Computational and Systems Biology

Structure, dynamics, coding and optimal biophysical parameters of efficient excitatory-inhibitory spiking networks

Reviewed Preprint
v1 • July 26, 2024
Not revised

Veronika Koren ✉, **Simone Blanco Malerba, Tilo Schwalger, Stefano Panzeri**

Institute of Neural Information Processing, Center for Molecular Neurobiology (ZMNH), University Medical Center Hamburg-Eppendorf (UKE), 20251 Hamburg, Germany • Institute of Mathematics, Technische Universität Berlin, 10623 Berlin, Germany • Bernstein Center for Computational Neuroscience Berlin, 10115 Berlin, Germany

Abstract

The principle of efficient coding posits that sensory cortical networks are designed to encode maximal sensory information with minimal metabolic cost. Despite the major influence of efficient coding in neuro-science, it has remained unclear whether fundamental empirical properties of neural network activity can be explained solely based on this normative principle. Here, we rigorously derive the structural, coding, biophysical and dynamical properties of excitatory-inhibitory recurrent networks of spiking neurons that emerge directly from imposing that the network minimizes an instantaneous loss function and a time-averaged performance measure enacting efficient coding. The optimal network has biologically-plausible biophysical features, including realistic integrate-and-fire spiking dynamics, spike-triggered adaptation, and a non-stimulus-specific excitatory external input regulating metabolic cost. The efficient network has excitatory-inhibitory recurrent connectivity between neurons with similar stimulus tuning implementing feature-specific competition, similar to that recently found in visual cortex. Networks with unstructured connectivity cannot reach comparable levels of coding efficiency. The optimal biophysical parameters include 4 to 1 ratio of excitatory vs inhibitory neurons and 3 to 1 ratio of mean inhibitory-to-inhibitory vs. excitatory-to-inhibitory connectivity that closely match those of cortical sensory networks. The efficient network has biologically-plausible spiking dynamics, with a tight instantaneous E-I balance that makes them capable to achieve efficient coding of external stimuli varying over multiple time scales. Together, these results explain how efficient coding may be implemented in cortical networks and suggests that key properties of biological neural networks may be accounted for by efficient coding.

eLife assessment

This study offers a **useful** treatment of how the population of excitatory and inhibitory neurons integrates principles of energy efficiency in their coding strategies. The analysis provides a comprehensive characterisation of the model, highlighting the structured connectivity between excitatory and inhibitory neurons. However, the manuscript provides an **incomplete** motivation for parameter choices. Furthermore, the work is insufficiently contextualized within the literature, and some of the findings appear overlapping and incremental given previous work.

https://doi.org/10.7554/eLife.99545.1.sa4

Introduction

Information about the sensory world is represented in the brain through the dynamics of neural population activity[1,2]. One prominent theory about the principles that may guide the design of neural computations for sensory function is efficient coding[3,4,5]. This theory posits that neural computations are optimized to maximize the information that neural systems encode about features of sensory stimuli while at the same time limiting the metabolic cost. Efficient coding has been highly influential, especially in visual neuroscience and computational vision[6,7,8,9], and has been developed to become a normative theory of how networks are organized and designed to optimally process natural sensory stimuli in visual [10,11], auditory[12] and olfactory sensory pathways[13].

The first normative neural network models[4,11] designed with efficient coding principles had at least two major levels of abstractions. First, information was assumed to be processed in a purely feedforward manner, whereas information processing in real neural circuits often involves recurrent or feedback computations. Second, neural dynamics was greatly simplified, ignoring the spiking nature of neural activity. Instead, in biological networks considerable amount of information are encoded or transmitted only through the millisecond-precise timing of spikes[14,15,16,17,18,19,20]. Also, these earlier works mostly considered encoding of static sensory stimuli, whereas the sensory environment changes continuously at multiple timescales and the dynamics of neural networks encodes these temporal variations of the environment [21,22,23,24].

Recent years have witnessed a considerable effort and success in laying down the mathematical tools and methodology to understand how to formulate efficient coding theories of neural networks with much more biological realism[25]. This work has established the incorporation of recurrent connectivity[26,27], of spiking neurons, and of time-varying stimulus inputs[28,29,30,31,32,33,34,35]. In these models, the efficient coding principle has been implemented by designing networks whose activity minimizes the encoding accuracy (the error between a desired representation and a linear readout of network's activity) subject to a constraint on the metabolic cost of processing (proportional to the total number of spikes fired by a population of neurons). This double objective is captured by a loss function that trades-off encoding accuracy and metabolic cost. The minimization of the loss function is performed through a greedy approach, by assuming that a neuron will emit a spike only if this will decrease the loss. This, in turn, yields a set of leaky integrate-and-fire (LIF) neural equations which govern the network dynamics[28], which can also include biologically plausible non-instantaneous synaptic delays[36,35,34]. These previous implementations, however, had neurons that did not respect Dale's law. In recent work[37], we further extended the biological plausibility of these models by

analytically deriving how to implement efficient coding in networks of spiking neurons that respect Dale's law. These networks take the form of generalized leaky integrate-and-fire (gLIF) models of excitatory (E) and inhibitory (I) neurons endowed with spike-triggered adaptation[38,39,40], which can provide highly accurate predictions of spike times in biological networks[41]. Efficient spiking models have thus the potential to provide a unifying theory of neural coding through spiking dynamics of E-I circuits [42,37] with elements that are fully biologically plausible and potentially interpretable as biophysical variables.

However, despite the major progress described above, as well the progress provided by other studies of efficient coding with spikes [31,43,44,33,29], we still lack a thorough characterization of which structural, coding, biophysical and dynamical properties of excitatory-inhibitory recurrent networks of spiking neurons are directly related to efficient coding principles. Previous studies only rarely made predictions that could be quantitatively compared against experimentally measurable properties of biological neural networks. As a consequence, we still do not know which, if any, fundamental properties of cortical networks emerge directly from imposing efficient coding.

To address the above questions, we analyze systematically our biologically plausible efficient coding model of E and I neurons that respect Dale's law[37] to make concrete predictions about experimentally measurable structural, coding and dynamical features of neural activity that arise from efficient coding. We systematically investigated how experimentally measurable emergent dynamical properties, such as firing rates, trial-to-trial spiking variability of single neurons, E-I balance[45] and noise correlations, relate to optimally-efficient coding. We further analyze how the organization of the connectivity arising by imposing efficient coding relates to the anatomical and effective connectivity recently reported in visual cortex, which suggests competition between excitatory neurons with similar stimulus tuning. We found that several key and robustly found empirical properties of cortical circuits match the predictions of our efficient coding network, lending support to the notion that efficient coding may be a design principle that has shaped the evolution of cortical circuits and that may be used to conceptually understand and interpret them.

Results

Assumptions and emergent properties of the efficient E-I network derived from first principles

We study the properties of a spiking neural network in which the dynamics and structure of the network are analytically derived starting from first principles of efficient coding of sensory stimuli. The model relies on a number of assumptions, described next.

The network responds to M time-varying features of a sensory stimulus, $s_k(t)$ (e.g., for a visual stimulus, contrast, orientation, etc) received as inputs from an earlier sensory area (e.g., retina). We model features as independent Ornstein–Uhlenbeck (OU) processes (see Methods). The network's objective is to compute a leaky integration of sensory features, a relevant computation of cortical sensory areas[46]. The target representations of the network, $x_k(t)$, are defined as

$$\frac{dx_k(t)}{dt} = -\frac{1}{\tau}x_k(t) + s_k(t). \quad (1)$$

with τ a characteristic integration time-scale (**Fig. 1A**).

