Response sub-additivity and variability quenching in visual cortex

Robbe L. T. Goris[1†], Ruben Coen-Cagli[2], Kenneth D. Miller[3], Nicholas J. Priebe[4], Máté Lengyel[5,6]

[1]Center for Perceptual Systems, University of Texas at Austin, Austin, Texas, USA [2]Department of Systems and Computational Biology, Albert Einstein College of Medicine, Bronx, NY, USA [3]Center for Theoretical Neuroscience, Swartz Program in Theoretical Neuroscience, Kavli Institute for Brain Science, and Dept. of Neuroscience, College of Physicians and Surgeons and Morton B. Zuckerman Mind Brain Behavior Institute, Columbia University, NYC, NY, USA [4]Center for Learning and Memory, University of Texas at Austin, Austin, Texas, USA [5] Computational and Biological Learning Lab, Department of Engineering, University of Cambridge, Cambridge, UK [6] Center for Cognitive Computation, Department of Cognitive Science, Central European University, Budapest, Hungary

† e-mail: Robbe.Goris@utexas.edu

Abstract | Sub-additivity and variability are ubiquitous response motifs in primary visual cortex (V1). Response sub-additivity provides a sign of the brain processes that enable us to construct useful interpretations of the visual environment (*i.e.*, nonlinear input transformations), while response variability provides a sign of the brain processes that limit the precision with which we can do this (*i.e.*, neural information loss). Historically, these two motifs have been studied independently of each other. Yet, there is increasing evidence that experimental manipulations that elicit response sub-additivity often also quench response variability. Here we provide a unifying review of these phenomena, suggesting that response sub-additivity and variability quenching may have a common origin. We review empirical findings as well as recent model-based insights into the functional operations, computational objectives, and circuit mechanisms underlying V1 activity. Although these modeling approaches address different aspects of cortical activity, they all predict that response sub-additivity and variability quenching will often co-occur. Response sub-additivity and variability quenching are not limited to V1 but are widespread cortical phenomena. Many of the insights we review generalize to other cortical areas, suggesting that the connection between response sub-additivity and variability quenching may be a canonical motif across cortex.

The primary visual cortex (V1) has long been a key model system for studying cortical circuitry and computations. In recent decades, two major foci of study in V1 have been response sub-additivity and response variability. Response sub-additivity involves phenomena in which neuronal responses to two simultaneously presented stimuli are less than the sum of the responses to the two stimuli presented independently (sometimes also referred to as sublinear response summation). For example, V1 cells have distinct spatial receptive fields, classically defined as the locations in visual space where a stimulus elicits an increase in activity (Fig 1a, green circle). Beyond this classical receptive field lies the receptive field surround. Although ineffective by itself, stimulation of the surround often suppresses the response to an effective stimulus within the receptive field (Fig. 1a, grey vs black). Similarly, for many neurons, doubling the contrast of a 50% contrast stimulus confined to the classical receptive field does not double the neural response. Indeed, sub-additive effects occur beyond [1–4] as well as within [5–9] the classical receptive field, are broadly tuned in the orientation domain [3,6,7,9], and are better described by division than by subtraction [10,11]. Response variability involves phenomena in which repeated presentations of the same stimulus elicit variable responses in cortical cells (Fig. 1a, rasters). This variability is evident both in cells' membrane potential [12–14] and in their spiking activity [15–17]. Response variability in visual cortex appears largely random [18,19], exhibits strong stimulus dependence [20,21], has non-trivial spatiotemporal structure [22–27], and is often well described by a doubly stochastic process of spike generation [28–33].

Response sub-additivity and variability feature prominently in the literature because they provide directly observable indications of the brain's processes that enable us to perform natural perceptual tasks on the one hand (*i.e.*, nonlinear neural transformations [34,35]), and of the processes that are traditionally believed to limit our ability to do so on the other hand (*i.e.*, neural information loss [36,37]). They have largely been studied independently from each other. Yet, it has become apparent that experimental manipulations that elicit sub-additivity often change variability too. In particular, response sub-additivity often co-occurs with variability quenching. This has been observed for experimental manipulations within [21,33] and beyond the classical receptive field [38–41] (Fig 1b,c). Is this mere coincidence? We propose it is not. Recent theoretical studies of the functional operations, representational objectives, and circuit mechanisms underlying V1 activity all suggest that response sub-additivity and variability quenching are intimately connected. The aim of this review is to describe what is known about this connection. We first discuss the connection between sub-additivity and quenching through the lens of models that seek to offer an economic description of the transformations that govern stimulus-response relations in V1. Next, we consider this connection from the viewpoint of models that seek to identify the computational and representational goals that shape V1 activity. Finally, we discuss this connection from a mechanistic perspective on cortical circuitry and computation. Note that we primarily review data collected

in cat and monkey. Recent work in rodent V1 has revealed similar sub-additive phenomena (and is rapidly advancing our understanding of the underlying circuit mechanisms[42–48]), but studies of response variability in rodent V1 have thus far been less extensive than in these other species (but see[49–51]).

Experimental observations: response sub-additivity

Phenomena of sub-additivity challenge an influential model of cell function. In cat and monkey V1, layer 4 neurons exhibit numerous response properties that are fundamentally different from their feedforward thalamic inputs. This includes selectivity for orientation, direction of motion, and depth[54,55]. A longstanding view, pioneered by Hubel and Wiesel, holds that this selectivity arises from the alignment of the receptive fields of presynaptic thalamic relay cells[54]. In its simplest form, this framework predicts that V1 receptive fields perform a linear filtering operation in space and time, which is followed by a thresholding operation to transform intracellular signals into spikes. This model for V1 responses is at once simple, elegant, and powerful. It enabled neuroscientists to approach a fundamental biological question (*"How do cortical sensory circuits transform their input into a novel representation of the visual environment?"*) through the principled abstraction of a linear system at a time when little was known about the cortical representation of visual information. The chief benefit of this approach is that it readily generates quantitative predictions for arbitrary visual stimuli. And to a first approximation, these predictions are quite good. A model that includes a linear receptive field and a static threshold nonlinearity can explain V1 selectivity for elementary stimulus attributes such as position, scale, orientation, speed, and direction of motion[56]. However, cortical cells also exhibit clear violations of linearity beyond a static threshold nonlinearity, which often manifest through the phenomenology of sub-additivity.

One prominent example of sub-additivity arises when a masking stimulus is superimposed on a cell's preferred stimulus (Fig 1c). Masking stimuli exert a suppressive influence across a broad range of spatial frequencies[6,8], orientations[6,7], and temporal frequencies[7,57,58]. Failures of responses to sum linearly are not restricted to the classical receptive field. Responses of V1 neurons can also be substantially diminished by stimuli outside the receptive field[2,3,11,59]. The strength of this "surround suppression" depends on the stimulus' exact position (the larger the distance from the receptive field center, the weaker the suppression[60]), and its similarity to the stimulus within the classical receptive field (the larger the resemblance, the stronger the suppression[4,11,61,62]).

A different example of sub-additivity comes from contrast summation experiments (Fig 1c). In these experiments, the same stimulus is presented at various contrasts, ranging from low to high. Scaling the contrast of an effective stimulus presented within the classical receptive field will scale the response of a linear system by the same factor, a property known as 'response homogeneity' in linear systems analysis. But this is not what happens in visual cortex. With increasing contrast, responses of V1 cells typically grow faster than linearly for low contrasts, but grow more slowly than linearly above some low level of contrast and may approach saturation at higher contrasts[5,9]. This sublinear or saturated response is immediately present at response onset and occurs for preferred as well as non-preferred stimuli[63]. Thus, as soon as stimulus contrast exceeds a low level, sub-additivity is a general property of cortical contrast responses.

sub-additivity induced by stimuli in the receptive field center and surround share important properties, but also differ in critical regards. In both cases, the masking stimulus tends to act as if it divides the contrast of the driving stimulus by a constant fraction[3,9,64]. This observation motivates the attempt to build stimulus-response models of V1 activity that capture both types of sub-additive effects with a single model component. However, surround suppression is partially delayed relative to response onset[65–67], exhibits interocular transfer[66], and is modified by contrast adaptation[66]. None of this is true for within-receptive-field violations of linearity[63,66,68] (though see[69]). This discrepancy suggests that the two types of sub-additive effects have a distinct mechanistic origin.

V1 response sub-additivity might arise in retina, LGN, or V1. Recent studies have explored the mechanistic origin of V1 response sub-additivity by combining visual stimuli with direct optogenetic stimulation of the visual cortex (Fig 1c). In macaque and marmoset V1, responses to optogenetic and visual stimulation combine sub-additively[41,70], suggesting that cortical circuits contribute to at least some forms of sub-additivity. Note however that in these studies the illuminated patch of cortex was large enough to engage lateral connections thought to be involved in surround suppression[59].

Experimental observations: response variability

Response variability in visual cortex may impose a fundamental limit on perception. Sensory neurons transmit information about the external world with sequences of action potentials that are inherently variable. This is also true of area V1: repeated presentations of identical visual stimuli elicit different patterns of spiking activity[16,17,20]. The origins of this variability are still unknown, but its consequences may be profound. If this variability represents irreducible noise, then it will limit the reliability with which neural populations can represent sensory events, and ultimately the capacity of the organism to perform perceptual tasks. Motivated by this insight, physiologists have made considerable efforts to directly compare neuronal and psychophysical sensitivity – an enterprise pioneered by Werner and Mountcastle (1963)[71]. This comparison is most meaningful for neurons suitably tuned for the task under consideration. And ideally, sensitivity estimates should be based on physiological and behavioral data obtained in the same animals, from the same set of trials[72]. Studies that meet these criteria have consistently reported that some individual sensory neurons rival the behavioral capacity of well-trained macaque monkeys. For example, the ability of V1 cells to signal changes in stimulus orientation closely approximates macaques' perceptual orientation acuity[73–75], while MT cells exhibit sensitivity for visual motion that is very similar to perceptual motion sensitivity[76]. These findings sparked an enormous interest in the origins and role of neural response variability. Here too, many found it fruitful to approach these

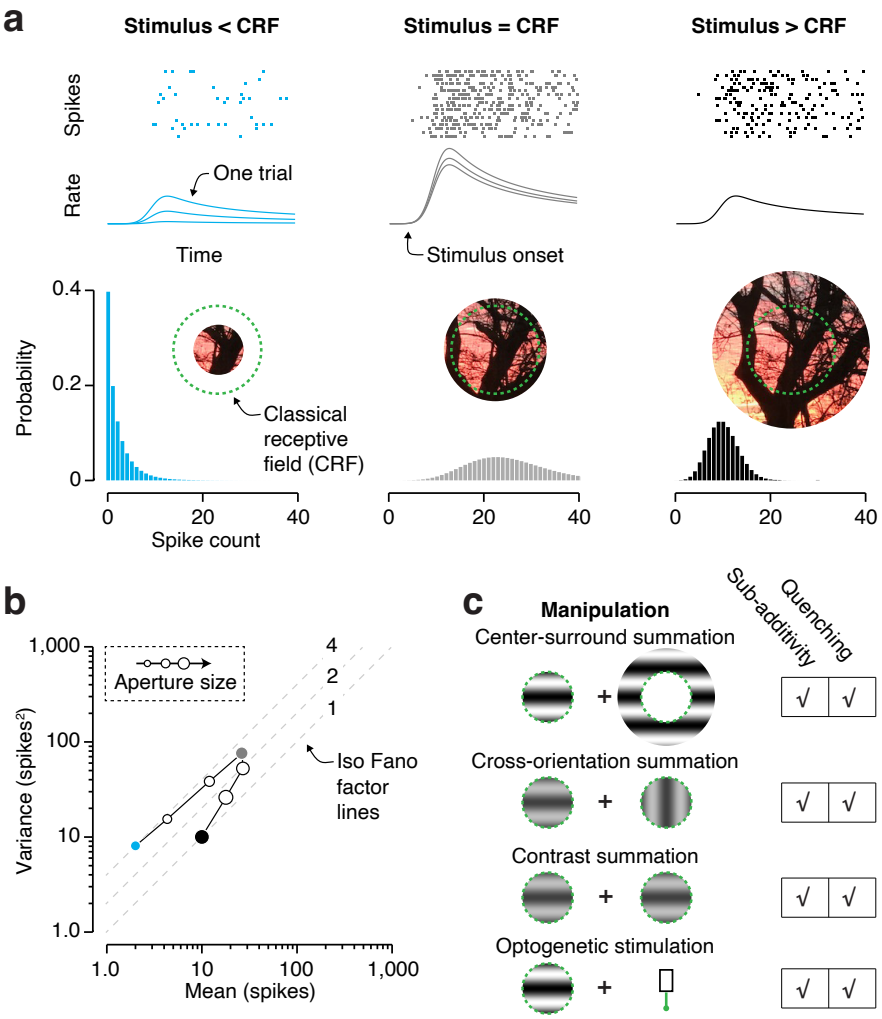

Figure 1 Response sub-additivity and variability quenching co-occur under various experimental manipulations. (a) Simulated responses of a V1 neuron to three patches cropped from the same image using differently sized apertures. (Top) Spike times are plotted as a raster, with one tick per spike. Each row depicts a repeated presentation of the same stimulus. Different stimuli are associated with different levels of responsiveness[3,4]. Within each raster, there is considerable variability in spiking activity across trials[12,17]. (Middle) These spikes were generated by simulating a Poisson process with an underlying rate that is multiplied by a response gain that varies across trials[28]. The rate variability decreases with stimulus size, as is the case for real V1 cells both at the level of spike counts[39] and, relatedly[52,53], membrane potentials[13,21]. (Bottom) The associated spike count distributions, obtained by using a counting window whose length matches the total stimulus duration. In experimental studies, response mean and response variance are computed from this distribution, not from the generative process itself. The dotted green line indicates the neuron's classical receptive field. Photo taken by R.L.T.G. (b) Spike count variance plotted as a function of spike count mean for the simulated size tuning experiment shown in panel a. Larger symbol sizes indicate larger stimulus apertures. Responses initially increase with stimulus size, but begin to decrease as the stimulus engages the suppressive surround. This suppressive effect co-occurs with a change in the relative amount of response variability, measured as the variance-to-mean ratio (*i.e.*, the Fano factor)[39]. The blue, grey, and black circle indicate the three conditions shown in panel a. (c) Summary of some classical experimental manipulations that elicit response sub-additivity[3,5,6,41] and variability quenching[21,33,39,41]. Details of model simulation described in Supplementary Information.

fundamental biological questions through a principled abstraction that generates quantitative predictions for arbitrary stimuli: the Poisson point process.

