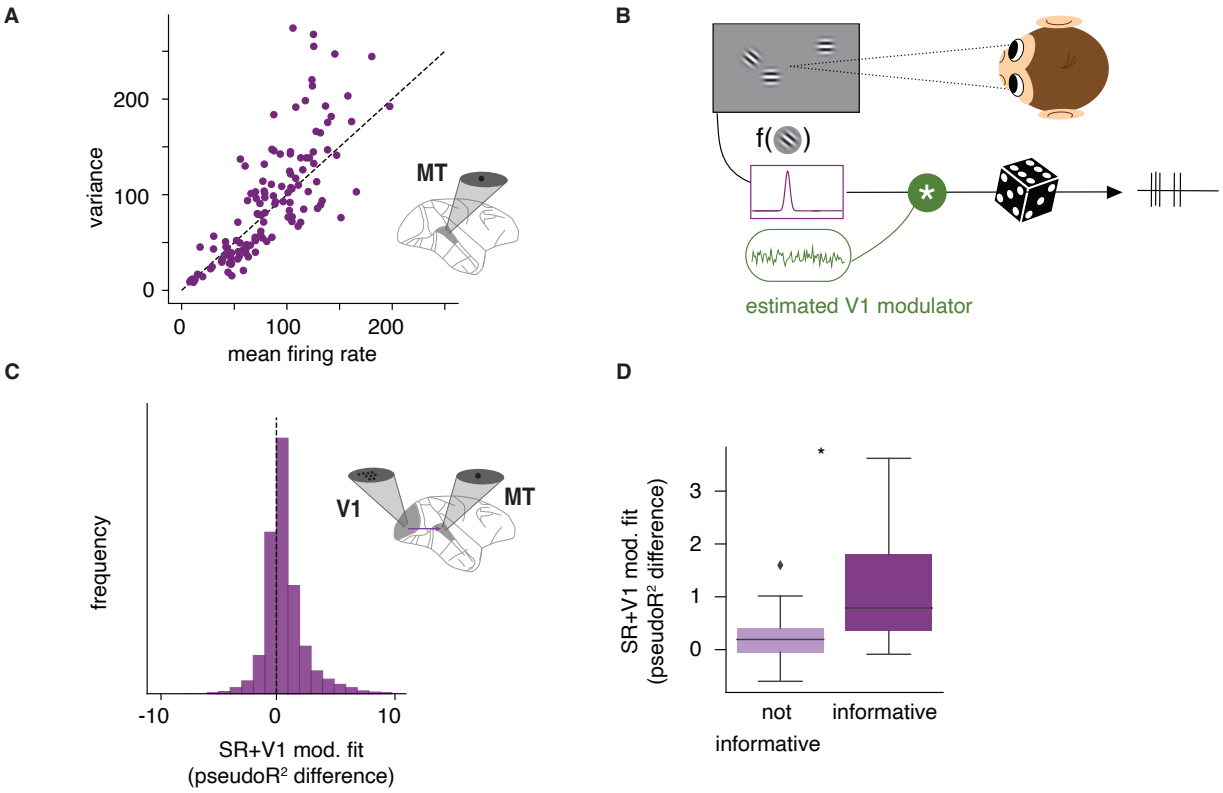

Figure 4.12 Effects of V1 modulator on simultaneously recorded MT units. A) Stimulus response variance as a function of mean firing for all MT units, and stimulus presentations. B) Schematic of the model; the spiking of each MT unit is specified by a tuning function potentially multiplicatively gated by the modulator estimated from V1 activity, with Poisson noise. C) Distribution of model fit (pseudo-R^2) values obtained by comparing the log-likelihood of the SR model that includes the V1 modulator as an additional dimension (SR+V1 modulation model) against the SR model. D) Improvement in fit quality for the SR+V1 modulation model, grouping MT units into those with high informativeness values (50% with highest $|d'|$) and those uninformative. Boxplot shows median and interquartile range. Black star indicates significant difference (t-test, $p = 0.01$).

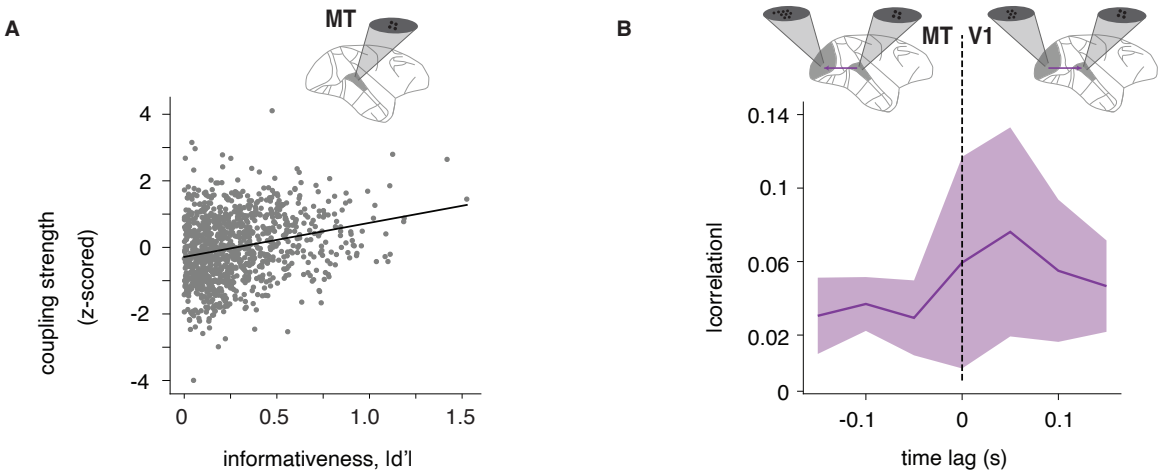

Figure 4.13 Targeted modulation in populations of MT units. A) A modulator is extracted from a population of MT cells. Shown are modulator couplings over informativeness in MT units over all 43 blocks. B) Correlations of the extracted V1 and MT modulators with positive (V1 before MT) and negative (MT before V1) time lag in seconds.

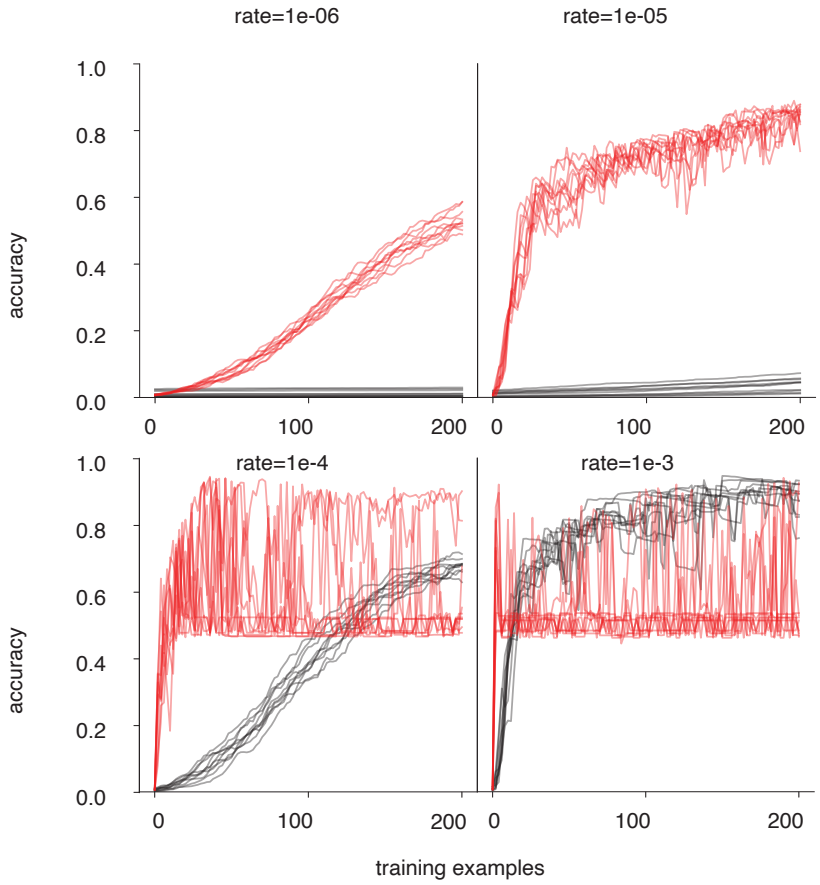

Figure 4.14 Retraining (red) and modulator coupling training (black) on a specific task (digit 1 vs. 7) with different learning rates for 50 batches of 2 images each.

