LETTER **Communicated by Aapo Hyvarinen**

Bridging the Functional and Wiring Properties of V1 Neurons Through Sparse Coding

Xiaolin Hu*
xlhu@tsinghua.edu.cn
Department of Computer Science and Technology, State Key Laboratory of Intelligent Technology and Systems, BNRist, Tsinghua Laboratory of Brain and Intelligence, and IDG/McGovern Institute for Brain Research, Tsinghua University, Beijing 100084, China

Zhigang Zeng
zgzeng@hust.edu.cn
School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China, and Key Laboratory of Image Processing and Intelligent Control, Education Ministry of China, Wuhan 430074, China

The functional properties of neurons in the primary visual cortex (V1) are thought to be closely related to the structural properties of this network, but the specific relationships remain unclear. Previous theoretical studies have suggested that sparse coding, an energy-efficient coding method, might underlie the orientation selectivity of V1 neurons. We thus aimed to delineate how the neurons are wired to produce this feature. We constructed a model and endowed it with a simple Hebbian learning rule to encode images of natural scenes. The excitatory neurons fired sparsely in response to images and developed strong orientation selectivity. After learning, the connectivity between excitatory neuron pairs, inhibitory neuron pairs, and excitatory-inhibitory neuron pairs depended on firing pattern and receptive field similarity between the neurons. The receptive fields (RFs) of excitatory neurons and inhibitory neurons were well predicted by the RFs of presynaptic excitatory neurons and inhibitory neurons, respectively. The excitatory neurons formed a small-world network, in which certain local connection patterns were significantly overrepresented. Bidirectionally manipulating the firing rates of inhibitory neurons caused linear transformations of the firing rates of excitatory neurons, and vice versa. These wiring properties and modulatory effects were congruent with a wide variety of data measured in V1, suggesting that the sparse coding principle might underlie both the functional and wiring properties of V1 neurons.

*Corresponding author.

Neural Computation **34, 104–137** (2022)
https://doi.org/10.1162/neco_a_01453

1 Introduction

Revealing the functional properties of visual neurons and their wiring pattern is key to understanding the working mechanisms of the visual system. One of the greatest discoveries about the functions of neurons in the mammalian primary visual cortex (V1) comes from Hubel and Wiesel's experiments in which a large portion of them were found to be orientation or direction selective (Hubel, 1959; Hubel & Wiesel, 1962). This has been attributed to the sparse activity of neurons in response to visual stimuli (Olshausen & Field, 1996, 1997).

Over the past 20 years, technical advances have enabled researchers to probe the topology of the networks formed by V1 neurons. We now know that the wiring pattern of layer 2/3 neurons in the rodent V1 is highly nonrandom (Alonso & Martinez, 1998; Bock et al., 2011; Cossell et al., 2015; Hofer et al., 2011; Ko et al., 2011; Yoshimura, Dantzker, & Callaway, 2005). For instance, the connection probability between two pyramidal (PYR) excitatory neurons depends on their preferred orientation difference (Hofer et al., 2011; Ko et al., 2011), and the connection strength between two PYR neurons correlates with their response similarity and receptive field (RF) similarity (Cossell et al., 2015). However, it remains unclear how these wiring properties emerge and how they are related to the functions of neurons

To address these unanswered questions, it would be helpful to have a simple, biologically plausible, and sensitive learning model that is based on few assumptions and is capable of unifying previous data while predicting new connectivity patterns. As a preliminary requirement, such a model should incorporate the emergence of orientation selectivity of V1 neurons. Sparse coding models (Olshausen & Field, 1996, 1997) and related independent component analysis models (Bell & Sejnowski, 1997) are able to replicate this functional property. They often have a two-layer structure where the first layer contains visible units corresponding to image pixels and the second layer contains latent units corresponding to V1 neurons. Some of these models either do not consider the dependence between latent units (Bell & Sejnowski, 1997; Lee, Battle, Raina, & Ng, 2006; Olshausen & Field, 1996, 1997) or do not model the dependence by explicit connections (Garrigues & Olshausen, 2010; Hyvärinen, Hoyer, & Inki, 2001), which makes it impossible to compare the models with cortical circuits in terms of structure. A nonfactorial sparse coding model (Garrigues & Olshausen, 2008) considers lateral connections between neurons, but the model lacks biological plausibility. A more biologically plausible model assuming lateral connections between neurons refers to the locally competitive network (LCN) (Rozell, Johnson, Baraniuk, & Olshausen, 2008). A combination of Hebb rule and anti-Hebb rule can be used by the LCN to learn the oriented bar-like RFs of V1 neurons (Brito & Gerstner, 2016). A spiking model equipped with local plasticity rules, named SAILnet (Zylberberg, Murphy, & DeWeese, 2011)

was shown to be able to not only replicate the bar-like RFs of V1 neurons but also log-normal-like distributions of inhibitory connection weight between excitatory neurons. Similar distributions of synaptic efficacy have been observed in the brain (Buzsaki & Mizuseki, 2014; Song, Sjostrom, Reigl, Nelson, & Chklovskii, 2005).

The aforementioned models do not differentiate between excitatory neurons and inhibitory neurons. Some excitatory-inhibitory models (Brunel, 2016; Carlson, Richert, Dutt, & Krichmar, 2013; Miner & Triesch, 2016; Montangie, Miehl, & Gjorgjieva, 2020) have been shown to be able to replicate either the orientation selectivity of V1 neurons or wiring features of cortical neuron—for example, log-normal-like distribution of the connection strength (Song et al., 2005) and network motifs (Perin, Berger, & Markram, 2011; Song et al., 2005) among PYR neurons, but not both. We are aware of only one excitatory-inhibitory model (King, Zylberberg, & DeWeese, 2013) that related the orientation selectivity of V1 neurons to their connectivity by taking structured visual input. But in that study, only the connections between excitatory-inhibitory pairs were analyzed with respect to the RFs of neurons. In addition, the model assumes no lateral connections between excitatory neurons and therefore cannot be used directly to study the wiring features among excitatory neurons reported in animal studies (Cossell et al., 2015; Perin et al., 2011). In fact, many computational studies, including most of those already noted (Brunel, 2016; Carlson et al., 2013; King et al., 2013; Miner & Triesch, 2016), do not assume existence and plasticity of all types of connections: excitatory-to-excitatory (E-to-E), excitatory-to-inhibitory (E-to-I), inhibitory-to-excitatory (I-to-E), and inhibitory-to-inhibitory (I-to-I) connections. It is yet to be known if the wiring features of V1 neurons could emerge in a fully learnable model.

We extended the LCN (Rozell et al., 2008) into an excitatory–inhibitory network (see Figure 1A) and adopted the Hebb rule (Brito & Gerstner, 2016) to learn all types of connections given natural images as stimuli. The model replicated numerous wiring properties of V1 neurons discovered in recent years and made many interesting predictions that can now be tested experimentally.

2 Results

2.1 Model Structure and Learning. We aimed to construct a simple model based on very few reasonable biological assumptions. First, the ratio of excitatory neurons to inhibitory neurons is approximately 4:1 (Markram et al., 2004; Pfeffer, Xue, He, Huang, & Scanziani, 2013; Sillito, 1975). Second, the firing rates of most inhibitory neurons are higher than those of excitatory neurons (Atallah, Bruns, Carandini, & Scanziani, 2012; Hofer et al., 2011; Kerlin, Andermann, Berezovskii, & Reid, 2010; Tateno, Harsch, & Robinson, 2004). Third, the sum of the incoming connection strengths to any excitatory neuron is approximately the same. Fourth, the average

connection strengths from excitatory neurons to inhibitory neurons and from inhibitory neurons to excitatory neurons are similar (Holmgren, Harkany, Svennenfors, & Zilberter, 2003; Pfeffer et al., 2013) Fifth, the connection probability, calculated as the number of detected connections divided by the number of potential connections assayed, between excitatory neurons was about 20% and the connection probability from excitatory neurons to inhibitory neurons was about 90% (Cossell et al., 2015; Hofer et al., 2011; Ko et al., 2011; Yoshimura et al., 2005).

Starting from these assumptions, we constructed a network consisting of 1000 excitatory neurons and 250 inhibitory neurons (see Figure 1A). The dynamic equations of the neurons are

$$\tau_{\mathrm{E}} \frac{d\boldsymbol{z}_{\mathrm{E}}}{dt} = -\boldsymbol{z}_{\mathrm{E}} + \boldsymbol{M}_{\mathrm{EE}}\boldsymbol{r}_{\mathrm{E}} - \boldsymbol{M}_{\mathrm{EI}}\boldsymbol{r}_{\mathrm{I}} + \boldsymbol{W}_{\mathrm{E}}\boldsymbol{x} + \boldsymbol{c}, \tag{2.1}$$

$$\tau_{\mathrm{I}} \frac{d\boldsymbol{z}_{\mathrm{I}}}{dt} = -\boldsymbol{z}_{\mathrm{I}} + \boldsymbol{M}_{\mathrm{IE}}\boldsymbol{r}_{\mathrm{E}} - \boldsymbol{M}_{\mathrm{II}}\boldsymbol{r}_{\mathrm{I}} + \boldsymbol{W}_{\mathrm{I}}\boldsymbol{x} + \boldsymbol{c}, \tag{2.2}$$

where $\boldsymbol{z}_{\mathrm{E}}$ and $\boldsymbol{z}_{\mathrm{I}}$ denote the membrane potentials of the excitatory neurons and inhibitory neurons, respectively; $\boldsymbol{r}_{\mathrm{E}}$ and $\boldsymbol{r}_{\mathrm{I}}$ denote the firing rates of excitatory neurons and inhibitory neurons, respectively; $\boldsymbol{x}$ denotes visual stimuli (small patches extracted from natural images in our experiment that contain both positive and negative values); and $\boldsymbol{c}$ denotes input from other brain areas. The firing rates $\boldsymbol{r}_{\mathrm{E}}$ and $\boldsymbol{r}_{\mathrm{I}}$ are determined by the activation functions $\boldsymbol{f}_{\mathrm{E}}(\boldsymbol{z}_{\mathrm{E}})$ and $\boldsymbol{f}_{\mathrm{I}}(\boldsymbol{z}_{\mathrm{I}})$, respectively (see Figure 1B). $\boldsymbol{M}_{\mathrm{PQ}}$ denotes the lateral connections from neuron set Q to neuron set P (red and blue arrows in Figure 1A); P and Q can take two values, E and I, denoting excitatory neurons and inhibitory neurons, respectively. $\boldsymbol{W}_{\mathrm{E}}$ and $\boldsymbol{W}_{\mathrm{I}}$ denote the feedforward connections from visual stimuli $\boldsymbol{x}$ to excitatory neurons and inhibitory neurons, respectively (gray arrows in Figure 1A). Therefore, each row of $\boldsymbol{W}_{\mathrm{E}}$ denotes the RF of an excitatory neuron, and each row of $\boldsymbol{W}_{\mathrm{I}}$ denotes the RF of an inhibitory neuron. All elements in $\boldsymbol{M}_{\mathrm{PQ}}$ are constrained to be nonnegative, whereas the elements in $\boldsymbol{W}_{\mathrm{E}}$ and $\boldsymbol{W}_{\mathrm{I}}$ do not have this constraint. It is possible to model the neural projection from the lateral geniculate nucleus to V1 and separate $\boldsymbol{W}_{\mathrm{E}}$ and $\boldsymbol{W}_{\mathrm{I}}$ into matrices with nonnegative elements, but that is not the focus of this study. The time constants τ_{E} and τ_{I} govern the evolving speeds of $\boldsymbol{z}_{\mathrm{E}}(t)$ and $\boldsymbol{z}_{\mathrm{I}}(t)$, respectively. Given the stimuli $\boldsymbol{x}$ and external input $\boldsymbol{c}$, the states of the neurons evolve over time according to equations 2.1 and 2.2 (see Figure 1C). Throughout this letter, the characters in bold denote vectors or matrices, and the letters in plain type denote scalars.