Which aspects of a spike train are signal and which are noise? One extreme possibility is that only the number of spikes realized during a temporal interval matters, and that there is no information in the exact timing of each spike[18]. This concept is formalized by the Poisson process. It is the simplest stochastic point process, fully characterized by a single firing rate parameter that represents a reproducible response to a sensory stimulus. If the rate is fixed over time, the process is said to be homogeneous; if it varies over time, it is inhomogeneous. Both variants give rise to Poisson-distributed spike counts. A hallmark of this distribution is that the spike count variance across repeated measurements matches the spike count mean, regardless of the length or placement of the time interval over which spikes are counted. In other words, the ratio of the variance to the mean, a statistic known as the Fano factor, is always one. In visual cortex, this prediction enjoys some support. Spike count variance often approximately equals the mean[17,77]. However, cortical spiking statistics also exhibit clear deviations from a Poisson distribution. This most commonly manifests in the form of excess variance. When the mean count is high, either due to a high firing rate, or due to the use of a long counting window, super-Poisson variability (*i.e.*, more variability than expected from a Poisson process) becomes apparent[17,28,31,77]. Statistically, both behaviors can be explained by expanding the Poisson process with a slowly fluctuating gain that modulates the rate and varies from trial to trial[28,29] (*i.e.*, a doubly stochastic process known as the 'modulated Poisson model', Fig. 1a, traces). Empirically, these behaviors make it difficult to identify changes in response variability that are not simply a consequence of changes in response mean. One popular approach to overcome this challenge is to estimate Fano factor using an analysis procedure that corrects for differences in mean response level[22] (but see[78]). This statistic is called the mean-matched Fano factor. It generally exceeds 1 in cortex, and, under the Poisson assumption, represents a measure of cross-trial variability in firing rate.

Firing-rate variability is stimulus dependent in a manner that resembles phenomena of sub-additivity. Across cortex, it is maximal in the absence of stimulation and decreases rapidly following stimulus onset[22]. The magnitude of the decrease depends on the amount of stimulus energy. For example, in area V1, low contrast stimuli placed within a neuron's receptive field are associated with stronger rate fluctuations than high-contrast stimuli[21,33]. Variability quenching occurs for preferred as well as non-preferred stimuli that drive a neuron[21,22,33], thus resembling response saturation in contrast summation experiments. It also occurs for stimuli that do not drive a neuron[22], thus resembling the broad tuning of suppressive effects in masking experiments. Finally, stimuli outside of the receptive field can quench neural response variability beyond the reduction of variability caused by increasing stimulation inside the receptive field[39] (Fig 1b, black vs grey symbol) – thus resembling surround suppression. The strength of this effect weakly depends on the similarity between the center and surround stimulus[39]. The effect also depends on cortical layer[79] and on the size and location of the surround stimulus[40,79].

The co-occurrence of response sub-additivity and variability quenching can be illustrated by plotting the variance vs. mean relationship and graphically illustrating stimulus size[33] (Fig. 1b). Increases in stimulus size initially increase both the variance and mean of the spike count response, but decrease their ratio (the Fano factor)[39]. Further increases in size cause surround suppression, reducing both the variance and the mean and continuing to decrease their ratio towards the line of Fano factor equal to 1[39].

Is it mere coincidence that experimental manipulations that elicit response sub-additivity often quench variability too (Fig. 1c)? Are these phenomena partly related? Or might they be distinct manifestations of shared underlying mechanisms? These questions are difficult to answer because we cannot directly observe the signals of interest. There is no empirical measurement that directly reveals the strength of a cell's "sub-additive signal". And firing rate variability is a statistical construct that cannot be mapped onto an observable biophysical quantity. Answering these questions requires a theoretical exploration of the issues at stake.

Models of V1 activity

Hubel and Wiesel's trail blazing work inspired many to build, test, and refine models of V1 activity. There is a great deal of diversity among these models, reflecting differences in the underlying aspirations. For example, some models seek to explain V1 responses in a manner that remains faithful to known physiological mechanisms, thus revealing how the structure of neural circuits gives rise to their function[56,80–83]. We will refer to such models as "mechanistic" accounts of V1 activity. Other models seek to explain V1 responses on the basis of theoretical coding principles[21,84–89], thereby revealing the computational objectives that shape neural function ("normative" accounts). And yet other models aspire to describe quantitatively the transformation of visual stimuli into neural responses using a limited set of operations and parameters that can be fit to neural data[9,90–94] ("descriptive" accounts). Models of this latter type are useful to simulate V1 activity and hence can provide insight into V1's representation of visual information beyond experimentally feasible measurements[95]. They are also an essential point of comparison for studies that aim to connect V1 representations to downstream transformations[96], perceptual capabilities[97,98], other sensory modalities[99], and artificial visual systems[100].

Descriptive accounts of response sub-additivity and variability quenching in V1

We will focus here on one prominent descriptive framework, the "normalization" model[91,99,101]. This model describes the firing rate of V1 neurons as the ratio of a narrowly tuned excitatory channel and a broadly tuned inhibitory channel. The excitatory channel usually consists of a linear spatio-temporal filtering operation followed by a nonlinear pooling operation and determines the stimulus selectivity of the neuron (Fig. 2a). The in-

hibitory channel is built from the same operations, but has much weaker tuning. Specifically, it extends over a larger area of visual space than the excitatory channel, and is weakly tuned for orientation and spatial phase (Fig. 2a). Because this model incorporates a divisive operation, its responses saturate with increasing stimulus contrast – that is, once stimulus contrast exceeds a low level, they are sub-additive[91,101]. This occurs for preferred as well as non-preferred stimuli, as is the case for real V1 cells[5,9]. Moreover, because the inhibitory channel is broadly tuned, the normalization model also exhibits cross-orientation suppression and surround suppression[3,87,91]. Critically, in the normalization framework, all sub-additive effects arise from a single operation. This is an extreme proposition, yet it provides a remarkably accurate description of classic sub-additive phenomena[3,9].

As discussed above, V1 neurons often exhibit super-Poisson variability in a manner that resembles the effects of a noisy response gain. This behavior naturally arises in the normalization model when we assume spikes arise from a Poisson process[94] and allow for noise in the normalization signal[33] (Box 1). As shown in ref.[33], including noise in the normalization signal has almost no effect on the mean responses of the model but affects response variability in a number of ways. First, since the firing rate now varies across repeated presentations of the same stimulus, spike generation results from a doubly stochastic process, yielding super-Poisson variability. Second, because the denominator rescales the output of the excitatory channel, this additive noise has a multiplicative effect on firing rate, *i.e.*, it introduces gain fluctuations. Third, the strength of these gain fluctuations depends on the output of the inhibitory channel (Fig. 2b). As is evident from the expressions that govern the model's behavior[33], excitatory drive and normalization noise increase excess response variance, while inhibitory drive has the opposite effect (Box 1, equation 6). For this reason, the stochastic normalization model predicts that response sub-additivity and variability quenching will often co-occur, as illustrated for some classical experimental manipulations in Fig. 2c. These predictions have not yet been tested in great quantitative detail. It is likely that additional model complexity will be required to capture the exact relationship between response sub-additivity and variability quenching. Note for example that in Fig. 2c, Fano factor initially increases with stimulus contrast (bottom left). The available evidence suggests that this occurs for some cortical cells, but in most cases, Fano factor decreases monotonically with stimulus contrast[32]. The model also predicts that Fano factor initially increases with increasing stimulus size (Fig. 2c, bottom right), which appears true for the majority of V1 cells ([79], see also[39]). It is possible that simple variation in parameter values can account for this cross-neural diversity[32]. However, it is also possible that additional model components are required to fully capture the diverse empirical behaviors.

An alternative version of the stochastic normalization model replaces the Poisson point process with a Gaussian noise source in the excitatory channel[32]. Stimulus-response relations are governed by different quantitative expressions, but qualitatively behave in a very similar manner. Most importantly, this variant also predicts a general quenching effect of normalization on neural response variability. With the additional flexibility afforded by the stochastic numerator, this model can capture empirical deviations from a modulated Poisson process. Another important feature is that this model can be inverted to estimate the single-trial strength of the normalization signal (which, as discussed above, is a statistical construct that cannot be measured empirically), from the measured neural activity. Using this method, it has been shown that even when the stimulus is constant, normalization strength fluctuates substantially across trials, and the variability of V1 responses is more strongly quenched during trials with stronger normalization[32].

In summary, from the vantage point of this descriptive model of V1 activity, many of the classical phenomena of sub-additivity appear to result from the same operation (divisive normalization) and neural responses appear to contain two layers of variability, variability of spiking and of rate. The second layer (variability in firing rate) is quenched by the suppressive signal.

Computational and representational objectives that shape V1 activity

The operations implemented by sensory systems are shaped through processes of evolution, development, and learning, and are adapted to the tasks the organism must perform in its natural environment. This raises the question of whether key features of sensory systems can be derived from studying artificial systems designed to either optimally perform such tasks[104,105] or to realize computational goals essential to these tasks[106,107]. In this section, we focus on theoretical coding principles that provide such normative insight into response sub-additivity and variability quenching in V1.

The most successful theoretical proposal concerning the goal of early sensory processing is the efficient coding hypothesis[106,107]. Applied to V1, this hypothesis states that the goal of V1 activity is to represent natural inputs with less statistical redundancy than present in those inputs. This notion enjoys strong empirical support. When a linear filter basis is optimized such that responses to a generic ensemble of natural stimuli are both as informative and as statistically independent as possible, the resulting basis functions resemble the visual receptive field structure of V1 simple cells[84,85], particularly if the stimuli are natural movies rather than static natural images[108].