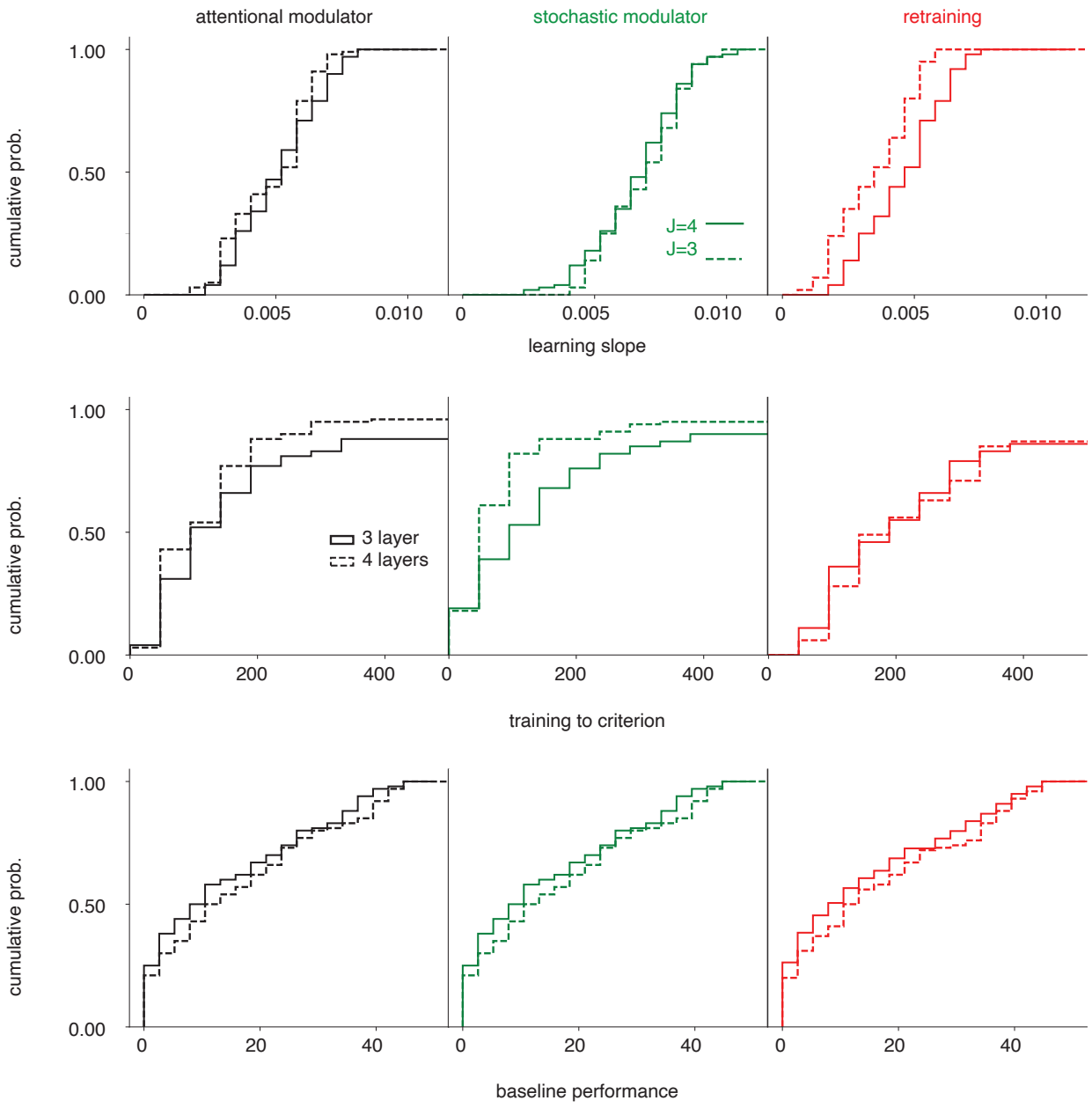

Figure 4.15 The cumulative distributions of learning slope, training to criterion and baseline performance for the 3-layer network (dashed line) and the 4-layer network (continuous line).

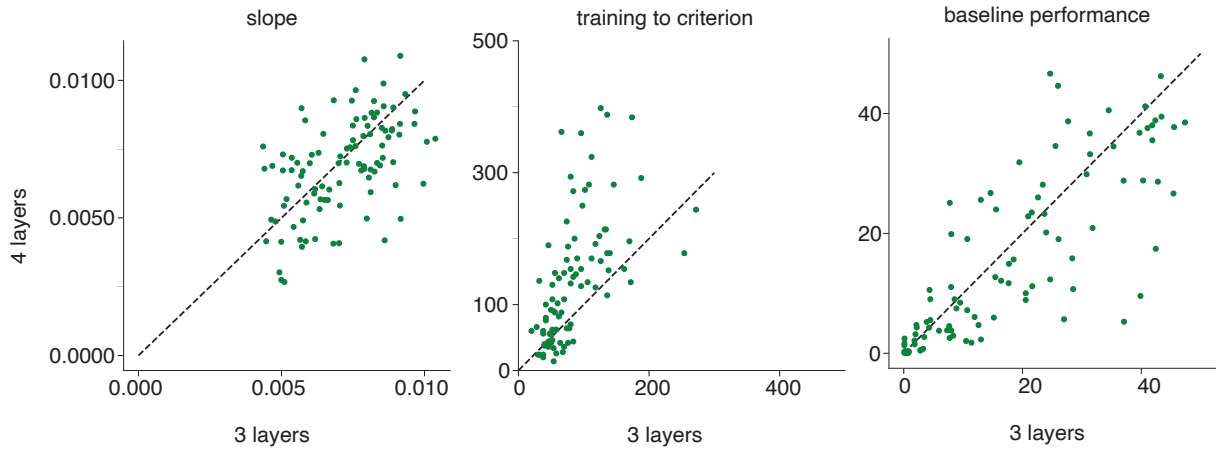

Figure 4.16 Comparing the 4-layer network against the 3-layer network with respect to their learning slope, training to criterion and baseline performance in 10 different tasks and 10 simulations each. Learning slope is measured for the first 50 training examples. Training to criterion is the number of training trials necessary to reach a minimum of 70% performance. Baseline performance is the performance of the pretrained network before any task-specific learning has happened.

