PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

# Constrained inference in sparse coding reproduces contextual effects and predicts laminar neural dynamics

**Federica Capparelli*, Klaus Pawelzik, Udo Ernst**

Institute for Theoretical Physics, University of Bremen, Bremen, Germany

* federica@neuro.uni-bremen.de





**Citation:** Capparelli F, Pawelzik K, Ernst U (2019) Constrained inference in sparse coding reproduces contextual effects and predicts laminar neural dynamics. PLoS Comput Biol 15(10): e1007370. https://doi.org/10.1371/journal.pcbi.1007370

**Editor:** Blake A. Richards, University of Toronto at Scarborough, CANADA

**Received:** February 19, 2019

**Accepted:** September 2, 2019

**Published:** October 3, 2019



**Data Availability Statement:** All relevant data are within the manuscript and its Supporting Information files.

**Funding:** This work was supported by the BMBF (Bernstein Award, UE, Grant no. 01GQ1106) and by the University of Bremen, Creative Unit "I-See". The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing interests:** The authors have declared that no competing interests exist.

## Abstract

When probed with complex stimuli that extend beyond their classical receptive field, neurons in primary visual cortex display complex and non-linear response characteristics. Sparse coding models reproduce some of the observed contextual effects, but still fail to provide a satisfactory explanation in terms of realistic neural structures and cortical mechanisms, since the connection scheme they propose consists only of interactions among neurons with overlapping input fields. Here we propose an extended generative model for visual scenes that includes spatial dependencies among different features. We derive a neurophysiologically realistic inference scheme under the constraint that neurons have direct access only to local image information. The scheme can be interpreted as a network in primary visual cortex where two neural populations are organized in different layers within orientation hypercolumns that are connected by local, short-range and long-range recurrent interactions. When trained with natural images, the model predicts a connectivity structure linking neurons with similar orientation preferences matching the typical patterns found for long-ranging horizontal axons and feedback projections in visual cortex. Subjected to contextual stimuli typically used in empirical studies, our model replicates several hallmark effects of contextual processing and predicts characteristic differences for surround modulation between the two model populations. In summary, our model provides a novel framework for contextual processing in the visual system proposing a well-defined functional role for horizontal axons and feedback projections.

## Author summary

An influential hypothesis about how the brain processes visual information posits that each given stimulus should be efficiently encoded using only a small number of cells. This idea led to the development of a class of models that provided a functional explanation for various response properties of visual neurons, including the non-linear modulations observed when localized stimuli are placed in a broader spatial context. However, it remains to be clarified through which anatomical structures and neural connectivities a network in the cortex could perform the computations that these models require. In this

paper we propose a model for encoding spatially extended visual scenes. Imposing the constraint that neurons in visual cortex have direct access only to small portions of the visual field we derive a simple yet realistic neural population dynamics. Connectivities optimized for natural scenes conform with anatomical findings and the resulting model reproduces a broad set of physiological observations, while exposing the neural mechanisms relevant for spatio-temporal information integration.

## Introduction

Single neurons in the early visual system have direct access to only a small part of a visual scene, which manifests in their 'classical' receptive field (cRF) being localized in visual space. Hence for understanding how the brain forms coherent representations of spatially extended components or more complex objects in our environment, one needs to understand how neurons integrate local with contextual information represented in neighboring cells. Such integration processes already become apparent in primary visual cortex, where spatial and temporal context strongly modulate a cell's response to a visual stimulus inside the cRF. Electrophysiological studies revealed a multitude of signatures of contextual processing, leading to an extensive literature about these phenomena which have been termed 'non-classical' receptive fields (ncRFs) (for a review, see [1, 2]). ncRF modulations have a wide spatial range, extending up to a distance of 12 degrees of visual angle [3] and are tuned to specific stimulus parameters such as orientation [4]. Modulations are mostly suppressive [5], although facilitatory effects are also reported, especially for collinear arrangements where the center-stimulus is presented at low contrast [6] and for cross-orientation configurations [7, 8]. However, there is also a considerable variability in the reported effects, even in experiments where similar stimulation paradigms were used: for example, [6] found iso-orientation facilitation for low center stimulus contrasts, whereas another study [9] did not report facilitation at all, regardless of the contrast level. A further example [7] found strong cross-orientation facilitation, while [8] reports only moderate levels of cross-orientation facilitation, if at all. These discrepancies might be rooted in differences between the experimental setups, such as the particular choice of center/surround stimulus sizes, contrasts, and other parameters like the spatial frequency of the gratings, but might also be indicative of different neurons being specialized for different aspects of information integration.

From the observed zoo of different effects in conjunction with their apparent variability, the question arises if explanations based on a unique functional principle could provide a unifying explanation of the full range of these phenomena.

Even though the circuits linking neurons in visual cortex are still a matter of investigation, the nature of their properties suggest that the emergence of nCRF phenomena is a consequence of the interplay between different cortical mechanisms [10] that employ orientation-specific interactions between neurons with spatially separate cRFs. Anatomical studies have established that long-range horizontal connections in V1 have a patchy pattern of origin and termination, link preferentially cortical domains of similar functional properties, such as orientation columns, ocular dominance columns and CO compartments [11–13] and extend up to 8 mm [11, 14]. Although the functional specificity of feedback connections from extrastriate cortex is more controversial, some studies [15, 16] have reported that terminations of V2-V1 feedback projections are also clustered and orientation-specific, providing input from regions that are on average 5 times larger than the cRF. These results make both horizontal and

feedback connections well-suited candidates for mediating contextual effects, potentially with different roles for different spatio-temporal integration processes.

Is it possible to interpret the structure of these connections in terms of the purpose they serve?

For building a model of visual information processing from first principles, a crucial observation is that visual scenes are generated by a mixture of elementary causes. Typically, in any given scene, only *few* of these causes are present [17]. Hence, for constructing a neural explanation of natural stimuli, sparseness is likely to be a key requirement. Indeed electrophysiological experiments have demonstrated that stimulation of the nCRF increases sparseness in neural activity and decorrelates population responses, in particular under natural viewing conditions [18–20]. Perhaps the most influential work that linked sparseness to a form of neural coding that could be employed by cortical neurons was the paradigm introduced by Olshausen and Field [21]. After it was shown that sparseness, combined with unsupervised learning using natural images, was sufficient to develop features which resemble receptive fields of primary visual cortex [21–24], a number of extensions have been proposed that have successfully explained many other aspects of visual information processing, such as complex cell properties [25] and topographic organization [26]. Moreover, a form of code based on sparseness has many potential benefits for neural systems, being energy efficient [27], increasing storage capacity in associative memories [28, 29] and making the structure of natural signals explicit and easier to read out at subsequent level of processing [30]. Particularly noteworthy is the fact that these statistical models can be reformulated as dynamical systems [31], where processing units can be identified with real neurons having a temporal dynamics that can be implemented with various degrees of biophysical plausibility: using local learning rules [32], spiking neurons [33, 34] and even employing distinct classes of inhibitory neurons [35, 36]. In summary, sparse coding models nicely explain fundamental properties of vision such as classical receptive fields.

But can these models also explain signatures of contextual processing, namely non-classical receptive fields?

Recently, Zhu and Rozell reproduced a variety of key effects such as surround suppression, cross-orientation facilitation, and stimulus contrast-dependent ncRF modulations [37]. In their framework, small localized stimuli are best explained by activating the unit whose input field ('dictionary' vector) best matches the stimulus. If the stimuli grow larger, other units become also activated and compete for representing a stimulus, thus inducing ncRF modulations. This mechanism is similar to Bayesian models in which contextual effects are caused by surround units 'explaining away' the sensory evidence provided to a central unit [38]. The necessary interactions between neural units are mediated by couplings whose strengths are anti-proportional to the overlaps of the units' input fields. However, most of the effects observed in experiments are caused by stimuli extending far beyond the range of the recorded neuron's input fields [3, 5, 6]. Hence the mechanism put forward by this model [37] can only be a valid explanation for a small part of these effects, covering situations in which the surround is small and in close proximity to the cRF. This observation raises the important question, how sparse coding models have to be extended to better reflect cortical dynamics and anatomical structure. In particular, such models would have to allow for direct interactions between non-overlapping input fields.

If these models are then learned from natural images, which local and global coupling structures emerge, how do they compare to anatomical findings, and do they still exhibit the expected cRF properties? Can inference and learning dynamics be implemented in a biophysically realistic manner? Are such models capable of providing satisfactory explanations of ncRF

phenomena, and what are the underlying mechanisms? And finally, which predictions emerge from modeling and simulation for experimental studies?

In this paper, we address the above questions by building a novel framework to better capture contextual processing within the sparse coding paradigm. In particular, we define a generative model for visual scenes that takes into account spatial correlations in natural images. To perform inference in this model, we derive a biologically inspired dynamics and a lateral connection scheme that can be mapped onto a neural network of populations of neurons in visual cortex. We show that the emerging connectivity structures have similar properties to the recurrent interactions in cortex. Finally, we evaluate the model's ability to predict empirical findings reported in a set of electrophysiological experiments and we show that it replicates several hallmark effects of contextual processing. In summary, our model provides a unifying framework for contextual processing in the visual system proposing a well-defined functional role for horizontal axons.

## Results

### Extended generative model

The low-level, pixel representation of a natural image is multidimensional and complex. However, the corresponding scene can often be described by a much smaller number of high-level, spatially extended components such as textures, contours or shapes, which in turn are composed of more elementary, localized features such as oriented lines or grating patches. Standard sparse coding posits that images can be generated from linear combinations of such elementary features. In particular, it proposes that an image patch $\boldsymbol{s} \in \mathbb{R}^M$ can be written as

$$\boldsymbol{s} = \Phi \boldsymbol{a}, \tag{1}$$

where the feature vectors $\phi_i \in \mathbb{R}^M$ are arranged in a $M \times N$ matrix $\Phi$ often called 'dictionary' and the vector $\boldsymbol{a} \in \mathbb{R}^N$ contains the cofficients with which a particular image can be represented in feature space. An implicit assumption made by many sparse coding models (e.g. [21, 24, 39, 40]) is that the features are localized and thus have a limited spatial extent. Such assumption is plausible when features are interpreted as the synaptic input fields of cortical neurons, nevertheless it restricts sparse coding models to encoding only *small* patches of much larger images.

For constructing an extended generative model for natural scenes, we want to take into account that the presence of objects in the scene typically induces long-range dependencies among the elementary features—for instance, an oriented edge that belongs to a contour entails the presence of a co-aligned edge in its proximity [41, 42]. We start by considering a discretization of a (potentially large) visual scene. The simplest scenario that still allows to capture dependencies between features situated in different, non-overlapping locations consists in having two adjacent image patches, as the two horizontally aligned square regions indexed by $u$ and $v$ in Fig 1**A**. Next, we assume that the presence of a feature $i$ at one particular location $u$ can be 'explained' by the presence of features $j$ at other locations $v$ via coefficients $C_{ij}^{uv}$. We illustrate this in Fig 1**A**, where we have highlighted pairs of oriented edge that belong to the same object and that are thus present in both locations of the visual field. With such matrices $C^{uv}$ and $C^{vu}$ that capture the co-occurrence of features in different locations, we can then define the following feature representations

$$\boldsymbol{b}^u = \boldsymbol{a}^u + C^{uv}\boldsymbol{a}^v, \tag{2}$$

$$\boldsymbol{b}^v = \boldsymbol{a}^v + C^{vu}\boldsymbol{a}^u. \tag{3}$$

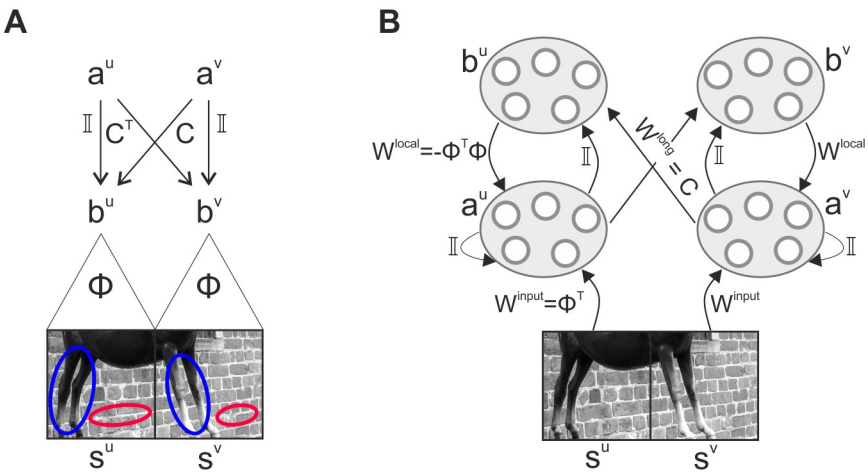


**Fig 1. Simplified generative model and neural inference network.** (**A**) In a simplified model, we consider visual scenes composed of two horizontally aligned, separate image patches which are encoded by their sparse representation $\boldsymbol{a}^u$, $\boldsymbol{a}^v$ via local features $\Phi$ and non-local dependencies $C$. The highlighted regions indicate how particular pairs of local features may co-occur due to the long-range dependencies induced by spatially extended objects. (**B**) Inference in the simplified generative model can be performed by a neural population dynamics (22) whose activities represent the coefficients $\boldsymbol{a}^u$, $\boldsymbol{a}^v$ and $\boldsymbol{b}^u$, $\boldsymbol{b}^v$. The corresponding neural circuit involves feedforward, recurrent, and feedback interactions which are functions of the dictionary $\Phi$ and of the long-range dependencies $C$.

https://doi.org/10.1371/journal.pcbi.1007370.g001

Furthermore, we assume a *reversal symmetry* $C^{uv}_{ij} = C^{vu}_{ji}$ for all $i, j$, which implies $C^{vu} = (C^{uv})^\top$: if the presence of a feature $i$ at location $u$ implies the presence of a feature $j$ at location $v$, then the presence of a feature $j$ at location $v$ should imply the presence of a feature $i$ at location $u$ to the same extent [41]. This allows us to drop the indexes $u$, $v$ from Eqs (2) and (3) and write $C^{uv} = C$ and $C^{vu} = C^\top$ with which, finally, we get

$$\boldsymbol{b}^u = \boldsymbol{a}^u + C\boldsymbol{a}^v \tag{4}$$

$$\boldsymbol{b}^v = \boldsymbol{a}^v + C^T\boldsymbol{a}^u \tag{5}$$

$$\boldsymbol{s}^u = \Phi\boldsymbol{b}^u \tag{6}$$

$$\boldsymbol{s}^v = \Phi\boldsymbol{b}^v. \tag{7}$$

In what follows, we will interpret the two patches as a 'central' and 'contextual' stimulus. The extension to more than two patches is straightforward and is presented in the supplement (S1 Text).