However, a simple linear transformation is insufficient to produce fully independent responses to natural images. Removing the remaining statistical dependencies requires an additional nonlinear response transformation. In particular, dividing each filter's response by the weighted sum of the rectified responses of neighboring filters increases response independence[87]. When optimized for natural image statistics, the resulting divisive normalization model exhibits response sub-additivity reminiscent of cortical cells. For example, simulations of classical contrast summation, cross-orientation summation, and size tuning experiments all yield model responses that qualitatively match the behavior of V1 cells[87]. This framework can also account for the intricate cortical suppression phenomena elicited by natural im-

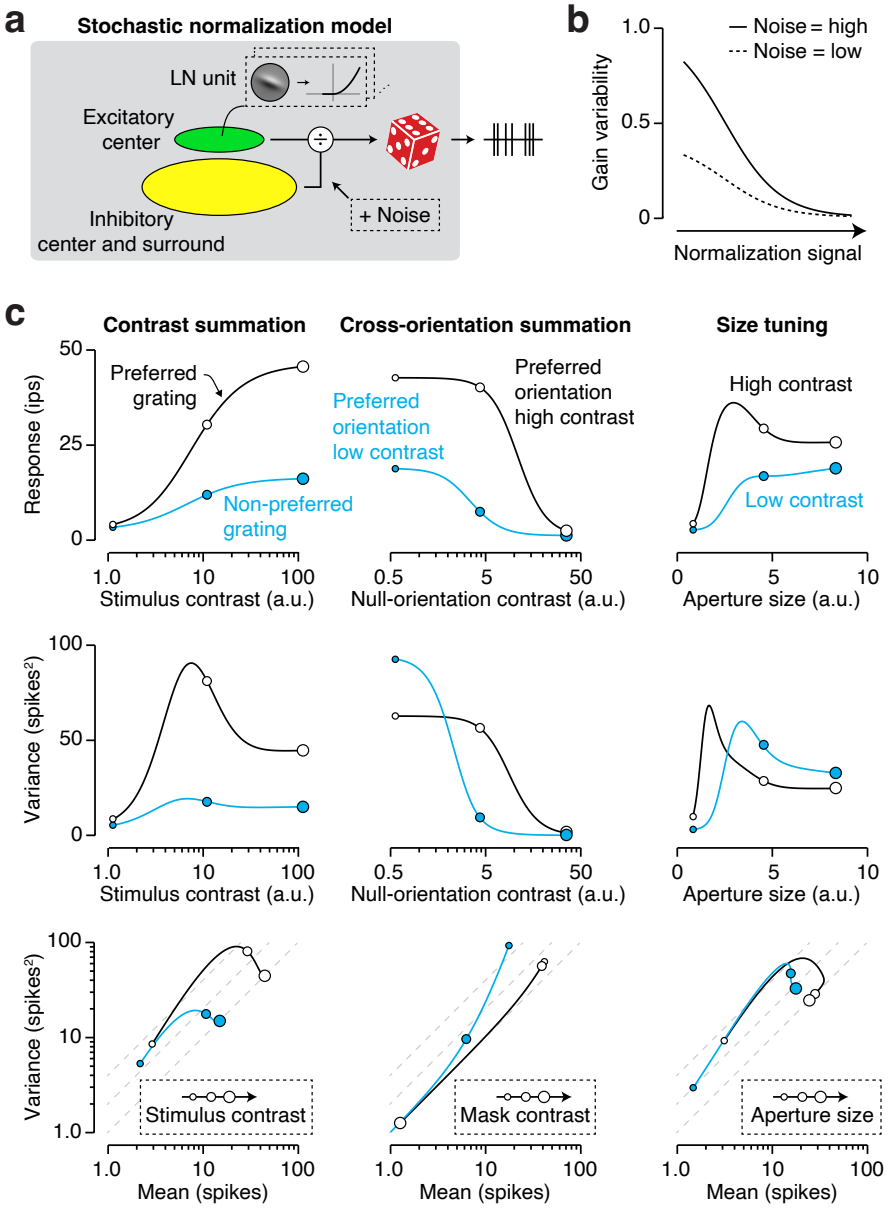

Figure 2 Relationship between response sub-additivity and variability quenching under a stochastic normalization model. (a) Model schematic. Stimuli are first processed by a bank of linear-nonlinear units, whose responses are pooled to form a narrowly tuned excitatory channel and a broadly tuned inhibitory channel which acts divisively[3,9,87,91,94,103]. The normalization signal provided by the inhibitory channel is subject to stimulus-independent, additive noise and spikes are generated by a Poisson process[33]. (b) Under the stochastic normalization model, gain variability depends on the normalization signal (X-axis) and on the level of normalization noise (full vs dotted line). This plot graphically illustrates the relationship derived in ref.[33], and shown in Box 1 (eq. 5). (c) In this model, response sub-additivity and variability quenching often co-occur. Top: simulated mean model response for some classical experimental manipulations (from left to right: contrast summation, cross-orientation summation, and size tuning). These model predictions provide a good quantitative account for the behavior of V1 cells[3,5,91]. Middle: the associated predicted response variance, obtained by applying Box 1's eq. 5, taken from ref.[33]. Bottom: the variance-to-mean relationships directly illustrate the joint occurrence of response sub-additivity and variability quenching in the stochastic normalization model.

Consider the simplest instantiation of the normalization framework[99]:

$$\lambda(S) = \frac{E(S)}{\beta + I(S)} \tag{1}$$

where λ is firing rate, S an image, E the excitatory drive obtained by measuring the local energy in a narrow range of orientations and spatial frequencies, I the inhibitory drive obtained by measuring global energy across a broad range of orientations and spatial frequencies, and β a stimulus-independent constant[91]. Equation 1 specifies a deterministic relation between stimulus and firing rate, a statistic that is not directly observable but inferred from trial-averaged measurements. The simplest way to obtain a full generative model of spiking activity is to include a Poisson point process[94,96]. Together, these model components suffice to express the probability of every possible spike count for arbitrary visual stimuli:

$$p(N|S, \Delta t) = \frac{(\lambda(S)\,\Delta t)^N}{N!}\, e^{-\lambda(S)\,\Delta t} \tag{2}$$

where N is spike count and Δt duration of the counting window. Under this model, response variance simply equals the response mean:

$$\mathrm{Var}[N|S, \Delta t] = \mathrm{Mean}[N|S, \Delta t] = \lambda(S)\,\Delta t \tag{3}$$

Ref.[33] assumed that the normalization signal is not deterministic, but subject to additive Gaussian noise with zero mean and variance σ_N^2. On a single trial, spikes are generated from a Poisson process with firing rate

$$\lambda_i(S) = \frac{E(S)}{\beta + I(S) + \epsilon_i} \tag{4}$$

where subscript i is a trial index and ϵ_i is the Gaussian noise. Because the denominator in equation 4 rescales the output of the excitatory channel, this additive noise has a multiplicative effect on firing rate, *i.e.*, it introduces gain fluctuations[33]. The strength of these gain fluctuations depends on the output of the inhibitory channel and is well approximated by:

$$\sigma_\mathrm{G} = \frac{\sigma_\mathrm{N}}{\beta + I(S)} \tag{5}$$

where σ_G expresses the standard deviation of the gain[33]. The suppressive signal reduces gain fluctuations, and hence reduces response variability. The stochastic normalization model describes a doubly stochastic process. It follows from the law of total variance that spike count variance is composed of the sum of the expected Poisson variance (equation 3) and a term that represents the contribution of rate variability[102]. This term is the product of the variance of the gain signal and the squared mean response[28]:

$$\mathrm{Var}[\lambda(S)\,\Delta t] = \frac{(\sigma_\mathrm{N}\,E(S))^2}{(\beta + I(S))^4}\,\Delta t^2 \tag{6}$$

Box 1 Inhibitory drive suppresses response strength and quenches response variability in the stochastic normalization model. Based on ref.[33].

ages[4]. Phenomena of sub-additivity can thus be understood as a direct consequence of the visual system's attempt to efficiently encode the statistical structure of natural images. In summary, this line of work demonstrates that the nonlinear response properties of sensory neurons captured by divisive normalization-based descriptive models (described in the previous section) are well predicted by the normative principle of efficient coding. In other words, divisive normalization is "not an accident of biological implementation, but has an important functional role"[87].

Although redundancy reduction leads to normalization, which can account for response sub-additivity, it does not provide a normative justification for variability, or its quenching, *per se*. In fact, neural variability seems at odds with the normative goal of efficient coding as response variability limits coding capacity and is therefore undesirable for any pure stimulus encoding system[109]. But sensory systems seek to do more than just representing sensory input. Ultimately, they must construct perceptual interpretations of the environment that facilitate behavioral tasks. The most relevant aspects of the environment (*e.g.*, the presence of potential prey or a potential predator) typically have a complex and ambiguous relationship with raw sensory input. These aspects thus need to be inferred. Inevitably, these inferences have varying degrees of certainty. To achieve optimal behavioral outcomes, the uncertainty of perceptual inferences needs to be taken into account[110–112]. How neural circuits do so is debated and an important topic of modern research[21,27,33,39,113–120].

A prominent hypothesis concerning the role of neural response variability in sensory cortex is that its structure may facilitate the assessment of perceptual uncertainty by downstream circuits. Theorists have proposed several variants of this

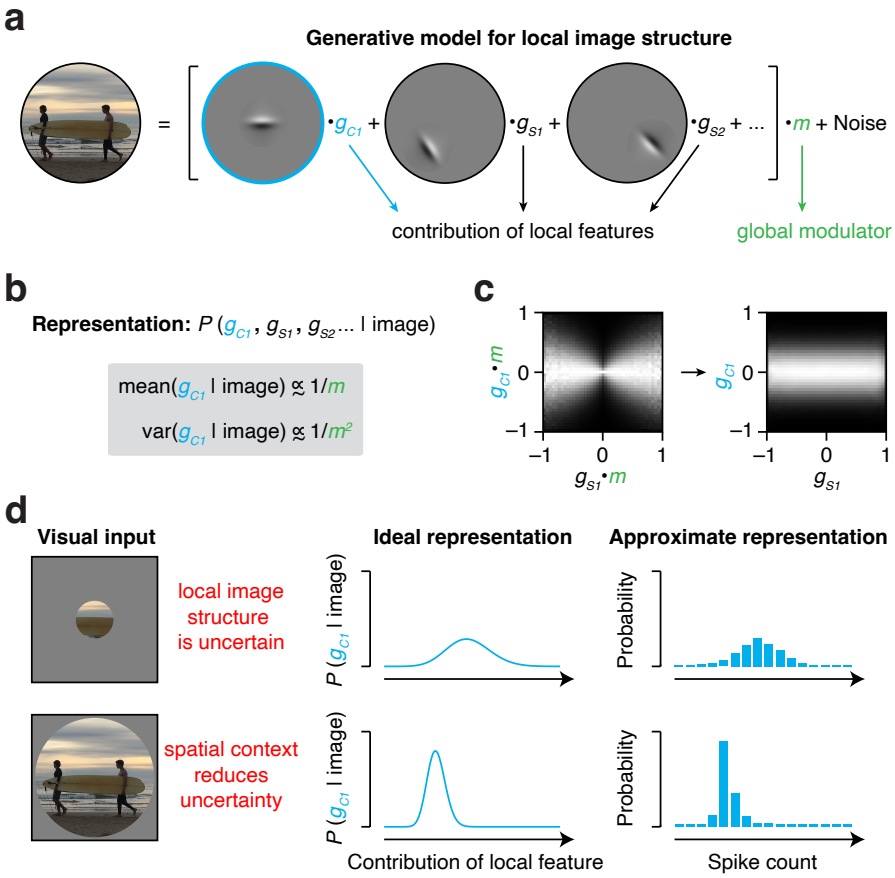

Figure 3 A normative account for the relationship between response sub-additivity and variability quenching in area V1. (a) The local structure of natural images is well described as a linear combination of a set of spatially localized image features that is subject to global modulation and noise (*i.e.*, a Gaussian scale mixture model)[121]. Photo taken by R.C.-C. (b) Encoding image information by inferring the contribution of each local feature naturally results in divisive response suppression and response variability quenching for higher values of the global modulator[39] (*e.g.*, for images with higher contrast levels). (c) Joint histograms of the simulated responses of a nearby pair of local image filters before (left) and after (right) normalizing by the global modulator. Responses were random samples from a Gaussian scale mixture model[121]. Pixel intensity is proportional to the bin count, rescaled per column to fill the range of intensities. Normalization reduces response redundancy among these filters[87]. (d) Schematic representation of sampling based inference in the Gaussian scale mixture model[39]. If the inferred contribution of local features is represented probabilistically, informative image content in the surround will lower the peak and narrow the width of the posterior belief in the contribution of a local feature positioned at the center of the image (center panels). If neural responses represent samples from the posterior distribution, this will manifest as response suppression and variability quenching (right panels).

idea[117,122]. In particular, the Neural Sampling hypothesis proposes that neural responses in sensory cortex represent samples from a probabilistic model of the environment[21,122,123]. It follows that neural response variability reflects uncertainty about the inferred stimulus feature. Consistent with this idea, factors that improve the quality of perceptual orientation estimates such as image contrast and aperture size[124] also quench response variability of V1 neurons[21,33,39].

From a computational perspective, if sensory systems must take uncertainty into account, the optimal way to do so is to learn the causes of sensory inputs, by analyzing the statistical regularities of sensory inputs and forming a so–called generative model that reproduces those statistics. Probabilistic inference consists of inverting this generative model, to correctly map an observed input onto a probability distribution of the causes of that input (the so-called posterior distribution)[125].