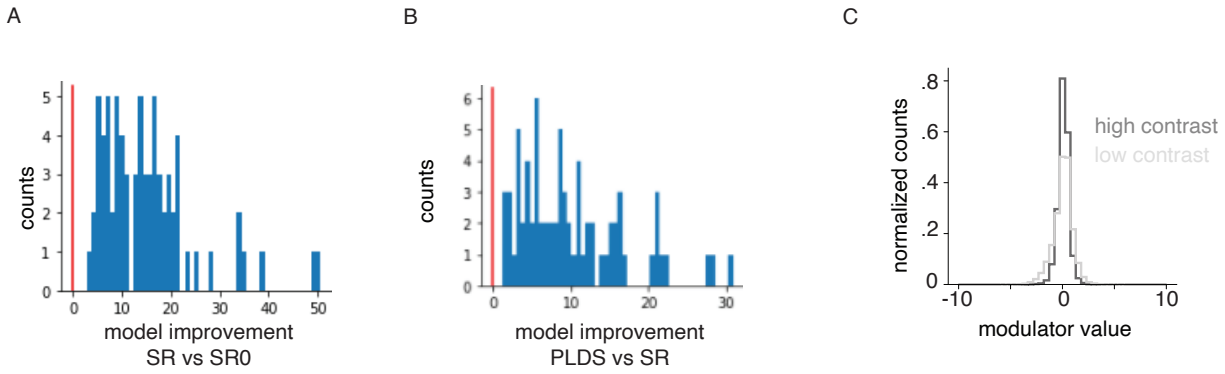

Figure 4.17 Model fit for population MT recordings. A) Average log-likelihood fit for the SR model for each block population compared to a constant rate model. B) Average log-likelihood fit for the modulated SR model for each block population compared to the SR model. C) The distribution of MT modulator values during high/low contrast stimulus presentations.

Chapter 5

Conclusion

Constructing neural representations is a fine balancing act between stability and plasticity: the brain is able to quickly adapt to new task demands, while maintaining performance in previous contexts. How is this accomplished? Resource constraints prevent the construction of *de novo* representations for each new task, while reorganization of existing synapses runs the risk of catastrophic loss of previous capabilities (Fusi et al., 2005; Kirkpatrick et al., 2017; Masse et al., 2018). Instead, the brain seems to achieve its balance by dynamically altering the flow of information through circuits, while keeping plastic changes to a minimum. The neural mechanisms of this process are not fully understood, but dynamic gain modulation is a ubiquitous aspect of neural activity (Carandini and Heeger, 2012; McCormick et al., 2020) and seems likely to play a critical role (see Chapter 2).

In Chapter 3 we introduced a novel theory where a stochastic top-down modulatory signal induces shared variability in neural responses and provides an information channel that carries a label for task relevance. This task-relevance channel is separate from but coexists with the stimulus-information channel carried by the strength of activity. By separating the encoded information from the task relevance, both can propagate through several pro-

cessing stages and provide the components required for decision making downstream, as illustrated in Chapter 4. We uncovered evidence for this labeling scheme in neural recordings obtained from non-human primate areas V1 and MT, while the animals switch between local orientation discrimination tasks at different spatial locations (Ruff and Cohen, 2016a,b). In Chapter 2, we found that recorded population activity in V1 exhibits fluctuations consistent with a shared modulator that preferentially targets task-informative neurons. In Chapter 3, we demonstrated through simulations that this can act as a functional label that guides decoding. We show in the V1 data that such "modulator-guided decoding" can be learned within just a handful of trials, facilitating fast readout from the population. By studying stochastic modulation in an artificial neural network model of the visual hierarchy in Chapter 4, we demonstrate that task information can be read out using the modulator label after multiple stages of processing and with minimal amounts of task-specific feedback. As predicted by the theory, we find that the modulatory signal extracted from the V1 population also modulates MT units and that task-informative MT units are most strongly modulated. These results support the hypothesis that the task-specific labeling is propagated through the visual hierarchy, facilitating downstream decisions and actions.

5.1 Outlook on future experimental work

The proposed role for modulation in flexible information routing presented here still leaves several open questions to be explored in future experimental and theoretical work. In order to shed light on the origin and mechanistic details of the modulation and its task-specific targeting structure, additional experiments are required. We here presented results that suggested that the modulator might propagate feedfoward from V1 to MT (see Chap-

ter 4). However, whether the modulator fluctuations themselves originate in a primary sensory area like V1, due to local circuit dynamics, or whether they are caused by a different modulatory brain area, such as thalamic nuclei that integrate sensory and top-down information (Purushothaman et al., 2012; Sampathkumar et al., 2021), requires careful experimental evaluation. For instance, Zénon and Krauzlis (2012) showed that superior colliculus (SC) inactivation disrupts the behavioral benefits of attention while keeping mean rate increases intact. One prediction of our theory would be that SC instead is involved in setting up the modulator label. Evidence for this hypothesis would be given if the inactivation of SC does cause interference with the modulator, for instance a decrease or strong increase in strength, or a weakening of its targeting structure (see theoretical results in Chapter 3). This would then suggest a key role of superior colliculus in the modulatory tagging mechanism. Once a potential source for the modulator is identified, theoretical predictions regarding the overall modulator strength could be tested. Specifically, the finding that both too weak or too strong a modulation negatively impacts performance by disrupting decoding/encoding precision (see Fig. 3.1) predicts negative behavioral effects when experimentally manipulating the modulation (e.g. shutting it off or strongly increasing it).

Learning the task-specific targeting structure likely requires input from brain areas involved in reward evaluation, additionally to local circuitry. One possible regulatory knob may be inhibition to the circuitry, which has been shown to change low-dimensional dynamics in a population (Huang et al., 2019). However, it is unclear whether these dynamics can create modulation that is task-specific and targeted. Specifically, two dimensions remain to be explored to understand modulator targeting, the spatial scale and the temporal scale. Regarding the spatial dimension, here we evaluated targeting at the single neuron level, but neural modulation tends to be broader in space (Shine et al., 2021),

suggesting a smoother targeting profile, so that groups of spatially localized neurons are modulated similarly. This could be a limitation for the precision with which information can be labeled. However, it could also act as a spatial regularizer to support the learning process, for instance if the spatial smoothness is matched to the spatial scale of the feature map present in the targeted brain area (such as orientation and space in V1). Regarding the temporal dimension, in order to better understand the dynamics of learning the targeting, how quickly targeting can emerge and change with task demands, future experiments could provide insight by tracking large populations across learning. Finally, there may be interesting interactions between spatial and temporal scale, for instance, it may be that a coarse reorganization of the modulator targeting happens fast, while a finer reorganization requires more extensive learning in the task. One important consideration when studying targeting is the importance of diverse populations as the precise scale of the targeting structure can only be studied if a variety of neurons (both informative and uninformative, strongly and weakly active) are included (for a more detailed discussion on the importance of diverse population recordings see Chapter 3).

Here we have provided evidence for targeted modulation in V1 and MT. Future experiments may address the question of whether this label is still found in higher order areas and how it's strength and precision changes. Experimental manipulation of shared noise structure at varying stages of processing may shed light on the behavioral importance of the modulator label. Specifically manipulating at early versus late stages could help to determine, how early in the hierarchy the label for a particular task first emerges. In a follow-up, there may be primary and secondary behavioral effects of neurally disrupting the modulator label, for instance, if the label is applied early but reinforced/reapplied later.

5.2 Outlook on future theoretical work

Similarly, on the theoretical side many open questions remain. Chapter 4 illustrated how the modulator label may propagate across several nonlinear processing stages, but it remains to be shown what the limits are on the number of stages that can be passed and what type of processing conserves or breaks the fluctuations. Here we showed that the optimal placement of the modulator in the network here is closely linked to the emergence of sparsely informative representations and an information bottleneck. If the modulator targeting strongly relies on this type of representations, two predictions can be formed. First, it may be limited or altogether impossible to apply a label in tasks for which there exists no sparse representation in the brain. Second, different tasks relying on different feature maps may require the modulator to be targeted to different brain areas. In order to study modulation in a network with different representations at different stages, we need to move to more complex, naturalistic tasks, that justify and can train deeper architectures.