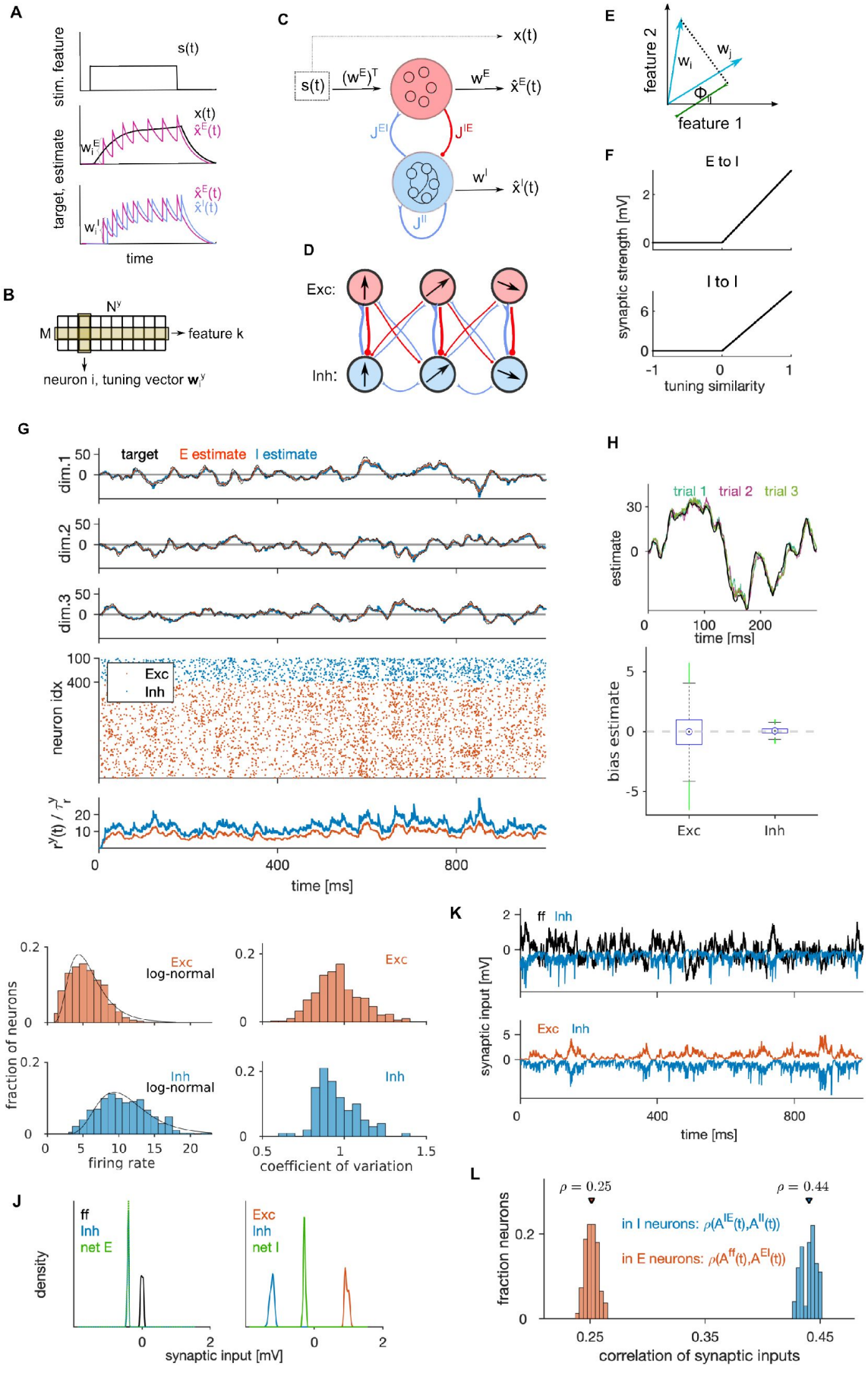

A
stim. feature
s(t)
x(t)
target, estimate
time
B
feature k
neuron i, tuning vector
C
x(t)
s(t)
D
Exc:
Inh:
E
feature 2
feature 1
F
E to I
I to I
synaptic strength [mV]
tuning similarity
G
target
E estimate
I estimate
dim.1
dim.2
dim.3
neuron idx
Exc
Inh
time [ms]
H
trial 1
trial 2
trial 3
estimate
time [ms]
bias estimate
I
Exc
log-normal
Inh
log-normal
fraction of neurons
firing rate
coefficient of variation
J
ff
Inh
net E
Exc
Inh
net I
density
synaptic input [mV]
K
ff
Inh
Exc
Inh
synaptic input [mV]
time [ms]
L
fraction neurons
correlation of synaptic inputs
in I neurons:
in E neurons:

Figure 1.

Structural and dynamical properties of the efficient E-I spiking network.

(A) Encoding of a target signal representing the evolution of a stimulus feature (top) with one E (middle) and one I spiking neuron (bottom). The target signal $x(t)$ integrates the input signal $s(t)$. The readout of the E neuron tracks the target signal and the readout of the I neuron tracks the readout of the E neuron. Neurons spike to bring the readout of their activity closer to their respective target. Each spike causes a jump of the readout, with the sign and the amplitude of the jump being determined by neuron's tuning parameters.

(B) Schematic of the matrix of tuning parameters. Every neuron is selective to all stimulus features (columns of the matrix), and all neurons participate in encoding of every feature (rows).

(C) Schematic of the network with E (red) and I (blue) cell type. E neurons are driven by the stimulus features while I neurons are driven by the activity of E neurons. E and I neurons are connected through recurrent connectivity matrices.

(D) Schematic of E (red) and I (blue) synaptic interactions. Arrows represent the direction of the tuning vector of each neuron. Only neurons with similar tuning are connected.

(E) Schematic of similarity of tuning vectors (tuning similarity) in a 2-dimensional space of stimulus features.

(F) Synaptic strength as a function of tuning similarity.

(G) Coding and dynamics in a simulation trial. Top three rows show the signal (black), the E estimate (red) and the I estimate (blue) in each of the three stimulus dimensions. Below are the spike trains. In the bottom row, we show the average instantaneous firing rate (in Hz).

(H) Top: Example of the target signal (black) and the E estimate in 3 simulation trials (colors) in one signal dimension. Bottom: Distribution (across time) of the time-dependent bias of estimates in E and I cell type.

(I) Left: Distribution of time-averaged firing rates in E (top) and I neurons (bottom). Black traces are fits with log-normal distribution. Right: Distribution of coefficients of variation of interspike intervals for E and I neurons.

(J) Distribution (across neurons) of time-averaged synaptic inputs to E (left) and I neurons (right). In E neurons, the distribution of inhibitory and of net synaptic inputs overlap.

(K) Sum of synaptic inputs over time in a single E (top) and I neuron (bottom) in a simulation trial.

(L) Distribution (across neurons) of Pearson's correlation coefficients measuring the correlation of synaptic inputs in single E (red) and I (blue) neurons. For model parameters, see **Table 1**.