This foundational idea has recently been adapted to provide a unified account for the phenomenologies of response sub-additivity and response variability in V1. The theory, which combines critical elements of efficient coding and neural sampling, proposes that V1 activity represents approximate probabilistic inferences based on a generative model of local image structure[21] (Fig. 3a,b). The model postulates that images are generated by a combination of local features, and a global modulator representing luminance or contrast (Fig. 3a). The theory assumes that the computational goal of V1 is to represent local image features by undoing the effect of nuisance variables such as the global modulator, a computation termed marginalization. This has the effect of removing redundancies that are present in the raw visual inputs due to the nuisance variables (Fig. 3c). The theory additionally assumes that V1 activity represents samples from the inferred posterior probability distribution of the feature coefficients, *i.e.* a neural sampling-based representation (Fig. 3d). In this way, the average neural response (*i.e.* the sample mean) represents the mean of the posterior distribution, that is, the estimate of the feature coefficients that is expected to have minimal squared-error, while neural variability (sample variance) represents uncertainty about this estimate (posterior variance).

When this model is optimized for natural image statistics, response sub-additivity and variability quenching often co-occur[21,39,126]. The intuition for this connection is that marginalization results in more certain inferences about local image features and is achieved via divisive normalization (Fig. 3b). Thus, the average neural response (representing the posterior mean) in this model inherits – at least qualitatively – the sub-additive effects predicted by the divisive normalization model of efficient coding discussed above[21,87]. For example, responses of nearby linear filters representing a neuron's receptive field and its surround are typically informative about the global modulator. Homogeneous images that extend beyond the receptive field elicit similar responses of nearby filters, suggesting a large value of the global modulator, and therefore evoking strong normalization from the surround[4,127]. Critically, variability quenching is a consequence of the same computation: observing the image content in the surround lowers the estimation uncertainty about local image structure inside the receptive field, and thus results in a smaller amount of response variability[39]. Conversely, when the image is confined to the receptive field (Fig. 3d), or when the surround image is part of a different object (termed heterogeneous center-surround configuration[4]), responses of nearby linear filters are less redundant (resulting in weaker normalization) and uncertainty about local image features is higher (resulting in higher response variability).

In summary, studies of the computational and representational objectives underlying V1 activity offer a parsimonious explanation for the co-occurrence of response sub-additivity and variability quenching: Divisive normalization in V1 serves to compute probabilistic inferences about visual inputs, relating sub-additive phenomena that maximize coding efficiency and quenching phenomena that express uncertainty about inferred image features.

Circuit mechanisms that govern V1 activity

V1 activity is shaped by retinal, thalamic, and cortical circuit mechanisms. How do cortical response sub-additivity and response variability arise mechanistically from the interplay of these distinct anatomical components? The descriptive and normative modeling approaches discussed thus far offer little insight into this. These models are formulated in terms of normative principles and phenomenological operations – *e.g.*, linear filtering, divisive normalization, noisy spike generation – whose biophysical and anatomical substrates are not specified. To address this question, we turn to neural circuit models.

Response sub-additivity and variability quenching need not be produced by the V1 circuit; they could instead arise from the inputs to V1. This is likely the case for sub-additivity induced by stimuli in the receptive field center[56,57,128–132]; but see[69]. Likewise, variability in firing rate of LGN cells decreases with stimulus contrast[13,133]. Thus, some forms of response sub-additivity and variability quenching in cortex have, at least in part, a feedforward origin.

How might cortical circuitry further contribute to response sub-additivity and variability quenching? This question is an active topic of contemporary research. The proposed models that have received the most attention, *e.g.*[53,134,135], share two common features. First, excitatory cortical circuitry is structured: excitatory neurons with shared selectivity are more strongly coupled whereas those with distinct preferences exhibit weak coupling. Second, inhibitory connections are strong enough to stabilize the network despite excitatory connectivity being strong enough to potentially cause instability. These features are apparent in cortex[45,136–140], although there is as yet limited direct evidence for their role in response sub-additivity or variability quenching (but see[136]).

Cortical neurons receive inputs from many cells. A neuron receiving a large number of excitatory inputs, without compensating inhibition, would have a large mean input and exhibit regular clock-like spiking, inconsistent with the variable firing of cortical cells[18,141]. Therefore a number of additional mechanisms have been proposed to account for spiking variability. Two pos-

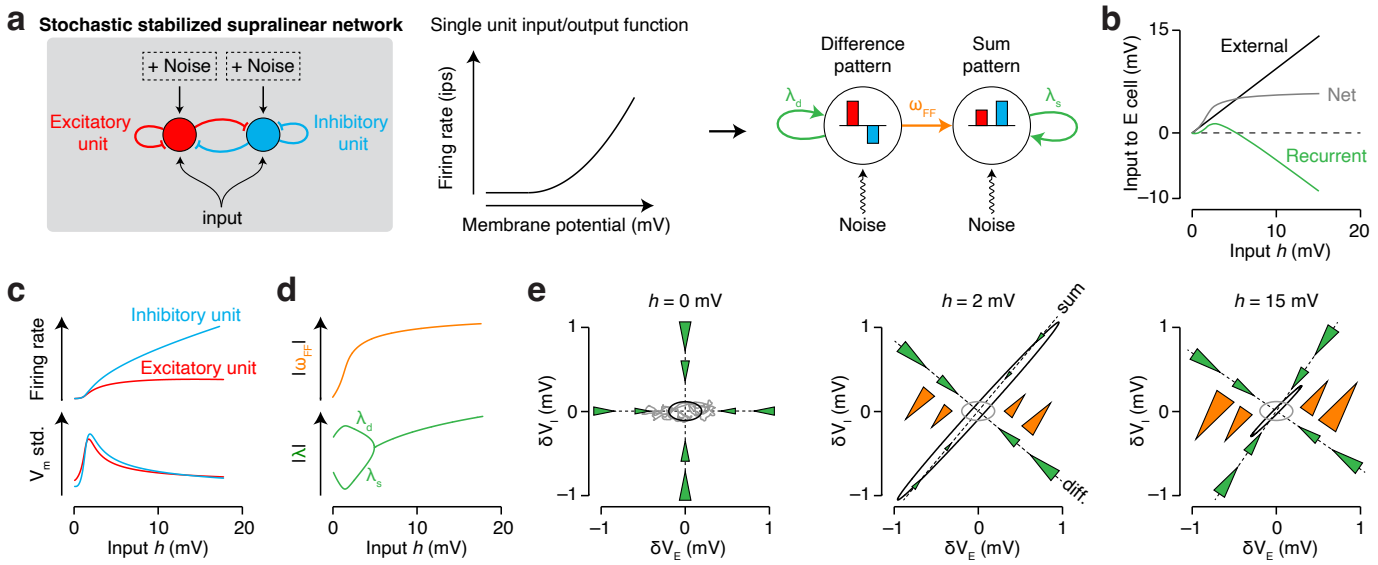

Figure 4 Response sub-additivity and variability quenching co-occur in the stochastic stabilized supralinear network model. (a) Left: Model network, studied in[53]: two recurrently connected units, representing activity of populations of excitatory (E, red) and inhibitory (I, blue) neurons. The units receive private input noise and a common mean input. Middle: A supralinear (threshold-quadratic) neural input/output function determines instantaneous firing rate as a function of membrane potential. Right: The dynamics can be equivalently expressed in terms of two activity patterns: a "difference pattern", in which E and I activities (red and blue bars) have opposite signs; and a "sum pattern", in which they have the same sign. The patterns inhibit themselves with weights λ_d and λ_s, respectively, while the difference pattern excites the sum pattern with weight w_{FF}. (b) For weak external input h, external input (x-axis and black) dominates recurrent input (green). For stronger external input, the SSN dynamics leads the recurrent input to largely cancel the external input, so that net input (gray; black plus green) grows sub-linearly as a function of external input. Over the dynamic range of this model, balance is "loose" (net input is similar in size to the other two, cancelling, inputs), but would become "tight" (net input much smaller than the others) for very strong (likely non-physiological) input. (c) Firing rate (top) and voltage (bottom) for the E (red) and I (blue) units, as a function of mean input strength h. (d) Dependence of the weights λ_d, λ_s, and w_{FF} on input strength. (e) How these weights alter the dynamics and variability of the fluctuations $(\delta V_E, \delta V_I)$ about the mean rates. Black ovals represent 1-standard-deviation contours of the fluctuations. For h=0, the E and I units are effectively uncoupled, so oval is just the external input noise filtered by each isolated unit's time constant (repeated in other panels in grey, for comparison). The green arrows represent the effect on the flow of λ_d (diff) and λ_s (sum), which drive fluctuations toward the origin. Orange arrows represent the effects of w_{FF}, which drives positive (negative) differences in the positive (negative) sum direction. Area of arrows indicates strength of influence on the flow. See text for further details. Right panel of (a), (c)-(e) adapted from[53].

sible sources are stochasticity in cellular and synaptic mechanisms[142–144] and input correlations which prevent the variability of individual inputs from being averaged out[145,146]. Spiking variability can also arise from network dynamics, if inhibition balances excitation sufficiently that the mean input to a cell is sub- or peri-threshold, so that the neuron is driven to fire by input fluctuations – brief imbalances in excitation and inhibition – which occur at random times[147–150]. The "balanced network" model[148] demonstrated that network dynamics can automatically yield such balancing in a broad parameter regime, without requiring fine tuning of parameters. This model produced "tight balance", meaning that the excitation and inhibition that cancel are much larger than the net input remaining after cancellation[151]. However, tightly balanced network models do not naturally produce nonlinear input-output transformations that could give rise to response sub-additivity. They also do not generate super-Poisson variability or variability quenching. These problems can be solved by considering more loosely balanced models[151] and/or structured[134] or heterogenous[152] connectivity, as we now discuss.

The above mechanisms predict Fano factors below 1. What additional sources of variability yield super-Poisson variability characteristic of cortical neurons, and why is this additional variability quenched by a stimulus? In one family of models, this additional variability arises during spontaneous activity because the network is wandering among many states, *e.g.* corresponding to possible responses to many different stimuli, with neurons firing at different rates in different states[134,153]. A stimulus "pins" the network to one state, quenching the variability. Some of these models depend on specific connectivity. For example, given stronger connections within and weaker connections between distinct clusters of excitatory units in an otherwise balanced network, with one cluster's activity inhibiting the others, network activation can be largely restricted to one cluster at a time and wander between clusters over time[134]. A similar mechanism could apply if neurons are most strongly connected to neurons with similar response properties, forming a continuum of clusters rather than discrete clusters. Consistent with this idea, spontaneous activity in V1 wanders through states resembling stimulus responses, both to laboratory[26,154,155] and natural[21,27] stimuli, more often than expected by chance. Note that a stimulus that pins the wandering reduces the variability of all neurons, including those not driven by the stimulus[156,157]. A related proposal is that a network's variability is generated by chaotic dynamics of spontaneous activity, as occurs in tightly balanced networks with sufficient variability in their weights[152]. A stimulus can then suppress variability by suppressing the chaos[158].

Note that in these models, there is no connection between the mechanisms that alter variability and sub-additivity of responses. If such a model shows response sub-additivity, it will be due to mechanisms distinct from the stimulus-induced pinning of network state that quenches variability. However, in these models, changes in firing rates (which need not involve sub-additivity) are naturally coupled to changes in variability: for example, a decrease in stimulus strength decreases firing rates and increases variability at the same time[157]. Such an increase in variability accompanying a decrease in firing rates has been observed in some cases of surround suppression[79], and it was suggested[79,157] that in these cases, surround suppression arises primarily from suppression of feedforward inputs (consistent with experimental evidence for feedforward contributions to surround suppression: [66,159] in macaque V1, and [160,161] in mouse V1). However, in most cases, a decrease in variability accompanies surround suppression[39,79] (consistent with experimental evidence for cortical contributions to surround suppression[66,159]). To explain this phenomenon, other mechanisms relying on the intrinsic dynamics of V1 are necessary.

An alternative model of V1 dynamics posits that the network randomly fluctuates about a single steady state for a given fixed external input, including the input driving spontaneous activity; but the amplitude of the fluctuations decreases with external input strength, quenching variability. In this model, the fluctuations are due to external input noise amplified by an excitatory network that is stabilized by inhibitory cells[53]. This inhibitory stabilization and its increasing strength with increasing external input drives both response sub-additivity and variability quenching. A key ingredient of this "stabilized supralinear network" (SSN) model[53,82,135] is that neuronal input/output (I/O) functions are supralinear (Fig. 4a, middle). This means that neuronal *gain* – the change in output per change in input, *i.e.* the I/O function's slope – increases with neuronal activation. The result is that "effective connection strengths" – the change in postsynaptic firing rate per change in presynaptic firing rate – increase with increasing strength of the network's external input[135] (Box 2). This increase plays a central role in both response sub-additivity and variability quenching in the SSN.