Deeper hierarchical networks would also allow studying dependencies between the networks parameters and the depth of the network. For instance, it is possible that the optimal modulator mean and variance varies with the number of processing layers that need to be passed. Similarly, the effect of multiple nonlinear processing stages on the effective informativeness of neurons requires further study. We used $|d'|$ as a measure for neural informativeness, but the hierarchical setting requires taking into account the processing stages that follow an area. This complicates the quantification of informativeness due to the interactions between many neurons.

Finally, continual learning and biologically realistic learning of the targeting structure are two important directions for future work. Studying the modulator targeting in an online

learning context, where the network is fine-tuned for one task after another, will test the capacity and continuous flexibility of the mechanism. In Chapter 3 we outlined how the modulator-guided decoding could be learned with more biologically plausible eligibility traces, however, learning the modulator targeting in Chapter 4 still relies on backpropagation. Different approaches to learning modulation terms without detailed error propagation have been proposed in the literature. For instance, Stroud et al. (2018) learn a neuron-specific modulation via reward-based learning rules to modulate M1 movement controls, and Naumann et al. (2021) use a recurrent neural network trained on the feedforward network's input and errors to modulate the readout. Additionally local approximations to backpropagation may be employed by the brain to change weights in deep networks (Lillicrap et al., 2020). In the case of our targeted stochastic modulation, identifying more biologically plausible learning mechanisms will likely require experimental data that can provide additional constraints on possible origins of the modulator and on the temporal and spatial scale of learning the targeting.

5.3 Limitations and challenges

5.3.1 Beyond binary discrimination tasks

Our decoders are designed for a discrimination (as opposed to estimation) task because the experiments showing task-specific modulation have been done with binary discrimination tasks. In principle, it might be possible to extend this framework to estimation, which also entails learning to appropriately weight informative neurons while ignoring uninformative ones. Modulator-labeling should prove useful in this context, although the details of the decoder will likely change. More generally, labeling of task-information could facilitate

decoding whenever there exists a small subpopulation of neurons that carry this information. A single modulator's induced variability, however, may not be sufficient if information is broadly distributed across a population. For instance, one may imagine a multi-category classification task where informativeness of a neuron is not as clearly defined (a neuron's response could strongly signal the boundaries of one category to another but not show differences for other boundaries). Informativeness therefore is more diffuse, the more categories are involved and the advantage of a single label may decrease. Another theoretical strategy could be to employ different modulators, either per category or per category comparison. Several challenges may emerge here. First, there may be cross-talk between modulators and a biological limit to how many modulators may coexist and still allow reliable targeting that can be read out at later stages. Second, the targeting structure of several modulators quickly increases the dimensionality of the learning problem. Thirdly, different modulators would potentially need to be available to the read-out area in order to identify labeled neurons.

5.3.2 Fine versus coarse discrimination

The experiments in Chapter 2 involved fine discrimination of only a few 10s of degrees difference in orientation of a small localized grating. A coarse discrimination of orientation differences or substantially larger stimuli may recruit larger subpopulations of informative neurons and allow simpler decoding mechanisms such as the sign-only decoder or the rate-guided decoder to work well enough (see Chapter 3).

5.3.3 Cued tasks

Here we studied a blocked task design, which suggests a training pattern where the subject learns a task A, followed by task variations B, C, etc. and then learns to switch between them. A mixed task design where a cue indicates quick contextual task changes could answer questions on the capacity of circuits for building up and maintaining multiple task-specific modulators in parallel, and whether they could be employed as needed and directed by the cueing signal.

5.3.4 Task hierarchies

In Chapter 4 we considered learning the modulator's task-specific targeting structure for a general network with (fixed) pre-trained connectivity. We also gave an example of learning several tasks in continuation, one after the other, if those tasks share a common feature (here one digit and the task-relevant location). While here we found that learning the modulator targeting profited from previous task-learning, this does not necessarily have to be the case if the tasks have no similarity or even employ orthogonal populations of units. More broadly, given a hierarchy of tasks that are more or less similar with respect to the neural subpopulations that they engage, a prediction of our theory could be that those tasks that share informative neurons in the targeted population should have lower switching costs. Tasks that require very different feature maps, represented by different neural populations or brain areas, would even require a change in modulator targeting across areas. Given the unclear source of the modulator, it is not possible to make precise predictions regarding difficulty of such across-area changes, compared to within area changes. However, psychophysical experiments of switching between task-features represented by

different areas vs switching between task-features represented within an area by different subpopulations, could be a helpful direction for constraining the mechanisms of the modulator and how targeting may be learned/switched.

5.4 Broader impact

To conclude, the lack of a biologically plausible theory of neural decoding strongly limits our understanding of neural computation. Resolving the puzzle of how sensory information is routed through brain regions and extracted to perform specific tasks is critical for the study of sensory and cognitive dysfunction that involve decision making or attention, as well as clinical applications such as brain-computer interfaces (BCI) (Andersen et al., 2004). For example, our theory predicts that the strength of modulation is key for whether it can facilitate decoding or be detrimental (see Chapter 3). Hence, there might be a sensitive equilibrium, and abnormalities in shared variance (e.g. abnormal strength of oscillatory activity) could result in dysfunctions. The theory outlined here could potentially guide the study of such dysfunctions in a new direction, taking into account shared variability. Moreover, flexible task-dependent information routing in hierarchical networks is an unsolved problem in the rapidly-expanding field of machine learning and poses a fundamental obstacle for achieving adaptive artificial systems. Here we identified several key ingredients for labeling and using task-information in a hierarchical network, that may support flexibility in artificial systems as well. Our work provides a new conceptual framework for flexible decoding and information routing, and proposes a plausible solution for the brain, supported by both physiological data and computational theory.

References

Abbott, L. F. and Dayan, P. (1999). The Effect of Correlated Variability on the Accuracy of a Population Code. *Neural Computation*, 11(1):91–101.

Akam, T. and Kullmann, D. M. (2014). Oscillatory multiplexing of population codes for selective communication in the mammalian brain. *Nature Reviews Neuroscience*, 15(2):111–122.

Akam, T. E. and Kullmann, D. M. (2012). Efficient "communication through coherence" requires oscillations structured to minimize interference between signals. *PLoS Computational Biology*, 8(11).

Allen, C. and Stevens, C. F. (1994). An evaluation of causes for unreliability of synaptic transmission. *Proceedings of the National Academy of Sciences*, 91(22):10380–10383.

Altmann, E. M. (2004). The preparation effect in task switching: Carryover of soa. *Memory & Cognition*, 32(1):153–163.

Andersen, R. A., Musallam, S., and Pesaran, B. (2004). Selecting the signals for a brain–machine interface. *Current opinion in neurobiology*, 14(6):720–726.

Andrychowicz, M., Denil, M., Gomez, S., Hoffman, M. W., Pfau, D., Schaul, T., Shillingford, B., and De Freitas, N. (2016). Learning to learn by gradient descent by gradient descent. *Advances in neural information processing systems*, 29.

Archer, E. W., Koster, U., Pillow, J. W., and Macke, J. H. (2014). Low-dimensional models of neural population activity in sensory cortical circuits. *Advances in Neural Information Processing Systems 27*, 27:343–351.

Armbruster, D. J., Ueltzhöffer, K., Basten, U., and Fiebach, C. J. (2012). Prefrontal cortical mechanisms underlying individual differences in cognitive flexibility and stability. *Journal of cognitive neuroscience*, 24(12):2385–2399.