By distinguishing excitatory neurons and inhibitory neurons, the two equations are essentially in the form of LCN (Rozell et al., 2008). This form can trace back to the continuous Hopfield network (Hopfield, 1984). In biologically detailed neuronal models, the synaptic inputs $\boldsymbol{z}_{\mathrm{E}}$, $\boldsymbol{z}_{\mathrm{I}}$, and $\boldsymbol{c}$ usually denote potential, but here for convenience we assume that they have been

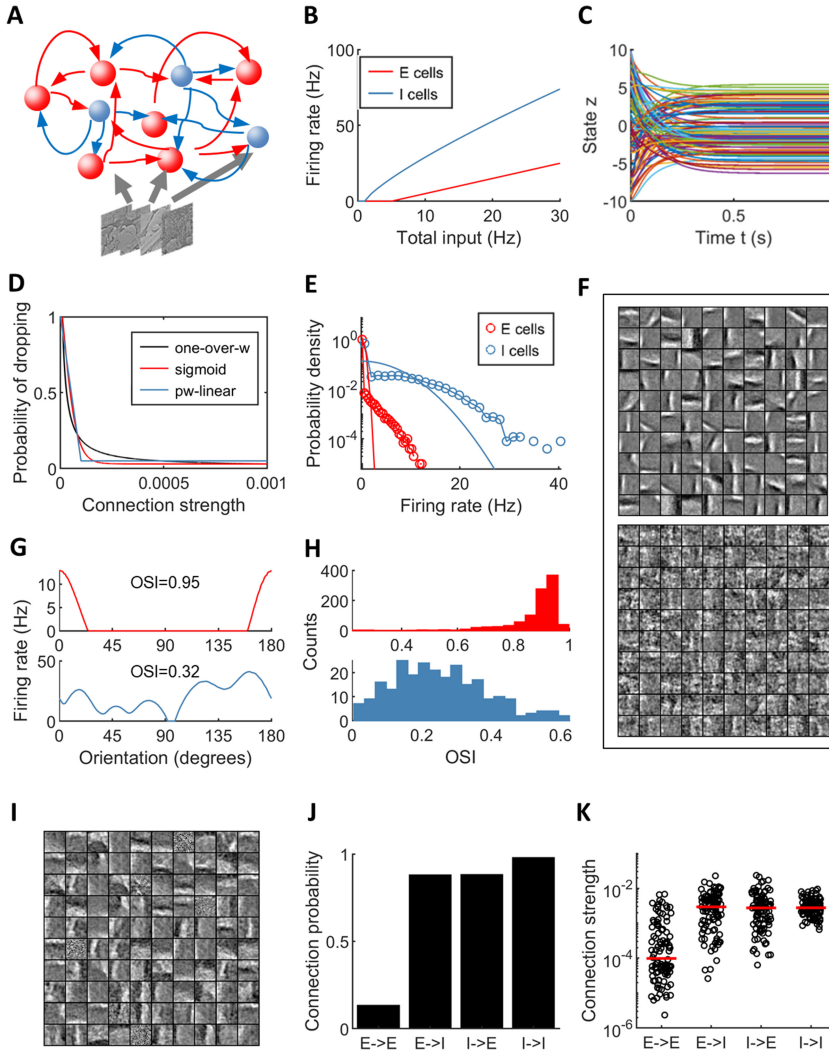


Figure 1: The model and the overall results. (A) The model with several representative excitatory (red) and inhibitory (blue) neurons. The red and blue arrows represent excitatory and inhibitory synapses between neurons, respectively. The bottom shows images input to neurons through feedforward synapses (gray arrows). (B) Activation functions of the excitatory neurons and inhibitory neurons. (C) States of 100 randomly selected sample neurons over time. (D) Probability of connection dropping (to model synaptic pruning) as a function of connection strength. Three functions are plotted: one-over-w, sigmoid and piecewise-linear. (E) Probability density of the firing rates across 100 randomly sampled image patches after learning. The two solid curves denote

multiplied by a constant to convert potential to firing rate and are therefore measured in units of Hz (Dayan & Abbott, 2001). This makes the synaptic weights dimensionless. Note that this view does not imply that z_E, z_I, and c could only be nonnegative.

The neurons start to fire when the inputs exceed certain thresholds (Desai, Rutherford, & Turrigiano, 1999; Fuortes & Mantegazzini, 1962; Tateno et al., 2004). Specifically, the activation functions for the excitatory and inhibitory neurons are defined as

$$f_E(z_E) = \max(0, z_E - \lambda_E), \tag{2.3}$$

$$f_I(z_I) = a \times \max(0, z_I - \lambda_I)^b, \tag{2.4}$$

where λ_E and λ_I denote the spiking thresholds, and a and b specify the shape of $f_I(z_I)$. Unless otherwise stated, all results reported in the letter were obtained with $\lambda_E = 5$ Hz $\lambda_I = 1$ Hz, $a = 5$ and $b = 0.8$. Equations 2.3 and 2.4 model the response properties of the PYR excitatory neurons and the fast-spiking interneurons in the visual cortex (presumably parvalbumin, or PV, neurons, a subtype of GABAergic neurons), respectively. We tried different parameters in equations 2.3 and 2.4 and obtained qualitatively similar results as reported in this letter, if only the inhibitory neurons had higher firing rates than the excitatory neurons.

All connection weights in equations 2.1 and 2.2 were updated according to the Hebb rule after z_E and z_I converged to steady state. Let θ_{pq} denote the connection weight from a unit q to a unit p. For lateral connections (red and blue arrows in Figure 1A), p and q can be excitatory or inhibitory neurons. After simulating equations 2.1 and 2.2, θ_{pq} was updated

$$\theta_{pq} \leftarrow \theta_{pq} + \Delta_{pq}, \quad \Delta_{pq} = \eta \langle r_p r_q \rangle, \tag{2.5}$$

two half-normal distributions fitted to the firing rates of excitatory neurons and inhibitory neurons. (F) Connection weights from stimuli to 100 randomly selected excitatory neurons (left) and 100 randomly selected inhibitory neurons (right), respectively. Every patch corresponds to one row of W_E or W_I, with dimension 20×20. (G) Orientation tuning curves for two sample neurons. Top: The first cell in the top of panel F. Bottom: The first cell in the bottom of panel F. (H) The distribution of the OSI of excitatory (top) and inhibitory (bottom) neurons. The median OSI was 0.95 for excitatory neurons and 0.32 for inhibitory neurons. (I) RFs of 100 randomly selected inhibitory neurons when the firing thresholds of excitatory and inhibitory neurons were increased from 5 Hz and 1 Hz to 7 Hz and 8 Hz, respectively. In every simulation trial, about 7% excitatory neurons and 6% inhibitory neurons were active. (J, K) Connection probability and strengths, respectively, between neurons. One hundred samples are shown for each type of connection in (K). Red lines indicate medians. Best viewed in color.

where r_p and r_q denote the firing rates of the two neurons and η denotes the learning rate. The feedforward connections (gray arrows in Figure 1A) were updated in the same way. The only difference is that q indexes an input pixel and r_q denotes its value. We implemented the leaning algorithm in the minibatch mode, and $\langle\cdot\rangle$ denotes the average over a minibatch of stimuli presented to the model. To prevent the weights from increasing without bound, after every update of the weights according to equation 2.5, a divisive normalization method was used (see section 4). Similar rules to equation 2.5 have been used to learn RFs of V1 neurons (Brito & Gerstner, 2016) in the LCN (Rozell et al., 2008).

In a biological system, connected neurons can become disconnected when the synapses are weak. To simulate the process, one could use a small threshold to prune all connections whose strengths are below it (Miner & Triesch, 2016). But that would only prune the E-to-E connections because the other three types of lateral connections should be much stronger according to physiological data (Hofer et al., 2011; Holmgren et al., 2003) and therefore would not be affected. A better strategy is to define a probability function depending on the connection strength w for randomly dropping connections. Three nonincreasing probability functions $p(w)$ were tested, called one-over-w, sigmoid and piecewise-linear (see Figure 1D) The parameters of the functions were set such that after convergence of the learning algorithm, the probability of connection between excitatory neurons was about 20% and the probability of connection from excitatory neurons to inhibitory neurons was about 90% according to experimental observations in layer 2/3 of rodents' V1 area (Cossell et al., 2015; Hofer et al., 2011; Ko et al., 2011; Yoshimura et al., 2005). We empirically found that to satisfy the above conditions, $p(w)$ should first decrease quickly to a low probability then stay there or decrease slowly. We did not find qualitatively different results by using different functions or different parameters of a particular function if only the above conditions were satisfied. All results presented in this letter were obtained with $p(w)$ being the one-over-w function unless otherwise specified. Without connection dropping, the results presented in this letter, except those depending on disconnections between neurons (e.g., connection probability and network topology), did not change significantly.

The model and the learning algorithm used in the study are quite conventional and it is very likely that other excitatory-inhibitory models (King et al., 2013; Miner & Triesch, 2016) equipped with similar learning algorithms could lead to similar results. The aim of this study is not to propose a novel model or learning algorithm, but to investigate the wiring properties of these kinds of models and further verify their validity as V1 models.

2.2 Overall Firing and Connection Patterns. The distributions of firing rates of the excitatory neurons and inhibitory neurons showed a higher

peak at zero and a heavier "tail" than the fitted half-normal distributions (see Figure 1E), indicating sparseness of the neural activities. This was due to the threshold activation functions of the neurons. Consistent with the sparse coding theory (Olshausen & Field, 1996, 1997), the RFs of the excitatory neurons resembled simple oriented bars, but the RFs of inhibitory neurons were much more complex (see Figure 1F). Using gratings with different orientations (0°–180°) as the input, we calculated the preferred orientations and the orientation selectivity indexes (OSIs) of all neurons (see Figures 1G and 1H): 84.7% of excitatory neurons had an OSI larger than 0.8, but 85.6% of inhibitory neurons had an OSI less than 0.4, consistent with physiological data (Atallah et al., 2012; Hofer et al., 2011; Kerlin et al., 2010). We attribute these differences to different levels of sparseness in the activity of the excitatory neurons and inhibitory neurons. Given a set of stimuli, the percentage of active excitatory neurons was approximately 5%, while the percentage of active inhibitory neurons was approximately 35%. When inhibitory neurons were made to fire more sparsely, which was achieved by increasing their firing thresholds, many of their RFs also resembled oriented bars (see Figure 1I). This is consistent with a previous computational study (King et al., 2013) in which both excitatory neurons and inhibitory neurons fired sparsely to image patches and the obtained RFs all resembled oriented bars.

The connection probabilities for E-to-E, E-to-I, I-to-E and I-to-I connections were 14.3%, 88.2%, 88.5% and 97.6%, respectively (see Figure 1J). The strength of E-to-E connections was significantly lower than those of the other three types of connections (median of E-to-E, 9.8×10^{-5}; median of E-to-I, 2.9×10^{-3}; median of I-to-E, 2.7×10^{-3}; median of I-to-I, 2.8×10^{-3}; $P = 1.6 \times 10^{-18}$, 3.0×10^{-19} and 1.8×10^{-23}, respectively, rank sum test with 100 random samples in each set; see Figure 1K). The conclusion about the comparison of connection strength between excitatory-excitatory pairs and excitatory-inhibitory pairs was consistent with the data obtained in V1 layer 2/3 of rats (Holmgren et al., 2003) and mice (Hofer et al., 2011). The strengths of the E-to-I, I-to-E and I-to-I connections were not significantly different ($P > 0.33$, rank sum test with 100 random samples in each set; see Figure 1K).