The network is composed of two neural populations of excitatory (E) and inhibitory (I) neurons, defined by their postsynaptic action which respects Dale's law. For each population, $y \in \{E, I\}$, we define a population readout of each feature, $\hat{x}_k^y$, as a filtered weighted sum of spiking activity of neurons in the population,

$$\frac{d\hat{x}_k^y(t)}{dt} = -\frac{1}{\tau}\hat{x}_k^y(t) + \sum_{i=1}^{N^y} w_{ki}^y f_i^y(t), \tag{2}$$

parameter	notation	value
number of E neurons	N^E	400
ratio of E to I neuron numbers	$N^E : N^I$	4:1
number of the input features	M	3
time constant of the population readout (E and I)	τ	10 ms
time constant of the single neuron readout	$\tau_r^E = \tau_r^I$	10 ms
noise intensity (non-specific current)	σ	5.0 $(\mathrm{mV})^{1/2}$
heterogeneity factor of tuning parameters in E	σ_w^E	1.0 $(\mathrm{mV})^{1/2}$
ratio of mean E-I to I-I synaptic connectivity	mean E-I : mean I-I	3:1
metabolic constant	β	14 mV
threshold constant	$c/2$	18 mV
distance between firing threshold and reset potential (E neurons)	$\lvert\vartheta^E - V_{\mathrm{rest}}^E\rvert$	19 mV
distance between firing threshold and reset potential (I neurons)	$\lvert\vartheta^I - V_{\mathrm{rest}}^I\rvert$	21 mV
connection probability (recurrent synapses)	$p^{IE} = p^{II} = p^{EI}$	0.5
mean E-I synaptic weight (EPSP to I at max)	$\langle J_{ij}^{IE} \rangle$	0.75 mV
mean I-E synaptic weight (IPSP to E at max)	$\langle J_{ij}^{EI} \rangle$	0.75 mV
mean I-I synaptic weight (IPSP at max)	$\langle J_{ij}^{II} \rangle$	2.25 mV

Table 1.

Table of default model parameters for the efficient E-I network

Parameters above the double horizontal line are the minimal set of parameters needed to simulate model equations (**Eqs. 30a**-**30h** in Methods). Parameters below the double horizontal line are biophysical parameters, derived from the same model equations and from model parameters listed above the horizontal line. Parameters N^E, M, τ and σ_w^E were chosen for their biological plausibility and computational simplicity. Parameters N^I, τ_r^E, τ_r^I, σ, ratio of mean E-I to I-I synaptic connectivity and β are parameters that maximize network efficiency (see the section "Criterion for determining model parameters" in Methods). The metabolic constant β and the noise intensity σ are interpreted as global network parameters and are for this reason assumed to be the same across the E and I population, e.g., $\beta^E = \beta^I = \beta$ and $\sigma^E = \sigma^I = \sigma$ (see **Eq. 3**). The connection probability of $p^{xy} = 0.5$ is the consequence of rectification of the connectivity (see **Eq. 25** in Methods).

where $f_i^y(t)$ is the spike train of neuron i of type y and w_{ki}^y are its decoding weights for features $k = 1, \ldots, M$. As a result of the optimization, the decoding weights of the neurons are equivalent to neuron's tuning parameters to the stimulus features (see Methods;[42]). We draw tuning parameters from a normal distribution with zero mean and SD σ_w^E for E neurons and σ_w^I for I neurons. We assume that every neuron encodes multiple ($M > 1$) stimulus features[47] and define the vector $\boldsymbol{w}_i^y = [w_{1i}^y, \ldots, w_{Mi}^y]^\top$ as the tuning vector of neuron i (**Fig. 1B**).

Unlike previous approaches[28,48], we hypothesize that E and I neurons have distinct normative objectives and define cell-type specific loss functions relative to the activity of the E and I neuron types. To implement at the same time, as requested by efficient coding, the constraints of faithful stimulus representation with limited computational resources[49], we define the loss functions of E and I population as a weighted sum of a time-dependent encoding error and time-dependent metabolic cost:

$$\mathrm{L}^y(t) = \epsilon^y(t) + \beta\kappa^y(t), \qquad y \in \{E, I\}. \tag{3}$$

We refer to β, the parameter controlling the relative importance of the metabolic cost over the encoding error, as the metabolic constant of the network. We hypothesize that population readouts of E neurons, $\hat{x}_k^E(t)$, track the target representations, $x_k(t)$, and the population readouts of I neurons, $\hat{x}_k^I(t)$, track the population readouts of E neurons $\hat{x}_k^E(t)$, by minimizing the squared error between these quantities[37] (see also[28,48] for related approaches). Furthermore, we hypothesize the time-resolved metabolic cost to be proportional to the estimate of a momentary firing rate of the neural population. We thus define the variables of loss functions in Eq. 3 as

$$\begin{aligned} \epsilon^E(t) &= \sum_{k=1}^{M} \left[x_k(t) - \hat{x}_k^E(t)\right]^2, \qquad \kappa^E(t) = \sum_{i=1}^{N^E} [r_i^E(t)]^2, \\ \epsilon^I(t) &= \sum_{k=1}^{M} \left[\hat{x}_k^E(t) - \hat{x}_k^I(t)\right]^2, \qquad \kappa^I(t) = \sum_{i=1}^{N^I} [r_i^I(t)]^2, \end{aligned} \tag{4}$$

where $r_i^y\ y \in \{E, I\}$ is the low-pass filtered spike train of neuron i (single neuron readout) with time constant τ_r^y. We then impose the following condition for spiking: a neuron shall emit a spike only if this decreases the loss function of its population in the immediate future. The condition for spiking also includes a noise term accounting for sources of stochasticity in spike generation[50], including the effect of unspecific inputs from the rest of the brain.

By assuming that each neuron emits a spike at time t only if this decreases the loss function of its population (Eq. 3), we derived the dynamics and network structure of a spiking network that instantiates efficient coding (**Fig. 1C**, see Methods). The derived dynamics of the subthreshold membrane potential $V_i^E(t)$ and $V_i^I(t)$ obey the equations of the generalized leaky integrate and fire (gLIF) neuron

$$\tau \dot{V}_i^y(t) = -\left(V_i^y(t) - V_{\mathrm{rest}}^y\right) + R_m\left(I_i^{\mathrm{syn},y}(t) - I_i^{\mathrm{ad},y}(t) + I_i^{\mathrm{ext},y}(t)\right), \qquad y \in \{E, I\}, \tag{5}$$

where $I_i^{\mathrm{syn},y}$, $I_i^{\mathrm{ad},y}$, and $I_i^{\mathrm{ext},y}$ are synaptic current, spike-triggered adaptation current and non-specific external current, respectively, R_m is the membrane resistance and V_{rest}^y is the resting potential. This dynamics is complemented with a fire-and-reset rule: when the membrane potential reaches the firing threshold ϑ^y, a spike is fired and $V_i^y(t)$ is set to the reset potential

$V^{\text{reset},y}$. The analytical solution in **Eq. (5)** holds for any number of neurons (with at least 1 neuron in each population) and predicts an optimal spike pattern to encode the presented external stimulus.

The synaptic currents in E neurons, $I_i^{\text{syn},E}$, consist of feedforward currents, obtained as stimulus features $s_k(t)$ weighted by the tuning weights of the neuron, and of recurrent inhibitory currents. Synaptic currents in I neurons, $I_i^{\text{syn},I}$, consist of recurrent excitatory and inhibitory currents (**Fig. 1C**).

The optimization of the loss function also yields structured recurrent connectivity (**Fig. 1D**). The synaptic strength between two neurons is proportional to their tuning similarity if the tuning similarity is positive; otherwise the synaptic weight is set to zero (**Fig. 1E,F**) to ensure that Dale's law is respected. This also sets the overall connection probability to 0.5. (For a study of how efficient coding would be implemented if the above Dale's law constraint was removed and each neuron is free to have either an inhibitory or excitatory effect depending on the postsynaptic target, see Supplementary Text 1). Neurons with opposite tuning have low connection probability, consistent with experimental results[51,52,53] (**Fig. 1D**). Note that the structured recurrent connectivity leads to both E and I cells being stimulus-tuned, even though I cells do not receive feedforward inputs (**Fig. 1C**). The spike-triggered adaptation current of neuron i in population y, $I_i^{\text{ad},y}$, has the dynamics of its low-pass filtered spike train $r_i^y(t)$. This current realizes spike-frequency adaptation or facilitation depending on the difference between the time constants of population and single neuron readout (see Results section "Weak spike-triggered adaptation optimizes network efficiency"). Finally, external currents have a constant mean, that depends on the parameter β, plus fluctuations that depend on the noise in the condition for spiking with intensity σ. Importantly, the relative weight of the metabolic cost over the encoding error controls the operating regime of the network biophysically plausibly, by modulating the mean of the external current. (Previous studies interpreted changes of the metabolic constant β as changes to the firing thresholds, which has less biophysical plausibility[36,33]) (see section "State-dependent coding and dynamics are controlled by the metabolic cost on spiking").