Averbeck, B. B., Latham, P. E., and Pouget, A. (2006). Neural correlations, population coding and computation. *Nature Reviews Neuroscience*, 7(5):358–366.

Beck, J. M., Ma, W. J., Pitkow, X., Latham, P. E., and Pouget, A. (2012). Not Noisy, Just Wrong: The Role of Suboptimal Inference in Behavioral Variability. *Neuron*, 74(1):30–39.

Benjamin, A. S., Fernandes, H. L., Tucker, T., Ramkumar, P., VerSteeg, C., Raeed, C., Lee, M., and Kording, K. P. (2017). Modern machine learning outperforms GLMs at predicting spikes. *bioRxiv*.

Berens, P., Ecker, A. S., Gerwinn, S., Toliasa, A. S., and Bethge, M. (2011). Reassessing optimal neural population codes with neurometric functions. *PNAS*, 108(11):4423–4428.

Berens, P., Ecker, A. S., James Cotton, R., Ma, W. J., Bethge, M., and Tolias, A. S. (2012). A fast and simple population code for orientation in primate V1. *Journal of Neuroscience*, 32(31):10618–10626.

Bertelson, P. and Aschersleben, G. (1998). Automatic visual bias of perceived auditory location. *Psychonomic Bulletin & Review*, 5(3):482–489.

Biró, S., Lasztóczi, B., and Klausberger, T. (2019). A visual two-choice rule-switch task for head-fixed mice. *Frontiers in Behavioral Neuroscience*, 13.

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

Born, R. T. and Bradley, D. C. (2005). Structure and function of visual area mt. *Annu. Rev. Neurosci.*, 28:157–189.

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

Bronkhorst, A. W. (2015). The cocktail-party problem revisited: early processing and selection of multi-talker speech. *Attention, Perception, and Psychophysics*, 77(5):1465–1487.

Buzsaki, G. and Draguhn, A. (2004). Neuronal oscillations in cortical networks. *science*, 304(5679):1926–1929.

Carandini, M., Demb, J. B., Mante, V., Tolhurst, D. J., Dan, Y., Olshausen, B. A., Gallant, J. L., and Rust, N. C. (2005). Do we know what the early visual system does? *Journal of Neuroscience*, 25(46):10577–10597.

Carandini, M. and Heeger, D. J. (2012). Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, 13(1):51–62.

Carrasco, M. (2011). Visual attention: The past 25 years. *Vision research*, 51(13):1484–1525.

Caruana, R. (1997). Multi-task learning. 28:75–88.

Cheng, K. Y. and Frye, M. A. (2020). Neuromodulation of insect motion vision. *Journal of Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioral Physiology*, 206(2):125–137.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the acoustical society of America*, 25(5):975–979.

Churchland, A. K., Kiani, R., Chaudhuri, R., Wang, X.-J., Pouget, A., and Shadlen, M. N. (2011). Variance as a signature of neural computations during decision making. *Neuron*, 69(4):818–831.

Cohen, M. R. and Kohn, A. (2011). Measuring and interpreting neuronal correlations. *Nature neuroscience*, 14(7):811–819.

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

Crick, F. (1989). The recent excitement about neural networks. 337(6203):129–132.

Cunningham, J. P. and Yu, B. M. (2014). Dimensionality reduction for large-scale neural recordings. *Nature Neuroscience*, 17(11):1500–1509.

Dayan, P. and Abbott, L. F. (2005). *Theoretical neuroscience*. MIT Press, Cambridge, MA.

Denfield, G. H., Ecker, A. S., Shinn, T. J., Bethge, M., and Tolias, A. S. (2018). Attentional fluctuations induce shared variability in macaque primary visual cortex. *Nature Communications*, 9(1).

DiCarlo, J. J. and Cox, D. D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences*, 11(8):333–341.

Drummond, N. and Niv, Y. (2020). Model-based decision making and model-free learning. *Current Biology*, 30(15):R860–R865.

Duncker, L., Driscoll, L., Shenoy, K. V., Sahani, M., and Sussillo, D. (2020). Organizing recurrent network dynamics by task-computation to enable continual learning. *Advances in Neural Information Processing Systems*, 33.

Duncker, L. and Sahani, M. (2021). Identifying computational dynamical activity from neural population recordings. *Current Opinion in Neurobiology*, 70:1–13.

Ecker, A. S., Berens, P., Cotton, R. J., Subramaniyan, M., Denfield, G. H., Cadwell, C. R., Smirnakis, S. M., Bethge, M., and Tolias, A. S. (2014). State dependence of noise correlations in macaque primary visual cortex. *Neuron*, 82(1).

Ecker, A. S., Denfield, G. H., Bethge, M., and Tolias, A. S. (2016). On the structure of population activity under fluctuations in attentional state. *Journal of Neuroscience*, 0(5):1–21.

Engel, T. A., Chaisangmongkon, W., Freedman, D. J., and Wang, X.-J. (2015). Choice-correlated activity fluctuations underlie learning of neuronal category representation. *Nature communications*, 6:6454.

Engel, T. A., Steinmetz, N. A., Gieselmann, M. A., Thiele, A., Moore, T., and Boahen, K. (2016). Selective modulation of cortical state during spatial attention. *Science*, 354(6316).

Felleman, D. J. and Van Essen, D. C. (1991a). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex (New York, NY: 1991)*, 1(1):1–47.

Felleman, D. J. and Van Essen, D. C. (1991b). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1(1):1–47.

Ferguson, K. A. and Cardin, J. A. (2020). Mechanisms underlying gain modulation in the cortex. *Nature Reviews Neuroscience*, 21(2):80–92.

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

Finn, C., Abbeel, P., and Levine, S. (2017a). Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR.

Finn, C., Abbeel, P., and Levine, S. (2017b). Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR.

Fiscella, M., Franke, F., Farrow, K., Müller, J., Roska, B., da Silveira, R. A., and Hierlemann, A. (2015). Visual coding with a population of direction-selective neurons. *Journal of Neurophysiology*, 114(4):2485–2499.

Franke, F., Fiscella, M., Sevelev, M., Roska, B., Hierlemann, A., and Azeredo da Silveira, R. (2016). Structures of Neural Correlation and How They Favor Coding. *Neuron*, 89(2):409–422.

Froudarakis, E., Berens, P., Ecker, A. S., Cotton, R. J., Sinz, F. H., Yatsenko, D., Saggau, P., Bethge, M., and Tolias, A. S. (2014). Population code in mouse V1 facilitates readout of natural scenes through increased sparseness. *Nature Neuroscience*, 17(6):851–857.

Fusi, S., Drew, P. J., and Abbott, L. F. (2005). Cascade models of synaptically stored memories. *Neuron*, 45(4):599–611.

Gao, Y., Buesing, L., Shenoy, K. V., and Cunningham, J. P. (2015). High-dimensional neural spike train analysis with generalized count linear dynamical systems. *Advances in Neural Information Processing System*, pages 1–9.

Geisler, W. S. and Albrecht, D. G. (1997). Visual cortex neurons in monkeys and cats: Detection, discrimination, and identification. *Visual Neuroscience*, 14(5):897–919.

Gerstner, W., Lehmann, M., Liakoni, V., Corneil, D., and Brea, J. (2018). Eligibility traces and plasticity on behavioral time scales: experimental support of NeoHebbian Three-Factor Learning Rules. *Frontiers in Neural Circuits*, 12(July):1–16.

Gilbert, C. D. and Li, W. (2013). Top-down influences on visual processing. *Nature Reviews Neuroscience*, 14(5):350–363.

Goris, R. L., Movshon, J. A., and Simoncelli, E. P. (2014). Partitioning neuronal variability. *Nature Neuroscience*, 17(6):858–865.