Note that such model might be considered as sparse coding with additional wiring constraint. In fact, substituting Eqs (4) and (5) into (6) and (7) and defining

$$\boldsymbol{s} = \begin{bmatrix} \boldsymbol{s}^u \\ \boldsymbol{s}^v \end{bmatrix}, \ \boldsymbol{a} = \begin{bmatrix} \boldsymbol{a}^u \\ \boldsymbol{a}^v \end{bmatrix} \text{ and } F = \begin{bmatrix} \Phi & \Phi C \\ \Phi C^\top & \Phi \end{bmatrix}$$

yields the classic linear mixture models used to investigate sparse coding of natural scenes [22, 43]

$$\boldsymbol{s} = F\boldsymbol{a}. \tag{8}$$

In fact, for $C = 0$, Eq (8) becomes exactly equivalent to the standard sparse coding model,

where image patches would be encoded independently without using the potential benefits of long-range dependencies.

**Learning visual features and their long-range dependencies.** To fully define the coding model we posit an objective function, used for optimization of the latent variables and the parameters. In our scheme, it allows to learn which fundamental features $\phi_i$ are best suited to encode an ensemble of images, and to derive a suitable inference scheme for the latent variables $\boldsymbol{a}^u$, $\boldsymbol{a}^v$ such that they optimally explain a given input image $(\boldsymbol{s}^u, \boldsymbol{s}^v)$ given the constraints. Most importantly, it allows to determine the spatial relations $C$ between pairs of features.

The objective function $E$ consists of four terms. The first two quantify how well the two image patches are represented, by means of computing the quadratic error between the patches and their reconstruction. The third and fourth terms require the representation in the coefficients $\boldsymbol{a}$'s and the matrix $C$ to be sparse, which is crucial for our assumption that only few non-zero coefficients are necessary to represent a complex image $(\boldsymbol{s}^u, \boldsymbol{s}^v)_\mu$ from an ensemble of images $\mu = 1, \ldots, P$. Mathematically, it is defined as

$$E_\mu(\boldsymbol{a}^u, \boldsymbol{a}^v, \Phi, C) = \left\| \boldsymbol{s}^u_\mu - \Phi(\boldsymbol{a}^u + C\boldsymbol{a}^v) \right\|_2^2 + \left\| \boldsymbol{s}^v_\mu - \Phi(\boldsymbol{a}^v + C^\top \boldsymbol{a}^u) \right\|_2^2 + \lambda_a(\|\boldsymbol{a}^u\|_1 + \|\boldsymbol{a}^v\|_1) + \lambda_C \|C\|_2^2. \tag{9}$$

The parameters $\lambda_a$ and $\lambda_C$ are sparseness constants, with larger values implying sparser representations. To obtain the matrices $\Phi$ and $C$ we used a gradient descent with respect to $\boldsymbol{a}^u$, $\boldsymbol{a}^v$, $\Phi$ and $C$ on the objective function defined by Eq (9). As image patches $(\boldsymbol{s}^u, \boldsymbol{s}^v)$ we used pairs of neighboring quadratic patches (aligned either horizontally or vertically) extracted from natural images (McGill data set [44]) after applying a whitening procedure as described in [22]. Our optimization scheme consisted of two alternating steps: First, we performed inference for an ensemble of image patches by iterating, for each image $\mu$,

$$\boldsymbol{a}^{\text{new}}_\mu = \boldsymbol{a}^{\text{old}}_\mu - \eta_a \frac{\partial E_\mu}{\partial \boldsymbol{a}} \tag{10}$$

until convergence to a steady state while holding $\Phi$ and $C$ fixed. Then, we updated $\Phi$ or $C$ by computing

$$\Phi^{\text{new}} = \Phi^{\text{old}} - \eta_\Phi \left\langle \frac{\partial E}{\partial \Phi} \right\rangle_\mu \tag{11}$$

or

$$C^{\text{new}} = C^{\text{old}} - \eta_C \left\langle \frac{\partial E}{\partial C} \right\rangle_\mu \tag{12}$$

with learning rates $\eta_\Phi$ and $\eta_C$, respectively. Angle brackets $\langle \cdots \rangle_\mu$ denotes the average over the image ensemble while keeping the $\boldsymbol{a}$'s at the steady states (for details, see Methods). This learning schedule reflects the usual assumption that inference and learning take place at different time scales. For increasing computational efficiency, we performed optimization in two phases. First, using only Eqs (10) and (11), we learned the dictionary $\Phi$ assuming $C = 0$, and second, using only Eqs (10) and (12), we obtained the long-range dependencies $C$ while holding $\Phi$ fixed.

## Inference with a biologically plausible dynamics

While in theory inference and learning can be realized by the general optimization scheme presented above, in the brain inference needs to respect neurobiological constraints. In what

follows, we derive a dynamics where the mixture coefficients $\boldsymbol{a}^u$, $\boldsymbol{a}^v$ and $\boldsymbol{b}^u$, $\boldsymbol{b}^v$ are activities of populations of neurons which we hypothesize to realize the necessary computations in cortical hyper-columns connected by local and long-range recurrent interactions (see Fig 1B). Hereby we require populations to have direct access only to 'local' image information, conveyed by their synaptic input fields.

For inference, we assume the quantities $\Phi$ and $C$ to be given and we associate each feature $i$ to one neural population having an internal state (e.g. an average membrane potential) and an activation level (i.e., its average firing rate). Following the approach of Rozell et al. [31], we define the population activities $\boldsymbol{a}^X = (a_j^X)_{j=1,\dots,N}$ as the thresholded values of the internal states $\boldsymbol{h}^X = (h_j^X)_{j=1,\dots,N}$ by setting

$$\boldsymbol{a}^X = [\boldsymbol{h}^X - \lambda_a]^+, \text{ for } X = u, v, \tag{13}$$

using the sparseness constant $\lambda_a$ as a threshold, and we let $\boldsymbol{h}^X$ evolve according to

$$\tau_h \dot{\boldsymbol{h}}^X = -\frac{\partial E}{\partial \boldsymbol{a}^X}. \tag{14}$$

The linear threshold operation ensures the positivity of $\boldsymbol{a}$, which is a necessary requirement for a neural output. Writing (14) explicitly leads to

$$\tau_h \dot{\boldsymbol{h}}^u = -\boldsymbol{h}^u + \Phi^\top \boldsymbol{s}^u - (\Phi^\top\Phi - \mathbb{I}_N + C\Phi^\top\Phi C^\top)\boldsymbol{a}^u - (C\Phi^\top\Phi + \Phi^\top\Phi C)\boldsymbol{a}^v + C\Phi^\top \boldsymbol{s}^v \tag{15}$$

$$\tau_h \dot{\boldsymbol{h}}^v = -\boldsymbol{h}^v + \Phi^\top \boldsymbol{s}^v - (\Phi^\top\Phi - \mathbb{I}_N + C^\top\Phi^\top\Phi C)\boldsymbol{a}^v - (C^\top\Phi^\top\Phi + \Phi^\top\Phi C^\top)\boldsymbol{a}^u + C^\top\Phi^\top \boldsymbol{s}^u. \tag{16}$$

Interpreting these equations in a neural context reveals one problem: The dynamics of the populations at location $u$ explicitly depends on the 'stimulus' (image patch) at location $v$—and vice versa (last terms on the r.h.s of Eqs (15) and (16)). This dependency violates our assumption of populations having access to only local image information. One way to get rid of this dependence is to approximate the input by its reconstruction suggested by the generative model, that is $\boldsymbol{s}^u = \Phi(\boldsymbol{a}^u + C\boldsymbol{a}^v)$ and $\boldsymbol{s}^v = \Phi(\boldsymbol{a}^v + C^T\boldsymbol{a}^u)$, which leads to

$$\tau_h \dot{\boldsymbol{h}}^u = -\boldsymbol{h}^u + \Phi^\top \boldsymbol{s}^u - (\Phi^\top\Phi - \mathbb{I}_N)\boldsymbol{a}^u - \Phi^\top\Phi C\boldsymbol{a}^v \tag{17}$$

$$\tau_h \dot{\boldsymbol{h}}^v = -\boldsymbol{h}^v + \Phi^\top \boldsymbol{s}^v - (\Phi^\top\Phi - \mathbb{I}_N)\boldsymbol{a}^v - \Phi^\top\Phi C^\top\boldsymbol{a}^u. \tag{18}$$

These two equations can be further simplified by extending the dynamical reformulation to include the coefficients $\boldsymbol{b}$ using Eqs (4) and (5). For this, we define another set of internal variables $\boldsymbol{k}^X$ satisfying

$$\boldsymbol{b}^X = [\boldsymbol{k}^X]^+ \tag{19}$$

and let them evolve according to a similar relaxation equation (i.e. leaky integration):

$$\tau_k \dot{\boldsymbol{k}}^u = -\boldsymbol{k}^u + \boldsymbol{a}^u + C\boldsymbol{a}^v \tag{20}$$

$$\tau_k \dot{\boldsymbol{k}}^v = -\boldsymbol{k}^v + \boldsymbol{a}^v + C^\top\boldsymbol{a}^u. \tag{21}$$

The final model is thus given by the following four differential equations

$$\begin{cases} \tau_h \dot{\boldsymbol{h}}^u = -\boldsymbol{h}^u + \Phi^\top \boldsymbol{s}^u - \Phi^\top \Phi \boldsymbol{b}^u + \boldsymbol{a}^u & = -\boldsymbol{h}^u + W^{\text{input}} \boldsymbol{s}^u + W^{\text{local}} \boldsymbol{b}^u + \boldsymbol{a}^u \\ \tau_h \dot{\boldsymbol{h}}^v = -\boldsymbol{h}^v + \Phi^\top \boldsymbol{s}^v - \Phi^\top \Phi \boldsymbol{b}^v + \boldsymbol{a}^v & = -\boldsymbol{h}^v + W^{\text{input}} \boldsymbol{s}^v + W^{\text{local}} \boldsymbol{b}^v + \boldsymbol{a}^v \\ \tau_k \dot{\boldsymbol{k}}^u = -\boldsymbol{k}^u + \boldsymbol{a}^u + C\boldsymbol{a}^v & = -\boldsymbol{k}^u + \boldsymbol{a}^u + W^{\text{long}} \boldsymbol{a}^v \\ \tau_k \dot{\boldsymbol{k}}^v = -\boldsymbol{k}^v + \boldsymbol{a}^v + C^\top \boldsymbol{a}^u & = -\boldsymbol{k}^v + \boldsymbol{a}^v + (W^{\text{long}})^\top \boldsymbol{a}^u \end{cases} \tag{22}$$

and by the linear threshold operations of Eqs (13) and (19).

This temporal dynamics can be implemented in a network of four neural populations organized in two cortical columns (Fig 1**B**). Specifically, populations $\boldsymbol{a}^u$ and $\boldsymbol{a}^v$ in the two columns receive *feed-forward* input $W^{\text{input}} = \Phi^\top$ from two different locations in the visual field. The input is then processed by a set of recurrent *local* connections that couple population $\boldsymbol{a}^u$ to $\boldsymbol{b}^u$ and $\boldsymbol{a}^v$ to $\boldsymbol{b}^v$ within the same column (matrices $\mathbb{I}$ and $W^{\text{local}} = -\Phi^\top \Phi$). The two populations $\boldsymbol{b}^u$ and $\boldsymbol{b}^v$ are also targets of *long-range* connections $W^{\text{long}} = C$ and $(W^{\text{long}})^\top$ originating from populations $\boldsymbol{a}^v$ and $\boldsymbol{a}^u$ in the neighboring column, respectively. For example, the two populations $\boldsymbol{a}$ and $\boldsymbol{b}$ inside a column could be interpreted as neural ensembles located in different cortical layers, or alternatively as two subpopulations in the same layer, but with different connection topologies. Note that the term 'long-range' not necessarily relates to long-ranging horizontal interactions—different anatomical interpretations are possible, and we will speculate on two alternative explanations in the Discussion.

The computation performed within single columns implements a competition based on tuned inhibition between units that code for similar features—which is a typical characteristics of sparse coding models—and it produces a sparse representation of the incoming stimulus. The interactions conveyed by horizontal connections between columns can induce modulatory effects on such a representation. All these connection patterns are completely determined by the matrices $\Phi$ and $C$.

Since the representations $\boldsymbol{a}$ and $\boldsymbol{b}$ can contain both positive and negative entries, each original unit can be realized by two neural populations which we will term 'ON' and 'OFF' units. Hereby ON-units represent positive activations of the original units, while OFF-units represent negative activations of the original units through positive neural activities. Accordingly, OFF-units are assigned the same shape of the synaptic input field, but with opposite polarity. With this necessary extension, Eq (14) implies that $\boldsymbol{a}$ will minimize the energy function $E$: even though the dynamics does not follow the gradient along the direction of its steepest slope, it still performs a gradient descent, since $\boldsymbol{a}$ is a monotonously increasing function of $\boldsymbol{h}$. We note here that, despite converging to the same fixed point, the dynamics defined by Eq (22) is not equivalent to performing standard gradient descent as in Eq (10) (see Discussion for a more complete explanation).

## Connection patterns and topographies

The link between the formal generative model and its realization as a cortical network allows to interpret $\Phi$ and $C$ (shown in Figs 2 and 3) in terms of the connection matrices $W^{\text{input}}$, $W^{\text{local}}$ and $W^{\text{long}}$.