To summarize, the analytical derivation of an optimally efficient network includes gLIF neurons[54,41,40,55,56], a distributed code with mixed selectivity to the input stimuli, spike-triggered adaptation current, structured synaptic connectivity, and an operating regime controlled by the metabolic constant β.

The equations for the E-I network of gLIF neurons in Eq. (5) optimize the loss functions at any given time and for any set of parameters. In particular, the network equations have the same analytical form for any positive value of the metabolic constant β. To find a set of parameters that optimizes the overall performance, we defined a performance measure as the average over time and trials of the loss function. We then optimized the parameters by setting the metabolic constant β such that the encoding error weights 70 % and the metabolic error weights 30 % of the total performance, and by choosing all other parameters such as to minimize numerically our network performance measure (see Methods). The numerical optimization was performed by simulating a model of 400 E and 100 I units, a network size relevant for computations within one layer of a cortical microcolumn[57]. The set of model parameters that optimized network efficiency is detailed in **Table 1**. Unless otherwise stated, in all simulations we will use the optimal parameters of **Table 1** and only vary those parameters detailed in the figure axes.

With optimally efficient parameters, population readouts closely tracked the target signals (**Fig. 1G**, M=3, R^2 = [0.95, 0.97] for E and I neurons, respectively). When stimulated by our 3-dimensional time-varying feedforward input, the optimal E-I network provided a precise and unbiased estimator of the multi-dimensional and time-dependent target signal (**Fig. 1H**).

Next, we examined the emergent dynamical properties of an optimally efficient E-I network. The distribution of firing rates was well described by a log-normal distribution (**Fig. 1I**, left). Neurons fired irregularly, with mean coefficient of variation (CV) slightly smaller than 1 (**Fig. 1I**, right; CV= [0.97, 0.95] for E and I neurons, respectively). We assessed E-I balance in single neurons through two complementary measures. First, we calculated the *average* (global) balance of E-I currents by taking the time-average of the net sum of currents[58]. Second, we evaluated the *instantaneous*[59] (also termed detailed[45]) E-I balance using the Pearson correlation (ρ) of E and I currents received by a single neuron over time (see Methods).

We observed a strong average E-I balance (indicated by a net sum of synaptic inputs close to zero, with only a weak residual of inhibition in both E and I cells (**Fig. 1J**). Furthermore, we found a moderate instantaneous balance, stronger in I compared to E cell type (**Fig. 1K-L**, ρ = [0.44, 0.25], for I and E neurons, respectively). The presence of instantaneous balance between E and I synaptic currents within single neurons has been reported in cortical data[59,60].

Competition across neurons with similar stimulus tuning emerging in efficient spiking networks

We next explored coding properties emerging from recurrent synaptic interactions between E and I populations in the optimally efficient networks.

An approach that has recently provided empirical insight into local recurrent interactions between neurons is measuring the effective connectivity with cellular resolution, by photostimulating individual neurons and measuring the effect of such perturbation on other neurons in the network. Recent effective connectivity experiments photostimulated single E neurons in primary visual cortex and measured its effect on neighbouring neurons, finding that the photostimulation of an E neuron led to a decrease in firing rate of similarly tuned close-by neurons[61]. This effective lateral inhibition[26] between E neurons with similar tuning to the stimulus implements competition between neurons for the representation of stimulus features (termed feature-specific competition[61]).

To assess how E-I interactions shape coding in efficient networks, we simulated photostimulation experiments in these networks. We performed such experiments in the absence of the feedforward input to insure all effects are only due to the recurrent processing and not to feedforward processing. We stimulated a randomly selected single “target” E neuron and measured the change in the instantaneous firing rate from the baseline firing rate, $\Delta z_i(t)$, in all the other I and E neurons (**Fig. 2A**, left). The photo-stimulation was modeled as an application of a constant depolarising current with a strength parameter, a_p, proportional to the distance between the resting potential and the firing threshold (a_p = 0 means no stimulation, while a_p = 1 indicates photostimulation at the firing threshold). We quantified the effect of the simulated photostimulation of a target E neuron on other E and I neurons, distinguishing neurons with either similar or different tuning with respect to the target neuron (**Fig. 2A**, right; **Supplementary Fig. S2**).

The photostimulation of the target E neuron increased the instantaneous firing rate of similarly-tuned I neurons and reduced that of other similarly-tuned E neurons (**Fig. 2B**, **Supplementary Fig. S2**). We quantified the effective connectivity as the difference between the time-averaged firing rate of the recorded cell in presence or absence of the photostimulation of the targeted cell, measured during perturbation and up to 50 ms after. We found positive effective connectivity on I and negative effective connectivity on E neurons with similar tuning to the stimulated neuron, with a positive correlation between tuning similarity and effective connectivity on I neurons and a negative correlation on E neurons (**Fig. 2C**). As we varied the strength of the photostimulation, the firing rate of the target neuron increased proportionally to the photostimulation strength, as did the effect of the perturbation on I and E neurons with similar tuning to the target neuron (**Fig.**

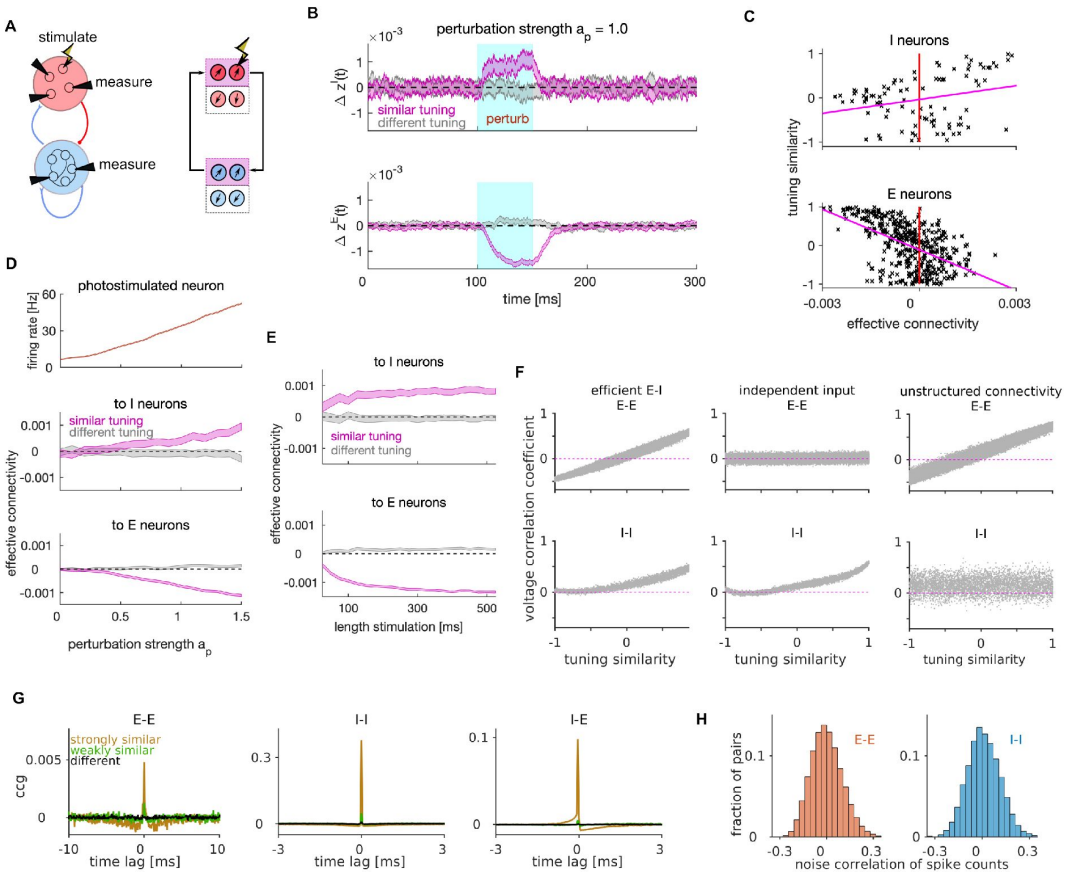


Figure 2.