It has been reported that the connection strength between pyramid neurons in different regions in the brain follows lognormal distribution (Ikegaya et al., 2013; Lefort, Tomm, Sarria, & Petersen, 2009; Song et al., 2005). We found that the strengths of all four types of lateral connections in our model followed approximate lognormal distributions (see Figure 2A), although the distributions of the strengths of E-to-I and I-to-E connections were fitted with exponential functions a little bit better. Since the connection dropping operation may influence the distributions, we also plot the results obtained with the other two probability functions in Figure 2. The same conclusions were obtained when the sigmoid function was used (see Figure 2B). When the piecewise linear function was used, the log-normal

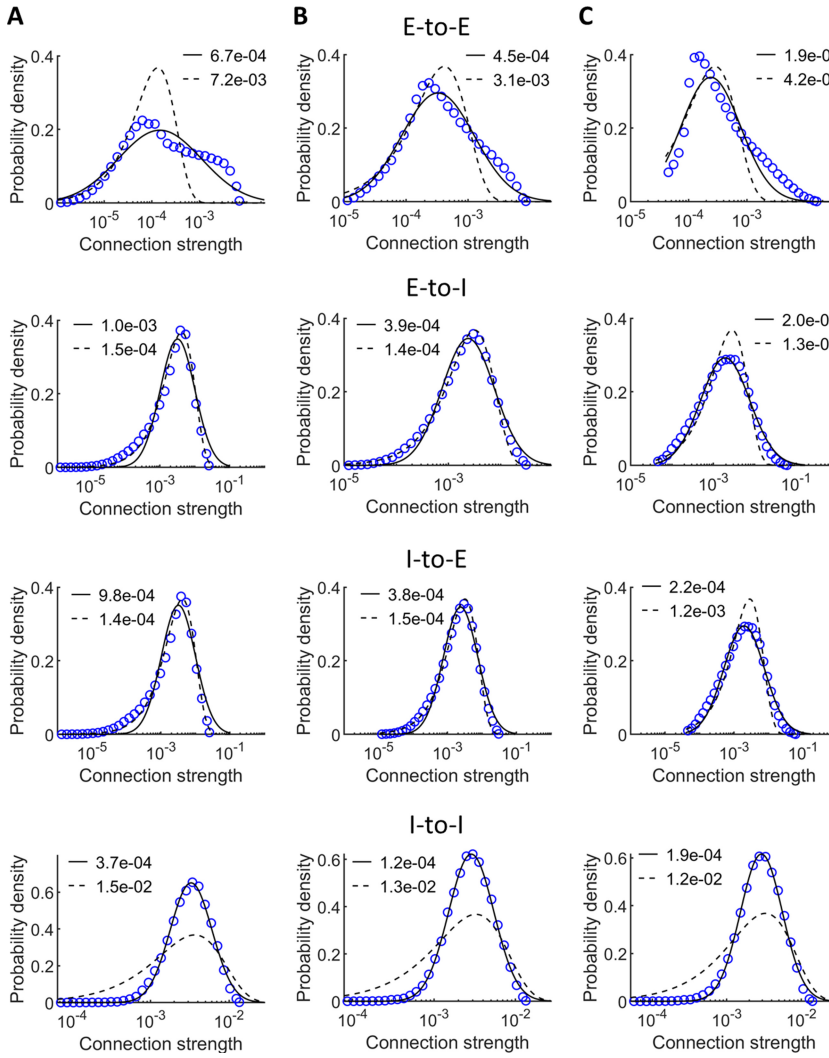


Figure 2: Distributions of the strengths of E-to-E, E-to-I, I-to-E and I-to-I connections with the probability function of dropping connections as one-over-w function (A), sigmoid function (B) and piecewise linear function (C), respectively. In each panel, the circles denote the probability distribution of the connection strength. The continuous curve and the dashed curve are the log-normal fitting and exponential fitting of the circles, respectively. The legend shows the mean squared errors of the two fitting results.

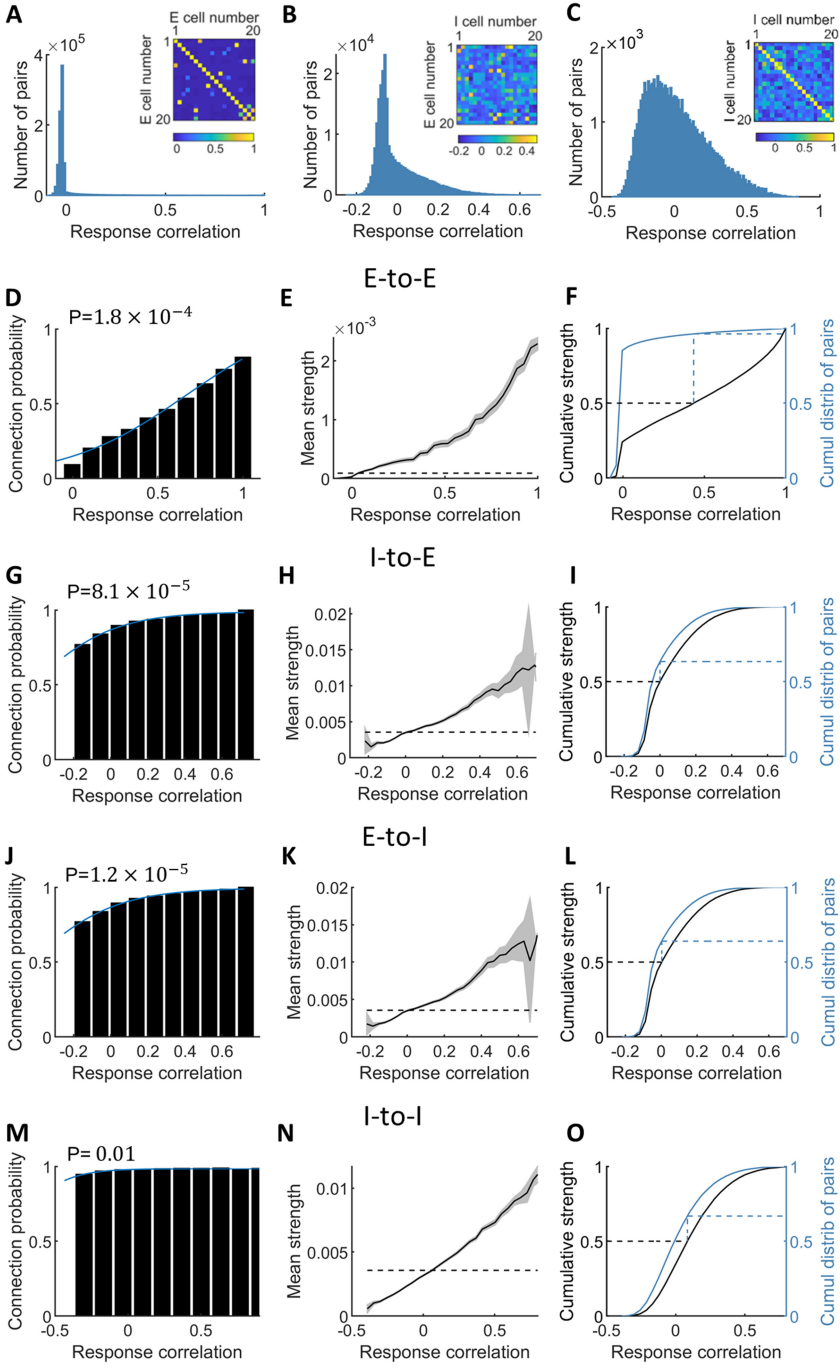
fitting was always better than the exponential fitting for all four types of connections (see Figure 2C).

2.3 Neurons with Similar Responses Form Strong Connections. Because the learning rule was based on the Hebb rule, we expected that the strengths of connections between neurons would correlate with their responses to natural images. We therefore calculated the correlation coefficients of responses of pairs of neurons to 100 randomly sampled image patches (see Figures 3A–3C). The distribution of the coefficients between excitatory neurons was quite sparse, and most coefficients were small (see Figure 3A). Similar results were also produced in a previous computational model (Zylberberg et al., 2011). Excitatory neurons were more likely to be connected if their responses were more similar (see Figure 3D). In addition, they tended to form strong connections if their response correlation coefficient was large but tended weak connections if the coefficient was small or negative (see Figure 3E). In fact, 3.7% of the most correlated E-to-E pairs accounted for 50% of the total strength of all E-to-E connections (see Figure 3F), highlighting the nonuniform distribution of the connection strengths between excitatory neurons. These results are consistent with experimental data obtained from the mouse V1 (Cossell et al., 2015).

The distributions of the correlation coefficients between excitatory-inhibitory pairs (see Figure 3B) and between inhibitory pairs (see Figure 3C) were peaked at points below zero, wider than that between excitatory pairs (see Figure 3A). The connection probability between excitatory-inhibitory pairs was higher if their responses were positively correlated and lower if their responses were negatively correlated (see Figures 3G and 3J). The I-to-I connection probability was always close to one and had weak dependence on the response correlation (see Figure 3M). Similar to the E-to-E connections, the I-to-E, E-to-I and I-to-I connections were stronger between neurons with more similar responses (Figures 3E, 3H, 3K, and 3N). The nonuniformity of the strength distributions was not as high as that of the E-to-E connection strength distribution. The 36.6% most correlated I-to-E pairs (see Figure 3I), 36.0% most correlated E-to-I pairs (see Figure 3L) and 33.2% of most correlated I-to-I pairs (see Figure 3O) accounted for 50% of the total strength of all I-to-E, E-to-I and I-to-I connections, respectively.

Taken together, these results indicated that the model neurons, regardless of type, were more strongly connected if their responses were more similar.

2.4 Neurons with Similar RFs Form Strong Connections. Besides response correlation, RF correlation, quantified using the pixel-to-pixel correlation coefficient between two RFs, can be also used to measure the functional similarity between neurons (see Figure 4). The distributions of RF correlation coefficients were more symmetric than those of response correlation coefficients (compare Figures 4A–4C with Figures 3A–3C). This



difference was observed between the two distributions calculated between PYR neurons in mouse V1 (Cossell et al., 2015). We investigated whether the connection probability and the strength of connections between neurons correlated with the similarity between the RFs of the neurons (see Figures 4D–4O). All the conclusions based on the response similarity were also obtained based on the RF similarity. In fact, 1.7% of strong E-to-E connections accounted for 50% of the total strength (see Figure 4F), while considerably larger portions of strong connections of other types (I-to-E, 24.4%; E-to-I, 24.3%; I-to-I, 29.2%) were required to account for 50% of the total strength (see Figures 4I, 4L, and 4O).

Consistent with the results obtained between PYR/PYR pairs in the mouse V1 (Cossell et al., 2015), we found that the connections between bidirectionally connected excitatory neuron pairs were stronger than the connections between unidirectionally connected excitatory pairs, and the RFs of bidirectionally connected neuron pairs were more similar than the RFs of unidirectionally connected pairs (see Figure 5).

We also investigated the relationships between the RFs of postsynaptic neurons and the RFs of presynaptic neurons using the method described in a published physiological study (Cossell et al., 2015). For each postsynaptic neuron, we calculated the weighted sum of the RFs of all presynaptic excitatory or inhibitory neurons using the corresponding connection strengths as the weighting coefficients and then compared this sum with the actual RF of the postsynaptic neuron. The RFs of the excitatory neurons were well predicted by their presynaptic excitatory neurons (see Figure 6A) but not

Figure 3: Connection strength reflects the similarity of neural responses. (A–C) The distributions of the response correlation coefficient between all excitatory-excitatory pairs, excitatory-inhibitory pairs, and inhibitory-inhibitory pairs, respectively. The inset shows the response correlation coefficients between 20 example neurons. (D) The connection probability between excitatory pairs plotted against the pairwise response correlation coefficient. The curve shows the logistic regression result $y = L/(1 + \exp(-\alpha x - \beta))$, where L, α, and β are fitting parameters. The p-value corresponds to the slope parameter α. (E) The mean connection strength between excitatory pairs plotted against the pairwise response correlation coefficient. Shaded region, 99% confidence level. Dashed line, mean connection strength between all pairs of excitatory neurons. (F) The cumulative distribution of connection strength between excitatory pairs with respect to response correlation (black curve), and the cumulative distribution of response correlation coefficients (blue curve). Dashed lines illustrate how the percentage of pairs accounting for 50% of the total connection strength is determined. (G–I) As for panels D to F except that the results of I-to-E connections are presented. (J–L) As for panels D to F except that the results of E-to-I connections are presented. (M–O) As for panels D to F except that the results of I-to-I connections are presented. Best viewed in color.