After convergence of the training procedure (see Methods) our model produces feature vectors that resemble Gabor filters (Fig 2**A**), having spatial properties similar to those of V1 receptive fields. This result is a consequence of the sparseness constraint and does not come as a surprise, since it was obtained in a number of studies before [21, 23, 24], but verifies that our extended framework produces meaningful results by being able to learn similar features. The

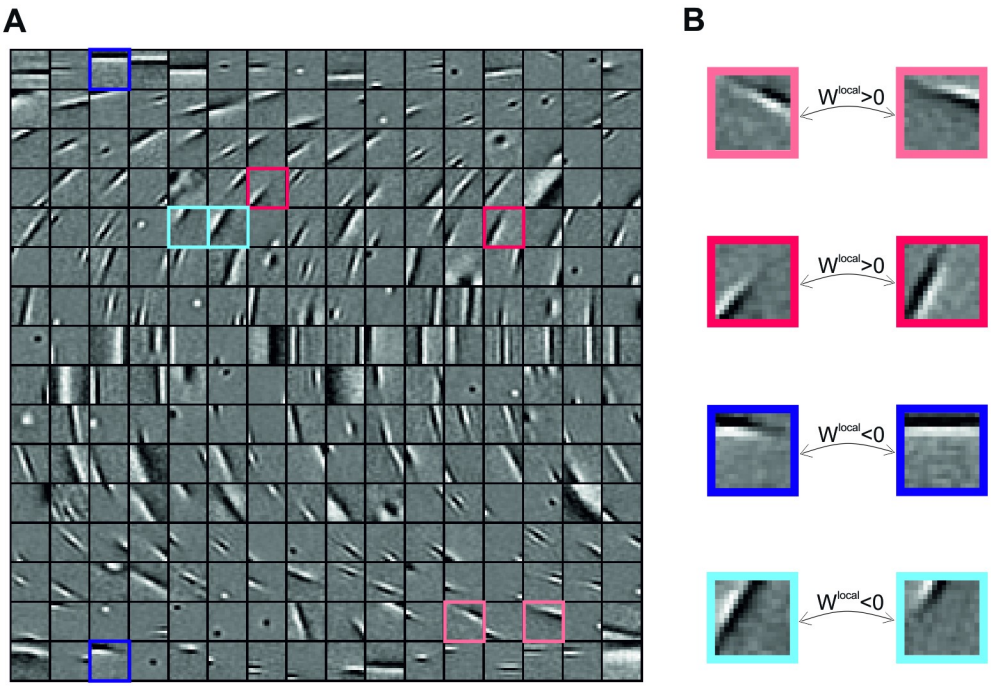


**Fig 2. Dictionary Φ and local connections.** (**A**) Feature vectors learned by training the model on natural images resemble localized Gabor filters. Features are ordered according to their orientation, which was estimated by fitting a Gabor function. Only a subset of the total set of $N = 1024$ dictionary elements is shown. (**B**) Units with overlapping input fields have strong short-range connections. The sign of the coupling is determined by the arrangement of on/off regions of the input fields: opposite phases correspond to excitatory connections (red) and matching phases to inhibitory connections (blue).

https://doi.org/10.1371/journal.pcbi.1007370.g002

variety of the dictionary elements is represented in Fig 2**A** and contains examples of localized and oriented Gabor-like patches, concentric shapes, and structures with multiple, irregularly shaped subfields. Each of the dictionary elements represents the synaptic input field of a neural unit and typically shows up as its classical RF when mapped with localized random stimuli through a reverse correlation procedure [22]. For further analysis, we extracted parameters that characterize the cell's tuning properties—namely its orientation preference, spatial frequency preference, RF center and size—by fitting a Gabor filter to each feature vector (see Methods). Typically, all feature vectors taken together build a complete representation for all orientations (and other stimulus features), thus the columns indicated in Fig 1**B** are similar to orientation hypercolumns found in primary visual cortex [45]. The distribution of orientation preferences exhibits a bias for cardinal orientations as observed in physiological studies [46].

As previously mentioned, short-range interactions are specified by the dictionary matrix through the equation $W^{\text{local}} = -\Phi^{\top}\Phi$. This implies that the absolute strength with which two units are locally connected is proportional to how closely their respective input fields match. In particular, as it is illustrated in Fig 2**B**, units with similar orientation preference and opposite phase are excitatorily connected, while units with similar orientation and similar phase are inhibitorily connected. Support for such like-to-like suppression can be found in a recent experiment [47], where optogenetic stimulation of mice V1 revealed a prominent inhibitory influence between neurons with similar tuning, suggesting that feature competition is indeed implied in sensory coding.

Together with the dictionary, we also learn the long-range feature dependencies $C$ (Fig 3 shows results relative to the horizontal configuration). To investigate which pattern of

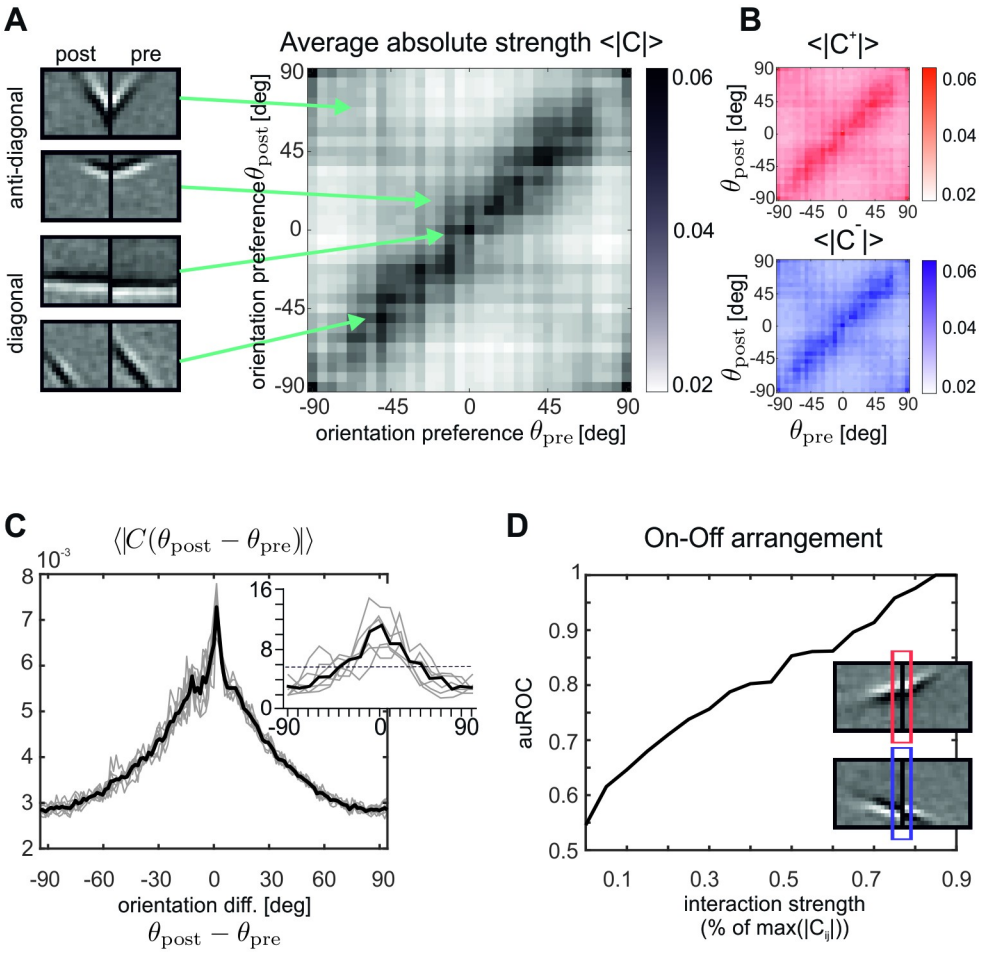


**Fig 3. Correlation matrix *C* and long-range interactions.** (**A**) Average absolute strength of long-range connections $W^{long} = C$ as a function of the orientation preferences of the pre- and postsynaptic units. Each point in the graph represents a connection from a unit responsive to the right portion of the visual field to a unit responsive to the left portion of the visual field (see four examples on the left). (**B**) Average absolute strength of excitatory (top, red color scale) and inhibitory (bottom, blue color scale) long-range connections as a function of the pre- and postsynaptic orientation preferences as in (**A**). (**C**) Average absolute strength of long-range connections as a function of the difference in orientation preference of the connected units. For comparison, data from the primary visual cortex of tree shrews are shown in the inset. The graph displays the percentage of boutons contacting postsynaptic sites that differ in orientation preference by a specified amount from the presynaptic injection site of a biocytin tracer. Individual cases are shown in gray and the median is shown in black. The dashed line reflects the percentage of boutons expected in each orientation difference bin if the boutons were distributed evenly over the map of orientation preference (redrawn from [13]). (**D**) Long-range interactions between units having positive correlations between the adjacent borders of their synaptic input fields tend to be excitatory (red frame in upper input fields example), while units having negative correlations tend to be inhibitory. This effect increases with increasing absolute coupling strengths $|C_{ij}|$, as indicated by the area under the ROC curve (auROC) computed from the corresponding correlation distributions for positive and negative connections.

https://doi.org/10.1371/journal.pcbi.1007370.g003

connections is induced, we computed the average absolute connection strength $\langle |W^{long}(\theta_{post}, \theta_{pre})| \rangle$ as a function of the orientation preferences $\theta_{post}$ and $\theta_{pre}$ of the units they connect (Fig 3A). The highest absolute connection strengths appear along the diagonal, indicating that pairs of units with similarly oriented input fields tend to be more strongly connected via long-range interactions. The distribution contains another structure, although more faint, located along the anti-diagonal, indicating that pairs of units whose orientations sum up to 0 degrees

are also strongly connected. Particular examples of units that have strong long-range coupling are shown in Fig 3A.

This result is consistent with anatomical measurements taken in primary visual cortex of mammals. Several experiments [11–13, 48, 49] report that horizontal long-range connections in V1 show a 'patchy' pattern of origin and termination, linking preferentially cortical domains responding to similar features. We quantified such a tendency in our model by computing the average connection strength as a function of the orientation preference difference $\Delta\theta = \theta_{\text{post}} - \theta_{\text{pre}}$ between pre- and post-synaptic cell. The corresponding graph is shown in Fig 3C, and a similar distribution obtained from anatomical measurements is reported for comparison in the inset.

In three shrew [13], cat [50] and monkeys [51], it has been shown that long-range connections between neurons of similar orientation selectivity exist primarily for neurons that are retinotopically aligned along the direction of their cells' preferences. We computed average absolute coupling strength between populations with aligned cRFs (i.e., 0 ± 15 degrees), and between populations with parallel cRFs (i.e., 90 ± 15 degrees), revealing that aligned couplings were indeed 26% percent stronger on average.

When splitting long-range interactions into negative and positive weights, we do not find any significant difference between their dependency on pre- and postsynaptic orientation preference (Fig 3B). However, a different pattern emerges when we take the polarities or phases of the synaptic input fields into account: For this purpose we measured the correlation $\rho$ between the right border of the left input field, and the left border of the right input field (colored frames in inset of Fig 3D), which are adjacent in visual space. Excitatory connections tend to exhibit positive correlations, while inhibitory connections tend to exhibit negative correlations. The stronger the couplings, the more pronounced this effect becomes. To quantify this effect, we compared the distributions $p(\rho \mid W_{ij}^{\text{long}} > \delta)$ for positive couplings larger than $\delta$ with the distributions $p(\rho \mid W_{ij}^{\text{long}} < -\delta)$ for negative couplings smaller than $-\delta$ by computing a receiver-operator characteristics ROC. Consistently, we find that separability as quantified by the area under ROC (auROC) increases with $\delta$ (Fig 3D). This effect is opposite to what we have (by construction) for the short-range connections: while units with similar cRFs *within a column* compete with each other, units with similar cRFs *across two columns* facilitate each other.

## Contextual effects

With the input fields (dictionary) and the long-range interactions obtained from a representative ensemble of natural images, the connectivity of the network represented in Fig 1 is completely specified. We can then subject the model to arbitrary stimulus configurations and investigate how well the dynamics described by Eqs (13), (19) and (22) predicts key effects exhibited by real neurons when processing contextual visual stimuli, and whether it can offer a coherent explanation to experimentally established context effects. For this purpose, we first selected units that were well driven and well tuned to the orientation $\theta_c$ of small patches $s_c$ of drifting sinusoidal gratings positioned at the center $\boldsymbol{r}^u$ of the left input region (cf. Fig 2B),

$$\begin{aligned} s_c(\boldsymbol{r}, t) &= k_c \gamma_c(\boldsymbol{r}) \sin\left(\omega_c(\boldsymbol{r} - \boldsymbol{r}^u)\boldsymbol{e}_{\theta_c} + \omega_t t\right), \\ \gamma_c(\boldsymbol{r}) &= \frac{1}{2}\left(1 + \tanh\left(\beta(r_c - |\boldsymbol{r} - \boldsymbol{r}^u|)\right)\right). \end{aligned} \tag{23}$$

Here $k_c$ denotes grating contrast, $r_c$ the radius of the patch, $\omega_c$ its spatial frequency, $\omega_t$ the drifting frequency and $\beta$ controls the steepness of the transition between stimulus und background. Thereby we mimic the situation in experiments in which typically also time-dependent stimuli

are used. Subsequently, these selected units were subjected to contextual stimulation, and the induced modulation by the context quantified.

In the following, we will focus on three exemplary stimulation paradigms in contextual processing, assessing size tuning, orientation-contrast effects, and luminance contrast effects.

**Size tuning.** Experiments in monkey and cat [5, 52] have shown that the stimulation of visual space surrounding the classical receptive field often has a suppressive influence on neurons in V1. Stimuli typically used to reveal this effect consist of a moving grating or an oscillating Gabor patch having the cell's preferred orientation, and being positioned at the center of its cRF. Recording the neural response while increasing the size of the grating yields the size tuning curve which exhibits two characteristic response patterns [5], as indicated in Fig 4**A**: After an initial increase in firing rate with increasing stimulus size, either the cell's response becomes suppressed and firing rate decreases (upper panel), or firing rate increases further and finally saturates (lower panel). In our model we realized a similar stimulation paradigm by using an optimally oriented grating (Eq (23)) and increasing its size $r_c$. Hence the stimulus first grows towards the border of the input field in which it is centered, and then extends into the neighboring fields. From all selected units, we show the size tuning curves of two exemplary cells in Fig 4**B**, demonstrating that the model can capture both qualitative behaviors known from cortical neurons.