For very weak external input – around spontaneous levels – effective synaptic strengths are weak. As a result, monosynaptic pathways (the external input) are much stronger than the di-and poly-synaptic pathways they evoke (recurrent input). Thus, responses largely follow the supralinear input/output function of decoupled cells, and hence sum supralinearly. With increasing input strength, the relative contribution of network drive increases, eventually exceeding a point at which the excitatory subnetwork alone would become unstable, and so the network enters an inhibition-stabilized regime[45,135,136,162]. Stabilization occurs through "loose balancing"[135,151] (Fig. 4b), meaning (1) recurrent input largely cancels the external input, so that the net input grows sublinearly as a function of the external input; and (2) the net input is comparable in size to the factors that cancel, *i.e.* the balance is "loose". This loose balance is sufficient to yield irregular spiking[163], yet allows nonlinear behaviors such as response sub-additivity that are absent when balance is tight. In particular, when a second stimulus is added to a first, most of the extra feedforward input is cancelled. The result is (a) there are two parameter regimes, in only one of which contrast saturation occurs[135] but (b) for most parameters, when two different stimuli are added, response summation is sublinear, whether the second stimulus is added within the receptive field or in the surround[82,135].

In the SSN, at steady state, the contribution of a particular presynaptic cell to the input of a postsynaptic cell, I, is simply given by the firing rate of the presynaptic cell, r_{pre}, scaled by the strength of the connection between the two cells, W:

$$I = W\, r_{\mathrm{pre}} + \ldots \tag{7}$$

where $\ldots$ denotes terms that are independent of the presynaptic neuron, such as recurrent inputs from other neurons in the network, as well as external, feedforward inputs from upstream areas. The firing rate of the postsynaptic neuron is then determined by the neuronal input/output function:

$$r_{\mathrm{post}} = f(I) \tag{8}$$

Combining Equations 7 and 8 yields a self-consistent equation that expresses the (steady state) relationship between the firing rates of the pre- and postsynaptic neuron as

$$r_{\mathrm{post}} = f(W\, r_{\mathrm{pre}} + \ldots) \tag{9}$$

This means that for a sufficiently small deviation in the activity of a presynaptic neuron (*e.g.* due to a change in its external drive, or random fluctuations), δr_{pre}, the change in the response of the postsynaptic neuron, δr_{post}, is given by

$$\delta r_{\mathrm{post}} = f'(I)\; W\, \delta r_{\mathrm{pre}} \tag{10}$$

where $f'(I)$ is the "neural gain", the slope of the input/output function at the steady state input of the postsynaptic neuron. Thus, the effective connection strength between the two neurons is the actual connection strength scaled by the neuronal gain:

$$W_{\mathrm{eff}} = \frac{\delta r_{\mathrm{post}}}{\delta r_{\mathrm{pre}}} = f'(I)\; W \tag{11}$$

The supralinearity of f means that not only the rate, $f(I)$, grows with the input, I, but so too does the gain, $f'(I)$, and thus the effective connection strength, W_{eff}.

Box 2 The scaling of effective connectivity with input strength in the SSN. Based on ref. [135], see also [53,82].

This mechanism also creates and quenches super-Poisson variability, as illustrated for a simple model of one E and one I population [53] in Fig. 4. With both strong amplifying excitatory and strong stabilizing inhibitory connections, the network shows "balanced amplification" [164]: small input imbalances favoring E (or I) strongly drive both E and I cells up (or down). This can be mathematically summarized by formulating the dynamics in terms of the strengths of two *patterns* of activity: a difference (D) pattern and a sum (S) pattern, in which E and I activities have opposite signs (D) or the same signs (S) (Fig. 4a, right). Any actual pattern of E and I activities can be expressed as a linear combination of these two patterns. Each pattern effectively inhibits or damps its own activity, with weights λ_D and λ_S. The difference pattern excites the sum pattern with a weight w_{FF}, but there is no connection in the opposite direction (a feedforward connection pattern [164]).

In this model, as external input h increases from 0, the variability, as measured by voltage standard deviation, first increases to a peak before thereafter being suppressed (Fig. 4C). The peak occurs around the transition between the external-input-dominated and recurrently-dominated regimes, where the recurrent input "turns around" and starts balancing the external input (Fig. 4B). As h increases from zero, effective connection weights rapidly increase and w_{FF}, the feedforward drive from difference to sum, rapidly grows (Fig. 4D). Thus, variability is increased by increasingly strong balanced amplification – small E/I differences in the external noise driving large joint fluctuations of E and I. The decrease of the sum pattern's self-inhibition λ_S also contributes, decreasing the damping of fluctuations of the sum pattern. Beyond the regime transition, the growth of w_{FF} greatly slows, while λ_S, and later λ_D, grow. This represents increasingly strong inhibitory stabilization, which damps fluctuations and so quenches variability. The net result (Fig. 4e) is that the fluctuations (black ovals), initially driven by the input noise ($h = 0$), are greatly amplified in the sum direction, producing the peak in voltage variability ($h = 2$), before being quenched by the increasingly strong inhibitory damping ($h = 15$).

A signature of the SSN is a non-monotonic dependence of variability on stimulus strength (Fig. 4c), in agreement with the stochastic normalization model (Fig. 2c, bottom left, bottom right). This has been studied in the SSN for changes in contrast [53], but experimental data is currently not available for sufficiently fine manipulations of contrast to test this prediction comprehensively. In particular, a decrease in variability with increasing contrast for larger contrasts was robustly seen [21,22,32,165], but very low contrasts, for which an increase in variability would be expected, have not been carefully studied (an increase in

variability with increasing stimulus contrast at moderately lower contrasts was seen in a minority of cells[32]). Instead, experiments have seen a clear non-monotonic change in variability with increasing stimulus size at least in some layers of V1 – variability increasing for the smallest sizes, then decreasing for larger sizes[39,79]. The SSN mechanism reproduces surround suppression of firing rates[82] but the dependence of variability on stimulus size has not been explicitly studied. Nevertheless, we expect the same non-monotonic dependence as for stimulus contrast, as the same mechanisms should apply in both cases.

Experimental data also suggests that multiple surround mechanisms are engaged in different layers and by different spatiotemporal stimulus configurations[59,66,79,159]. Modulation of variability might also vary accordingly[39,40,79], including the possibility, described above, that when surround suppression is inherited from the feedforward inputs, it should act like a decrease in contrast and thus increase variability, while when it derives from recurrent cortical mechanisms, it represents increasingly strong inhibitory stabilization which decreases variability. Similar considerations apply to masking suppression, which includes a weaker cortical component[69] that can be described by the SSN[82], and a stronger component due to masking effects on the feedforward inputs to cortex[56,57,128–132].

In summary, in mechanistic models of V1 activity, response sub-additivity and variability quenching can both arise via a common mechanism: network effects that yield increasingly strong inhibitory stabilization. Variability quenching can also arise through stimulus pinning of wandering network activity, without any necessary connection to response sub-additivity. In all of these models, suppression of feedforward input will suppress responses and is expected to increase variability. Particular forms of sub-additivity or suppression may occur through different sets of these mechanisms in different locations. Thus, mechanistic models suggest that sub-additivity will often, but not always, co-occur with variability suppression.

Conclusions

We have seen that response sub-additivity in V1 often co-occurs with variability quenching. Response sub-additivity arises from nonlinear input transformations while response variability results from the accumulation and amplification of small amounts of noise as signals flow through neural circuits. It is therefore not obvious that both types of phenomena should have common origins. Yet that is exactly what we propose. This proposal is motivated by recent model-based insights into the functional operations, computational objectives, and circuit mechanisms that govern V1 activity. Although these modeling approaches address different aspects of cortical activity and rely on very different model architectures, they all predict that response sub-additivity and variability quenching will often co-occur. We do not wish to suggest that a single circuit mechanism underlies this relationship – different forms of response sub-additivity and variability quenching likely arise from distinct circuit mechanisms. Moreover, more work is needed to establish whether the discussed models are rich enough to account for the diversity of neural behaviors seen within the same experimental paradigm. That said, the converging insights naturally raise new questions. We end this review by considering three that seem particularly important to us: *"Can the modeling insights be unified?","Is the connection between response sub-additivity and variability quenching a canonical motif across cortex?"*, and *"Do specific model components map onto specific subtypes of neurons?"*

Descriptive, normative, and mechanistic modeling approaches offer different levels of explanation, but they are not mutually exclusive enterprises. Progress at one level can spark progress at another level. For example, refining descriptive models to better capture the diverse effects of surround stimulation on response suppression[3,11] has provided critical guidance for normative models of V1 activity[39,87]. Likewise, descriptive accounts of variability quenching across cortex[22] inspired progress in mechanistic models of spiking activity[53,134]. More direct examples of cross-level interactions are offered by recent attempts to combine different levels of explanation in a single model[83,126]. One study[126] bridged normative and mechanistic levels by optimizing the connectivity of the SSN architecture for probabilistic inference, so that SSN response variability closely matched the variability produced by a sampling-based normative model for stimuli with a cross-orientation mask. The network optimized for this variability structure was precisely in the SSN loosely balanced regime described above that shows response sub-additivity and variability quenching. The study also showed that this SSN regime produced other phenomena not previously studied, including contrast-controlled oscillations (see also[166]) and stimulus-onset transients, each of which played a functionally well-defined role in network computations.

Another recent study developed a model that bridges descriptive and a mechanistic levels. Specifically, the family of dynamic circuit models called Oscillatory Recurrent Gated Neural Integrator Circuits (ORGaNICs[83]) was explicitly designed to produce a steady state exactly described by the equations of divisive normalization. Similar to the SSN framework, recurrent inhibition in ORGaNICs stabilizes the network when recurrent excitation would otherwise make it unstable, producing both sub-additivity and variability quenching in stochastic variants of ORGaNICs (Martiniani and Heeger, personal communication). ORGaNICs relies on recurrent amplification through a multiplicative interaction between recurrent drive and recurrent gain that can be regarded as a phenomenological description of actual circuit mechanisms. One advantage of this model family is that its steady state and its variability and covariability can all be computed analytically, simplifying the study of large-scale and multi-area networks.

Looking forward, training deep neural networks whose connectivity resembles visual cortical circuitry to either perform visual tasks[167–170] or to predict responses of visual neurons[171–173] holds promise as a powerful approach to build bridges between descriptive, normative, and mechanistic approaches. However, thus far, this approach has not yielded any insight into neural response variability or its quenching – this is an important open challenge for future research.

Response sub-additivity and variability quenching are not lim-

ited to V1. Suppression of neural responses to a preferred stimulus by the simultaneous presentation of a non-preferred stimulus has been documented for many sensory [174–177] and non-sensory brain areas [178,179]. Likewise, the quenching of response variability by stimulus onset is thought to be a general property of cortical neurons [22]. This raises the question of whether the connection between response sub-additivity and variability quenching is a canonical motif across cortex. The insights provided by the V1 models we reviewed suggest that this may be the case. Specifically, these models suggest that both phenomena result from the neural mechanisms that implement the operation of divisive normalization. This operation is considered a canonical neural computation that is repeated modularly in many distinct brain systems through a variety of circuits and mechanisms [99]. Determining the generality of the co-occurrence of response sub-additivity and variability quenching may reveal a lawful aspect of neural activity and as such represents a crucial step for developing a principled understanding of cortical computation.

The models we discussed offer abstracted descriptions of neural stimulus-response transformations. As we have highlighted throughout this article, such abstractions can provide valuable insight into brain function *even* if the model components cannot be mapped onto biophysical substrates. Notwithstanding this, establishing such mapping is a quintessential goal of systems neuroscience. The recent advent of circuit-dissection tools capable of distinguishing the functional role of specific sub-types of cortical neurons [41,70,180] brings this goal within experimental reach.

In this article, we have focused on sub-additivity of firing rate and on response variability, both single-neuron response statistics. We have reviewed modeling frameworks that suggest unified descriptions and explanations for those phenomena, but that also help us distinguish separate mechanisms underlying similar phenomena. A natural and important extension of this work is to additionally consider pairwise and population-level response statistics (*e.g.* pairwise noise correlations and the geometry of population activity). These statistics have been studied extensively in cortical areas [181–183], are influenced by similar factors as those that elicit response sub-additivity and variability quenching [38,51,184–188], and further constrain models of neural activity.

Acknowledgements

We thank David Heeger, Stefano Martiniani, and Yashar Ahmadian for helpful discussions. This work was supported by US National Institutes of Health grants EY032999 (R.L.T.G.), EY030578 and DA056400 (R.C.-C.), EY025102, EY024071 and NS120562 (N.J.P), and U01NS108683 and U19NS107613 (K.D.M.), a CAREER award #2146369 (R.L.T.G.) and award DBI-1707398 (K.D.M.) from the National Science Foundation, a Wellcome Trust Investigator Award in Science 212262/Z/18/Z (M.L.), and Simons Foundation award 543017 and the Gatsby Charitable Foundation (K.D.M.).

Author contributions

All authors contributed to discussion, writing, and editing of this manuscript.

Competing interests statement

The authors declare no competing interests.