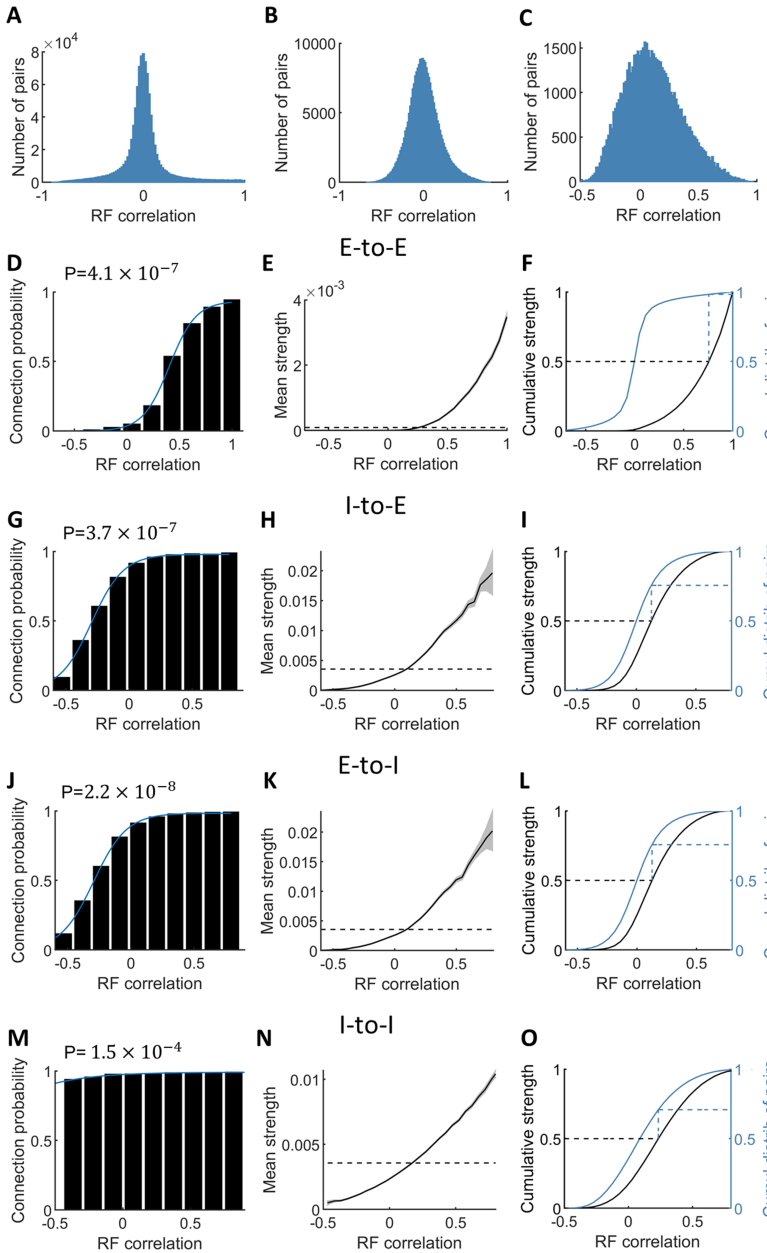


Figure 4: Connection strength reflects the similarity of RFs. All of the details are as described for Figure 3, except that the RF correlation coefficients rather than the response correlation coefficients are shown on the horizontal axes. Best viewed in color.

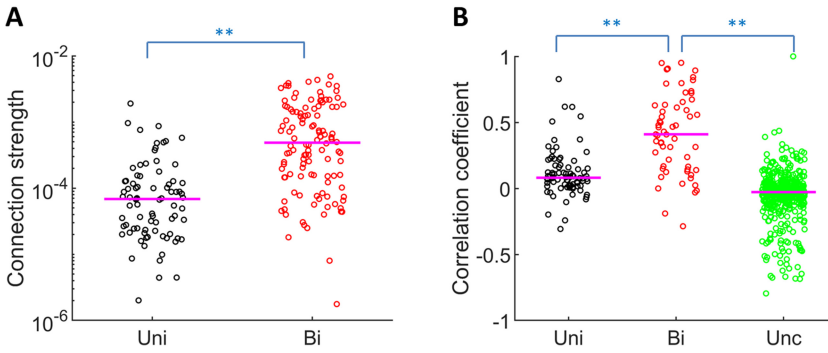


Figure 5: Comparison of connections between unidirectionally connected excitatory neuron pairs and between bidirectionally connected excitatory pairs. (A) Distribution of the strengths of connections between unidirectionally connected pairs (black) and bidirectionally connected pairs (red) in 600 randomly selected excitatory neuron pairs. (B) Distribution of the RF correlation coefficients between unidirectionally connected neurons (black), bidirectionally connected neurons (red), and unconnected neurons (green) in the same set of 600 excitatory neuron pairs. Magenta lines denote the medians of the populations. Uni: unidirectional. Bi: bidirectional. Unc: unconnected. $^{**}P < 10^{-4}$, ranksum test. Best viewed in color.

by the presynaptic inhibitory neurons (see Figure 6B). The RFs of the inhibitory neurons were predicted by their presynaptic excitatory neurons to some extent (see Figure 6C) and well predicted by their presynaptic inhibitory neurons (see Figure 6D).

For each neuron, we sorted the presynaptic neurons into descending order of connection strength and then divided them into four quarters. In the case of E-to-E connections, the first two quarters of presynaptic neurons accounted for 81.2% and 14.0% of the total connection strength on average, and the predicted RFs of the postsynaptic neurons based on these neurons were highly correlated with the actual RF (median correlations were 0.94 and 0.78, respectively; see Figure 6A). These findings are consistent with the results obtained between PYR/PYR pairs in the mouse V1 (Cossell et al., 2015). In the cases of E-to-I and I-to-I connections, only the first quarter of presynaptic neurons could predict the RFs of postsynaptic neurons to some extent (median correlations were 0.64 and 0.83, respectively; see Figures 6C and 6D). In the case of I-to-E connections, none of the quarters of presynaptic neurons could predict the RFs of postsynaptic neurons well (median correlations were smaller than 0.5; see Figure 6B). In all four cases, the median correlations decreased with increasing quarter number. Finally, as expected, the unconnected neurons in all four cases failed to predict the RFs of the postsynaptic neurons.

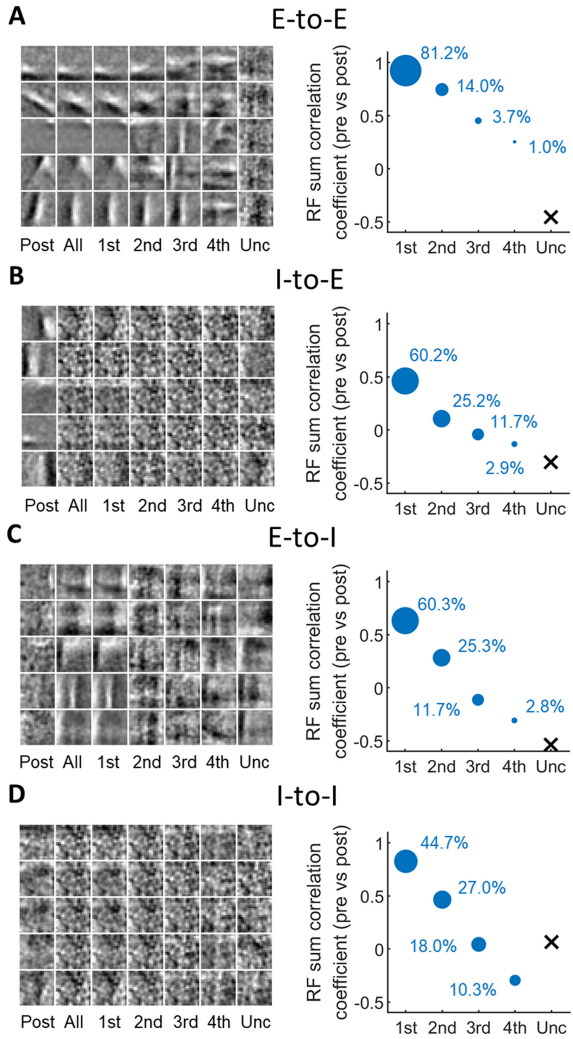


Figure 6: Prediction of neuron RFs using presynaptic neuron RFs. (A, B) Prediction of excitatory neurons' RFs using presynaptic excitatory neurons' RFs and presynaptic inhibitory neurons' RFs, respectively. (C, D) Prediction of inhibitory neurons' RFs using presynaptic excitatory neurons' RFs and presynaptic inhibitory neurons' RFs, respectively. The left portion of each panel shows five examples of the postsynaptic neurons' RFs (the first column) and the synthesized RFs by weighted sums of the RFs of all presynaptic neurons (the second column), the first to the fourth quarter of presynaptic neurons (the third to sixth columns), and unconnected neurons (the last column). The weighting coefficients for presynaptic neurons were the corresponding connection strengths. The weighting coefficients for unconnected neurons were all set to 1. Each

These analyses indicate that all neurons in the model with more similar RFs were more strongly connected, and the presynaptic neurons played an important role in shaping the RFs of the postsynaptic neurons.

2.5 Excitatory Neurons Form a Small-World Network. The previous analyses suggested that the learned network was not randomly wired. We therefore systematically evaluated the nonrandomness of the network. We investigated only the subnetwork consisting of the excitatory neurons, because we aimed to analyze the connectivity pattern, which was meaningful only in a population with sparse connections while the inhibitory neurons were densely connected to each other and to the excitatory neurons. We randomly selected 10,000 small subnets consisting of K excitatory neurons, with K equal to 3 to 8, and then calculated the number of connections between all neurons in each subnet. For comparison, we constructed 100 random networks whose unidirectional and bidirectional connection probabilities matched those of the learned network. A higher number of connections in a K-cell subnet tended to be found more frequently in the learned network than in random networks (see Figures 7A and 7B), implying that certain local connection patterns (or motifs) (Perin et al., 2011; Song et al., 2005) were significantly overrepresented. The neurons with more similar RFs tended to form overrepresented motifs (see Figure 7C). Moreover, pairs of neurons tended to share more common neighbors than expected, and with more common neighbors, neurons were more likely to be connected (see Figure 8). Similar motif patterns and common neighbor effect were observed among somatosensory neurons of rats (Perin et al., 2011).

Many types of networks can exhibit the local clustering effect described above. Next, we tested two most common hypotheses in describing nonrandom complex networks: the scale-free network (Barabasi & Albert, 1999) and small-world network (Watts & Strogatz, 1998) hypotheses. These are global characteristics of complex networks. The hallmark of the scale-free network is that its node degree follows a power-law distribution. In other words, the distribution of the node degree is approximately a straight line with a negative slope in the log-log plane; so is its cumulative distribution

20×20 image patch was normalized independently by dividing its maximum absolute value for display. The presynaptic and postsynaptic neurons are indicated on the top. The right portion of each panel shows the median correlation coefficients between the RFs of the postsynaptic neurons and the RFs synthesized by different groups of presynaptic neurons, as well as unconnected neurons. The size of the dot is proportional to the percentage of the sum of the connection strengths from the group of presynaptic neurons over the sum of the connection strengths from all presynaptic neurons and is indicated by the blue numbers. Best viewed in color.