For quantifying the degree of suppression and the extent to which this effect is present at the population level, we computed for all selected units a suppression index (SI) defined as

$$\mathrm{SI} = 1 - a_{\mathrm{full}} / \max_{r_c}(a(r_c)) \,,$$

where $a_{\mathrm{full}}$ was the response to a stimulus fully covering the input field. The SI indicates how much, in percentage, the response of a unit at largest stimulus size is reduced with respect to its maximum response, with 0 meaning no suppression and 1 meaning total suppression. The distribution of the SI across all the simulated cells is plotted in Fig 4**C**. For population $\boldsymbol{a}$, we find values comparable to what has been found experimentally: [5] reports that 44% of cells had less than 10% suppression and in the model the percentage of cells with SI$< 0.1$ is 38%. In general, the model shows less suppression (i.e., lower SI values) for population $\boldsymbol{b}$.

Since surround suppression was already observed in sparse coding models without long-range interactions [37], we expect this effect to stem from a combination of local and long-range connections. To quantify their roles in producing surround suppression, we simulated a version of the model without long-range interactions by setting $C = 0$. The resulting distribution of changes in SI is shown in Fig 4**D** and displays a mean increase of the SI for population $\boldsymbol{a}$ when including long-range connections, indicating that they contribute considerably to suppressive modulation induced by stimuli in the surround. In fact, without long-range interactions the percentage of cells with SI$< 0.1$ becomes 64%, which is quite far from the experimental result reported above. Conversely, the effect of including long-range connections is predominantly facilitatory for population $\boldsymbol{b}$, leading to a decrease in the observed SI's.

**Cross-orientation modulation.** Contextual processing is often probed by combining a central grating patch inside the cRF with a surround annular grating outside the cRF. For such configurations, the influence of the surround annulus on the response to an optimally oriented center stimulus was found to be orientation selective. When center and surround have the same orientation, the firing rate modulation is mostly suppressive, as we already know from studying size tuning (previous subsection).

If the surround strongly deviates from the orientation of the center, suppression becomes weaker [4, 8, 53, 54] and in some cases even facilitation with respect to stimulation of the center alone is observed [7, 55]. In particular, one study in cats [4] reports three typical response

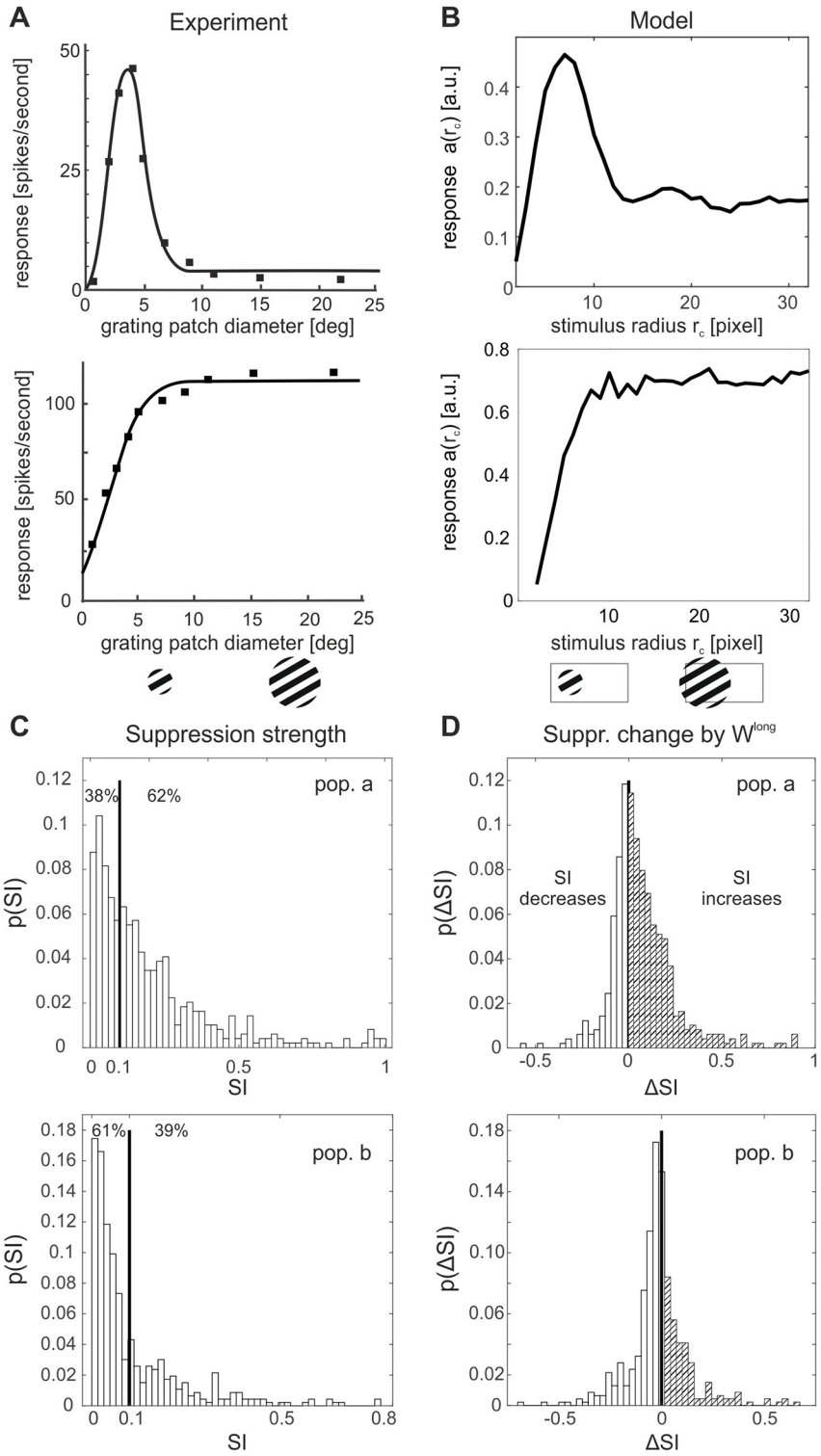


**Fig 4. Size tuning and surround suppression.** Dependence of neural responses on the size of a circular moving grating presented at the cell's preferred orientation. (**A**) Single-cell size tuning curves in primary visual cortex of cat exhibiting surround suppression (top) or saturation (bottom). Redrawn from [5]. (**B**) Size tuning curve of exemplary units in the model showing similar behaviour as in (**A**). (**C**) Distribution of suppression indices SI for the full model with long-range interactions. Values of 0 correspond to no suppression, values of 1 to full suppression. (**D**) Change in SI ($\Delta SI = SI^{with\ long} - SI^{without}$) induced by long-range connections. Enhanced suppression occurs more frequently than facilitation in population ***a***, while in population ***b*** one observes the opposite effect.

https://doi.org/10.1371/journal.pcbi.1007370.g004

patterns: (I) equal suppression regardless of the orientation of the surround, (II) suppression which decays with increasing difference between the orientations of center and surround, and (III) suppression that is strongest for small differences between orientations of center and surround, and weaker for large orientation differences and orientation differences close to zero. In the literature, the last effect is also termed 'iso-orientation release from suppression' (see Fig 5A for examples).

We realized this experimental paradigm in our model by combining a central grating patch (Eq (23)) with a surround annulus

$$\begin{aligned} s_a(\boldsymbol{r}, t) &= k_a \gamma_a(\boldsymbol{r}) \sin\left(\omega_a(\boldsymbol{r} - \boldsymbol{r}^u)\boldsymbol{e}_{\theta_a} + \omega_t t\right), \\ \gamma_a(\boldsymbol{r}) &= \frac{1}{4}(1 + \tanh\left(\beta(|\boldsymbol{r} - \boldsymbol{r}^u| - r_i)\right))(1 + \tanh\left(\beta(r_a - |\boldsymbol{r} - \boldsymbol{r}^u|)\right)) \end{aligned} \tag{24}$$

having orientation $\theta_a$, spatial frequency $\omega_a = \omega_c$, inner radius $r_i = r_c$, outer radius $r_a$, and grating contrast $k_a = k_c$. For each neural unit we investigated, the center stimulus had an optimal size defined by the radius $r_c$ for which we obtained the maximum response in the unit's size tuning curve. The surround annulus had the same parameters as the center patch and extended from the radius of the center patch to the whole input space (as displayed in Fig 5, stimulus icons in the legends). While the center orientation was held at the unit's preferred orientation, the surround orientation $\theta_a$ was systematically varied between 0 and $\pi$. For this experiment, we selected all units for which their optimal size was not larger than 21 pixels, to ensure that there was still space for a surround annulus in the restricted input space.

The three distinct behaviors observed in the experiments are qualitatively captured by the model: in Fig 5B (dashed lines) we show the orientation tuning curve of selected units of the model. Adding an annular surround stimulus to an optimally oriented center induces modulations which are mostly suppressive and tuned to the orientation of the surround (Fig 5B, solid lines). Cross-orientation modulations are summarized across the investigated model subpopulation in Fig 5C and 5D, where responses of cells exhibiting the same qualitative behavior are averaged together, as in the experiment (cf. panel **A**, see Methods for a detailed description of the pooling procedure). We distinguish, from top to bottom, untuned suppression, iso-orientation suppression, and iso-orientation release from suppression.

To assess the contributions of long-range connections to these effects, we repeated the experiment with $C = 0$. The population averages over the same categories of behaviors are overlaid in Fig 5C and 5D in gray. A comparison between the results of the model with and without $C$ shows that long-range interactions induce two different effects: enhancing responses for large orientation differences for cells with untuned surround suppression, and increasing maximum suppression for cells with tuned surround suppression in population ***a***. In particular, we observe strong facilitatory effects in population ***b***. This difference between the two populations might explain an apparent contradiction in experimental data where in a similar orientation contrast tuning paradigm one study exhibited strong facilitation [7], while a different investigation found only moderate release from suppression [8].

**Luminance-contrast effects.** In addition to orientation, also the relative contrast between the brightness of the center and the surround can be varied. In particular, such stimuli often reveal facilitatory effects, which are more frequently observed when the cRF is weakly activated, for example by presentation of a low-contrast visual stimulus. For many cells in V1 ($\approx$ 30%, in [6, 56]), collinear configurations of center-surround stimuli induce both facilitation and suppression. Here the visual contrast of the center stimulus in comparison to a fixed-contrast surround controls the sign of the modulation, and the point of crossover between suppression and facilitation is related to the cell's contrast threshold [3, 4, 6, 8, 57]. The

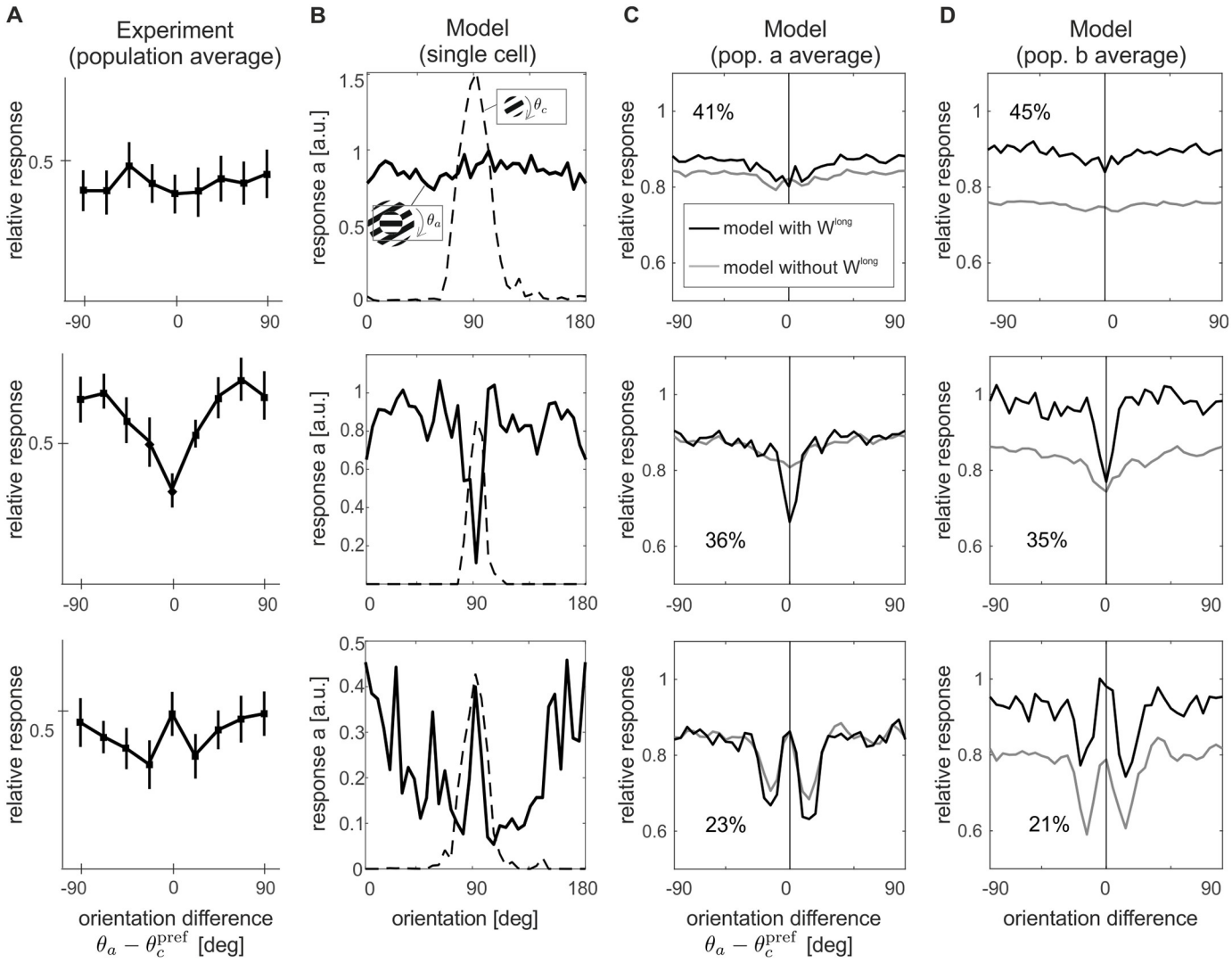


**Fig 5. Orientation-contrast modulations.** A center stimulus with preferred orientation is combined with an annulus of varying orientations (see icons in column (**B**)). (**A**) In experiments three response patterns are observed, namely, from top to bottom, untuned suppression, iso-orientation suppression and iso-orientation release from suppression (data replotted from [4]). The model reproduces these three response patterns both at the single cell level (**B**) and at the population level for ***a*** (**C**) and ***b*** (**D**). For comparison, orientation tuning for a center-alone stimulus is shown by the dashed line in (**B**). In (**C, D**), the gray lines display orientation-contrast tuning of the same ensembles without long-range interactions. Note that in (**A**) and (**C, D**), responses are shown normalized by the response to the center alone at the preferred orientations of the units. Percentages indicate the proportion of cells that fall in the same orientation-modulation class.

https://doi.org/10.1371/journal.pcbi.1007370.g005

characteristics of differential modulation is exemplified in Fig 6A where the contrast response function of a single cell in cat V1 (filled circles) is plotted together with the response of the same cell to the compound stimulus (empty circles). The graph shows that the same surround stimulus can enhance the response to a low-contrast center stimulus and reduce the response to a high-contrast center stimulus.