Graf, A. B. A., Kohn, A., Jazayeri, M., and Movshon, J. A. (2011). Decoding the activity of neuronal populations in macaque primary visual cortex. *Nature Neurosci.*, 14(2):239–245.

Gupta, S. K., Zhang, M., Wu, C.-C., Wolfe, J., and Kreiman, G. (2021). Visual search asymmetry: Deep nets and humans share similar inherent biases. *Advances in Neural Information Processing Systems*, 34.

Haefner, R. M., Berkes, P., and Fiser, J. (2016). Perceptual Decision-Making as Probabilistic Inference by Neural Sampling. *Neuron*, 90(3):649–660.

Haefner, R. M., Gerwinn, S., Macke, J. H., and Bethge, M. (2013). Inferring decoding strategies from choice probabilities in the presence of correlated variability. *Nature neuroscience*, 16(2):235–242.

Haimerl, C., Ruff, D. A., Cohen, M. R., Savin, C., and Simoncelli, E. P. (2021). Targeted comodulation supports flexible and accurate decoding in V1. *bioRxiv*.

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

Hair, J. F., Black, W. C., Babin, B. J., and Anderson, R. E. (2014). *Multivariate Data Analysis*. Pearson Education Limited, Essex, 7th edition.

Hartline, H. K. (1938). The response of single optic nerve fibers of the vertebrate eye to illumination of the retina. *American Journal of Physiology-Legacy Content*, 121(2):400–415.

Hayes, J. and Allinson, C. W. (1998). Cognitive style and the theory and practice of individual and collective learning in organizations. *Human relations*, 51(7):847–871.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer.

Hecht, S., Shlaer, S., and Pirenne, M. H. (1942). Energy, quanta, and vision. *The Journal of general physiology*, 25(6):819–840.

Hénaff, O. J., Boundy-Singer, Z. M., Meding, K., Ziemba, C. M., and Goris, R. L. (2020). Representation of visual uncertainty through neural gain variability. *Nature Communications*, 11(1).

Herzog, M. H. and Clarke, A. M. (2014). Why vision is not both hierarchical and feedforward. *Frontiers in Computational Neuroscience*, 8.

Hong, H., Yamins, D. L., Majaj, N. J., and DiCarlo, J. J. (2016). Explicit information for category-orthogonal object properties increases along the ventral stream. *Nature neuroscience*, 19(4):613–622.

Huang, C., Ruff, D., Pyle, R., Rosenbaum, R., Cohen, M., and Doiron, B. (2019). Circuit models of low-dimensional shared variability in cortical networks. *Neuron*, 101:1–12.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.

Hubel, D. and Wiesel, T. (1959). Receptive Fields of Single Neurones in the Cat's Striate Cortex. *Journal of Physiology*, 148:574–591.

Hutmacher, F. (2019). Why Is There So Much More Research on Vision Than on Any Other Sensory Modality? *Frontiers in Psychology*, 10(October).

Ionescu, T. (2012). Exploring the nature of cognitive flexibility. *New Ideas in Psychology*, 30(2):190–200.

Itti, L. and Koch, C. (2001). Computational modelling of visual attention. *Nature reviews neuroscience*, 2(3):194–203.

Izhikevich, E. M. (2007). Solving the distal reward problem through linkage of STDP and dopamine signaling. *Cerebral Cortex*, 17(10):2443–2452.

Jaderberg, M., Czarnecki, W. M., Osindero, S., Vinyals, O., Graves, A., Silver, D., and Kavukcuoglu, K. (2017). Decoupled neural interfaces using synthetic gradients. In *International Conference on Machine Learning*, pages 1627–1635. PMLR.

Jaegle, A., Mehrpour, V., Mohsenzadeh, Y., Meyer, T., Oliva, A., and Rust, N. (2019). Population response magnitude variation in inferotemporal cortex predicts image memorability. *eLife*, 8:1–12.

Jazayeri, M. and Movshon, J. A. (2006). Optimal representation of sensory information by neural populations. *Nature Neuroscience*, 9(5):690–696.

Jazayeri, M. and Movshon, J. A. (2007). A new perceptual illusion reveals mechanisms of sensory decoding. *Nature*, 446(7138):912–915.

Johansen-Berg, H. and Lloyd, D. (2000). The physiology and psychology of selective attention to touch. *Frontiers in bioscience : a journal and virtual library*, 5:D894–904.

Johnson, K. (1980). Sensory discrimination: neural processes preceding discrimination decision. *Journal of Neurophysiology*, 43(6):1793–1815.

Kang, I. and Maunsell, J. H. (2020). The correlation of neuronal signals with behavior at different levels of visual cortex and their relative reliability for behavioral decisions. *Journal of Neuroscience*, 40(19):3751–3767.

Kang, J., Wu, J., Smerieri, A., and Feng, J. (2010). Weber's law implies neural discharge more regular than a poisson process. *European Journal of Neuroscience*, 31(6):1006–1018.

Kanitscheider, I., Coen-Cagli, R., Kohn, A., and Pouget, A. (2015a). Measuring Fisher Information Accurately in Correlated Neural Populations. *PLoS Computational Biology*, 11(6):1–27.

Kanitscheider, I., Coen-Cagli, R., and Pouget, A. (2015b). Origin of information-limiting noise correlations. *Proceedings of the National Academy of Sciences of the United States of America*, 112(50):E6973–82.

Kiesel, A., Steinhauser, M., Wendt, M., Falkenstein, M., Jost, K., Philipp, A. M., and Koch, I. (2010). Control and interference in task switching-a review. *Psychological Bulletin*, 136(5):849–874.

Kingma, D. P. and Ba, J. L. (2015). Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, pages 1–15.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., and Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

Kobak, D., Brendel, W., Constantinidis, C., Feierstein, C. E., Kepecs, A., Mainen, Z. F., Qi, X. L., Romo, R., Uchida, N., and Machens, C. K. (2016). Demixed principal component analysis of neural population data. *eLife*, 5(APRIL2016):1–36.

Koren, V., Andrei, A. R., Hu, M., Dragoi, V., and Obermayer, K. (2020). Pairwise Synchrony and Correlations Depend on the Structure of the Population Code in Visual Cortex. *Cell Reports*, 33(6):108367.

Kriegeskorte, N. (2015). Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual review of vision science*, 1:417–446.

Kuchibhotla, K. V., Gill, J. V., Lindsay, G. W., Papadoyannis, E. S., Field, R. E., Sten, T. A. H., Miller, K. D., and Froemke, R. C. (2017). Parallel processing by cortical inhibition enables context-dependent behavior. *Nature neuroscience*, 20(1):62–71.

Lakshminarasimhan, K. J., Pouget, A., DeAngelis, G. C., Angelaki, D. E., and Pitkow, X. (2018). Inferring decoding strategies for multiple correlated neural populations. *PLOS Computational Biology*, 14(9):e1006371.

Lamme, V. A. and Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in neurosciences*, 23(11):571–579.

LeCun, Y. and Cortes, C. (2010). MNIST handwritten digit database.

Lee, I. and Lee, C. H. (2013). Contextual behavior and neural circuits. *Frontiers in Neural Circuits*, 7(APR 2013):1–21.

Liefooghe, B., Demanet, J., and Vandierendonck, A. (2009). Is advance reconfiguration in voluntary task switching affected by the design employed? *Quarterly journal of experimental psychology (2006)*, 62(5):850–857.

Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J., and Hinton, G. (2020). Backpropagation and the brain. *Nature Reviews Neuroscience*, 21(6):335–346.

Lin, I. C., Okun, M., Carandini, M., and Harris, K. D. (2015). The nature of shared cortical variability. *Neuron*, 87(3):644–656.