Mechanism of lateral excitation/inhibition in the efficient spiking network.

(A) Left: Schematic of the E-I network and of the stimulation and measurement in a perturbation experiment. Right: Schematic of the propagation of the neural activity between E and I neurons with similar tuning.
(B) Trial and neuron-averaged deviation of the firing rate from the baseline, for the population of I (top) and E (bottom) neurons with similar (magenta) and different tuning (gray) to the target neuron. The stimulation strength corresponded to an increase in the firing rate of the stimulated neuron by 28.0 Hz.
(C) Scatter plot of the tuning similarity vs. effective connectivity to the target neuron. Red line marks zero effective connectivity and magenta line is the least-squares line. Stimulation strength was a_p = 1.
(D) Top: Firing rate of the photostimulated neuron as a function of the photostimulation strength. Middle: Effective connectivity with I neurons with similar and different tuning to the target neuron. Bottom: Effective connectivity with E neurons.
(E) Effective connectivity with I (top) and E neurons (bottom) while varying the length of the stimulation window. The window for measuring the effective connectivity was always 50 ms longer than the stimulation window.
(F) Correlation of membrane potentials vs. the tuning similarity in E (top) and I cell type (bottom), for the efficient E-I network (left), for the network where each E neuron receives independent instead of shared stimulus features (middle), and for the network with unstructured connectivity (right). In the model with unstructured connectivity, elements of each connectivity matrix were randomly shuffled. We quantified voltage correlation using the (zero-lag) Pearson's correlation coefficient, denoted as $\rho(V_i^y(t), V_j^y(t))$, for each pair of neurons.
(G) Average cross-correlogram (CCG) of spike timing with strongly similar (orange), weakly similar (green) and different tuning (black).
(H) Distribution of noise correlations across neuronal pairs. The correlation coefficient was measured in bins of 30 ms.

2D, **Supplementary Fig. S2**). As we varied the time window of photostimulation, we found that the effective connectivity converges within a time window of about 300 ms (**Fig. 2E**). We confirmed these effects of photostimulation in presence of a weak feedforward input (**Supplementary Fig. S2**), similar to the experiments of Ref[61] in which photostimulation was applied during the presentation of visual stimuli with weak contrast.

In summary, lateral excitation of I neurons and lateral inhibition of E neurons with similar tuning is an emerging coding property of the efficient E-I network. Lateral excitation and inhibition leads to competition between neurons with similar tuning to stimulus features, comparable to that found in the visual cortex[61,62]. An intuitive summary of how this mechanism is implemented is that the E neuron that fires first activates I neurons with similar tuning. In turn, these I neurons inhibit all similarly tuned E neurons (**Fig. 2A**, right), preventing them to generate redundant spikes and encoding the sensory information that has already been encoded by the first spike. Suppression of redundant spiking allows efficient coding because it reduces the metabolic cost without compromising on encoded information[36].

To explore further the consequences of E-I interactions for stimulus encoding, we next investigated the dynamics of lateral inhibition in a network driven by the feed-forward sensory input but without perturbing neurons. In this case, shared feedforward inputs $s_k(t)$ create a particular pattern of voltage correlations in E-E neuronal pairs, where voltage correlations linearly depend on the tuning similarity (**Fig. 2F**, left). The feedforward inputs are shared across neurons and weighted by the tuning parameters of E neurons. For this reason, they cause strong positive voltage correlations between E-E neuronal pairs with very similar tuning and strong negative correlations between pairs with very different (opposite) tuning (**Fig. 2F**, top-left). Voltage correlations between E-E pairs vanished regardless of tuning similarity when we made the inputs independent across neurons (**Fig. 2F**, top-middle), showing the relation between tuning similarity and voltage correlation occurs because of shared feedforward inputs. In contrast to E neurons, I neurons do not receive feedforward inputs and are driven only by similarly tuned E neurons (**Fig. 2A**, right). This causes positive voltage correlations in I-I neuronal pairs with similar tuning and vanishing correlations in neurons with different tuning (**Fig. 2F**, bottom-left). Such dependence of voltage correlations on tuning similarity disappears when removing the structure from the E-I synaptic connectivity (**Fig. 2F**, bottom-right).

Although membrane potentials could be strongly correlated or anti-correlated depending on tuning similarity (**Fig. 2F**, left), the coordination of spike timing of pairs of E neurons (measured with cross-correlograms or CCGs) was very weak (**Fig. 2G-H**). For I-I neuronal pairs, the peaks of CCGs were stronger than those observed in E-E pairs, but they were present only at very short lags (lags < 1 ms), and the same was true for E-I pairs. Additionally, noise correlations measured as Pearson correlation on spike counts in trials with the same stimulus (r_{SC}) had values distributed around zero (**Fig. 2H**). These findings lead to two conclusions. First, recurrent interactions of the efficient E-I network wipe away the effect of membrane potential correlations to produce largely uncorrelated spiking output, consistently with the efficient coding hypothesis of reducing redundancy in cases of low noise[3,6]. Second, such precise cancelling of correlations between voltages and the spiking output reflects the millisecond precision of information processing in efficient E-I networks.

The effect of structured connectivity on coding efficiency and neural dynamics

The analytical solution of the optimally efficient E-I network predicts that recurrent synaptic weights are proportional to the tuning similarity between neurons. We here investigated the role of such efficient connectivity structure by comparing the behavior of an efficiently structured network with a similar but randomly structured E-I network of the type studied in previous works[63,64,23]. We removed the connectivity structure by randomly permuting synaptic

weights across neuronal pairs. We either randomized connections within a single connectivity type (E-I, I-I or I-E) or within all these three connectivity types at once ("all"). Such procedure destroys the relationship between tuning similarity and synaptic strength as in **Fig. 1F** while it preserves Dale's law and the overall distribution of connectivity weights. We found that randomizing the connectivity structure significantly altered neural dynamics and coding (**Fig. 3A-H**). The structure in E-I and in I-E connectivity has a major effect on efficient coding. Randomizing E-I and I-E connectivity led to several-fold increases in the encoding error as well as to significant increases in the metabolic cost (**Fig. 3A-B**). In particular, with unstructured E-I connectivity the network failed completely to encode the target with I population (**Fig. 3C**).

Unstructured E-I and I-E connectivity also yielded an increase of the variance in the membrane potentials (**Fig. 3D**) and firing rate in E neurons (**Fig. 3E**), while pulling the average net synaptic inputs towards inhibition (**Fig. 3F**) and removing the instantaneous balance (**Fig. 3G**). Together, these findings suggest a shift from mean-driven to fluctuation-driven spiking activity as the connectivity structure is removed. The structure of E-I connectivity was also found to be crucial for the linear relation between voltage correlations and tuning similarity in pairs of I neurons (**Fig. 3H**, magenta). Interestingly, we found no effect of connectivity structure on the variability of spiking of single neurons, with both structured and unstructured networks showing strong variability (**Supplementary Fig. S3**), suggesting that the variability of spiking is independent of the connectivity structure.