For obtaining corresponding contrast response curves in our model, we presented each selected unit with a center stimulus of optimal orientation and size of which we varied its contrast $k_c$ (Eq (23)). To mimic the collinear configuration of the compound stimulus, we then placed a surround annulus (Eq (24)) at high contrast $k_a = 1$, iso-oriented with the center patch (see stimulus icons in Fig 6), and again varied the contrast of the center patch. The resulting

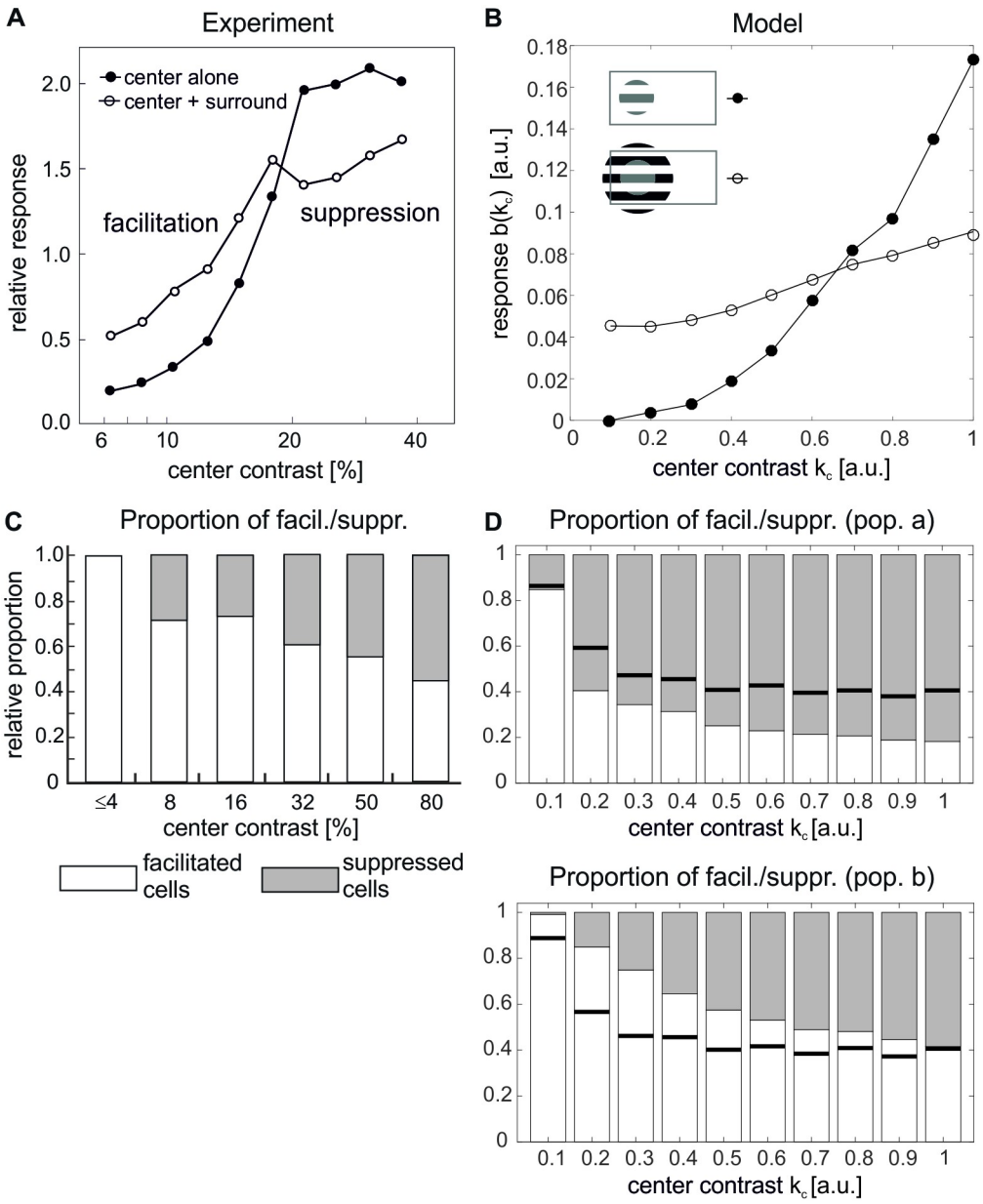


**Fig 6. Luminance contrast tuning. (A, B)** Single-cell responses to a center stimulus of varying contrast without flanking surround stimuli (filled circles) are compared to responses to the same center stimulus combined with high-contrast flanking surround stimuli of the same preferred orientation (open circles) in experiment **(A)** (redrawn from [6]) and model **(B)**. The stimulus configurations are indicated inside the graphs. **(C, D)** Population statistics, detailing the proportion of cells showing facilitation (light bars) or suppression (gray bars) in dependence on center stimulus contrast. Experimental data in **(C)** is redrawn from [6]. In the model **(D)**, cells were judged to be significantly facilitated (suppressed) if their activation ratio between center-surround and center alone stimulation $b^{\mathrm{sur}}(k_c)/b^{\mathrm{cen}}(k_c)$ at contrast $k_c$ was larger than $1 + \varepsilon$ (smaller than $1 - \varepsilon$), with $\varepsilon = 0.01$. Solid black lines indicate proportion of cells showing facilitation without long-range interactions. The top plot in **(D)** shows the statistics for population ***a*** and the bottom plot for populations ***b***.

https://doi.org/10.1371/journal.pcbi.1007370.g006

switch from facilitation to suppression, apparent by the crossing of the two response curves, is well captured by the model and illustrated for an example unit in population ***b*** in Fig 6B.

As in previous examples, differential modulation shows considerable variability across recorded cells. In particular, there are V1 neurons which exclusively show suppressive effects,

while other neurons exclusively exhibit facilitatory effects. The corresponding statistics is displayed in Fig 6C: For each value of contrast that was tested in [6], the bars show the proportion of cells that exhibit either facilitation or suppression. In particular, suppression becomes increasingly more common as the contrast of the center stimulus increases. The same analysis applied to our model reveals an identical result (Fig 6D), thus indicating that the model also captures the diversity of behaviors observed in electrophysiology. For population $\boldsymbol{b}$, the model statistics matches experimental findings also quantitatively. In particular, we observed that the increase in numbers of suppressed cells with increasing center contrast is mainly caused by the long-range connections, since this effect largely disappeared when we set $C = 0$ (horizontal lines in Fig 6D).

## Discussion

The pioneering work of Olshausen and Field [21] demonstrated that simple cell responses in primary visual cortex can be understood from the functional requirement that natural images should be represented efficiently by optimally coding an image with sparse activities. Since then, there have been many attempts to derive also other neuronal response properties in visual cortex from first principles. Common to these models is the framework of generative models, where the activities in an area are considered to represent the results of inference in the spirit of Helmholtz [58, 59]. Most of these investigations concentrate on local receptive field properties [22–26]. More recently, formal models were introduced that can qualitatively reproduce also several established non-classical receptive field effects [38, 60–63] and/or predict interactions resembling features of long-ranging horizontal and feedback connections in cortex [61, 64].

It is, however, unclear how the networks in cortex might perform the inference these models hypothesize given the neurobiological constraints on anatomy and neuronal dynamics. In this regard, the neural implementation proposed by [31] provided a significant advance, since it can explain a range of contextual effects [37] with a neural population dynamics that requires only synaptic summation and can also be extended to obey Dale's law [36]. But this model still presents a fundamental, conceptual difference to visual cortex: there are no interactions between neurons with non-overlapping input fields and thus the model can not account for the long-range modulatory influences from far outside the classical receptive field.

Here we propose a generative model for sparse coding of spatially extended visual scenes that includes long-range dependencies between local patches in natural images. An essential ingredient is the inclusion of plausible neural constraints by limiting the spatial extent of elementary visual features, thus mirroring the anatomical restrictions of neural input fields in primary visual cortex.

### Relations to standard sparse coding

As it becomes evident rearranging the equations that define the generative model (Eq (8)), our model offers an implementation of sparse coding that allows to encode spatially extended visual scenes. Although it might be tempting to consider it simply as a 'scaled-up' version of [31], we argue that this is indeed not the case. To demonstrate our reasoning, we consider the example of encoding a long horizontal bar. While a scaled-up-sparse-coding would have a specialized long horizontal feature to explain the stimulus (i.e. the sparsest representation), our model, by constraining the features to have a limited size, would require two separate horizontally aligned features to coactively form a representation of the stimulus; such collaborations between neighboring neurons are enforced by long-range connections.

## Connection structures

By optimizing model parameters via gradient descent it is possible to determine all connections in the network e.g. from the statistics of natural images. Synaptic input fields $\Phi$ resemble classical receptive fields of V1 neurons (Fig 2A). The structure of $C$ turns out to have similar characteristics as the anatomy of recurrent connections in visual cortex, exhibiting a preference to link neurons with similar orientation preferences via long-ranging horizontal axons [11, 50, 65] or via patchy feedback projections [15, 16]. Furthermore, we find a bias for collinear configurations being more strongly connected than parallel configurations, matching the observed elongation of cortical connection patterns along the axis of collinear configurations in the visual field in three shrew [13], cat [50] and monkeys [51]. These connection properties reflect regularities of the visual environment such as the edge co-occurrence observed in natural images [66].

The role of long-range connections in context integration was investigated also in a recent work [67]. Here the authors assume a neural code in which the firing rate of a neuron selective for a particular feature at a particular location is related to the probability of that feature to be present in an image, and influenced by the probability of other features being present in surrounding locations. In an analogous way as in our model, they assume that the only information a neuron tuned to a specific location in the visual field has about the stimulus context at neighboring locations comes from the neurons that are tuned to those neighboring locations (limited extent of the visual input). Thus, the lateral coupling scheme they obtain is also in good agreement with that observed in V1. Those connections are beneficial in increasing coding accuracy under the influence of noise, but the authors did not critically test their model with contextual stimulus configurations. Since their network does not implement competition, we expect their model to exhibit surround enhancement for co-aligned stimulus configurations, rather than the experimentally observed suppressive effects.

## Learning rules

To be a completely realistic model, still many details are missing. For example, the question of whether connectivity can be learned using realistic plasticity rules remains open. Currently our learning rules (11) and (12) require the change in single synapses to rely on information from *all* the neurons in the network. Moreover, the analytically derived formula for $W^{\text{local}}$ (Eq (22)) implies a pretty tight relation between the short-range interactions and the feed-forward weights and it is not clear which synaptic mechanisms could achieve it in parallel. The local plasticity rules used in [32] solved these issues in the context of the standard formulation of sparse coding, but it is not clear if a similar approach could be used to derive a learning rule also for $W^{\text{long}}$.

Finally, our model violates Dale's law, postulating direct inhibitory connections between excitatory cells (for both short- and long-range interactions). In the context of standard sparse coding, some work has been done to improve biological plausibility by implementing inhibition in a separate sub-population of neurons, both in spiking networks [35] and dynamical systems [36]. While the first model [35] consists in adding a second population of inhibitory units and then learning separately three sets of weights ($E - I$, $I - E$ and $I - I$), the second [36] relies on a low-rank decomposition of the recurrent connectivity matrix into positive and negative interactions. Both approaches were able to learn a sparse representation code and to develop Gabor-like input fields (notably, using the same E/I ratio observed in visual cortex). However, generalizing either one of them to our extended model might not be straightforward.

## Neural dynamics

Inference in the presented model is realized by a biologically realistic dynamics in a network of neural populations that are linked by short- and long-range connections. This implementation of a dynamics is close to the approach of Rozell [31] but additionally includes long-range interactions between units with non-overlapping input fields. Most importantly, the constraint that only local visual information is available to the units receiving direct input from the visual field implies, and predicts, that inference is performed by *two* separate neural populations with activities $\boldsymbol{a}$ and $\boldsymbol{b}$ and different connection structures.

It is worth to speculate about a direct relation to the particular properties of neurons and anatomical structures found in different layers and between areas of visual cortex: Physiological studies distinguish between the near ($< 2.5$ degrees) and far surround ($> 2.5$ degrees) in contextual modulation [10]. Taking into account the spread of long-range horizontal axons within V1, which is less than about three degrees in visual space [15], it seems likely that near surround effects are predominantly caused by horizontal interactions, while far surround effects are rather explained by feedback from higher visual areas. Assuming that one input patch in the model spans across 3 degrees in visual space, which is not implausible given the spatial extent of Gabor-like input fields shown in Fig 2**A** (up to 1 degree in cortex), we would therefore identify 'local' interactions $W^{\text{local}} = -\Phi^{\mathrm{T}}\Phi$ with horizontal axons within V1, while 'long-range' interactions $W^{\text{long}} = C$ would be mediated by the combination of feedforward and feedback connections between visual cortical areas. A possible circuit diagram emerging from this paradigm is depicted in the scheme in Fig 7**B**.