References

1. R. T. Born and R. B. Tootell. Single-unit and 2-deoxyglucose studies of side inhibition in macaque striate cortex. *PNAS*, 88(16):7071–7075, 1991.
2. M. P. Sceniak, M. J. Hawken, and R. Shapley. Visual spatial characterization of macaque v1 neurons. *Journal of neurophysiology*, 85(5):1873–1887, 2001.
3. J. R. Cavanaugh, W. Bair, and J. A. Movshon. Nature and interaction of signals from the receptive field center and surround in macaque v1 neurons. *Journal of neurophysiology*, 88(5):2530–2546, 2002.
4. R. Coen-Cagli, A. Kohn, and O. Schwartz. Flexible gating of contextual influences in natural vision. *Nature neuroscience*, 18(11):1648, 2015.
5. D. G. Albrecht and D. B. Hamilton. Striate cortex of monkey and cat: contrast response function. *Journal of neurophysiology*, 48(1):217–237, 1982.
6. M. C. Morrone, D. C. Burr, and L. Maffei. Functional implications of cross-orientation inhibition of cortical visual cells. i. neurophysiological evidence. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 216(1204): 335–354, 1982.
7. A. B. Bonds. Role of inhibition in the specification of orientation selectivity of cells in the cat striate cortex. *Visual Neuroscience*, 2(1):41–55, 1989.
8. G. C. DeAngelis, J. G. Robson, I. Ohzawa, and R. D. Freeman. Organization of suppression in receptive fields of neurons in cat visual cortex. *Journal of Neurophysiology*, 68(1): 144–163, 1992.
9. M. Carandini, D. J. Heeger, and J. A. Movshon. Linearity and normalization in simple cells of the macaque primary visual cortex. *Journal of Neuroscience*, 17(21):8621–8644, 1997.
10. D. Tolhurst and D. Heeger. Comparison of contrast-normalization and threshold models of the responses of simple cells in cat striate cortex. *Visual neuroscience*, 14 (2):293–309, 1997.
11. J. R. Cavanaugh, W. Bair, and J. A. Movshon. Selectivity and spatial distribution of signals from the receptive field surround in macaque v1 neurons. *Journal of neurophysiology*, 88(5):2547–2556, 2002.
12. R. Azouz and C. M. Gray. Cellular mechanisms contributing to response variability of cortical neurons in vivo. *Journal of Neuroscience*, 19(6):2209–2223, 1999.
13. S. Sadagopan and D. Ferster. Feedforward origins of response variability underlying contrast invariant orientation tuning in cat visual cortex. *Neuron*, 74(5):911–923, 2012.
14. S. Andoni, A. Tan, and N. J. Priebe. *The cortical assembly of visual receptive fields.* 2013.
15. G. J. Tomko and D. R. Crapper. Neuronal variability: non-stationary responses to identical visual stimuli. *Brain Research*, 79(3):405–418, 1974.
16. P. Heggelund and K. Albus. Response variability and orientation discrimination of single cells in striate cortex of cat.

Experimental Brain Research, 32(2):197–211, 1978.
17. R. Vogels, W. Spileers, and G. A. Orban. The response variability of striate cortical neurons in the behaving monkey. *Experimental Brain Research*, 77:432–436, 1989.
18. M. Shadlen and W. T. Newsome. The variable discharge of cortical neurons: implications for connectivity, computation, and information coding. *Journal of Neuroscience*, 18(10): 3870–3896, 1998.
19. B. C. Talluri, I. Kang, A. Lazere, K. R. Quinn, N. Kaliss, J. L. Yates, D. A. Butts, and H. Nienborg. Activity in primate visual cortex is minimally driven by spontaneous movements. *bioRxiv*, 2022.09. 08.507006, 2022.
20. D. J. Tolhurst, J. A. Movshon, and I. D. Thompson. The dependence of response amplitude and variance of cat visual cortical neurones on stimulus contrast. *Experimental brain research*, 41:414–419, 1981.
21. G. Orbán, P. Berkes, J. Fiser, and M. Lengyel. Neural Variability and Sampling-Based Probabilistic Representations in the Visual Cortex. *Neuron*, 92(2):530–543, 2016.
22. M. M. Churchland, Y. M. Byron, J. P. Cunningham, L. P. Sugrue, and et al. Stimulus onset quenches neural variability: a widespread cortical phenomenon. *Nature Neuroscience*, 13(3):369–378, 2010.
23. I. Arandia-Romero, S. Tanabe, J. Drugowitsch, A. Kohn, and R. Moreno-Bote. Multiplicative and additive modulation of neuronal tuning with population activity affects encoded information. *Neuron*, 89(6):1305–1316, 2016.
24. R. Rosenbaum, M. Smith, A. Kohn, J. Rubin, and B. Doiron. The spatial structure of correlated neuronal variability. *Nature Neuroscience*, 20(1):107–114, 2017.
25. Z. W. Davis, L. Muller, J. Martinez-Trujillo, T. Sejnowski, and J. H. Reynolds. Spontaneous travelling cortical waves gate perception in behaving primates. *Nature*, 587(7834):432–436, 2020.
26. T. Kenet, D. Bibitchkov, M. Tsodyks, A. Grinvald, and A. Arieli. Spontaneously emerging cortical representations of visual attributes. *Nature*, 425(6961):954–956, 2003.
27. P. Berkes, G. Orbán, M. Lengyel, and J. Fiser. Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. *Science*, 331(6013):83–87, 2011.
28. R. L. T. Goris, J. A. Movshon, and E. P. Simoncelli. Partitioning neuronal variability. *Nature Neuroscience*, 17(6): 858–865, 2014.
29. A. S. Ecker, P. Berens, R. J. Cotton, M. Subramaniyan, G. H. Denfield, C. R. Cadwell, S. M. Smirnakis, M. Bethge, and A. S. Tolias. State dependence of noise correlations in macaque primary visual cortex. *Neuron*, 82(1):235–248, 2014.
30. N. C. Rabinowitz, R. L. Goris, M. Cohen, and E. P. Simoncelli. Attention stabilizes the shared gain of V4 populations. *eLife*, 4:e08998, 2015.
31. R. L. Goris, C. M. Ziemba, J. A. Movshon, and E. P. Simoncelli. Slow gain fluctuations limit benefits of temporal integration in visual cortex. *Journal of vision*, 18(8):8–8, 2018.
32. R. Coen-Cagli and S. S. Solomon. Relating divisive normalization to neuronal response variability. *Journal of Neuroscience*, 39(37):7344–7356, 2019.
33. O. J. Hénaff, Z. M. Boundy-Singer, K. Meding, C. M. Ziemba, and R. L. T. Goris. Representation of visual uncertainty through neural gain variability. *Nature Communications*, 11(1):2513, 2020.
34. J. J. DiCarlo, D. Zoccolan, and N. C. Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, 2012.
35. X. Pitkow and D. E. Angelaki. Inference in the brain: statistics flowing in redundant population codes. *Neuron*, 94(5): 943–953, 2017.
36. R. Moreno-Bote, J. Beck, I. Kanitscheider, X. Pitkow, P. Latham, and A. Pouget. Information-limiting correlations. *Nature neuroscience*, 17(10):1410, 2014.
37. J. M. Beck, W. J. Ma, X. Pitkow, P. E. Latham, and A. Pouget. Not noisy, just wrong: the role of suboptimal inference in behavioral variability. *Neuron*, 74(1):30–39, 2012.
38. A. C. Snyder, M. J. Morais, A. Kohn, and M. A. Smith. Correlations in v1 are reduced by stimulation outside the receptive field. *Journal of Neuroscience*, 34(34):11222–11227, 2014.
39. D. Festa, A. Aschner, A. Davila, A. Kohn, and R. Coen-Cagli. Neuronal variability reflects probabilistic inference tuned to natural image statistics. *Nature communications*, 12(1):3635, 2021.
40. C. A. Henry and A. Kohn. Feature representation under crowding in macaque v1 and v4 neuronal populations. *Current Biology*, 32(23):5126–5137, 2022.
41. J. J. Nassi, M. C. Avery, A. H. Cetin, A. W. Roe, and J. H. Reynolds. Optogenetic activation of normalization in alert macaque visual cortex. *Neuron*, 86(6):1504–1517, 2015.
42. J. Isaacson and M. Scanziani. How inhibition shapes cortical activity. *Neuron*, 72(2):231–243, 2011.
43. H. Adesnik, W. Bruns, H. Taniguchi, Z. J. Huang, and M. Scanziani. A neural circuit for spatial summation in visual cortex. *Nature*, 490(7419):226–231, 2012.
44. B. V. Atallah, W. Bruns, M. Carandini, and M. Scanziani. Parvalbumin-expressing interneurons linearly transform cortical responses to visual stimuli. *Neuron*, 73(1):159–170, 2012.
45. A. Sanzeni, B. Akitake, H. C. Goldbach, C. E. Leedy, N. Brunel, and M. H. Histed. Inhibition stabilization is a widespread property of cortical networks. *Elife*, 9:e54875, 2020.
46. A. J. Keller, M. Dipoppa, M. M. Roth, M. S. Caudill, A. Ingrosso, K. D. Miller, and M. Scanziani. A disinhibitory circuit for contextual modulation in primary visual cortex. *Neuron*, 108(6):1181–1193, 2020.
47. D. J. Millman, G. K. Ocker, S. Caldejon, I. Kato, J. D. Larkin, E. K. Lee, J. Luviano, C. Nayan, T. V. Nguyen, and K. North. Vip interneurons in mouse primary visual cortex selectively enhance responses to weak but specific stimuli. *Elife*, 9: e55130, 2020.
48. J. Veit, G. Handy, D. P. Mossing, B. Doiron, and H. Adesnik. Cortical vip neurons locally control the gain but globally control the coherence of gamma band rhythms. *Neuron*, 111(3):405–417, 2023.
49. C. Stringer, M. Pachitariu, N. Steinmetz, C. B. Reddy, M. Carandini, and K. D. Harris. Spontaneous behaviors drive multidimensional, brainwide activity. *Science*, 364 (6437):eaav7893, 2019.