Randomizing I-I connectivity was less detrimental to the coding efficiency as it led to a slightly higher encoding error, but to a lower metabolic cost, and still allowed for a relatively good tracking of target signals in both cell types (**Fig. 3C**, "permuted I to I"). Contrary to randomization of the E-I and I-E connectivity, shuffling I-I connectivity decreased the variance of the membrane potential, decreased the firing rate in E neurons and increased instantaneous balance in E neurons. Thus it had opposite effects compared to shuffling of E-I and I-E connectivity. To understand if there was a minimal connectivity structure necessary for efficient coding, we also removed the connectivity structure only partially, keeping like-to-like connectivity structure and removing all structure beyond like-to-like. This manipulation only had very modest effects on network's coding and almost no effect on neural dynamics (**Supplementary Fig. S3**), thus showing that like-to-like structure of connectivity is largely sufficient to achieve efficient coding.

Finally, we analyzed how the structure in recurrent connectivity influences lateral inhibition that we observed in efficient (structured) networks (see **Fig. 2A-E**). We found that the dependence of lateral inhibition on tuning similarity vanish when the connectivity structure is fully removed (**Fig. 3I**, right), thus showing that connectivity structure is necessary for lateral inhibition. While networks with unstructured E-I and I-E connectivity still show inhibition in E neurons upon single neuron optostimulation (because of the net inhibitory effect of recurrent connectivity; **Supplementary Fig. S4**), this inhibition was largely unspecific to tuning similarity. Unstructured connectivity decreased the correlation between tuning similarity and effective connectivity from r = [0.31, − 0.54] in E and I neurons in a structured network to r = [0.02, − 0.13] and r = [0.57, 0.11] in networks with unstructured E-I and I-E connectivity, respectively (**Fig. 3I**, first and third from the left). Removing the structure in I-I connectivity, in contrast, increased the correlation between effective connectivity and tuning similarity in E neurons (r = [0.30, − 0.65], **Fig. 3I**, second from the left), showing that lateral inhibition takes place irrespective of the I-I connectivity structure. Furthermore, a partial removal of connectivity structure where we only removed the connectivity structure beyond like-to-like had smaller effects on lateral inhibition (**Supplementary Fig. S4**), thus confirming that like-to-like connectivity pattern is sufficient for lateral excitation/inhibition in I and E neurons.

While optimally structured connectivity predicted by efficient coding is biologically plausible, it may be difficult to realise it exactly on a synapse-by-synapse basis in biological networks. We verified the robustness of the model to small deviations from the optimal synaptic weights by

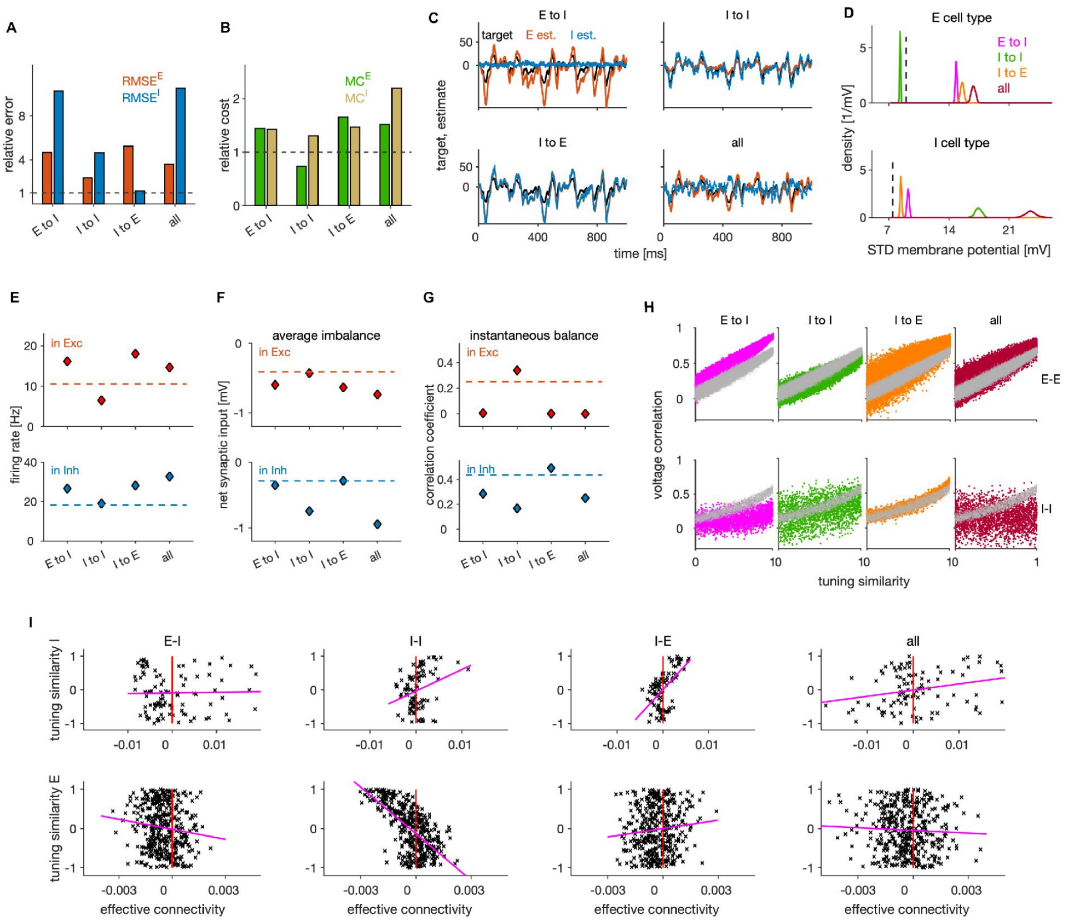


Figure 3.

Effects of connectivity structure on coding efficiency, neural dynamics and lateral inhibition.

(A) Relative error of networks with unstructured (shuffled) recurrent connectivity. The relative error is the RMSE of the unstructured network, relative to the RMSE of the structured network (dashed line). From left to right, we show the relative error for the unstructured E-I, I-I, I-E and all connectivities. **(B** Same as in **A**, showing the metabolic cost (MC) of unstructured networks relative to the metabolic cost of the structured network.
(C) Target signal (black), E estimate (red) and I estimate (blue) in one particular input dimension, for networks with unstructured connectivity.
(D) Standard deviation of the membrane potential (in mV) for networks with unstructured connectivity. Distributions are across neurons. The black vertical line marks the average SD of the structured network.
(E) Average firing rate of E neurons (top) and I neurons (bottom), for different cases of unstructured networks. Dashed lines show the same measures for the structured case.
(F) Same as in **E**, showing the average net synaptic input.
(G) Same as in **E**, showing the time-dependent correlation of synaptic inputs.
(H) Voltage correlation in E-E (top) and I-I neuronal pairs (bottom) for the four cases of unstructured connectivity (colored dots) and the equivalent result in the structured network (grey dots). We show the results for pairs with similar tuning.
(I) Scatter plot of effective connectivity in I (top) and E neurons (bottom) versus tuning similarity to the stimulated ("target") E neuron, for networks with unstructured connectivity. The magenta line is the least-squares regression line. The strength of the photostimulation is at threshold (a_p = 1.0). Other parameters for all plots are in **Table 1**.

adding a random jitter, proportional to the synaptic strength, to all synaptic connections (see Methods). The encoding performance and neural dynamics were barely affected by such perturbation, demonstrating that the network is robust against random perturbations of the optimal synaptic weights (**Supplementary Fig. S3**).