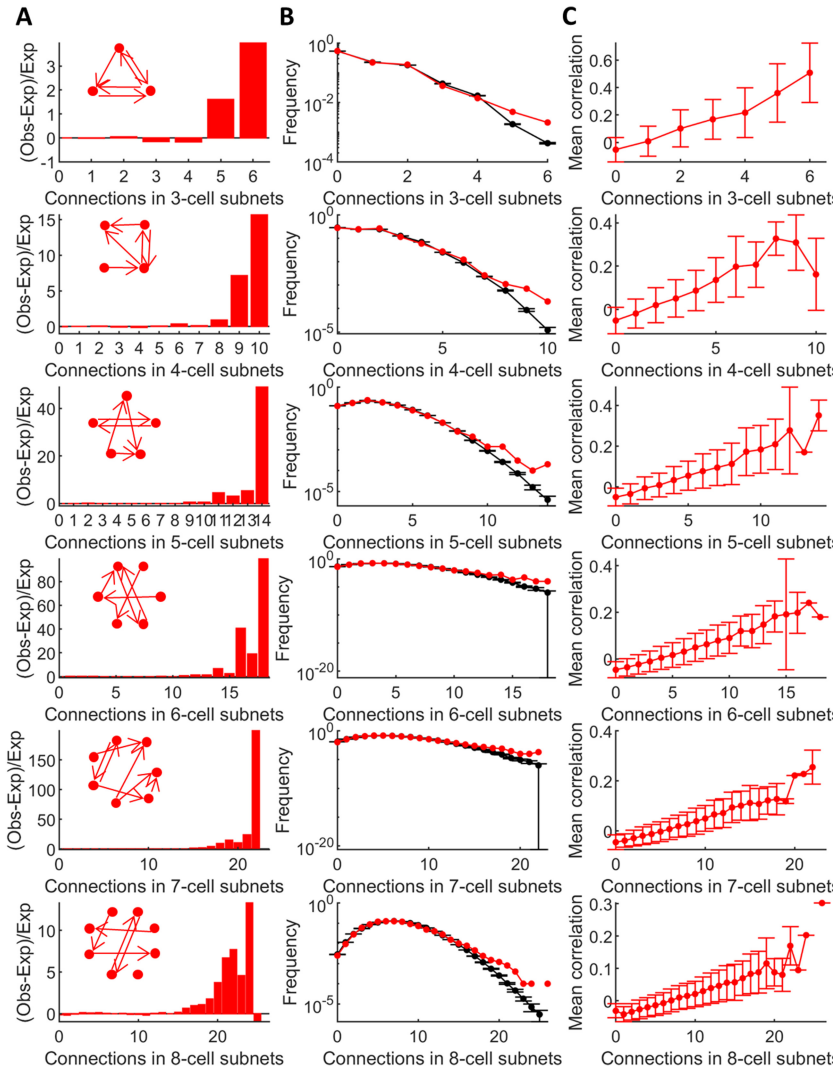


Figure 7: Motifs in the excitatory neural network. (A) The differences in the numbers of connections between observed and expected values divided by the expected value. The observed value was calculated in the learned network, and the expected value was the mean value calculated in 100 random networks. (B) The distributions of the number of connections in the learned network (red) and random networks (black) over 10,000 randomly selected K-cell subnets. Error bars are SEM. (C) Mean pairwise correlation of RFs of neurons in a subnet against the number of connections in the subnet. The results were averaged over 10,000 K-cell subnets. Error bars are standard deviations. Best viewed in color.

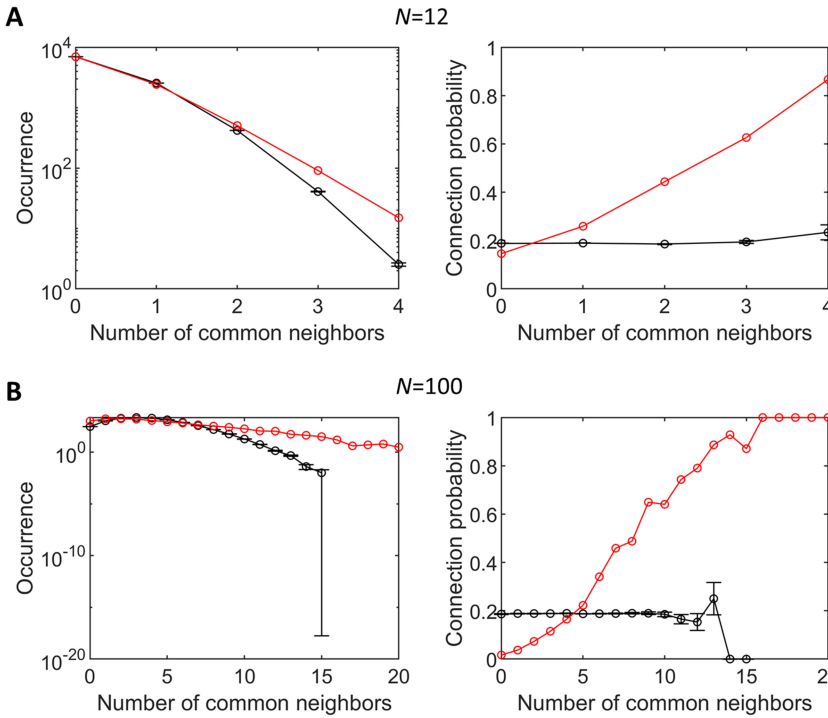


Figure 8: The common neighbor effect. Left: Distribution of the number of common neighbors for 10,000 pairs of neurons found in 10,000 N-cell subnets. Right: Connection probability of a pair as a function of the number of common neighbors. Red: Learned network. Black: Mean and SEM over 100 random networks with matched pairwise connection probability. (A) Results for $N = 12$. This represents a simulation of a physiological study (Perin et al., 2011) in which approximately 6 to 12 neurons in rat cortical slices were simultaneously recorded. (B) Results for $N = 100$. Best viewed in color.

(but with a different slope). We did not observe this pattern in the distribution of node in-degree, out-degree, or total degree (see Figure 9). To test the small-world hypothesis, we converted the excitatory neuronal network into an undirected and unweighted network: if there existed at least one connection between two neurons, they were said to be connected, and the connection weight was set to one. The average shortest path length between all pairs of excitatory neurons was 1.81 ± 0.39 (mean $\pm$ standard deviation), and the average clustering coefficient was 0.36 ± 0.7. For comparison, we constructed a random network with the same number of neurons and connections. This was achieved by randomly shuffling the connections in the undirected and unweighted excitatory neuronal network. The average

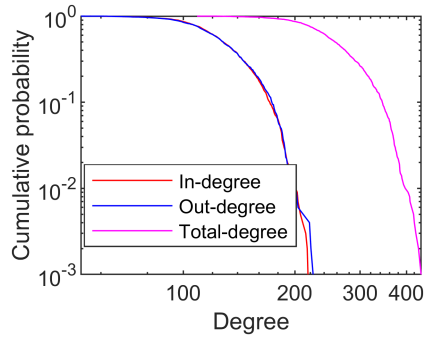


Figure 9: Cumulative probability of the node in-degree (red), out-degree (blue), and total degree (pink) of the learned excitatory neuronal network. Best viewed in color.

shortest path length in this random network was 1.81 ± 0.39, nearly the same as that of the learned network, but the average clustering coefficient was 0.18 ± 0.003, which is much smaller than that of the learned network, indicating the small-worldness of the learned network.

A key question is how this small-world network is wired. It is known that a small-world network can be obtained by randomly reconnecting some connections in a ring lattice (Watts & Strogatz, 1998). In many small-world networks, such as social networks and the World Wide Web, it is difficult to visualize this process because it is difficult to arrange the nodes into a ring lattice. In our network, the neurons can be ordered in a ring lattice according to their orientation preferences. We generated an example subnetwork consisting of 36 neurons whose preferred orientations increased from 0° to 180° (see Figure 10A). The subnetwork exhibited a clustering pattern on neighboring neurons as most of the output weights and input weights of each neuron were between neighboring neurons, though some distant connections were also present. We used a threshold to remove the half of the original weighted connections with the weakest weights and converted the connections to binary; we also observed a prominent clustering effect (see Figure 10B).

2.6 Modulatory Effects. One critical condition that makes the model stable is the balance between excitatory input and inhibitory input to each neuron. It is reasonable to infer that the tuning properties of excitatory neurons can be influenced by the activity level of inhibitory neurons, and vice versa. To investigate this effect, we first manipulated the firing rates of the inhibitory neurons by shifting their firing threshold λ_I (see Figure 1B) from 1 to -3, -1, 3, or 5, while keeping the firing threshold of the excitatory neurons unchanged. This operation emulates the optogenetic manipulation of

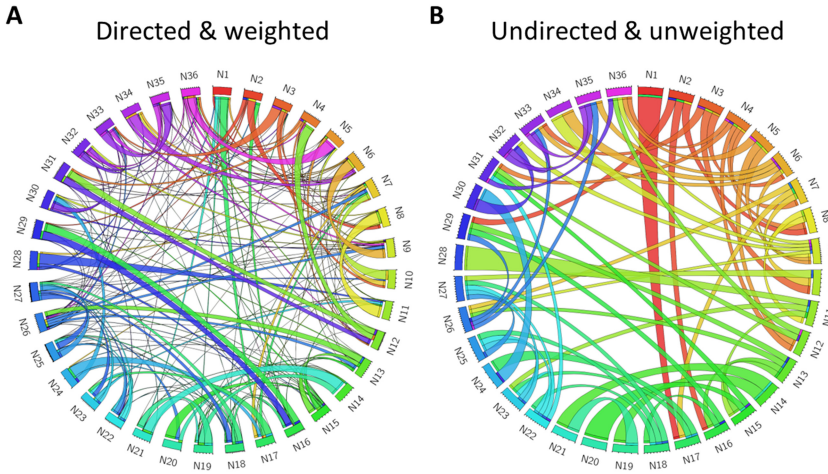


Figure 10: Visualization of the connection pattern of a 36-cell subnet. The neurons labeled N1 to N36 are presented according to their preferred orientations in 36 bins at 5° intervals across the range 0°–180°. (A) The original directed and weighted network. Every ribbon represents a directed connection from one neuron to another neuron. The thickness indicates the weight value. (B) The converted undirected and unweighted network. The half of connections with the weakest strengths were removed before binarization. Each ribbon represents a connection between two neurons. The same set of neurons was selected. Best viewed in color.

the activities of PV interneurons in the visual cortex (Atallah et al., 2012; Lee et al., 2012). We randomly selected 90 excitatory neurons whose tuning curves were well fitted with the gaussian curves for analysis. When the inhibitory neurons were activated ($\lambda_I < 1$), the average firing rate of the excitatory neurons decreased (see Figure 11A). Conversely, when the inhibitory neurons were suppressed ($\lambda_I > 1$), the average firing rate of the excitatory neurons increased (see Figure 11A). By aligning and normalizing the orientation tuning responses in all conditions, we found that a larger λ_I resulted in higher firing rates of the excitatory neurons (see Figure 11B, circles). Nevertheless, the preferred orientation and OSI of the excitatory neurons showed little change across modulated conditions (see Figures 11C and 11D). Across all modulated conditions, decreasing the firing rates of the excitatory neurons led to orientation sharpening as indicated by a decrease in the half-width at half height (HWHH) measure (see Figure 11E; Pearson's linear correlation coefficient: 0.87, $P < 10^{-50}$). Orientation sharpening was also observed in mouse V1 neurons when PV neurons were activated by injecting virus (Lee et al., 2012).

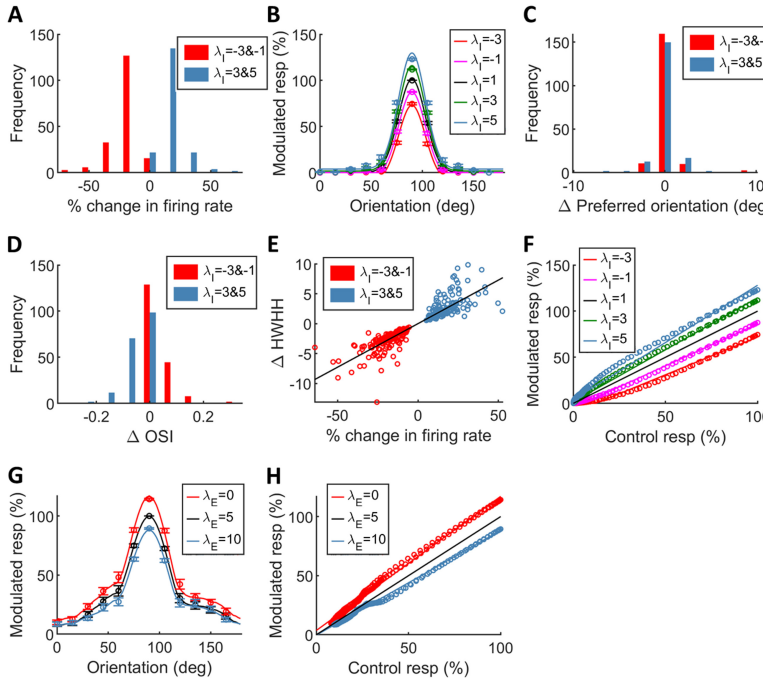


Figure 11: Modulatory effects on firing rates. (A–F) Impact of changing all inhibitory neurons' firing threshold λ_I on the firing rates of 90 selected excitatory neurons while λ_E was fixed at 5. (A) Histogram of percentage change in firing rate, defined as $(r - \tilde{r})/\tilde{r}$, where $\tilde{r}$ denotes the response of an excitatory neuron at its preferred orientation in the control condition and r denotes the response of the neuron at the same orientation in the modulated condition. (B) Aligned responses of the excitatory neurons against the orientation of gratings in different conditions. The responses of every neuron were shifted to peak at 90° and normalized such that the maximal response in the control condition was 100%. The circles denote the average responses of the neurons at 12 equidistant points, and error bars are SEM. The black curve represents the fitted gaussian function to the average responses in the control condition. The other curves were obtained by linearly transforming the black curve using the linear fitting parameters obtained in panel F while constraining the curves above zero. (C) Histogram of change in the preferred orientation. (D) Histogram of change in the OSI. (E) Change in the HWHH of the fitted gaussian function to the tuning curve against percentage change in firing rate. The black line is the linear regression of all circles. (F) Modulated responses of the excitatory neurons against control responses after firing rate normalization. The curves represent the threshold linear regression functions. (G, H) Impact of changing all excitatory neurons' firing threshold λ_E on the firing rates of the highly orientation-selective inhibitory neurons while λ_I was fixed at 2. Similar to panels B and F, respectively, but here the responses of the inhibitory neurons are shown. Best viewed in color.