50. C. Stringer, M. Michaelos, D. Tsyboulski, S. E. Lindo, and M. Pachitariu. High-precision coding in visual cortex. *Cell*, 184(10):2767–2778, 2021.
51. O. Weiss, H. A. Bounds, H. Adesnik, and R. Coen-Cagli. Modeling the diverse effects of divisive normalization on noise correlations. *bioRxiv*, pages 2022–06, 2022.
52. G. Hennequin and M. Lengyel. Characterizing variability in nonlinear recurrent neuronal networks. *arXiv preprint arXiv:1610.03110*, 2016.
53. G. Hennequin, Y. Ahmadian, D. B. Rubin, M. Lengyel, and K. D. Miller. The dynamical regime of sensory cortex: stable dynamics around a single stimulus-tuned attractor account for patterns of noise variability. *Neuron*, 98(4):846–860, 2018.
54. D. H. Hubel and T. N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, 160(1):106–154.2, 1962.
55. H. B. Barlow, C. Blakemore, and J. D. Pettigrew. The neural mechanism of binocular depth discrimination. *The Journal of physiology*, 193(2):327–342, 1967.
56. N. J. Priebe and D. A. Ferster. Inhibition, spike threshold, and stimulus selectivity in primary visual cortex. *Neuron*, 57(4):482–497, 2008.
57. N. J. Priebe and D. Ferster. Mechanisms underlying cross-orientation suppression in cat visual cortex. *Nature neuroscience*, 9(4):552–561, 2006.
58. T. C. Freeman, S. Durand, D. C. Kiper, and M. Carandini. Suppression without inhibition in visual cortex. *Neuron*, 35 (4):759–771, 2002.
59. A. Angelucci, M. Bijanzadeh, L. Nurminen, F. Federer, S. Merlin, and P. C. Bressloff. Circuits and mechanisms for surround modulation in visual cortex. *Annual review of neuroscience*, 40:425–451, 2017.
60. L. Nurminen and A. Angelucci. Multiple components of surround modulation in primary visual cortex: multiple neural circuits with multiple functions? *Vision research*, 104:47–56, 2014.
61. A. M. Sillito, K. L. Grieve, H. E. Jones, J. Cudeiro, and J. Davis. Visual cortical mechanisms detecting focal orientation discontinuities. *Nature*, 378(6556):492–496, 1995.
62. G. A. Walker, I. Ohzawa, and R. D. Freeman. Asymmetric suppression outside the classical receptive field of the visual cortex. *Journal of Neuroscience*, 19(23):10536–10553, 1999.
63. D. G. Albrecht, W. S. Geisler, R. A. Frazor, and A. M. Crane. Visual cortex neurons of monkeys and cats: temporal dynamics of the contrast response function. *Journal of neurophysiology*, 88(2):888–913, 2002.
64. W. S. Geisler and D. G. Albrecht. Cortical neurons: isolation of contrast gain control. *Vision research*, 32(8):1409–1410, 1992.
65. W. Bair, J. R. Cavanaugh, and J. A. Movshon. Time course and time-distance relationships for surround suppression in macaque v1 neurons. *Journal of Neuroscience*, 23(20): 7690–7701, 2003.
66. B. S. Webb, N. T. Dhruv, S. G. Solomon, C. Tailby, and P. Lennie. Early and late mechanisms of surround suppression in striate cortex of macaque. *Journal of Neuroscience*, 25(50):11666–11675, 2005.
67. C. A. Henry, S. Joshi, D. Xing, R. M. Shapley, and M. J. Hawken. Functional characterization of the extraclassical receptive field in macaque v1: contrast, orientation, and temporal dynamics. *Journal of Neuroscience*, 33(14): 6230–6242, 2013.
68. G. A. Walker, I. Ohzawa, and R. D. Freeman. Binocular cross-orientation suppression in the cat's striate cortex. *Journal of Neurophysiology*, 79(1):227–239, 1998.
69. F. Sengpiel and V. Vorobyov. Intracortical origins of interocular suppression in the visual cortex. *Journal of Neuroscience*, 25(27):6394–6400, 2005.
70. S. C.-Y. Chen, G. Benvenuti, Y. Chen, S. Kumar, C. Ramakrishnan, K. Deisseroth, W. S. Geisler, and E. Seidemann. Similar neural and perceptual masking effects of low-power optogenetic stimulation in primate v1. *Elife*, 11: e68393, 2022.
71. G. Werner and V. B. Mountcastle. The variability of central neural activity in a sensory system, and its implications for the central reflection of sensory events. *Journal of Neurophysiology*, 26(6):958–977, 1963.
72. A. J. Parker and W. T. Newsome. Sense and the single neuron: probing the physiology of perception. *Annual review of neuroscience*, 21(1):227–277, 1998.
73. R. Vogels and G. A. Orban. How well do response changes of striate neurons signal differences in orientation: a study in the discriminating monkey. *Journal of Neuroscience*, 10 (11):3543–3558, 1990.
74. R. L. Goris, C. M. Ziemba, G. M. Stine, E. P. Simoncelli, and J. A. Movshon. Dissociation of choice formation and choice-correlated activity in macaque visual cortex. *Journal of Neuroscience*, 37(20):5195–5203, 2017.
75. A. I. Jasper, S. Tanabe, and A. Kohn. Predicting perceptual decisions using visual cortical population responses and choice history. *Journal of Neuroscience*, 39(34):6714–6727, 2019.
76. K. H. Britten, M. N. Shadlen, W. T. Newsome, and J. A. Movshon. The analysis of visual motion: a comparison of neuronal and psychophysical performance. *Journal of Neuroscience*, 12(12):4745–4765, 1992.
77. D. J. Tolhurst, J. A. Movshon, and A. F. Dean. The statistical reliability of signals in single neurons in cat and monkey visual cortex. *Vision research*, 23(8):775–785, 1983.
78. A. S. Charles, M. Park, J. P. Weller, G. D. Horwitz, and J. W. Pillow. Dethroning the fano factor: A flexible, model-based approach to partitioning neural variability. *Neural Computation*, 30(4):1012–1045, 2018.
79. L. Nurminen, M. Bijanzadeh, and A. Angelucci. Size tuning of neural response variability in laminar circuits of macaque primary visual cortex. *bioRxiv*, page 2023.01. 17.524397, 2023.
80. T. W. Troyer, A. E. Krukowski, N. J. Priebe, and K. D. Miller. Contrast-invariant orientation tuning in cat visual cortex: Feedforward tuning and correlation-based intracortical connectivity. *J. Neurosci.*, 18:5908–5927, 1998.
81. D. McLaughlin, R. Shapley, M. Shelley, and D. J. Wielaard. A neuronal network model of macaque primary visual cortex (v1): Orientation selectivity and dynamics in the input layer 4cα. *Proceedings of the National Academy of Sciences*, 97(14):8087–8092, 2000.
82. D. B. Rubin, S. D. Van Hooser, and K. D. Miller. The stabilized supralinear network: a unifying circuit motif underly-

ing multi-input integration in sensory cortex. *Neuron*, 85(2): 402–417, 2015

This study introduced the stabilized supralinear network (SSN) as a unifying circuit model of normalization. Both excitatory and inhibitory neurons exhibit response suppression.

83. D. J. Heeger and K. O. Zemlianova. A recurrent circuit implements normalization, simulating the dynamics of v1 activity. *Proceedings of the National Academy of Sciences*, 117(36):22494–22505, 2020.
84. B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.
85. A. J. Bell and T. J. Sejnowski. The 'independent components' of natural scenes are edge filters. *Vision Research*, 37(23):3327–3338, 1997.
86. R. P. N. Rao and D. H. Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1): 79–87, 1999.
87. O. Schwartz and E. P. Simoncelli. Natural signal statistics and sensory gain control. *Nature Neuroscience*, 4:819–825, 2001.
88. L. Wiskott and T. J. Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural computation*, 14 (4):715–770, 2002.
89. R. Coen-Cagli, P. Dayan, and O. Schwartz. Statistical models of linear and nonlinear contextual interactions in early visual processing. *Advances in neural information processing systems*, 22, 2009.
90. J. A. Movshon, I. D. Thompson, and D. J. Tolhurst. Receptive field organization of complex cells in the cat's striate cortex. *The Journal of physiology*, 283(1):79–99, 1978.
91. D. J. Heeger. Normalization of cell responses in cat striate cortex. *Visual neuroscience*, 9(2):181–197, 1992.
92. N. C. Rust, O. Schwartz, J. A. Movshon, and E. P. Simoncelli. Spatiotemporal elements of macaque V1 receptive fields. *Neuron*, 46(6):945–956, 2005.
93. B. Vintch, J. A. Movshon, and E. P. Simoncelli. A convolutional subunit model for neuronal responses in macaque v1. *Journal of Neuroscience*, 35(44):14829–14841, 2015.
94. R. L. Goris, E. P. Simoncelli, and J. A. Movshon. Origin and function of tuning diversity in macaque visual cortex. *Neuron*, 88(4):819–831, 2015.
95. J. Freeman and E. P. Simoncelli. Metamers of the ventral stream. *Nature Neuroscience*, 14:1195–1201, 2011.
96. N. C. Rust, V. Mante, E. P. Simoncelli, and J. A. Movshon. How mt cells analyze the motion of visual patterns. *Nature neuroscience*, 9(11):1421–1431, 2006.
97. R. L. Goris, T. Putzeys, J. Wagemans, and F. A. Wichmann. A neural population model for visual pattern detection. *Psychological review*, 120(3):472, 2013.
98. O. J. Hénaff, R. L. Goris, and E. P. Simoncelli. Perceptual straightening of natural videos. *Nature neuroscience*, 22 (6):984–991, 2019.
99. M. Carandini and D. J. Heeger. Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, 13: 51–62, 2012.
100. S. A. Cadena, G. H. Denfield, E. Y. Walker, L. A. Gatys, A. S. Tolias, M. Bethge, and A. S. Ecker. Deep convolutional models improve predictions of macaque v1 responses to natural images. *PLoS computational biology*, 15(4):e1006897, 2019.
101. D. G. Albrecht and W. S. Geisler. Motion selectivity and the contrast-response function of simple cells in the visual cortex. *Visual neuroscience*, 7(6):531–546, 1991.
102. A. K. Churchland, R. Kiani, R. Chaudhuri, X.-J. Wang, A. Pouget, and M. N. Shadlen. Variance as a signature of neural computations during decision making. *Neuron*, 69(4):818–831, 2011.
103. W. S. Geisler and D. Albrecht. Bayesian analysis of identification performance in monkey visual cortex: nonlinear mechanisms and stimulus certainty. *Vision research*, 35 (19):2723–2730, 1995.
104. W. S. Geisler. Visual Perception and the Statistical Properties of Natural Scenes. *Annual Review of Psychology*, 59 (1):167–192, 2008.
105. D. L. K. Yamins and J. J. DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365, 2016.
106. H. Barlow. Possible principles underlying the transformations of sensory messages. *Sensory communication*, 1(1): 217–234, 1961.
107. E. P. Simoncelli and B. A. Olshausen. Natural image statistics and neural representation. *Annu Rev Neurosci*, 24: 1193–1216, 2001.
108. J. H. van Hateren and D. L. Ruderman. Independent component analysis of natural image sequences yields spatiotemporal filters similar to simple cells in primary visual cortex. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 265(1412):2315–2320, 1998.
109. C. E. Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 1948.
110. M. O. Ernst and M. S. Banks. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870):429–433, 2002.
111. D. C. Knill and A. Pouget. The bayesian brain: the role of uncertainty in neural coding and computation. *TRENDS in Neurosciences*, 27(12):712–719, 2004.
112. Z. M. Boundy-Singer, C. M. Ziemba, and R. L. T. Goris. Confidence reflects a noisy decision reliability estimate. *Nature Human Behaviour*, 7(1):142–154, 2023.
113. W. J. Ma, J. M. Beck, P. E. Latham, and A. Pouget. Bayesian inference with probabilistic population codes. *Nature neuroscience*, 9(11):1432–1438, 2006.
114. M. Jazayeri and J. A. Movshon. Optimal representation of sensory information by neural populations. *Nature Neuroscience*, 9(5):690–696, 2006.
115. L. Buesing, J. Bill, B. Nessler, and W. Maass. Neural dynamics as sampling: a model for stochastic computation in recurrent networks of spiking neurons. *PLoS computational biology*, 7(11):e1002211, 2011.
116. T. Lochmann and S. Deneve. Neural processing as causal inference. *Current opinion in neurobiology*, 21(5):774–781, 2011.
117. A. Pouget, J. M. Beck, W. J. Ma, and P. E. Latham. Probabilistic brains: knowns and unknowns. *Nature neuroscience*, 16(9):1170–1178, 2013.
118. F. Meyniel, M. Sigman, and Z. F. Mainen. Confidence as bayesian probability: From neural origins to behavior. *Neu-*