In summary, we found that some aspects of recurrent connectivity structure, such as the like-to-like organization of E-I and I-E connectivity, are crucial to achieve efficient coding. Instead, for other aspects there is considerable flexibility; the organization of I-I connectivity is less crucial, as is the connectivity structure beyond like-to-like, and adding small perturbations to optimal weights has only minor effects. Structured E-I and I-E, but not I-I connectivity, is necessary for a robust dependence of lateral inhibition on tuning similarity.

Weak spike-triggered adaptation optimizes network efficiency

We next investigated the role of spike-triggered adaptation current, $I_i^{\mathrm{ad},y}$, that emerges from the optimally efficient solution (**Eq. 5**). This current provides a within-neuron feedback triggered by each spike, with time constant equal to that of the single neuron readout τ_r^E (E neurons) and τ_r^I (I neurons). The strength of the current is proportional to the difference in inverse time constants of single neuron and population readouts, $1/\tau - 1/\tau_r^y$, and it is thus absent in previous studies assuming that these time constants are equal[29,28,33,31,44,42].

Depending on the sign of the difference of time constants, this spike-triggered current is negative, giving spike-triggered adaptation[39], if the single-neuron readout has longer time constant than the population readout $\left(\tau_r^y > \tau\right)$, or positive, giving spike-triggered facilitation, if the opposite is true $\left(\tau_r^y < \tau\right)$ (**Table 2**). We expected that network efficiency would benefit from spike-triggered adaptation (in short, adaptation), because accurate encoding requires fast temporal dynamics of the population readouts, to capture fast fluctuations in the target signal, while we expect a slower dynamics in the readout of single neuron's firing frequency, $r_i^y(t)$, a process that could be related to homeostatic regulation of single neuron's firing rate[65,66]. Measuring performance of a simulated E-I network, we indeed found that optimal coding efficiency is achieved with weak adaptation in both cell types, and in particular in regimes where the adaptation is stronger in E compared to I neurons (**Fig. 4A**). We note that adaptation in E neurons promotes efficient coding because it enforces every spike to be error-correcting, while a spike-triggered facilitation in E neurons would lead to additional spikes that might be redundant and reduce network efficiency. Contrary to previously proposed model of adaptation in LIF neurons[38], here strength and the time constant of adaptation are not independent, but they both depend on τ_i^y, with larger τ_i^y yielding both longer and stronger adaptation.

To gain further insights on how adaptation influences network performance, we set the adaptation in one cell type to 0 and vary the strength of adaptation in the other cell type by varying the time constant of the single neuron readout. In the absence of adaptation in I neurons $\left(\tau_r^I = \tau\right)$, adaptation in E neurons resulted in an increase of the encoding error in E neurons and a decrease in I neurons (**Fig. 4B**, top). Conversely, adaptation in I neurons (with no adaptation in E neurons) was harmful for the efficiency of the model, as it led to an increase in the encoding error in both cell types (**Fig. 4B**, bottom).

Firing rates and variability of spiking were sensitive to the strength of adaptation. As expected, adaptation in E neurons caused a decrease in the firing levels in both cell types (**Fig. 4C**). In contrast, adaptation in I neurons decreased the firing rate in I neurons, but increased the firing rate in E neurons, due to a decrease in the level of inhibition. Furthermore, adaptation decreased the variability of spiking, in particular in the cell type with strong adaptation (**Fig. 4D**), a well-known effect of spike-triggered adaptation in single neurons[67].

relative speed	relation of time constants	current
$\hat{x}^E$ faster than r^E	$\tau < \tau_r^E$	adaptation in E
$\hat{x}^E$ slower than r^E	$\tau > \tau_r^E$	facilitation in E
$\hat{x}^I$ faster than r^I	$\tau < \tau_r^I$	adaptation in I
$\hat{x}^I$ slower than r^I	$\tau > \tau_r^I$	facilitation in I

Table 2.

Relation of time constants of single-neuron and population readout set an adaptation or a facilitation current.

The population readout that evolves on a faster (slower) time scale than the single neuron readout determines a spike-triggered adaptation (facilitation) in its own cell type.

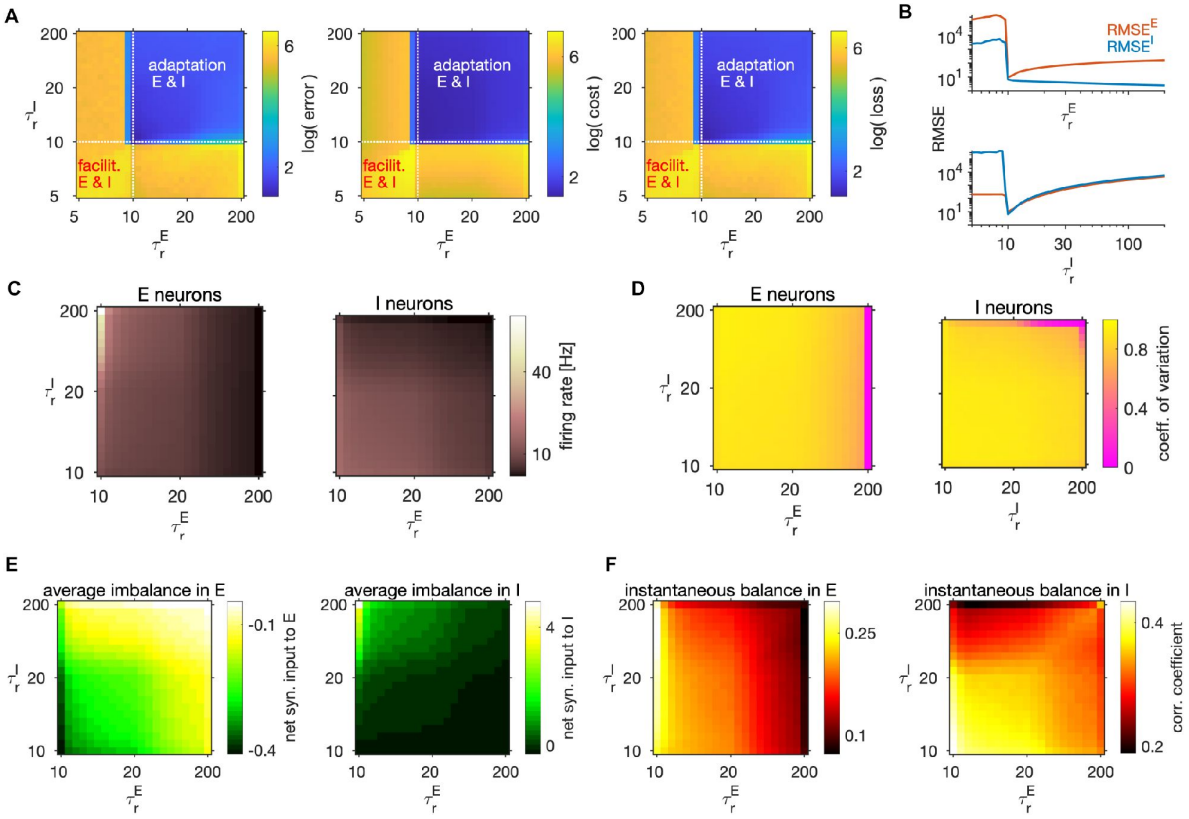


Figure 4.

Adaptation, network coding efficiency and excitation-inhibition balance.