An alternative picture evolves if we assume that input patches correspond to smaller regions in visual space. Now horizontal interactions within V1 would span over sufficiently long distances to mediate long-range interactions in the model ($W^{\text{long}} = C$), while local interactions $W^{\text{local}}$ would indeed be local to a cortical (hyper-)column, possibly realized by the dense network linking different cortical layers in a vertical direction (example circuit shown in Fig 7**C**).

In both discussed scenarios structure and polarity of cortical interactions are compatible with the model: horizontal and feedback connections are orientation-specific, and their effective interaction can be positive or negative [48, 71] since they have been found to target both, excitatory and inhibitory neurons [68]. It is more difficult, however, to identify the potential locations of populations $\boldsymbol{a}$ and $\boldsymbol{b}$ in the different cortical layers. Two possibilities are shown in Fig 7. The reason why this choice is ambiguous is because indirect input from LGN is provided via layer IV to both superficial and deeper layers [72], because horizontal axons exists in both layers II-III and layers V-VI [69, 73], and because feedback from higher visual areas also terminates in both superficial and deep layers [74].

Finally, the proposed neural dynamics presents several non-trivial computational aspects, who are essential for producing the contextual effects we obtained. Even though the gradient descent (Eq (10)) and the proposed inference scheme (Eq (22)) have the same fixed points, the latter is much richer in its dynamics, since each reconstruction coefficient is represented by two neural activities who are in addition subject to rectification, and since activities $\boldsymbol{a}$ and $\boldsymbol{b}$ are associated with different neural time constants. In consequence, the effects we describe are most probably caused by a combination of sparse constrained coding and the particular properties of its neural implementation. The fact that all the experimental paradigms we reproduce in our model employ time-varying stimuli makes it hard to disentangle these different factors, since the inference network does never reach a steady state and the largest differences between a 'classic' gradient descent and neural dynamics are expected to show up in those transient epochs.

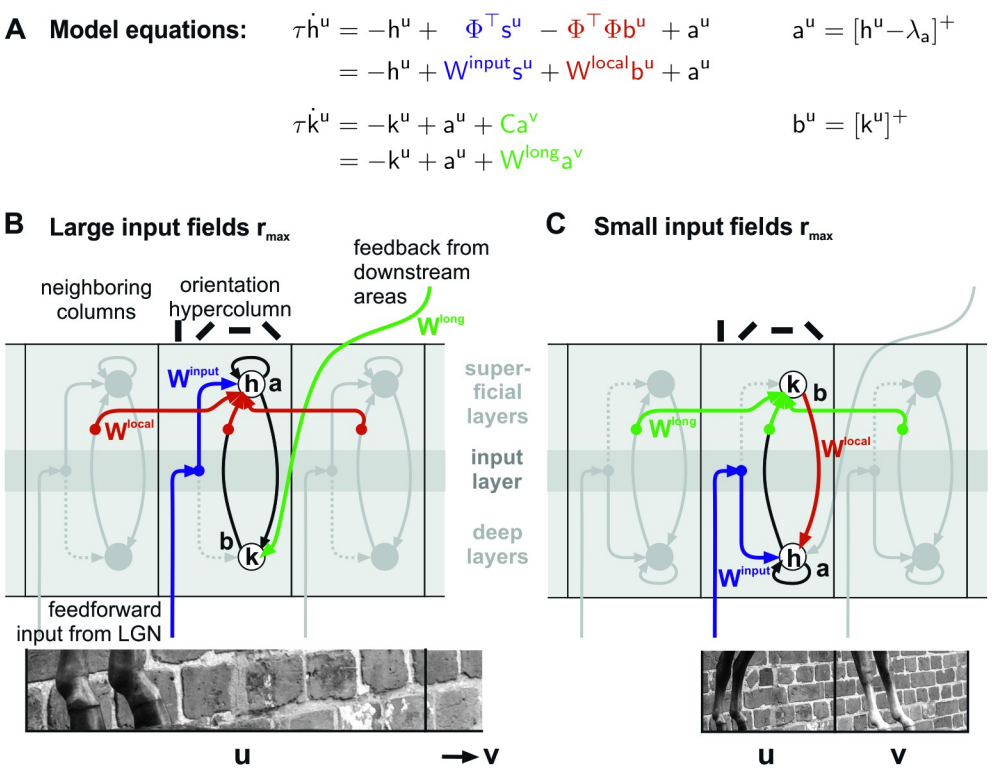


**Fig 7. Putative neural circuits performing inference in visual cortex.** (**A**) Equations that define the network dynamics. (**B, C**) Depending on the assumed spatial scale of input fields in the generative model, one distinguishes between cortical circuits where 'long-range' interactions $W^{long}$ would be mediated by recurrent loops between different cortical layers and 'local' interactions $W^{local}$ by long-ranging horizontal axons within primary visual cortex (**B**), or where long-range interactions $W^{long}$ would be mediated by long-ranging horizontal axons, and local interactions $W^{local}$ by the dense vertical/horizontal connection structures within a cortical hypercolumn (**C**). The length scales of input fields are indicated by the size of the image patch sections shown below. Interaction pathways associated with $W^{long}$, $W^{local}$ and $W^{input}$ are indicated in green, red and blue, respectively. Other links realizing different parts of the model equations (above the schemes) for column $u$ are drawn in black. The putative connection schemes are embedded into sections of primary visual cortex with light and dark gray shading indicating different layers. Note that in our scheme, horizontal interactions originate and terminate in different, but nearby layers as evident from anatomical evidence for layer II-III [68] and layer V-VI [69, 70] long-ranging axons, and that interactions might be indirect by being relayed over intermediary target populations (filled dots) such as inhibitory interneurons.

https://doi.org/10.1371/journal.pcbi.1007370.g007

## Contextual effects

Consistently the model reproduces a large variety of contextual phenomena, including size tuning, orientation-contrast effects and luminance-contrast modulations. In particular, all classical and non-classical receptive fields emerge in a fully unsupervised manner by training the model with ensembles of natural images. After training is finished, reproduction of all reported results is possible without change or fine-tuning of parameters, gains or thresholds—just by adhering to the exact visual stimulation procedures as used in the corresponding experimental studies. It is intriguing that also variability of the observed phenomena is reliably reflected in the statistics of model responses. Moreover, when we repeated the contextual-modulation experiments using a more general configuration of the visual field (i.e. using four surround patches instead of only one, as indicated in S1 Fig), we found that using a 'bigger' surround does not affect the agreement between our results and experimental data (the effects at the population-level are reported in S2, S3 and S4 Figs). This close match to experimental findings indicates that the assumed constraints from which dynamics and structure of the

model were derived are constructive for providing a comprehensive framework for contextual processing in the visual system.

The nature of the observed effects, being orientation-specific and exhibiting both enhancement and suppression (see Figs 4, 5 and 6), closely mirrors the structures and polarities of local and long-range interactions. Furthermore, they explicitly link functional requirements to the anatomy of the visual system: As already observed in [37], local interactions between similar features are strongly suppressive. They realize competition between alternative explanations of a visual scene which is related to 'explaining away' in Bayesian inference [38]. The effects of long-range interactions depend on the exact stimulus configuration, and on the balance between neural thresholds and the combination of all recurrent inputs in the inference circuit. They serve to integrate features across distances, leading to the enhancement of noisy evidence such as in low-contrast stimuli [6], but also to the suppression of activation by the model finding a simpler explanation for a complex stimulus configuration (i.e., by expressing the presence of multiple collinear line segments in terms of a single contour). This explicit link of natural statistics and cortical dynamics to function is also reflected in psychophysical studies: For example, in natural images an edge co-occurrence statistics being similar to the matrix $C$ was observed and used to quantitatively predict contour detection performance by human subjects via a local grouping rule [66]. High-contrast flankers aligned to a low-contrast center stimulus strongly modulated human detection thresholds [75], providing facilitation over long spatial and temporal scales of up to 16 seconds [76]. Also detection thresholds of 4-patch stimulus configurations are closely related to natural image statistics [77]. In both [75, 77], the interactions between feature detectors with similar cRF properties are inhibitory for near contexts, and exhibit disinhibitory or even facilitatory effects for far contexts—paralleling the differential effects that local and long-range interactions have in our model.

In parallel to sparse coding, hierarchical predictive coding has emerged as an alternative explanation for contextual phenomena [78]. The general idea is that every layer in a cortical circuit generates an error signal between a feedback prediction and feedforward inference, which is then propagated downstream in the cortical hierarchy. While being conceptually different on the inference dynamics, the corresponding hierarchical generative model of visual scenes is similar to our paradigm when subjected to spatial constraints.

Besides principled approaches, contextual processing has been investigated with models constructed directly from available physiological and anatomical evidence [79–81]. Core circuit of such models is often an excitatory-inhibitory loop with localized excitation and broader inhibition and different thresholds for the excitatory and inhibitory populations, which is similar to our proposed cortical circuits shown in Fig 7 with self-excitation of $\boldsymbol{a}$ and direct excitation on $\boldsymbol{b}$ and broader inhibition provided by $W^{\mathrm{local}}$ back onto $\boldsymbol{a}$. Such local circuits are connected by orientation-specific long-range connections, similar to the connections represented by $W^{\mathrm{long}}$, even though they are typically assumed to be more strongly tuned. From these structural similarities we would speculate that contextual effects are caused in both model approaches by similar effective mechanisms.

## Outlook

In summary, our paradigm provides a coherent, functional explanation of contextual effects and cortical connection structures from a first-principle perspective, which requires no fine-tuning to achieve a qualitative and quantitative match to a range of experimental findings. For future studies, the model has some important implications:

First, there are experimentally testable predictions. These include the strong dependency of local and long-range interactions on the relative phase of adjacent classical receptive fields.

Furthermore, we find two structures emerging in matrix $C$, namely a diagonal indicating stronger links between neurons with similar orientation preferences, as known from the literature, but also an anti-diagonal indicating enhanced links between neurons with opposite orientation preferences. Since connection probabilities were always reported w.r.t. orientation differences, the latter effect awaits experimental validation. Finally, we expect differences in the statistics of contextual effects between representations $\boldsymbol{a}$ and $\boldsymbol{b}$ to show up when information about the laminar origin of neural recordings is taken into account.

Second, it is formally straightforward to go back from the simplified model with just two separate input fields to the spatially extended, general scheme and subject it to much 'broader' visual scenes. Moreover, the neural dynamics allows also to address temporal contextual effects, or how neurons would respond to temporally changing contexts in the stimulus such as in 'natural' movies. For example, in simulations we observed strong transient effects shortly after stimulus onset, but a more thorough investigation and comparison to physiological findings is beyond the scope of this paper.

## Methods

### Learning and analysis of $\Phi$ and $C$

Variables $\Phi$ and $C$ were learned using the procedure outlined in the Results section (Eqs (10) and (11)). We sampled input patches of size $16 \times 32$ pixels (horizontal configuration) or $32 \times 16$ pixels (vertical configuration) from a database of natural images [44] from which we selected 672 images of size $576 \times 768$ pixels in uncompressed TIFF format. Images were first converted from RGB color space to grayscale values and then whitened using the method described in [22]. The optimization step for $\boldsymbol{a}$ (Eq (10)) was carried out for a batch of 100 image patches with a learning rate of $\eta_a = 0.01$. At the end of each update step for $\Phi$ (Eq (11)), the columns of $\Phi$ were normalized such that $||\phi_i||_2 = 1$. We learned $N = 1024$ feature vectors. Learning was performed with $10^4$ iterations each for $\Phi$ and $C$ (choosing as learning rates the values $\eta_\Phi = 0.05$ and $\eta_C = 0.01$), after which both dictionary and long-range dependencies matrices were stable. The parameters $\lambda_a$ and $\lambda_C$ were set to 0.5 and 0.02. To obtain a better statistics, we repeated learning of the dictionary and of the long-range interactions several times, initializing the simulations with different seeds. The results presented in Figs 3–6 are based on $N_{\text{seed}} = 8$ instances of the model.

To parametrize the feature vectors in terms of orientation, spatial frequency, size and location we fitted to each of them a Gabor function of the form

$$g(\theta, \lambda, \sigma_x, \sigma_y, x_0, y_0, \psi) = \kappa \exp\left(-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2}\right)\right) \cos\left(2\pi\frac{y}{\lambda} + \psi\right) + \kappa_0$$

$$x = \quad x_0 \cos(\theta) + y_0 \sin(\theta)$$

$$y = -x_0 \sin(\theta) + y_0 \cos(\theta),$$

where $\theta$ is the orientation of the sinusoidal carrier, $\lambda$ its wavelength, $\psi$ its phase, $\sigma_x$ and $\sigma_y$ are the standard deviations of the gaussian envelope, $\kappa > 0$ the contrast and $\kappa_0$ an offset. Fitting was done following a standard least square approach.

### Simulation of the neural model

The four differential equations that define the neural model (Eq (22)) were solved numerically with a Runge-Kutta method of order 4 for a time interval of $T = 600$ ms. The time constants $\tau_h$ and $\tau_k$ were chosen to be 10 ms, close to physiological values of neurons in cortex

[82]. For analyzing the responses, we discarded the initial transients and averaged over single cell activities over the last 333 ms, a period of time that allowed a complete cycle of the stimulus drifting with a temporal frequency of 3 Hz, being the average preferred speed for cortical neurons [83].