We next asked what the relationship is between the firing rates of excitatory neurons in the control condition and the modulated conditions. To address this question, we plotted the average modulated response against the average control response of the excitatory neurons in all orientations and found a linear relationship between them (see Figure 11F). Because the firing rate cannot be negative, we fitted a threshold linear function $y = \max(ax + b, 0)$ in every case. The optimal parameters (a, b) were (0.82, -10.00), (0.91, -4.80), (1.13, 1.72), and (1.24, 3.86) when λ_I was $-3, -1, 3$ and 5, respectively. We then used the optimal parameters to predict the modulated orientation tuning curves based on the gaussian curve fitted in the control condition. The predictions well matched the measured data (see Figure 11B). Similar firing rate changes in PYR neurons in layer 2/3 of V1 were reported in a previous study that modulated the activities of PV neurons (Atallah et al., 2012; Atallah, Scanziani, & Carandini, 2014).

Finally, we manipulated the firing rates of excitatory neurons by shifting their firing threshold λ_E while keeping the firing threshold of inhibitory neurons unchanged. We analyzed the responses of highly orientationselective inhibitory neurons and found that when the excitatory neurons were activated ($\lambda_E = 0$), the firing rates of the inhibitory neurons increased; when the excitatory neurons were suppressed ($\lambda_E = 10$), the firing rates of the inhibitory neurons decreased (see Figure 11G). The firing rates of the inhibitory neurons in the modulated conditions can also be well predicted by a threshold linear function $y = \max(ax + b, 0)$ based on their firing rates in the control condition (see Figures 11G and 11H). The optimal parameters (a, b) were (1.13, 3.62) and (0.89, -0.35) in the $\lambda_E = 0$ and 10 conditions, respectively.

Taken together, these results suggested that based on our model, modulating the activities of one type of neurons (excitatory or inhibitory) had predictable effects on the orientation tuning curves of the other type of neurons.

3 Discussion

We present an excitatory–inhibitory neural model endowed with a simple Hebb learning rule starting from a limited number of biologically reasonable assumptions. It replicated a wide variety of published results on the wiring properties of V1 neurons. Previous computational studies (Olshausen & Field, 1996, 1997) have suggested the critical role of the sparse coding principle in developing oriented bar-like RFs of V1 neurons, while our study suggests that this principle may underlie a huge number of wiring properties of the local circuitry of V1.

Two previous computational studies (King et al., 2013; Zylberberg et al., 2011) have already related the RFs of V1 neurons to some wiring properties of these neurons. Zylberberg et al. (2011) constructed a sparse coding model and reported that the strength of connection between neurons was

correlated with the RF similarity between neurons and followed a log normal-like distribution. But the connections are inhibitory connections between excitatory neurons, which violates Dale's law. Inhibitory neurons were introduced to resolve this problem (King et al., 2013). It was reported that the strength of connection between an excitatory-inhibitory neuron pair was correlated with the RF similarity between the two neurons. The same prediction was made in our study. However, in that model, excitatory neurons were not directly connected; therefore, the model cannot be used to study the wiring scheme of excitatory neurons about which there were abundant experimental data in recent decades. We extended these two studies by constructing an excitatory-inhibitory model with all four types of lateral connections (E-to-E, E-to-I, I-to-E, and I-to-I) and systematically analyzed the functional properties (including activities and RFs) of neurons and their wiring properties (including connection probabilities and strengths) and the relations between these properties. In addition, we analyzed local and global characteristics of the subnetwork consisting of excitatory neurons. We also analyzed modulatory effect by changing the firing rates of excitatory and inhibitory neurons.

Not all results presented in this letter have been verified in animals, and those results can be regarded as predictions for the local circuits in layer 2/3 of the rodent V1 area. The first set of predictions concerns the connection pattern between PYR/PYR pairs in layer 2/3 of the V1 area. One prediction is the existence of overrepresented network motifs in the network consisting exclusively of PYR neurons, and such motifs consist of neurons with similar RFs (see Figure 7). Previous studies revealed overrepresented network motifs in different sensory areas (Perin et al., 2011; Song et al., 2005) but did not investigate the properties (e.g., the RFs) of neurons in these motifs. Computational models have been proposed to unravel the underlying mechanism for the emergence of network motifs (Brunel, 2016; Druckmann & Chklovskii, 2012; Miner & Triesch, 2016; Montangie et al., 2020). However, these models take noise as input or do not take sensory input at all; therefore, they cannot relate the motifs related to the RFs of neurons in V1 either. Another prediction, which is closely related to the previous one, is that the PYR neuronal network in layer 2/3 of the rodent V1 area is a small-world network. Functional and anatomical studies have identified many small-world networks in the brain, including the network of all neurons of *Caenorhabditis elegans* (Watts & Strogatz, 1998), the medial reticular formation of the vertebrate brain (Humphries, Gurney, & Prescott, 2005), and a subnetwork of layer 5 PYR neurons in rat somatosensory cortex (Perin et al., 2011). But whether the layer 2/3 PYR neuronal network in V1 has this property and how this topology is related to the functional properties of neurons have not been studied yet.

A widely accepted assumption about the formation of small-world architecture is that this architecture minimizes the wiring cost (Bassett & Bullmore, 2006). However, it is difficult to conceptualize creating a model

specifically designed to result in small-worldness. Our findings demonstrate that this is not necessary. We have shown that the small-worldness emerged from a model based on more basic principles. If layer 2/3 PYR neurons in actual brain tissue are verified to form a small-world network, our model will establish a close link between the functional efficiency and structural efficiency of the V1 circuit.

The second set of predictions concerns the properties of connections between excitatory-inhibitory pairs and inhibitory-inhibitory pairs, which have not been assessed in animals to date. First, both the I-to-E and E-to-I connection probabilities increase with the similarity between the neurons' RFs (see Figures 4G and 4J), but the connection probability between inhibitory pairs has only weak dependence on the similarity between the neurons' responses (see Figure 3M) or RFs (see Figure 4M). Second, the I-to-E and E-to-I connection strengths increase with the similarity between the neurons' RFs (see Figures 4H and 4K). This was also predicted in a previous computational study (King et al., 2013). Third, the I-to-I connection strength increases with the similarity between the neurons' responses (see Figure 3N) and RFs (see Figure 4N). Fourth, the RF of an inhibitory neuron can somehow be predicted by its presynaptic excitatory and inhibitory neurons that are strongly connected to it (see Figures 6C and 6D). If these predictions are verified, our model would provide a more complete picture of the local circuits in layer 2/3 of V1.

The third set of predictions concerns the modulatory effects. One prediction is that if both excitatory neurons and inhibitory neurons in V1 are suppressed such that the inhibitory neurons' firing is sufficiently sparse, then after a long period of exposure to natural scenes, the RFs of many inhibitory neurons will also be oriented bars (see Figure 1I). This is also predicted in another study (King et al., 2013). An extension of the prediction is that certain subtypes of inhibitory neurons with sparser activities tend to have higher OSI than other subtypes of inhibitory neurons. Another prediction is closely related to a previous study (Atallah et al., 2012), which demonstrated that increasing (or decreasing) the firing rates of the PV neurons linearly decreased (or increased) the firing rates of the excitatory neurons. Besides reproducing these results, we predicted that if the activities of the excitatory neurons can be modulated, increasing (or decreasing) their firing rates will linearly increase (or decrease) the firing rates of PV neurons (see Figures 11G and 11H). These predictions suggest that excitatory neurons and inhibitory neurons in the visual cortex play complementary roles in encoding the visual stimuli.

Our model is an extension of a previous model LCN (Rozell et al., 2008) and the learning algorithm is based on the conventional Hebb rule, similar to what is described in a previous study (Brito & Gerstner, 2016). The aim of our study is not to propose novel models or learning algorithms, but to unify a number of functional and wiring properties of the V1 circuits in a single model, then make testable predictions.

We believe that some spiking models (Carlson et al., 2013; King et al., 2013; Miner & Triesch, 2016; Montangie et al., 2020) starting from the same set of assumptions as made in the this study could yield similar results as presented in this letter by taking natural images as input and setting all connections plastic and learnable. The reasons are as follows. First, by distinguishing excitatory and inhibitory neurons, spiking models can have sparse neural activities. This is mainly due to the balanced excitatory and inhibitory input to each neuron. Another factor contributing to the sparse activities of neurons in our model is the threshold firing property of neurons (see Figure 1B) because the neurons could not fire when the inputs did not exceed the thresholds. This is an inherent property of spiking models of neurons if the firing rate of a neuron is plotted against the input current (Dayan & Abbott, 2001). Second, the learning algorithms used in previous spiking models (Carlson et al., 2013; King et al., 2013; Miner & Triesch, 2016; Montangie et al., 2020) all follow the same spirit of the Hebb rule (neurons fire together, wire together) including Oja's rule, correlation measuring rule and spike-timing-dependent plasticity rule.

By using a firingrate model instead of a spiking model we aim to shed some light on the development of new artificial neural networks (ANN) considering that the mainstream ANNs are firing-rate models. The widely used multilayer perceptrons and convolutional neural networks in the artificial intelligence (AI) field originate from neuroscience, but they have little in common with the visual system of animals in terms of wiring patterns revealed in recent decades. A promising future direction is therefore to extend our proposed model to a new ANN by using deep learning techniques (LeCun, Bengio, & Hinton, 2015) and test its performance in AI tasks.

4 Materials and Methods

The source code can be found at http://xlhu.cn/codes/EInet.zip.

4.1 Stimuli. We downloaded 10 512×512 pixel grayscale images, used in a previous computational study (Olshausen & Field, 1996, 1997), that describe natural scenes (e.g., rocks, trees, and mountains).[1] The images were whitened such that the amplitudes of low-frequency and high-frequency components in the frequency domain were approximately the same (Olshausen & Field, 1997). The data set was augmented by rotating images 90 degrees. The stimuli consisted of 12,000 patches of 20×20 pixels extracted from the 20 images at random positions. The L2 norm of every patch was normalized to 800.

[1] The images can be found at http://xlhu.cn/codes/IMAGES.zip.

4.2 Model. We used the Matlab function ode45 to solve equations 2.1 and 2.2), which is based on an explicit Runge–Kutta (4,5) formula (Dormand & Prince, 1980). The external input $\boldsymbol{c}$ to every neuron for every presentation of stimuli was 2 Hz. Time constants were $\tau_E = 100$ ms and $\tau_I = 50$ ms. Different values did not affect learning results so long as the simulation time was long enough to allow the solution to converge to steady state. In our experiments, the simulation time was 1000 ms (see Figure 1C).