Lindsay, G. W. (2020). Attention in Psychology, Neuroscience, and Machine Learning. *Frontiers in Computational Neuroscience*, 14(April):1–21.

Lindsay, G. W. and Miller, K. D. (2018). How biological attention mechanisms improve task performance in a large-scale visual system model. *eLife*, pages 1–29.

Liu, T., Abrams, J., and Carrasco, M. (2009). Voluntary attention enhances contrast appearance. *Psychological science*, 20(3):354–362.

Ma, W. J., Beck, J. M., Latham, P. E., and Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, 9(11):1432–8.

Maboudi, K., Ackermann, E., de Jong, L. W., Pfeiffer, B. E., Foster, D., Diba, K., and Kemere, C. (2018). Uncovering temporal structure in hippocampal output patterns. *eLife*, 7:1–24.

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

Mante, V., Sussillo, D., Shenoy, K. V., and Newsome, W. T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, 503.

Markov, N. T., Vezoli, J., Chameau, P., Falchier, A., Quilodran, R., Huissoud, C., Lamy, C., Misery, P., Giroud, P., Ullman, S., et al. (2014). Anatomy of hierarchy: feedforward and feedback pathways in macaque visual cortex. *Journal of Comparative Neurology*, 522(1):225–259.

Marschall, O., Cho, K., and Savin, C. (2020). A unified framework of online learning algorithms for training recurrent neural networks. *Journal of Machine Learning Research*, 21(135):1–34.

Masse, N. Y., Grant, G. D., and Freedman, D. J. (2018). Alleviating catastrophic forgetting using context-dependent gating and synaptic stabilization. *Proceedings of the National Academy of Sciences*, 115(44):E10467—-E10475.

Maunsell, J. H. and Cook, E. P. (2002). The role of attention in visual processing. *Philos Trans R Soc Lond B Biol Sci.*, 357((1424)):1063–72.

Maunsell, J. H. R. and Van Essen, D. C. (1983). The connections of the middle temporal visual area (MT) and their relationship to a cortical hierarchy in the macaque monkey. *Journal of Neuroscience*, 3(12):2563–2586.

McAdams, C. J. and Maunsell, J. H. R. (1999). 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. *Neuron*, 23:765–773.

McCormick, D. A., Nestvogel, D. B., and He, B. J. (2020). Neuromodulation of brain state and behavior. *Annual review of neuroscience*, 43:391–415.

Meiran, N., Chorev, Z., and Sapir, A. (2000). Component processes in task switching. *Cognitive psychology*, 41(3):211–253.

Mitchell, J. F., Sundberg, K. A., and Reynolds, J. H. (2009). Spatial attention decorrelates intrinsic activity fluctuations in macaque area V4. *Neuron*, 63(6):879–888.

Mohan, K., Zhu, O., and Freedman, D. J. (2021). Interaction between neuronal encoding and population dynamics during categorization task switching in parietal cortex. *Neuron*, 109(4):700–712.e4.

Moran, J. and Robert, D. (1985). Selective attention gates visual processing in the extrastriate cortex. *Science*, 229(4715):782–784.

Moreno-Bote, R., Beck, J., Kanitscheider, I., Pitkow, X., Latham, P., and Pouget, A. (2014). Information-limiting correlations. *Nature Neuroscience*, 17(10):1410–1417.

Morris, R. (1984). Developments of a water-maze procedure for studying spatial learning in the rat. *Journal of neuroscience methods*, 11(1):47–60.

Movshon, J. A., Adelson, E. H., Gizzi, M. S., and Newsome, W. T. (1986). The analysis of moving visual patterns. In Chagas, C., Gattass, R., and Gross, C., editors, *Experimental Brain Research Supplementum II: Pattern Recognition Mechanisms*, pages 117–151. Springer-Verlag, New York.

Naumann, L. B., Keijser, J., and Sprekeler, H. (2021). Invariant neural subspaces maintained by feedback modulation. *bioRxiv*, page 2021.10.29.466453.

Ni, A. M., Ruff, D. A., Alberts, J. J., Symmonds, J., and Cohen, M. R. (2018). Learning and attention reveal a general relationship between neuronal variability and perception. *Science*, 465(January):1–28.

Niell, C. M. and Stryker, M. P. (2010). Modulation of visual responses by behavioral state in mouse visual cortex. *Neuron*, 65(4):472–479.

Nienborg, H. and Cumming, B. G. (2014). Decision-related activity in sensory neurons may depend on the columnar architecture of cerebral cortex. *Journal of Neuroscience*, 34(10):3579–3585.

Osborne, L. C., Lisberger, S. G., and Bialek, W. (2005). A sensory source for motor variation. *Nature*, 7057(437).

Pillow, J. W., Shlens, J., Paninski, L., Sher, A., Litke, A. M., Chichilnisky, E. J., and Simoncelli, E. P. (2008). Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, 454(7207):995–9.

Pitkow, X., Liu, S., Angelaki, D. E., DeAngelis, G. C., and Pouget, A. (2015). How Can Single Sensory Neurons Predict Behavior? *Neuron*, 87(2):411–423.

Posner, M. I. and Presti, D. E. (1987). Selective attention and cognitive control. *Trends in Neurosciences*, 10(1):13–17.

Purushothaman, G., Marion, R., Li, K., and Casagrande, V. A. (2012). Gating and control of primary visual cortex by pulvinar. *Nature neuroscience*, 15(6):905–912.

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

Ravizza, S. M. and Carter, C. S. (2008). Shifting set about task switching: Behavioral and neural evidence for distinct forms of cognitive flexibility. *Neuropsychologia*, 46(12):2924–2935.

Renart, A. and Machens, C. K. (2014). Variability in neural activity and behavior. *Current Opinion in Neurobiology*, 25:211–220.

Reynolds, J. H. and Chelazzi, L. (2004). Attentional modulation of visual processing. *Annual Review of Neuroscience*, (27):611–647.

Reynolds, J. H. and Heeger, D. J. (2009). The normalization model of attention. *Neuron*, 61(2):168–185.

Ritter, S., Faulkner, R., Sartran, L., Santoro, A., Botvinick, M., and Raposo, D. (2020). Rapid Task-Solving in Novel Environments. Technical report.

Ritter, S., Wang, J. X., Kurth-Nelson, Z., and Botvinick, M. (2018). Episodic Control as Meta-Reinforcement Learning.

Rodgers, C. C. and DeWeese, M. R. (2014). Neural correlates of task switching in prefrontal cortex and primary auditory cortex in a novel stimulus selection task for rodents. *Neuron*, 82(5):1157–1170.

Rogers, R. D. and Monsell, S. (1995). Costs of a Predictable Switch Between Simple Cognitive Tasks. *Journal of Experimental Psychology: General*, 124(2):207–231.

Roweis, S. and Ghahramani, Z. (1999). A unifying review of linear gaussian models. *Neural computation*, 11(2):305–345.

Ruff, D. A. and Cohen, M. R. (2014). Attention can either increase or decrease spike count correlations in visual cortex. *Nature neuroscience*, 17(11):1591–7.

Ruff, D. A. and Cohen, M. R. (2016a). Attention increases spike count correlations between visual cortical areas. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 36(28):7523–34.

Ruff, D. A. and Cohen, M. R. (2016b). Stimulus dependence of correlated variability across cortical areas. *Journal of Neuroscience*, 36(28):7546–7556.

Rumyantsev, O. I., Lecoq, J. A., Hernandez, O., Zhang, Y., Savall, J., Chrapkiewicz, R., Li, J., Zeng, H., Ganguli, S., and Schnitzer, M. J. (2020). Fundamental bounds on the fidelity of sensory cortical coding. *Nature*, 580(7801):100–105.