To ensure positivity of neural responses, in addition to the differential equations Eq (22) we had to introduce a linear threshold operation (Eqs (13) and (19)). In contrast, no constraint is imposed on the sign of $\boldsymbol{a}$ and $\boldsymbol{b}$ in the generative model (Eqs (4)–(7)), nor by the optimization Eq (10). To make the neural model consistent with the generative model, we therefore duplicated the number of neurons by introducing ON- and OFF units (see subsection Inference with a biologically plausible dynamics). In addition, we considered for all dictionary elements $\phi_i$ also their mirrored versions $-\phi_i$ and we split the long-range interactions into positive and negative contributions $C^+ = \max(C, 0)$ and $C^- = \min(C, 0)$ via

$$\Phi \leftarrow [\,\Phi \quad -\Phi\,] \in \mathbb{R}^{M\times 2N} \text{ and} \tag{25}$$

$$C \leftarrow \begin{bmatrix} C^+ & C^- \\ C^- & C^+ \end{bmatrix} \in \mathbb{R}^{2N\times 2N}. \tag{26}$$

For selecting cells well-tuned and well-responding to stimuli centered in one input patch (see Contextual effects), all units were first stimulated with a set of small drifting sinusoidal gratings centered at $\boldsymbol{r}^u$ with $r_c = 2$ pixels and $k_c = 1$. We varied $\theta_c$ from 0 to $\pi$ in steps of $\pi/N_\theta$ ($N_\theta = 36$) and the spatial frequency $f_c$ from 0.05 to 0.35 cycles/pixel in steps of 0.025. We then selected for each neuron the preferred orientation and preferred spatial frequency. A unit was said to be responsive if its peak response was at least 10% of the maximum recorded activity. We determined orientation selectivity by computing, for each unit $n$, the complex vector average

$$z_n = \sum_{k=0}^{N_\theta-1} a_n(\theta_k) e^{2i\theta_k} \Big/ \sum_{k=0}^{N_\theta-1} a_n(\theta_k), \text{ for } \theta_k = \frac{2\pi k}{N_\theta},$$

and we considered tuned those neurons for which it was $|z_n| > 0.85$, corresponding to a tuning width of approximately 20 degrees half-width. With these selection criteria, we were left with 490 cells from all $N_{\text{seed}}$ instantiations of the model.

## Selection of orientation contrast tuning classes

When we quantified the effect of cross-orientation stimulation, we pooled responses of units exhibiting the same qualitative behavior (Fig 5C and 5D). To determine which behavior a unit showed we first computed, for each unit $n$ with preferred orientation $\theta^\star$, the average response to the compound stimulus when the surround orientation was close to $\theta^\star$

$$\bar{a}_n^\star = \frac{1}{10^\circ} \int_{\theta^\star-5^\circ}^{\theta^\star+5^\circ} a_n(\theta_a) d\theta_a$$

and when the surround orientation was near-oblique

$$\bar{a}_n = \frac{1}{10^\circ} \int_{\theta^\star-20^\circ}^{\theta^\star-10^\circ} a_n(\theta_a) d\theta_a + \frac{1}{10^\circ} \int_{\theta^\star+10^\circ}^{\theta^\star+20^\circ} a_n(\theta_a) d\theta_a.$$

The unit was considered to show iso-orientation suppression if $\bar{a}_n - \bar{a}_n^\star > \varepsilon$, release from suppression if $\bar{a}_n^\star - \bar{a}_n > \varepsilon$ and untuned suppression in all other cases ($\varepsilon = 0.05$).

**Table 1. Parameter values.**

| | | | |
|---|---|---|---|
| **Size tuning** | | | |
| $\theta_c$ | preferred | $\theta_a$ | - |
| $\omega_c$ | preferred | $\omega_a$ | - |
| $r_c$ | from 2 to 32 in steps of 1 | $r_a$ | - |
| $k_c$ | 1 | $k_a$ | - |
| **Orientation-contrast (center-only)** | | | |
| $\theta_c$ | from 0 to $\pi$ in steps of $\pi/36$ | $\theta_a$ | - |
| $\omega_c$ | preferred | $\omega_a$ | - |
| $r_c$ | optimal | $r_a$ | - |
| $k_c$ | 1 | $k_a$ | - |
| **Orientation-contrast (center-surround)** | | | |
| $\theta_c$ | preferred | $\theta_a$ | from 0 to $\pi$ in steps of $\pi/36$ |
| $\omega_c$ | preferred | $\omega_a$ | preferred |
| $r_c$ | optimal | $r_a$ | $\infty$ |
| $k_c$ | 1 | $k_a$ | 1 |
| **Luminance-contrast (center-only)** | | | |
| $\theta_c$ | preferred | $\theta_a$ | - |
| $\omega_c$ | preferred | $\omega_a$ | - |
| $r_c$ | optimal | $r_a$ | - |
| $k_c$ | from 0.1 to 1 in steps of 0.1 | $k_a$ | - |
| **Luminance-contrast (center-surround)** | | | |
| $\theta_c$ | preferred | $\theta_a$ | preferred |
| $\omega_c$ | preferred | $\omega_a$ | preferred |
| $r_c$ | optimal | $r_a$ | $\infty$ |
| $k_c$ | from 0.1 to 1 in steps of 0.1 | $k_a$ | 1 |

https://doi.org/10.1371/journal.pcbi.1007370.t001

## Constants and parameters

Parameters used in numerical simulations are summarized in Table 1. The code to implement the model is available at https://github.com/FedericaCapparelli/ConstrainedInferenceSparseCoding.

## Supporting information

**S1 Text. In this document, we first outline how to extend the generative model to encode an arbitrary number *P* of patches and how to formulate it in terms of continuous variables for covering the full visual field and then we briefly report the results obtained performing the contextual-modulation experiments using a different, more general configuration of the visual field.**
(PDF)

**S1 Fig. Visual field (4–patch surround).** Structure of visual field used to investigate contextual phenomena, composed by one central and four surround patches. The same *cross* configuration is assumed for the cortical space, where $C^{uv}$ denotes long-range interactions between distant regions.
(TIF)

**S2 Fig. Size tuning and surround suppression (4–patch surround).** (**A**) Stimulus icons. (**B**) Distribution of suppression indices SI for the full model with long-range interactions.

Values of 0 correspond to no suppression, values of 1 to full suppression. (**C**) Change in SI ($\Delta SI = SI^{with\ long} - SI^{without}$) induced by long-range connections. Enhanced suppression occurs more frequently than facilitation in population ***a*** and, to a lesser extent, in population ***b***.
(TIF)

**S3 Fig. Orientation-contrast modulations (4–patch surround).** (**A**) Stimulus icons. (**B, C**) Response patterns observed experimentally reproduced by the model (from top to bottom, untuned suppression, iso-orientation suppression and iso-orientation release from suppression) in population ***a*** and ***b*** with (black curves) and without (gray curves) long-range interactions to an optimally oriented center stimulus combined with a concentric annulus of varying orientations. Note that responses are shown normalized by the response to the center alone at the preferred orientations of the units. Percentages indicate the proportion of cells that fall in the same orientation-modulation class.
(TIF)

**S4 Fig. Luminance contrast tuning (4–patch surround).** (**A**) Stimulus icons. (**B**) Population statistics, detailing the proportion of cells showing facilitation (light bars) or suppression (gray bars) in dependence on center stimulus contrast found in experiments (redrawn from [6]). (**C**) Population statistics computed from the model's responses of population ***a*** (top graph) and ***b*** (bottom graph). Cells were judged to be significantly facilitated (suppressed) if their activation ratio between center-surround and center alone stimulation $b^{sur}(k_c)/b^{cen}(k_c)$ at contrast $k_c$ was larger than $1 + \varepsilon$ (smaller than $1 - \varepsilon$), with $\varepsilon = 0.01$. Solid black lines indicate proportion of cells showing facilitation without long-range interactions.
(TIF)

## Acknowledgments

We would like to thank Andreas K. Kreiter for fruitful discussions about the potential links of our model framework to physiology and anatomy of the primate visual system.

## Author Contributions

**Conceptualization:** Federica Capparelli, Klaus Pawelzik, Udo Ernst.

**Formal analysis:** Federica Capparelli, Udo Ernst.

**Funding acquisition:** Klaus Pawelzik, Udo Ernst.

**Investigation:** Federica Capparelli.

**Methodology:** Federica Capparelli, Klaus Pawelzik, Udo Ernst.

**Project administration:** Udo Ernst.

**Software:** Federica Capparelli.

**Supervision:** Klaus Pawelzik, Udo Ernst.

**Validation:** Udo Ernst.

**Visualization:** Federica Capparelli, Udo Ernst.

**Writing – original draft:** Federica Capparelli, Udo Ernst.

**Writing – review & editing:** Klaus Pawelzik, Udo Ernst.

## References


1. Angelucci A, Shushruth S. Beyond the classical receptive field: surround modulation in primary visual cortex. In: The new visual neurosciences (Chalupa LM, Werner JS, eds), in press. Cambridge: MIT.
2. Series P, Lorenceau J, Frégnac Y. The "silent" surround of V1 receptive fields: theory and experiments. Journal of Physiology-Paris. 2003; 97(4-6):453–474. https://doi.org/10.1016/j.jphysparis.2004.01.023
3. Mizobe K, Polat U, Pettet MW, Kasamatsu T. Facilitation and suppression of single striate-cell activity by spatially discrete pattern stimuli presented beyond the receptive field. Visual neuroscience. 2001; 18 (3):377–391. https://doi.org/10.1017/s0952523801183045 PMID: 11497414
4. Sengpiel F, Sen A, Blakemore C. Characteristics of surround inhibition in cat area 17. Experimental Brain Research. 1997; 116(2):216–228. https://doi.org/10.1007/pl00005751 PMID: 9348122
5. Walker GA, Ohzawa I, Freeman RD. Suppression outside the classical cortical receptive field. Visual Neuroscience. 2000; 17(3):369–379. https://doi.org/10.1017/s0952523800173055 PMID: 10910105
6. Polat U, Mizobe K, Pettet MW, Kasamatsu T, Norcia AM. Collinear stimuli regulate visual responses depending on cell's contrast threshold. Nature. 1998; 391(6667):580. https://doi.org/10.1038/35372 PMID: 9468134
7. Sillito AM, Grieve KL, Jones HE, Cudeiro J, Davls J. Visual cortical mechanisms detecting focal orientation discontinuities. Nature. 1995; 378(6556):492. https://doi.org/10.1038/378492a0 PMID: 7477405
8. Levitt JB, Lund JS. Contrast dependence of contextual effects in primate visual cortex. Nature. 1997; 387(6628):73. https://doi.org/10.1038/387073a0 PMID: 9139823
9. Cavanaugh JR, Bair W, Movshon JA. Nature and interaction of signals from the receptive field center and surround in macaque V1 neurons. Journal of Neurophysiology. 2002; 88(5):2530–2546. https://doi.org/10.1152/jn.00692.2001 PMID: 12424292
10. Angelucci A, Bijanzadeh M, Nurminen L, Federer F, Merlin S, Bressloff PC. Circuits and mechanisms for surround modulation in visual cortex. Annual Review of Neuroscience. 2017; 40:425–451. https://doi.org/10.1146/annurev-neuro-072116-031418 PMID: 28471714
11. Gilbert CD, Wiesel TN. Columnar specificity of intrinsic horizontal and corticocortical connections in cat visual cortex. Journal of Neuroscience. 1989; 9(7):2432–2442. https://doi.org/10.1523/JNEUROSCI.09-07-02432.1989 PMID: 2746337
12. Malach R, Amir Y, Harel M, Grinvald A. Relationship between intrinsic connections and functional architecture revealed by optical imaging and in vivo targeted biocytin injections in primate striate cortex. Proceedings of the National Academy of Sciences. 1993; 90(22):10469–10473. https://doi.org/10.1073/pnas.90.22.10469
13. Bosking WH, Zhang Y, Schofield B, Fitzpatrick D. Orientation selectivity and the arrangement of horizontal connections in tree shrew striate cortex. Journal of Neuroscience. 1997; 17(6):2112–2127. https://doi.org/10.1523/JNEUROSCI.17-06-02112.1997 PMID: 9045738
14. Gilbert CD, Wiesel TN. Morphology and intracortical projections of functionally characterised neurones in the cat visual cortex. Nature. 1979; 280(5718):120. https://doi.org/10.1038/280120a0 PMID: 552600
15. Angelucci A, Levitt JB, Walton EJ, Hupe JM, Bullier J, Lund JS. Circuits for local and global signal integration in primary visual cortex. Journal of Neuroscience. 2002; 22(19):8633–8646. https://doi.org/10.1523/JNEUROSCI.22-19-08633.2002 PMID: 12351737
16. Shmuel A, Korman M, Sterkin A, Harel M, Ullman S, Malach R, et al. Retinotopic axis specificity and selective clustering of feedback projections from V2 to V1 in the owl monkey. Journal of Neuroscience. 2005; 25(8):2117–2131. https://doi.org/10.1523/JNEUROSCI.4137-04.2005 PMID: 15728852
17. Simoncelli EP, Olshausen BA. Natural image statistics and neural representation. Annual Review of Neuroscience. 2001; 24(1):1193–1216. https://doi.org/10.1146/annurev.neuro.24.1.1193 PMID: 11520932
18. Haider B, Krause MR, Duque A, Yu Y, Touryan J, Mazer JA, et al. Synaptic and network mechanisms of sparse and reliable visual cortical activity during nonclassical receptive field stimulation. Neuron. 2010; 65(1):107–121. https://doi.org/10.1016/j.neuron.2009.12.005 PMID: 20152117
19. Vinje WE, Gallant JL. Sparse coding and decorrelation in primary visual cortex during natural vision. Science. 2000; 287(5456):1273–1276. https://doi.org/10.1126/science.287.5456.1273 PMID: 10678835
20. Wolfe J, Houweling AR, Brecht M. Sparse and powerful cortical spikes. Current Opinion in Neurobiology. 2010; 20(3):306–312. https://doi.org/10.1016/j.conb.2010.03.006 PMID: 20400290
21. Olshausen BA, Field DJ. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature. 1996; 381(6583):607. https://doi.org/10.1038/381607a0 PMID: 8637596
22. Olshausen BA, Field DJ. Sparse coding with an overcomplete basis set: A strategy employed by V1? Vision Research. 1997; 37(23):3311–3325. https://doi.org/10.1016/s0042-6989(97)00169-7 PMID: 9425546

23. Bell AJ, Sejnowski TJ. The "independent components" of natural scenes are edge filters. Vision Research. 1997; 37(23):3327–3338. https://doi.org/10.1016/s0042-6989(97)00121-1 PMID: 9425547

24. Rehn M, Sommer FT. A network that uses few active neurones to code visual input predicts the diverse shapes of cortical receptive fields. Journal of Computational Neuroscience. 2007; 22(2):135–146. https://doi.org/10.1007/s10827-006-0003-9 PMID: 17053994