ron, 88(1):78–92, 2015.
119. A. Pouget, J. Drugowitsch, and A. Kepecs. Confidence and certainty: distinct probabilistic quantities for different goals. *Nature neuroscience*, 19(3):366–374, 2016.
120. R. D. Lange and R. M. Haefner. Task-induced neural covariability as a signature of approximate bayesian learning and inference. *PLoS computational biology*, 18(3):e21009557, 2022.
121. M. J. Wainwright and E. P. Simoncelli. Scale mixtures of gaussians and the statistics of natural images. In S. A. Solla, T. K. Leen, and K.-R. Müller, editors, *Adv. Neural Information Processing Systems (NIPS*99)*, volume 12, pages 855–861, Cambridge, MA, may 2000. MIT Press.
122. J. Fiser, P. Berkes, G. Orbán, and M. Lengyel. Statistically optimal perception and learning: from behavior to neural representations. *Trends in cognitive sciences*, 14(3):119–130, 2010.
123. P. Hoyer and A. Hyvarinen. Interpreting neural response variability as monte carlo sampling of the posterior. *Advances in neural information processing systems*, 15, 2002.
124. I. Mareschal and R. M. Shapley. Effects of contrast and size on orientation discrimination. *Vision research*, 44(1): 57–67, 2004.
125. P. Dayan and L. F. Abbott. *Theoretical neuroscience: computational and mathematical modeling of neural systems*. MIT press, 2005.
126. R. Echeveste, L. Aitchison, G. Hennequin, and M. Lengyel. Cortical-like dynamics in recurrent circuits optimized for sampling-based probabilistic inference. *Nature neuroscience*, 23(9):1138–1149, 2020.
127. R. Coen-Cagli, P. Dayan, and O. Schwartz. Cortical surround interactions and perceptual salience via natural scene statistics. *PLoS computational biology*, 8(3): e1002405, 2012.
128. D. Barbera, N. J. Priebe, and L. L. Glickfeld. Feedforward mechanisms of cross-orientation interactions in mouse v1. *Neuron*, 110(2):297–311, 2022.
129. T. Z. Lauritzen, A. E. Krukowski, and K. D. Miller. Local correlation-based circuitry can account for responses to multi-grating stimuli in a model of cat V1. *J. Neurophysiol.*, 86:1803–1815, 2001.
130. A. S. Kayser, N. J. Priebe, and K. D. Miller. Contrast-dependent nonlinearities arise locally in a model of contrast-invariant orientation tuning. *J. Neurophysiol.*, 85: 2130–2149, 2001.
131. M. Carandini, D. J. Heeger, and W. Senn. A synaptic explanation of suppression in visual cortex. *Journal of Neuroscience*, 22(22):10053–10065, 2002.
132. T. C. Freeman, S. Durand, D. C. Kiper, and M. Carandini. Suppression without inhibition in visual cortex. *Neuron*, 35 (4):759–771, 2002.
133. P. Kara, P. Reinagel, and R. C. Reid. Low response variability in simultaneously recorded retinal, thalamic, and cortical neurons. *Neuron*, 27(3):635–646, 2000.
134. A. Litwin-Kumar and B. Doiron. Slow dynamics and high variability in balanced cortical networks with clustered connections. *Nature neuroscience*, 15(11):1498–1505, 2012. **This study showed that clustering of excitatory connections in balanced E/I networks results in both fast spiking variability and slow firing rate fluctuations.**
135. Y. Ahmadian, D. B. Rubin, and K. D. Miller. Analysis of the stabilized supralinear network. *Neural computation*, 25(8): 1994–2037, 2013.
136. H. Ozeki, I. M. Finn, E. S. Schaffer, K. D. Miller, and D. Ferster. Inhibitory stabilization of the cortical network underlies visual surround suppression. *Neuron*, 62(4):578–592, 2009.
137. R. J. Douglas, C. Koch, M. Mahowald, K. A. Martin, and H. H. Suarez. Recurrent excitation in neocortical circuits. *Science*, 269:981–985, 1995.
138. W. H. Bosking, Y. Zhang, B. Schofield, and D. Fitzpatrick. Orientation selectivity and the arrangement of horizontal connections in tree shrew striate cortex. *J. Neurosci.*, 17: 2112–2127, 1997.
139. L. J. Borg-Graham, C. Monier, and Y. Frégnac. Visual input evokes transient and strong shunting inhibition in visual cortical neurons. *Nature*, 393:369–373, 1998.
140. C. D. Gilbert and T. N. Wiesel. Morphology and intracortical projections of functionally characterised neurones in the cat visual cortex. *Nature*, 280(5718):120–125, 1979.
141. W. R. Softky and C. Koch. The highly irregular firing of cortical cells is inconsistent with temporal integration of random epsps. *Journal of neuroscience*, 13(1):334–350, 1993.
142. Z. F. Mainen and T. J. Sejnowski. Reliability of spike timing in neocortical neurons. *Science*, 268:1503–1506, 1995.
143. E. Schneidman, B. Freedman, and I. Segev. Ion channel stochasticity may be critical in determining the reliability and precision of spike timing. *Neural Comput*, 10:1679–1703, 1998.
144. C. O'Donnell and M. C. van Rossum. Systematic analysis of the contributions of stochastic voltage gated channels to neuronal noise. *Front Comput Neurosci*, 8:105, 2014.
145. C. F. Stevens and A. M. Zador. Input synchrony and the irregular firing of cortical neurons. *Nature neuroscience*, 1 (3):210–217, 1998.
146. M. R. DeWeese and A. M. Zador. Non-gaussian membrane potential dynamics imply sparse, synchronous activity in auditory cortex. *Journal of Neuroscience*, 26(47):12206–12218, 2006.
147. M. V. Tsodyks and T. Sejnowski. Rapid state switching in balanced cortical network models. *Network*, 6:111–124, 1995.
148. C. Van Vreeswijk and H. Sompolinsky. Chaos in neuronal networks with balanced excitatory and inhibitory activity. *Science*, 274(5293):1724–1726, 1996.
149. T. W. Troyer and K. D. Miller. Physiological gain leads to high ISI variability in a simple model of a cortical regular spiking cell. *Neural Comput.*, 9:971–983, 1997.
150. D. Amit and N. Brunel. Dynamics of a recurrent network of spiking neurons before and following learning. *Network: Comput. Neural Syst.*, 8:373–404, 1997.
151. Y. Ahmadian and K. D. Miller. What is the dynamical regime of cerebral cortex? *Neuron*, 109:3373–3391, 2021.
152. J. Kadmon and H. Sompolinsky. Transition to chaos in random neuronal networks. *Physical Review X*, 5(4):041030, 2015.
153. G. Deco and E. Hugues. Neural network mechanisms underlying stimulus driven variability reduction. *PLoS computational biology*, 8(3):e1002395, 2012.
154. G. B. Smith, B. Hein, D. E. Whitney, D. Fitzpatrick, and

M. Kaschube. Distributed network interactions and their emergence in developing neocortex. *Nature neuroscience*, 21(11):1600–1608, 2018.

155. S. Trägenap, D. E. Whitney, D. Fitzpatrick, and M. Kaschube. Visual experience drives the development of novel and reliable visual representations from endogenously structured networks. *Journal of Vision*, 23(9):5225–5225, 2023.
156. A. Ponce-Alvarez, A. Thiele, T. D. Albright, G. R. Stoner, and G. Deco. Stimulus-dependent variability and noise correlations in cortical MT neurons. *Proc. Natl. Acad. Sci. U.S.A.*, 110:13162–13167, 2013.
157. P. C. Bressloff. Stochastic neural field model of stimulus-dependent variability in cortical neurons. *PLoS computational biology*, 15(3):e1006755, 2019.
158. K. Rajan, L. Abbott, and H. Sompolinsky. Stimulus-dependent suppression of chaos in recurrent neural networks. *Physical review e*, 82(1):011903, 2010.
159. C. A. Henry, M. Jazayeri, R. M. Shapley, and M. J. Hawken. Distinct spatiotemporal mechanisms underlie extra-classical receptive field modulation in macaque V1 microcircuits. *Elife*, 9, 2020.
160. D. P. Mossing, J. Veit, A. Palmigiano, K. D. Miller, and H. Adesnik. Antagonistic inhibitory subnetworks control cooperation and competition across cortical space. *BioRxiv*, 2021. doi: https://doi.org/10.1101/2021.03.31.437953.
161. S. Di Santo, M. Dipoppa, A. Keller, M. Roth, M. Scanziani, and K. D. Miller. Unifying model for three forms of contextual modulation including feedback input from higher visual areas. *bioRxiv*, 2022.
162. M. V. Tsodyks, W. E. Skaggs, T. J. Sejnowski, and B. L. McNaughton. Paradoxical effects of external modulation of inhibitory interneurons. *Journal of neuroscience*, 17(11): 4382–4388, 1997.
163. P. Ekelmans, N. Kraynyukova, and T. Tchumatchenko. Targeting operational regimes of interest in recurrent neural networks. *PLOS Computational Biology*, 19(5):e1011097, 2023.
164. B. K. Murphy and K. D. Miller. Balanced amplification: a new mechanism of selective amplification of neural activity patterns. *Neuron*, 61(4):635–648, 2009.
165. I. M. Finn, N. J. Priebe, and D. Ferster. The emergence of contrast-invariant orientation tuning in simple cells of cat visual cortex. *Neuron*, 54:137–152, 2007.
166. C. J. Holt, K. D. Miller, and Y. Ahmadian. The stabilized supralinear network accounts for the contrast dependence of visual cortical gamma oscillations. *bioRxiv*, pages 2023–05, 2023. doi: https://doi.org/10.1101/2023.05.11.540442.
167. C. J. Spoerer, T. C. Kietzmann, J. Mehrer, I. Charest, and N. Kriegeskorte. Recurrent neural networks can explain flexible trading of speed and accuracy in biological vision. *PLoS computational biology*, 16(10):e1008215, 2020.
168. K. Kar, J. Kubilius, K. Schmidt, E. B. Issa, and J. J. DiCarlo. Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature neuroscience*, 22(6):974–983, 2019.
169. A. Nayebi, D. Bear, J. Kubilius, K. Kar, S. Ganguli, D. Sussillo, J. J. DiCarlo, and D. L. Yamins. Task-driven convolutional recurrent models of the visual system. *Advances in neural information processing systems*, 31, 2018.
170. M. Miller, S. Chung, and K. D. Miller. Divisive feature normalization improves image recognition performance in alexnet. In *International Conference on Learning Representations*, 2021.
171. N. Maheswaranathan, L. T. McIntosh, H. Tanaka, S. Grant, D. B. Kastner, J. B. Melander, A. Nayebi, L. E. Brezovec, J. H. Wang, and S. Ganguli. Interpreting the retinal neural code for natural scenes: From computations to neurons. *Neuron*, 111(17):2742–2755, 2023.
172. M. F. Burg, S. A. Cadena, G. H. Denfield, E. Y. Walker, A. S. Tolias, M. Bethge, and A. S. Ecker. Learning divisive normalization in primary visual cortex. *PLOS Computational Biology*, 17(6):e1009028, 2021.
173. X. Pan, A. DeForge, and O. Schwartz. Generalizing biological surround suppression based on center surround similarity via deep neural network models. *PLoS Comput Biol*, page in press, 2023.
174. D. Zoccolan, D. D. Cox, and J. J. DiCarlo. Multiple object response normalization in monkey inferotemporal cortex. *Journal of Neuroscience*, 25(36):8150–8164, 2005.
175. S. R. Olsen, V. Bhandawat, and R. I. Wilson. Divisive normalization in olfactory population codes. *Neuron*, 66(2): 287–299, 2010.
176. N. C. Rabinowitz, B. D. Willmore, J. W. Schnupp, and A. J. King. Contrast gain control in auditory cortex. *Neuron*, 70 (6):1178–1191, 2011.
177. T. Ohshiro, D. E. Angelaki, and G. C. DeAngelis. A neural signature of divisive normalization at the level of multisensory integration in primate cortex. *Neuron*, 95(2):399–411, 2017.
178. K. Louie, L. E. Grattan, and P. W. Glimcher. Reward value-based gain control: divisive normalization in parietal cortex. *Journal of Neuroscience*, 31(29):10627–10639, 2011.
179. A. K. Churchland, R. Kiani, and M. N. Shadlen. Decision-making with multiple alternatives. *Nature Neuroscience*, 11 (6):693–702, 2008.
180. L. Fenno, O. Yizhar, and K. Deisseroth. The development and application of optogenetics. *Annual review of neuroscience*, 34:389–412, 2011.
181. M. R. Cohen and A. Kohn. Measuring and interpreting neuronal correlations. *Nature neuroscience*, 14(7):811, 2011.
182. S. Chung and L. Abbott. Neural population geometry: An approach for understanding biological and artificial neural networks. *Current opinion in neurobiology*, 70:137–144, 2021.
183. N. Kriegeskorte and X.-X. Wei. Neural tuning and representational geometry. *Nature Reviews Neuroscience*, 22 (11):703–718, 2021.
184. A. Kohn and M. A. Smith. Stimulus dependence of neuronal correlation in primary visual cortex of the macaque. *Journal of Neuroscience*, 25:3661–3673, 2005.
185. M. R. Cohen and J. H. Maunsell. Attention improves performance primarily by reducing interneuronal correlations. *Nature neuroscience*, 12(12):1594–1600, 2009.
186. J. F. Mitchell, K. A. Sundberg, and J. H. Reynolds. Spatial attention decorrelates intrinsic activity fluctuations in macaque area v4. *Neuron*, 63(6):879–888, 2009.
187. D. A. Ruff and M. R. Cohen. Global cognitive factors modulate correlated response variability between v4 neurons. *Journal of Neuroscience*, 34(49):16408–16416, 2014.

188. B.-E. Verhoef and J. H. Maunsell. Attention-related changes in correlated neuronal activity arise from normalization mechanisms. *Nature Neuroscience*, 20(7):969–977, 2017.

Supplementary information

To illustrate the phenomena of response sub-additivity, response variability, and variability quenching, we analyzed simulated neural responses in Fig. 1a,b. The simulation was conducted in the following way. First, we created a temporal response profile that unfolded over 1 second in which a fast response rise is followed by a slow decay by multiplying a cumulative Gaussian function with an exponentially decaying function. To capture the phenomenon of surround suppression, we created three different average levels of responsiveness (left, middle, and right panels) by multiplying this profile with three different numbers such that it peaked at a rate of 4, 50, and 20 spikes per second. To capture the phenomenon of variability quenching, we let spikes arise from a doubly stochastic process. We simulated multiple trials (*i.e.*, repeated stimulus presentations). Each trial, we multiplied the stimulus-specific response profile with a random gain value to obtain a trial-specific firing-rate profile (Fig. 1a, middle row). The gain values were drawn from a gamma-distribution with a mean value of 1 and a variance that depended on the stimulus condition[28] (set to 1.5, 0.07, and 0.001 for the left, middle, and right panels). Finally, we obtained spike times by using the trial-specific firing rate profile as input for an inhomogeneous Poisson process[125] (Fig 1a, top row). The spike count histograms (Fig. 1a, bottom row) illustrate the resulting cross-trial distributions of spike counts (using a 1 second counting window). The variance to mean plot (Fig. 1b) illustrates the mean and variance of these spike count distributions.