Rust, N. C. and Cohen, M. R. (2022). Priority coding in the visual system. *Nature Reviews Neuroscience*, 0123456789.

Rust, N. C. and DiCarlo, J. J. (2010). Selectivity and tolerance ("invariance") both increase as visual information propagates from cortical area V4 to IT. *Journal of Neuroscience*, 30(39):12978–12995.

Rust, N. C. and DiCarlo, J. J. (2012). Balanced increases in selectivity and tolerance produce constant sparseness along the ventral visual stream. *Journal of Neuroscience*, 32(30):10170–10182.

Salinas, E. and Abbott, L. (1997). Invariant visual responses from attentional gain fields. *Journal of Neurophysiology*, 77(6):3267–3272.

Salinas, E. and Thier, P. (2000). Gain modulation: a major computational principle of the central nervous system. *Neuron*, 27(1):15–21.

Sampathkumar, V., Miller-Hansen, A., Sherman, S. M., and Kasthuri, N. (2021). Integration of signals from different cortical areas in higher order thalamic neurons. *Proceedings of the National Academy of Sciences*, 118(30).

Scarpina, F. and Tagini, S. (2017). The stroop color and word test. *Frontiers in Psychology*, 8(APR):1–8.

Schneider, D. W. and Logan, G. D. (2009). Task Switching. *Encyclopedia of Neuroscience*, pages 869–874.

Semedo, J. D., Jasper, A. I., Zandvakili, A., Krishna, A., Aschner, A., Machens, C. K., Kohn, A., and Yu, B. M. (2022). Feedforward and feedback interactions between visual cortical areas use different population activity patterns. *Nature communications*, 13(1):1–14.

Shadlen, M. N., Britten, K. H., Newsome, W. T., and Movshon, J. A. (1996). A computational analysis of the relationship between neuronal and behavioral responses to visual motion. *Journal of Neuroscience*, 76(4):1486–1510.

Sherman, S. M. and Guillery, R. (1998). On the actions that one nerve cell can have on another: distinguishing "drivers" from "modulators". *Proceedings of the National Academy of Sciences*, 95(12):7121–7126.

Shine, J. M., Müller, E. J., Munn, B., Cabral, J., Moran, R. J., and Breakspear, M. (2021). Computational models link cellular mechanisms of neuromodulation to large-scale neural dynamics. *Nature Neuroscience*, 24(6):765–776.

Simoncelli, E. P. and Heeger, D. J. (1998). A model of neuronal responses in visual area MT. *Vision Research*, 38(5):743–761.

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

Snyder, D. L. and Miller, M. I. (2012). *Random point processes in time and space*. Springer Science & Business Media.

Spence, C. (2009). Explaining the colavita visual dominance effect. In Srinivasan, N., editor, *Attention*, volume 176 of *Progress in Brain Research*, pages 245–258. Elsevier.

Srivastava, R. K., Greff, K., and Schmidhuber, J. (2015). Training very deep networks. *arXiv preprint arXiv:1507.06228*.

Stein, R. B., Gossen, E. R., and Jones, K. E. (2005). Neuronal variability: Noise or part of the signal? *Nature Reviews Neuroscience*, 6(5):389–397.

Stringer, C., Pachitariu, M., Steinmetz, N., Carandini, M., and Harris, K. D. (2019). High-dimensional geometry of population responses in visual cortex. *Nature*, 571(7765):361–365.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions.

Stroud, J. P., Porter, M. A., Hennequin, G., and Vogels, T. P. (2018). Motor primitives in space and time via targeted gain modulation in cortical networks. *Nature Neuroscience*, 21(12):1774–1783.

Tingley, D., Alexander, A. S., Kolbu, S., de Sa, V. R., Chiba, A. A., and Nitz, D. A. (2014). Task-phase-specific dynamics of basal forebrain neuronal ensembles. *Frontiers in Systems Neuroscience*, 8(SEP):1–15.

Tinsley, J. N., Molodtsov, M. I., Prevedel, R., Wartmann, D., Espigulé-Pons, J., Lauwers, M., and Vaziri, A. (2016). Direct detection of a single photon by humans. *Nature Communications*, 7.

Tolman, E. C., Ritchie, B. F., and Kalish, D. (1946). Studies in spatial learning. i. orientation and the short-cut. *Journal of experimental psychology*, 36(1):13.

Treue, S. and Martínez Trujillo, J. C. (1999). Feature-based attention influences motion processing gain in macaque visual cortex. *Nature*, 399(6736):575–579.

Treue, S. and Maunsell, J. H. R. (1996). Attentional modulation of visual motion processing in cortical areas MT and MST. *Nature*, 382(1983):539–541.

Tsotsos, J. K., Kotseruba, I., and Wloka, C. (2019). Rapid visual categorization is not guided by early salience-based selection. *PLoS ONE*, 14(10):1–23.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Veuthey, T., Derosier, K., Kondapavulur, S., and Ganguly, K. (2020). Single-trial cross-area neural population dynamics during long-term skill learning. *Nature communications*, 11(1):1–15.

Vinck, M., Batista-Brito, R., Knoblich, U., and Cardin, J. A. (2015). Arousal and locomotion make distinct contributions to cortical activity patterns and visual encoding. *Neuron*, 86(3):740–754.

Wang, D., Shelhamer, E., Liu, S., Olshausen, B., and Darrell, T. (2021). TENT: Fully test-time adaptation by entropy minimization. *ICLR*.

Wang, J. X. (2017). Meta-learning in natural and artificial intelligence. *Current Opinion in Behavioral Sciences*, 38:90–95.

Whiteway, M. R., Averbeck, B., and Butts, D. A. (2020). A latent variable approach to decoding neural population activity. *bioRxiv*.

Willett, F. R., Avansino, D. T., Hochberg, L. R., Henderson, J. M., and Shenoy, K. V. (2021). High-performance brain-to-text communication via handwriting. *Nature*, 593(7858):249–254.

Woolgar, A., Afshar, S., Williams, M. A., and Rich, A. N. (2015). Flexible Coding of Task Rules in Frontoparietal Cortex: An Adaptive System for Flexible Cognitive Control. *Journal of Cognitive Neuroscience*, 27(10):1895–1911.

Wu, A., Roy, N. A., Keeley, S., and Pillow, J. W. (2017). Gaussian process based nonlinear latent structure discovery in multivariate spike train data. *Advances in neural information processing systems*, 30.

Yamins, D. L. and DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365.

Yang, G. R., Joglekar, M. R., Song, H. F., Newsome, W. T., and Wang, X. J. (2019). Task representations in neural networks trained to perform many cognitive tasks. *Nature Neuroscience*, 22(2):297–306.

Yu, B. M., Cunningham, J. P., Santhanam, G., Ryu, S. I., Shenoy, K. V., and Sahani, M. (2009). Gaussian-Process Factor Analysis for Low-Dimensional Single-Trial Analysis of Neural Population Activity. *Journal of Neurophysiology*, 102(April 2009):614–635.

Zénon, A. and Krauzlis, R. J. (2012). Attention deficits without cortical neuronal deficits. *Nature*, 489(7416):434–437.

Zhao, Y. and Park, I. M. (2017). Variational latent gaussian process for recovering single-trial dynamics from population spike trains. *Neural computation*, 29(5):1293–1316.

Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C., DiCarlo, J. J., and Yamins, D. L. (2021). Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3).

Zohary, E., Shadlen, M. N., and Newsome, W. T. (1994). Correlated neuronal discharge rate and its implications for psychophysical performance. *Nature*, 370(6485):140–143.