In accordance with physiological studies (Cossell et al., 2015; Hofer et al., 2011; Ko et al., 2011; Yoshimura et al., 2005), in our model network, the connection probability between a set of neurons was defined as the number of existing connections divided by the number of potential connections between the neurons. Since a pair of neurons can have reciprocal connections, the number of potential connections between N neurons is $N(N-1)$. If a neuron pair has reciprocal connections, they are said to be bidirectionally connected. If a neuron pair has one and only one connection, they are said to be unidirectionally connected. The bidirectional (or unidirectional) connection probability between a set of neurons was defined as the number of existing bidirectionally (or unidirectionally) connected neuron pairs divided by the maximum number of potential bidirectionally (or unidirectionally) connected pairs. The maximum numbers of potential bidirectionally and unidirectionally connected pairs between N neurons are both $N(N-1)/2$.

4.3 Responses of Neurons. The "response" or "firing rate" of an excitatory neuron and an inhibitory neuron to a stimulus (a natural image patch or an oriented grating) was defined as $\boldsymbol{r}_E(T)$ and $\boldsymbol{r}_I(T)$, respectively, where T denotes the end of the simulation time.

4.4 Learning Algorithm. To prevent the weights from increasing without bound, after every update of the weights (called "one iteration") according to equation 2.5, a simple normalization method was used. For lateral connections, the incoming connection strengths to each excitatory neuron (i.e., from all excitatory neurons and inhibitory neurons) was normalized to 1. Similar homeostatic plasticity mechanisms have been widely used in computational models to suppress runaway synaptic dynamics (Klos, Miner, & Triesch, 2018; Lazar, Pipa, & Triesch, 2009). Because $\boldsymbol{r}_I > \boldsymbol{r}_E$, the Hebb rule would make the elements in $\boldsymbol{M}_{EI}$ larger than those in $\boldsymbol{M}_{EE}$. Animal studies (Holmgren et al., 2003) indicated that the bidirectional connections between excitatory-inhibitory neuron pairs in layer 2/3 of rodent visual cortex had similar strengths. Therefore, we normalized $\boldsymbol{M}_{IE}$ by dividing a number such that its mean was equal to the mean of $\boldsymbol{M}_{EI}$. The mean of $\boldsymbol{M}_{II}$ was also set equal to the mean of $\boldsymbol{M}_{EI}$ (Pfeffer et al., 2013), but this is not essential, as we found that setting its mean to twice or half of the mean of $\boldsymbol{M}_{IE}$ produced qualitatively similar results to those reported in this letter so long as the threshold λ_I was adjusted such that the response sparseness of

excitatory and inhibitory neurons remained similar. These operations made the sum of the incoming connection strengths to each inhibitory neuron approximately the same during learning. Considering that $\boldsymbol{W}_{\mathrm{E}}$ and $\boldsymbol{W}_{\mathrm{I}}$ contained both positive and negative values, we normalized the L2-norm of each row to be constant such that these weights had magnitudes similar to the E-to-I and I-to-E lateral connection weights.

In our experiments, the minibatch size was 100. The learning rate η was 4×10^{-4} initially, 2×10^{-4} after 30 iterations, and 1×10^{-4} after 70 iterations. The learning stopped after all extracted image patches were input to the model once (120 iterations). The learning algorithm always converged before the last iteration.

We tried three probability functions for randomly dropping connections during learning. The first one is the *one-over-w function*,

$$p(w) = c_1 + \frac{1 - c_1}{b_1(w - a_1) + 1}, \tag{4.1}$$

where $a_1, b_1, c_1 > 0$. The second function is the *sigmoid function*,

$$p(w) = 2 - c_2 - \frac{2 - 2c_2}{1 + \exp(-b_2(w - a_2))}, \tag{4.2}$$

where $a_2, b_2, c_2 > 0$. In equations 4.1 and 4.2, the parameters $a_i (i = 1, 2)$ are thresholds at which $p(w) = 1$. If $w > a_i$, then $p(w) < 1$. If $w < a_i$, then $p(w) > 1$ indicating that all such connections should be dropped. The parameters $b_i (i = 1, 2)$ control the slope of the two monotonically decreasing functions. The parameters $c_i (i = 1, 2)$ denote the dropping probabilities when $w \to \infty$. The third function is the *piecewise-linear function*,

$$p(w) = \begin{cases} 1, & x < a_3 \\ 1 - \frac{1 - c_3}{b_3 - a_3}(w - a_3), & a_3 \le x \le b_3 \\ c_3, & x > b_3 \end{cases}, \tag{4.3}$$

where $a_3, b_3, c_3 > 0$. The parameters a_3 and b_3 are two thresholds for the strength w, between which the function decreases linearly. The parameter c_3 denotes the background dropping probability.

The parameters in equations 4.1 to 4.3 were set such that after convergence of the learning algorithm, the connection probability between excitatory neurons was about 20% and the connection probability from excitatory to inhibitory neurons was about 90%. If only these conditions were satisfied, different sets of parameters yielded qualitatively similar results. All results presented in this letter were obtained with $a_1 = 10^{-6}$, $b_1 = 3 \times 10^4$, and $c_1 = 0.01$ in equation 4.1; $a_2 = 10^{-5}$, $b_2 = 3 \times 10^4$, and $c_2 = 0.03$

in equation 4.2; and $a_3 = 10^{-6}$, $b_3 = 10^{-4}$, and $c_3 = 0.05$ in equation 4.3. We present only the results with the one-over-w function unless otherwise stated (see Figures 1D and 2).

We trained five different models from random initial points but did not observe significant differences in functional or structural properties. We present the results of a randomly selected model in the letter.

4.5 Orientation Tuning. To obtain the orientation tuning functions of the neurons, we exposed the model to gratings of size 20×20 pixels with different orientations and spatial phases. The gratings were parameterized as follows:

$$A \sin\left(2\pi f \frac{x}{20} \cos\left(\alpha - \frac{\pi}{2}\right) + 2\pi f \frac{y}{20} \sin\left(\alpha - \frac{\pi}{2}\right) + 2\pi\phi\right),$$

where A denotes amplitude, α denotes the orientation of the grating, (x, y) denotes the pixel location in the range of [1,20], f denotes the spatial frequency (cycles per image patch), and ϕ denotes the spatial phase. In our experiments, $A = 1$ and $f = 2$. Like natural image patches, every grating was normalized such that its L2 norm was 800. For each orientation α in the range $0 - \pi$, the grating shifted in the image patch by increasing ϕ from 0 to 1 in steps of 0.05. The "response" of a neuron to a grating with an orientation α was defined as the maximum response of the neuron over all ϕ.

The preferred orientation of a neuron was defined as the orientation of the grating that made the neuron fire most strongly. The OSI of a neuron has different definitions in the literature. In our study, the OSI of a neuron was defined as (Kerlin et al., 2010)

$$\left(\left(\sum r(\theta_i)\sin(2\theta_i)\right)^2 + \left(\sum r(\theta_i)\cos(2\theta_i)\right)^2\right)^{\frac{1}{2}} \Big/ \sum r(\theta_i),$$

where θ_i denotes the orientation of the grating and $r(\theta_i)$ denotes the firing rate of the neuron. Quantitatively similar results to Figures 1H and 11 were obtained when the following definition of OSI (Hofer et al., 2011) was used OSI $=(r_{\text{prefer}} - r_{\text{orth}})/(r_{\text{prefer}} + r_{\text{orth}})$, where r_{prefer} and r_{orth} denote the responses of the neuron at the preferred orientation and the orientation orthogonal to the preferred orientation, respectively. The major difference is that with the latter definition, neurons usually had larger OSIs.

The tuning curves of excitatory neurons to the orientation of gratings were fitted with gaussian functions. The tuning sharpness was measured as $(2\ln(2))^{0.5}\sigma$, that is, the half-width at halfheight (HWHH) (Atallah et al., 2012), where σ denotes the standard deviation of the fitted gaussian function.

4.6 Nonlinear Fitting of the Distribution of Connection Strength. We first calculated the probability distributions of the strengths of the four types of lateral connections in the log space (circles in Figure 2). They were calculated as the number of strengths belonging to bins with equal size Δ in the log space, divided by the sum of bin areas, $\sum_i h_i \Delta$, where h_i denotes the height of the ith bin. Let $f(x)$ denote a probability distribution function (PDF) in the linear space. We considered two typical PDFs. The first one is the log-normal PDF,

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right),$$

with parameters μ and σ. The second one is the exponential PDF,

$$f(x) = \frac{1}{\mu} \exp\left(-\frac{x}{\mu}\right),$$

with parameter μ. To fit data in the log space, we defined a new variable, $y = \ln x$, and denoted its PDF by $g(y)$, which satisfies $g(y)\mathrm{d}y = f(x)\mathrm{d}x$. Then $g(y) = xf(x)$. We used Matlab function nlinfit to fit the function $g(y)$ to the data points represented by circles in each panel of Figure 2 from multiple initial points and selected the best fitted function.

4.7 Local Connection Patterns in the Excitatory Neuronal Network. We generated 100 random networks whose bidirectional and unidirectional connection probabilities matched those of the learned excitatory neuronal network. From each network, we randomly selected 10,000 K-cell subnets and counted the number of connections in every K-cell subnet. Because the connections are directed, the possible number of connections in a K-cell subnet ranges between zero and $K(K-1)$. The numbers obtained in the learned network are called the *observed values*, and the numbers obtained in the random networks are called the *expected values* (see Figure 7).

Two neurons were said to be neighbors if a connection existed between them, regardless of its direction. We randomly selected an N-cell subnet from a network and recorded the number of common neighbors of two random neurons in the subnet. This process was repeated 10,000 times, yielding 10,000 numbers, for every network (i.e., the learned network or a random network with matched unidirectional and bidirectional probabilities). The distribution of the numbers is plotted in Figure 8, left. The connection probability of two neurons as a function of the number of common neighbors is plotted in Figure 8, right.

4.8 Characterization of the Small-Worldness. Two quantities are usually used to characterize a small-world network: the average shortest path

length over all pairs of nodes and the average clustering coefficient over all nodes (Watts & Strogatz, 1998). The shortest path length between two nodes is the minimum number of connections traveling from one node to the other. The clustering coefficient of a node is defined as the number of connections that actually exist between all its neighbors, divided by the maximum number of connections that can exist between all its neighbors. We only calculated the two quantities in the undirected and unweighted networks formed by excitatory neurons.

4.9 Visualization of the Connection Pattern among Excitatory Neurons. The connections among K excitatory neurons (see Figure 10) were visualized using the free software Circos (Krzywinski et al., 2009), available at http://mkweb.bcgsc.ca/tableviewer/visualize/. A $K \times K$ connection matrix was created in which the element at location (i, j) was the connection weight from the ith neuron to the jth neuron. The neurons were sorted according to their preferred orientations in both columns and rows. In the case of undirected and unweighted networks, the nonzero elements in the matrix were set to 1; only the upper diagonal matrix was shown because Circos treats every network as a directed graph, and there is no need to show the connections in the lower diagonal matrix, which are symmetric to the connections in the upper diagonal matrix.

4.10 Manipulation of Neural Activities. We first manipulated the firing rates of inhibitory neurons by shifting their firing threshold λ_I from 1 to -3, -1, 3, or 5 while keeping the firing threshold of excitatory neurons unchanged. We randomly selected 100 excitatory neurons and removed those whose maximum firing rate across gratings with all orientations were <1.0 or whose orientation tuning curves were not well fitted with gaussian functions ($R^2 < 0.8$) under the control condition ($\lambda_I = 1$). This resulted in 90 excitatory neurons for comparison of orientation tuning properties before and after changing λ_I. The results in Figures 11A to 11F were obtained in this manner.

We next manipulated the firing rates of excitatory neurons by shifting their firing threshold λ_E from 5 to 0 or 10 while keeping the firing threshold of inhibitory neurons unchanged. Because the orientation tuning curves of most inhibitory neurons were not gaussian-like, it is inappropriate to use the gaussian function fitting quality as a criterion for selecting neurons. For every modulated condition, we selected all 250 inhibitory neurons and removed those for which the maximum firing rate across gratings with all orientations was <1.0 and OSI <0.4 in both the control ($\lambda_E = 5$) and modulated conditions ($\lambda_E = 0$ or 10). This yielded 28 and 35 inhibitory neurons for studying the modulatory effect with $\lambda_E = 0$ and $\lambda_E = 10$, respectively. The results in Figures 11G and 11H were obtained in this way.