25. Hyvärinen A, Hoyer PO. A two-layer sparse coding model learns simple and complex cell receptive fields and topography from natural images. Vision Research. 2001; 41(18):2413–2423. https://doi.org/10.1016/s0042-6989(01)00114-6 PMID: 11459597

26. Hyvärinen A, Hoyer PO, Inki M. Topographic independent component analysis. Neural Computation. 2001; 13(7):1527–1558. https://doi.org/10.1162/089976601750264992 PMID: 11440596

27. Niven JE, Laughlin SB. Energy limitation as a selective pressure on the evolution of sensory systems. Journal of Experimental Biology. 2008; 211(11):1792–1804. https://doi.org/10.1242/jeb.017574 PMID: 18490395

28. Baum EB, Moody J, Wilczek F. Internal representations for associative memory. Biological Cybernetics. 1988; 59(4-5):217–228. https://doi.org/10.1007/BF00332910

29. Charles AS, Yap HL, Rozell CJ. Short-term memory capacity in networks via the restricted isometry property. Neural Computation. 2014; 26(6):1198–1235. https://doi.org/10.1162/NECO_a_00590 PMID: 24684446

30. Olshausen BA, Field DJ. Sparse coding of sensory inputs. Current Opinion in Neurobiology. 2004; 14 (4):481–487. https://doi.org/10.1016/j.conb.2004.07.007 PMID: 15321069

31. Rozell CJ, Johnson DH, Baraniuk RG, Olshausen BA. Sparse coding via thresholding and local competition in neural circuits. Neural Computation. 2008; 20(10):2526–2563. https://doi.org/10.1162/neco.2008.03-07-486 PMID: 18439138

32. Zylberberg J, Murphy JT, DeWeese MR. A sparse coding model with synaptically local plasticity and spiking neurons can account for the diverse shapes of V1 simple cell receptive fields. PLoS Computational Biology. 2011; 7(10):e1002250. https://doi.org/10.1371/journal.pcbi.1002250 PMID: 22046123

33. Hu T, Genkin A, Chklovskii DB. A network of spiking neurons for computing sparse representations in an energy-efficient way. Neural Computation. 2012; 24(11):2852–2872. https://doi.org/10.1162/NECO_a_00353 PMID: 22920853

34. Shapero S, Rozell C, Hasler P. Configurable hardware integrate and fire neurons for sparse approximation. Neural Networks. 2013; 45:134–143. https://doi.org/10.1016/j.neunet.2013.03.012 PMID: 23582485

35. King PD, Zylberberg J, DeWeese MR. Inhibitory interneurons decorrelate excitatory cells to drive sparse code formation in a spiking model of V1. Journal of Neuroscience. 2013; 33(13):5475–5485. https://doi.org/10.1523/JNEUROSCI.4188-12.2013 PMID: 23536063

36. Zhu M, Rozell CJ. Modeling inhibitory interneurons in efficient sensory coding models. PLoS Computational Biology. 2015; 11(7):e1004353. https://doi.org/10.1371/journal.pcbi.1004353 PMID: 26172289

37. Zhu M, Rozell CJ. Visual nonclassical receptive field effects emerge from sparse coding in a dynamical system. PLoS Computational Biology. 2013; 9(8):e1003191. https://doi.org/10.1371/journal.pcbi.1003191 PMID: 24009491

38. Lochmann T, Ernst UA, Deneve S. Perceptual inference predicts contextual modulations of sensory responses. Journal of Neuroscience. 2012; 32(12):4179–4195. https://doi.org/10.1523/JNEUROSCI.0817-11.2012 PMID: 22442081

39. Lewicki MS, Sejnowski TJ. Learning overcomplete representations. Neural Computation. 2000; 12 (2):337–365. https://doi.org/10.1162/089976600300015826 PMID: 10636946

40. Karklin Y, Lewicki MS. Learning higher-order structures in natural images. Network: Computation in Neural Systems. 2003; 14(3):483–499. https://doi.org/10.1088/0954-898X_14_3_306

41. Williams LR, Thornber KK. Orientation, scale, and discontinuity as emergent properties of illusory contour shape. Neural Computation. 2001; 13(8):1683–1711. https://doi.org/10.1162/08997660152469305 PMID: 11506666

42. Ernst UA, Mandon S, Schinkel-Bielefeld N, Neitzel SD, Kreiter AK, Pawelzik KR. Optimality of human contour integration. PLoS Computational Biology. 2012; 8(5):e1002520. https://doi.org/10.1371/journal.pcbi.1002520 PMID: 22654653

43. Hoyer PO. Non-negative sparse coding. In: Neural Networks for Signal Processing, 2002. Proceedings of the 12th IEEE Workshop on Neural Networks for Signal Processing 2002. p. 557–565.

44. Olmos A, Kingdom FA. A biologically inspired algorithm for the recovery of shading and reflectance images. Perception. 2004; 33(12):1463–1473. https://doi.org/10.1068/p5321 PMID: 15729913

45. Hubel DH, Wiesel TN. Uniformity of monkey striate cortex: a parallel relationship between field size, scatter, and magnification factor. Journal of Comparative Neurology. 1974; 158(3):295–305. https://doi.org/10.1002/cne.901580305 PMID: 4436457

46. Wang G, Ding S, Yunokuchi K. Difference in the representation of cardinal and oblique contours in cat visual cortex. Neuroscience letters. 2003; 338(1):77–81. https://doi.org/10.1016/s0304-3940(02)01355-1 PMID: 12565144

47. Chettih SN, Harvey CD. Single-neuron perturbations reveal feature-specific competition in V1. Nature. 2019; p. 1.

48. Weliky M, Kandler K, Fitzpatrick D, Katz LC. Patterns of excitation and inhibition evoked by horizontal connections in visual cortex share a common relationship to orientation columns. Neuron. 1995; 15(3):541–552. https://doi.org/10.1016/0896-6273(95)90143-4 PMID: 7546734

49. Yoshioka T, Blasdel GG, Levitt JB, Lund JS. Relation between patterns of intrinsic lateral connectivity, ocular dominance, and cytochrome oxidase-reactive regions in macaque monkey striate cortex. Cerebral Cortex. 1996; 6(2):297–310. https://doi.org/10.1093/cercor/6.2.297 PMID: 8670658

50. Schmidt KE, Goebel R, Löwel S, Singer W. The perceptual grouping criterion of colinearity is reflected by anisotropies of connections in the primary visual cortex. European Journal of Neuroscience. 1997; 9(5):1083–1089. https://doi.org/10.1111/j.1460-9568.1997.tb01459.x PMID: 9182961

51. Sincich LC, Blasdel GG. Oriented axon projections in primary visual cortex of the monkey. Journal of Neuroscience. 2001; 21(12):4416–4426. https://doi.org/10.1523/JNEUROSCI.21-12-04416.2001 PMID: 11404428

52. Sceniak MP, Ringach DL, Hawken MJ, Shapley R. Contrast's effect on spatial summation by macaque V1 neurons. Nature Neuroscience. 1999; 2(8):733. https://doi.org/10.1038/11197 PMID: 10412063

53. Walker GA, Ohzawa I, Freeman RD. Asymmetric suppression outside the classical receptive field of the visual cortex. Journal of Neuroscience. 1999; 19(23):10536–10553. https://doi.org/10.1523/JNEUROSCI.19-23-10536.1999 PMID: 10575050

54. Cavanaugh JR, Bair W, Movshon JA. Selectivity and spatial distribution of signals from the receptive field surround in macaque V1 neurons. Journal of Neurophysiology. 2002; 88(5):2547–2556. https://doi.org/10.1152/jn.00693.2001 PMID: 12424293

55. Jones H, Wang W, Sillito A. Spatial organization and magnitude of orientation contrast interactions in primate V1. Journal of Neurophysiology. 2002; 88(5):2796–2808. https://doi.org/10.1152/jn.00403.2001 PMID: 12424313

56. Chen CC, Kasamatsu T, Polat U, Norcia AM. Contrast response characteristics of long-range lateral interactions in cat striate cortex. Neuroreport. 2001; 12(4):655–661. https://doi.org/10.1097/00001756-200103260-00008 PMID: 11277558

57. Toth LJ, Rao SC, Kim DS, Somers D, Sur M. Subthreshold facilitation and suppression in primary visual cortex revealed by intrinsic signal imaging. Proceedings of the National Academy of Sciences. 1996; 93(18):9869–9874. https://doi.org/10.1073/pnas.93.18.9869

58. von Helmholtz H. Handbuch der physiologischen optik. 1860/1962. & Trans by JPC Southall Dover English Edition. 1962.

59. Doya K, Ishii S, Pouget A, Rao RP. Bayesian brain: Probabilistic approaches to neural coding. MIT Press; 2007.

60. Karklin Y, Lewicki MS. Emergence of complex cell properties by learning to generalize in natural scenes. Nature. 2009; 457(7225):83. https://doi.org/10.1038/nature07481 PMID: 19020501

61. Coen-Cagli R, Dayan P, Schwartz O. Cortical surround interactions and perceptual salience via natural scene statistics. PLoS Computational Biology. 2012; 8(3):e1002405. https://doi.org/10.1371/journal.pcbi.1002405 PMID: 22396635

62. Coen-Cagli R, Kohn A, Schwartz O. Flexible gating of contextual influences in natural vision. Nature Neuroscience. 2015; 18(11):1648. https://doi.org/10.1038/nn.4128 PMID: 26436902

63. Zetzsche C, Nuding U. Nonlinear and higher-order approaches to the encoding of natural scenes. Network: Computation in Neural Systems. 2005; 16(2-3):191–221. https://doi.org/10.1080/09548980500463982

64. Garrigues P, Olshausen BA. Learning horizontal connections in a sparse coding model of natural images. In: Advances in Neural Information Processing Systems; 2008. p. 505–512.

65. Kaschube M. Neural maps versus salt-and-pepper organization in visual cortex. Current Opinion in Neurobiology. 2014; 24:95–102. https://doi.org/10.1016/j.conb.2013.08.017 PMID: 24492085

66. Geisler WS, Perry JS, Super B, Gallogly D. Edge co-occurrence in natural images predicts contour grouping performance. Vision Research. 2001; 41(6):711–724. https://doi.org/10.1016/s0042-6989(00)00277-7 PMID: 11248261

67. Iyer R, Mihalas S. Cortical circuits implement optimal context integration. bioRxiv. 2017; p. 158360.

68. McGuire BA, Gilbert CD, Rivlin PK, Wiesel TN. Targets of horizontal connections in macaque primary visual cortex. Journal of Comparative Neurology. 1991; 305(3):370–392. https://doi.org/10.1002/cne.903050303 PMID: 1709953

69. Gilbert CD, Wiesel TN. Clustered intrinsic connections in cat visual cortex. Journal of Neuroscience. 1983; 3(5):1116–1133. https://doi.org/10.1523/JNEUROSCI.03-05-01116.1983 PMID: 6188819

70. Hübener M, Schwarz C, Bolz J. Morphological types of projection neurons in layer 5 of cat visual cortex. Journal of Comparative Neurology. 1990; 301(4):655–674. https://doi.org/10.1002/cne.903010412 PMID: 2177064

71. Hirsch JA, Gilbert CD. Synaptic physiology of horizontal connections in the cat's visual cortex. Journal of Neuroscience. 1991; 11(6):1800–1809. https://doi.org/10.1523/JNEUROSCI.11-06-01800.1991 PMID: 1675266

72. Callaway EM. Local circuits in primary visual cortex of the macaque monkey. Annual Review of Neuroscience. 1998; 21(1):47–74. https://doi.org/10.1146/annurev.neuro.21.1.47 PMID: 9530491

73. Rockland KS, Lund JS. Intrinsic laminar lattice connections in primate visual cortex. Journal of Comparative Neurology. 1983; 216(3):303–318. https://doi.org/10.1002/cne.902160307 PMID: 6306066

74. Rockland KS, Pandya DN. Laminar origins and terminations of cortical connections of the occipital lobe in the rhesus monkey. Brain research. 1979; 179(1):3–20. https://doi.org/10.1016/0006-8993(79)90485-2 PMID: 116716

75. Polat U, Sagi D. Lateral interactions between spatial channels: suppression and facilitation revealed by lateral masking experiments. Vision Research. 1993; 33(7):993–999. https://doi.org/10.1016/0042-6989(93)90081-7 PMID: 8506641

76. Tanaka Y, Sagi D. Long-lasting, long-range detection facilitation. Vision Research. 1998; 38(17):2591–2599. https://doi.org/10.1016/s0042-6989(97)00465-3 PMID: 12116705

77. Ernst UA, Schiffer A, Persike M, Meinhardt G. Contextual interactions in grating plaid configurations are explained by natural image statistics and neural modeling. Frontiers in Systems Neuroscience. 2016; 10:78. https://doi.org/10.3389/fnsys.2016.00078 PMID: 27757076

78. Rao RP, Ballard DH. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nature Neuroscience. 1999; 2:79–87. https://doi.org/10.1038/4580 PMID: 10195184

79. Stemmler M, Usher M, Niebur E. Lateral interactions in primary visual cortex: a model bridging physiology and psychophysics. Science. 1995; 269(5232):1877–1880.

80. Somers DC, Todorov EV, Siapas AG, Toth LJ, Kim DS, Sur M. A local circuit approach to understanding integration of long-range inputs in primary visual cortex. Cerebral cortex (New York, NY: 1991). 1998; 8(3):204–217.

81. Schwabe L, Obermayer K, Angelucci A, Bressloff PC. The role of feedback in shaping the extra-classical receptive field of cortical neurons: a recurrent network model. Journal of Neuroscience. 2006; 26(36):9117–9129. https://doi.org/10.1523/JNEUROSCI.1253-06.2006 PMID: 16957068

82. Dayan P, Abbott LF. Theoretical neuroscience. MIT Press, Cambridge, MA; 2001.

83. Foster K, Gaska JP, Nagler M, Pollen D. Spatial and temporal frequency selectivity of neurones in visual cortical areas V1 and V2 of the macaque monkey. The Journal of Physiology. 1985; 365(1):331–363. https://doi.org/10.1113/jphysiol.1985.sp015776 PMID: 4032318