Acknowledgments

We thank Sen Song, Jisong Guan, and the members of the IDG/McGovern Institute for Brain Research at Tsinghua University for their useful discussion and suggestions on this study. We thank Han Liu and Hang Chen for their technical support. This work was supported by the National Natural Science Foundation of China (62061136001, 61836014, and U19B2034) and the Tsinghua-Toyota Joint Research Fund.

References

Alonso, J. M., & Martinez, L. M. (1998). Functional connectivity between simple cells and complex cells in cat striate cortex. *Nature Neuroscience*, *1*(5), 395–403. 10.1038/1609, PubMed: 10196530

Atallah, B. V., Bruns, W., Carandini, M., & Scanziani, M. (2012). Parvalbumin-expressing interneurons linearly transform cortical responses to visual stimuli. *Neuron*, *73*(1), 159–170. 10.1016/j.neuron.2011.12.013

Atallah, B. V., Scanziani, M., & Carandini, M. (2014). Interneuron subtypes and orientation tuning. Reply. *Nature*, *508*(7494), E3–E3. 10.1038/nature13129, PubMed: 24695314

Barabasi, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*(5439), 509–512. 10.1126/science.286.5439.509, PubMed: 10521342

Bassett, D. S., & Bullmore, E. (2006). Small-world brain networks. *Neuroscientist*, *12*(6), 512–523. 10.1177/1073858406293182

Bell, A. J., & Sejnowski, T. J. (1997). The "independent components" of natural scenes are edge filters. *Vision Research*, *37*(23), 3327–3338. 10.1016/S0042-6989(97)00121-1, PubMed: 9425547

Bock, D. D., Lee, W. C. A., Kerlin, A. M., Andermann, M. L., Hood, G., Wetzel, A. W., . . . Reid, R. C. (2011). Network anatomy and in vivo physiology of visual cortical neurons. *Nature*, *471*(7337), 177–U159. 10.1038/nature09802, PubMed: 21390124

Brito, C. S. N., & Gerstner, W. (2016). Nonlinear Hebbian learning as a unifying principle in receptive field formation. *PLOS Computational Biology*, *12*(9). 10.1371/journal.pcbi.1005070

Brunel, N. (2016). Is cortical connectivity optimized for storing information? *Nature Neuroscience*, *19*(5), 749–755. 10.1038/nn.4286, PubMed: 27065365

Buzsaki, G., & Mizuseki, K. (2014). The log-dynamic brain: How skewed distributions affect network operations. *Nature Reviews Neuroscience*, *15*(4), 264–278. 10.1038/nrn3687, PubMed: 24569488

Carlson, K. D., Richert, M., Dutt, N., & Krichmar, J. L. (2013). *Biologically plausible models of homeostasis and STDP: Stability and learning in spiking neural networks.* Paper presented at the International Joint Conference on Neural Networks, Dallas, TX.

Cossell, L., Iacaruso, M. F., Muir, D. R., Houlton, R., Sader, E. N., Ko, H., . . . Mrsic-Flogel, T. D. (2015). Functional organization of excitatory synaptic strength in primary visual cortex. *Nature*, *518*(7539), 399–403. 10.1038/nature14182, PubMed: 25652823

Dayan, P., & Abbott, L. F. (2001). *Theoretical neuroscience*. Cambridge, MA: MIT Press.
Desai, N. S., Rutherford, L. C., & Turrigiano, G. G. (1999). BDNF regulates the intrinsic excitability of cortical neurons. *Learning and Memory*, *6*(3), 284–291. 10492010
Dormand, J. R., & Prince, P. J. (1980). A family of embedded Runge-Kutta formulae. *Journal of Computational and Applied Mathematics*, *6*(1), 19–26. 10.1016/0771-050X(80)90013-3
Druckmann, S., & Chklovskii, D. B. (2012). Neuronal circuits underlying persistent representations despite time varying activity. *Current Biology*, *22*(22), 2095–2103. 10.1016/j.cub.2012.08.058
Fuortes, M. G., & Mantegazzini, F. (1962). Interpretation of the repetitive firing of nerve cells. *Journal of General Physiology*, *45*(6), 1163–1179. 10.1085/jgp.45.6.1163
Garrigues, P., & Olshausen, B. A. (2008). Learning horizontal connections in a sparse coding model of natural images. In J. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.), *Advances in neural information processing systems*, *20*. Red Hook, NY: Curran.
Garrigues, P., & Olshausen, B. A. (2010). Group sparse coding with a laplacian scale mixture prior. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, & A. Culotta (Eds.), *Advances in neural information processing systems*, *23*. Red Hook, NY: Curran.
Hofer, S. B., Ko, H., Pichler, B., Vogelstein, J., Ros, H., Zeng, H. K., . . . Mrsic-Flogel, T. D. (2011). Differential connectivity and response dynamics of excitatory and inhibitory neurons in visual cortex. *Nature Neuroscience*, *14*(8), 1045–1052. 10.1038/nn.2876, PubMed: 21765421
Holmgren, C., Harkany, T., Svennenfors, B., & Zilberter, Y. (2003). Pyramidal cell communication within local networks in layer 2/3 of rat neocortex. *Journal of Physiology*, *551*(1), 139–153. 10.1113/jphysiol.2003.044784
Hopfield, J. J. (1984). Neurons with graded response have collective computational properties like those of 2-state neurons. In *Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences*, *81*(10), 3088–3092. 10.1073/pnas.81.10.3088
Hubel, D. H. (1959). Single unit activity in striate cortex of unrestrained cats. *Journal of Physiology–London*, *147*(2), 226–238. 10.1113/jphysiol.1959.sp006238
Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in cats visual cortex. *Journal of Physiology–London*, *160*(1), 106–154. 10.1113/jphysiol.1962.sp006837
Humphries, M. D., Gurney, K., & Prescott, T. J. (2005). The brainstem reticular formation is a small-world, not scale-free, network. In *Proceedings of the Royal Society B: Biological Sciences*, *273*(1585), 503–511. 10.1098/rspb.2005.3354
Hyvärinen, A., Hoyer, P. O., & Inki, M. (2001). Topographic independent component analysis. *Neural Computation*, *13*(7), 1527–1558.
Ikegaya, Y., Sasaki, T., Ishikawa, D., Honma, N., Tao, K., Takahashi, N., . . . Matsuki, N. (2013). Interpyramid spike transmission stabilizes the sparseness of recurrent network activity. *Cerebral Cortex*, *23*(2), 293–304. 10.1093/cercor/bhs006, PubMed: 22314044
Kerlin, A. M., Andermann, M. L., Berezovskii, V. K., & Reid, R. C. (2010). Broadly tuned response properties of diverse inhibitory neuron subtypes in mouse visual cortex. *Neuron*, *67*(5), 858–871. 10.1016/j.neuron.2010.08.002, PubMed: 20826316

King, P. D., Zylberberg, J., & DeWeese, M. R. (2013). Inhibitory interneurons decorrelate excitatory cells to drive sparse code formation in a spiking model of V1. *Journal of Neuroscience*, *33*(13), 5475–5485. 10.1523/JNEUROSCI.4188-12.2013, PubMed: 23536063

Klos, C., Miner, D., & Triesch, J. (2018). Bridging structure and function: A model of sequence learning and prediction in primary visual cortex. *PLOS Computational Biology*, *14*(6). 10.1371/journal.pcbi.1006187, PubMed: 29870532

Ko, H., Hofer, S. B., Pichler, B., Buchanan, K. A., Sjostrom, P. J., & Mrsic-Flogel, T. D. (2011). Functional specificity of local synaptic connections in neocortical networks. *Nature*, *473*(7345), 87–91. 10.1038/nature09880, PubMed: 21478872

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., . . . Marra, M. A. (2009). Circos: An information aesthetic for comparative genomics. *Genome Research*, *19*(9), 1639–1645. 10.1101/gr.092759.109, PubMed: 19541911

Lazar, A., Pipa, G., & Triesch, J. (2009). SORN: A self-organizing recurrent neural network. *Frontiers in Computational Neuroscience, 3*. 10.3389/neuro.10.023.2009

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. 10.1038/nature14539, PubMed: 26017442

Lee, H., Battle, A., Raina, R., & Ng, A. Y. (2006). Efficient sparse coding algorithms. In B. Schölkopf, J. Platt, & T. Hoffman (Eds.), *Advances in neural information processing systems*, *19*. Cambridge, MA: MIT Press.

Lee, S. H., Kwan, A. C., Zhang, S., Phoumthipphavong, V., Flannery, J. G., Masmanidis, S. C., . . . Dan, Y. (2012). Activation of specific interneurons improves V1 feature selectivity and visual perception. *Nature*, *488*(7411), 379–383. 10.1038/nature11312, PubMed: 22878719

Lefort, S., Tomm, C., Sarria, J.-C. F., & Petersen, C. C. (2009). The excitatory neuronal network of the C2 barrel column in mouse primary somatosensory cortex. *Neuron*, *61*(2), 301–316. 10.1016/j.neuron.2008.12.020, PubMed: 19186171

Markram, H., Toledo-Rodriguez, M., Wang, Y., Gupta, A., Silberberg, G., & Wu, C. (2004). Interneurons of the neocortical inhibitory system. *Nature Reviews Neuroscience*, *5*(10), 793–807. 10.1038/nrn1519, PubMed: 15378039

Miner, D., & Triesch, J. (2016). Plasticity-driven self-organization under topological constraints accounts for non-random features of cortical synaptic wiring. *PLOS Computional Biology*, *12*(2), e1004759. 10.1371/journal.pcbi.1004759

Montangie, L., Miehl, C., & Gjorgjieva, J. (2020). Autonomous emergence of connectivity assemblies via spike triplet interactions. *PLOS Computational Biology*, *16*(5). 10.1371/journal.pcbi.1007835, PubMed: 32384081

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*(6583), 607. 10.1038/381607a0, PubMed: 8637596

Olshausen, B. A., & Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, *37*(23), 3311–3325. 10.1016/S0042-6989(97)00169-7, PubMed: 9425546

Perin, R., Berger, T. K., & Markram, H. (2011). A synaptic organizing principle for cortical neuronal groups. In *Proceedings of the National Academy of Sciences of the United States of America*, *108*(13), 5419–5424. 10.1073/pnas.1016051108

Pfeffer, C. K., Xue, M., He, M., Huang, Z. J., & Scanziani, M. (2013). Inhibition of inhibition in visual cortex: The logic of connections between molecularly distinct interneurons. *Nature Neuroscience*, *16*(8), 1068. 10.1038/nn.3446, PubMed: 23817549

Rozell, C. J., Johnson, D. H., Baraniuk, R. G., & Olshausen, B. A. (2008). Sparse coding via thresholding and local competition in neural circuits. *Neural Computation*, *20*(10), 2526–2563. 10.1162/neco.2008.03-07-486, PubMed: 18439138

Sillito, A. (1975). The contribution of inhibitory mechanisms to the receptive field properties of neurones in the striate cortex of the cat. *Journal of Physiology*, *250*(2), 305–329. 10.1113/jphysiol.1975.sp011056

Song, S., Sjostrom, P. J., Reigl, M., Nelson, S., & Chklovskii, D. B. (2005). Highly nonrandom features of synaptic connectivity in local cortical circuits. *PLOS Biology*, *3*(3), 507–519. 10.1371/journal.pbio.0030068

Tateno, T., Harsch, A., & Robinson, H. P. (2004). Threshold firing frequency-current relationships of neurons in rat somatosensory cortex: Type 1 and type 2 dynamics. *Journal of Neurophysiology*, *92*(4), 2283–2294. 10.1152/jn.00109.2004, PubMed: 15381746

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of "small-world" networks. *Nature*, *393*(6684), 440–442. 10.1038/30918, PubMed: 9623998

Yoshimura, Y., Dantzker, J. L., & Callaway, E. M. (2005). Excitatory cortical neurons form fine-scale functional networks. *Nature*, *433*(7028), 868–873. 10.1038/nature03252, PubMed: 15729343

Zylberberg, J., Murphy, J. T., & DeWeese, M. R. (2011). 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*, *7*(10). 10.1371/journal.pcbi.1002250, PubMed: 22046123

Received April 23, 2021; accepted July 20, 2021.