(A) The encoding error (left), metabolic cost (middle) and average loss (right) as a function of single neuron time constants τ_r^E (E neurons) and τ_r^I (I neurons), in units of ms. These parameters set the sign, the strength, as well as the time constant of the feedback current in E and I neurons. Best performance is obtained in the top right quadrant, where the feedback current is spike-triggered adaptation in both E and I neurons. The performance measures are computed as a weighted sum of the respective measures across the E and I populations with equal weighting for E and I. All measures are plotted on the scale of the natural logarithm for better visibility.

(B) Top: Log-log plot of the RMSE of the E (red) and the I (blue) estimates as a function of the time constant of the single neuron readout of E neurons, τ_r^E. Feedback current in I neurons is set to 0. Bottom: Same as on the top, as a function of τ_r^I while the feedback current in E neurons is set to 0.

(C) Firing rate in E (left) and I neurons (right), as a function of τ_r^E and τ_r^I in the regime with spike-triggered adaptation.

(D) Same as in **(C)**, showing the coefficient of variation.

(E) Average net synaptic input in E neurons (left) and in I neurons (right) as a function of τ_r^E and τ_r^I.

(F) Correlation coefficient of synaptic inputs to E (left) and I neurons (right) as a function of τ_r^E and τ_r^I

Instantaneous balance of synaptic currents predicts network efficiency better than the average E-I balance

Next, we tested the capability of instantaneous and average E-I balance to predict the efficiency of the network. Measuring average balance and instantaneous balance of synaptic inputs from electrophysiology recordings is possible[59,60,58], while measuring efficiency from empirical data is challenging. The estimation of network efficiency requires the comparison between typically unknown network's target representations and the population readouts. The estimation of the population readout, in turn, requires an estimation of decoding weights and the knowledge of spiking dynamics from a complete neural network.

We focused the analysis on regimes with adaptation, because these regimes gave better performance. In regimes with adaptation, time constants of single neuron readout influenced the average imbalance (**Fig. 4E**) as well as the instantaneous balance (**Fig. 4F**) in E and I cell type. The average balance was precise (with the net synaptic current close to 0) with strong adaptation in E neurons, and it got weaker when increasing the adaptation in I neurons (**Fig. 4E**). However, regimes with precise average balance in both cell types coincided with suboptimal efficiency (compare **Fig. 4A**, right and E).

To test how well the average imbalance and the instantaneous balance of synaptic inputs predict network efficiency, we concatenated the column-vectors of the measured average loss and of the average imbalance in each cell type and computed the Pearson correlation between these quantities. The correlation between the average imbalance and the average loss was weak in the E cell type (R = 0.16) and close to zero in the I cell type (R = 0.02), suggesting almost no relation between efficiency and average imbalance in the E cell type. In contrast, the average loss was negatively correlated with the instantaneous balance in both E (R = − 0.35) and in I cell type (R = − 0.45), showing that instantaneous balance of synaptic inputs is positively correlated with network efficiency.

When measured for varying levels of spike-triggered adaptation, unlike the average balance of synaptic inputs, the instantaneous balance is therefore a reliable predictor of network efficiency.

State-dependent coding and dynamics are controlled by the metabolic cost on spiking

In our derivation of efficiency objectives, we obtained non-specific external current (in the following, non-specific current), described by the term $I_i^{\mathrm{ext},y}(t)$ and comprising mean and fluctuations (see Methods). Non-specific current captures the ensemble of all synaptic currents that are unrelated and un-specific with respect to the stimulus features. This non-specific term collates effects of synaptic currents from neurons untuned to the stimulus[68,69], as well as synaptic currents from other brain areas. This term can also be conceptualized as the "background" synaptic activity that is thought to provide a large fraction of all synaptic inputs to both E and I neurons in cortical networks[70], and which may modulate feedforward-driven responses by controlling how far is typically the membrane potential from the firing threshold[71]. Likewise, in our model, the external current does not directly convey information about the feedforward input features, but influences the operating regime of the network. The mean of the non-specific external currents is proportional to the metabolic constant β and its fluctuations reflect the noise that we assumed in the condition for spiking. Since β governs the trade-off between encoding error and metabolic cost (Eq. 3), higher values of β imply that more importance is assigned to the metabolic efficiency than to coding accuracy, yielding a reduction in firing rates. In the expression for the non-specific synaptic current, we found that the mean of the current is negatively proportional to the metabolic constant β (see Methods). The non-specific current is typically depolarizing, meaning that increasing β yields a weaker non-specific current

and increases the distance between mean membrane potential and the firing threshold. Thus, an increase of the metabolic constant is expected to create a network state that is less responsive to the feedforward signal.

We found the metabolic constant β to significantly influence the spiking dynamics (**Fig. 5A**). The optimal efficiency was achieved for non-zero levels of the metabolic constant (**Fig. 5B**). The metabolic constant modulated the firing rate as expected, with the firing rate decreasing with the increasing of the metabolic constant (**Fig. 5C**, top). It also modulated the variability of spiking, as increasing the metabolic constant decreased the variability of spiking in single neurons (**Fig. 5C**, bottom). Furthermore, it modulated the average imbalance and the instantaneous balance in opposite ways: larger values of β led to regimes that had stronger average balance, but weaker instantaneous balance (**Fig. 5D**). We note that, even with suboptimal values of the metabolic constant, the neural dynamics remained within biologically relevant ranges.

The fluctuation part of the non-specific current, modulated by the noise intensity σ, that we added in the definition of spiking rule for biological plausibility (see Methods), strongly affected the neural dynamics as well (**Fig. 5E**). The optimal performance was achieved with non-vanishing noise levels (**Fig. 5F**) and the beneficial effect of the noise in the non-specific current arose from its impact on the instantaneous E-I balance. While the average firing rate of both cell types, as well as the variability of spiking in E neurons, increased with noise variance (**Fig. 5G**), the average and instantaneous balance of synaptic currents exhibited a non-linear behavior as a function of noise variance (**Fig. 5H**). Due to decorrelation of membrane potentials by the noise, instantaneous balance decreased with increasing noise variance (**Fig. 5H**, bottom). Some level of noise in the non-specific inputs is therefore necessary to establish the optimal level of instantaneous E-I balance. Interestingly, single neurons manifest significant levels of spiking variability already in the absence of noise in the non-specific inputs (**Fig. 5H**, bottom), indicating that the recurrent network dynamics generates substantial variability even in absence of variability in the external current. Variability in absence of noise demonstrates the intrinsic chaotic behavior of the network[72].

In summary, non-specific external currents derived in our optimal solution have a major effect on coding efficiency and on neural dynamics. The noise in the external current is particularly important to obtain optimal levels of the instantaneous E-I balance in I neurons.

Optimal ratio of E-I neuron numbers and of the mean I-I to E-I synaptic efficacy coincide with biophysical measurements

Next, we investigated how coding efficiency and neural dynamics depend on the ratio of the number of E and I neurons ($N^E : N^I$ or E-I ratio) and on the relative synaptic strengths between E-I and I-I connections.

Efficiency objectives (Eq. 3) are based on population, rather than single-neuron activity. Our efficient E-I network thus realizes a computation of the target representation that is distributed across multiple neurons (**Fig. 6A**). We predict that, if number of neurons within the population decreases, neurons have to fire more spikes to achieve an optimal population readout because the task of tracking the target signal is distributed among fewer neurons. To test this prediction, we varied the number of I neurons while keeping the number of E neurons constant. As predicted, a decrease of the number of I neurons (and thus an increase in the ratio of the number of E to I neurons) caused a linear increase in the firing rate of I neurons, while the firing rate of E neurons stayed constant (**Fig. 6B**, top). However, the variability of spiking and the average synaptic inputs remained relatively constant in both cell types as we varied these ratios (**Fig. 6B**, bottom, C), indicating a compensation for the change in the ratio of E-I neuron numbers through

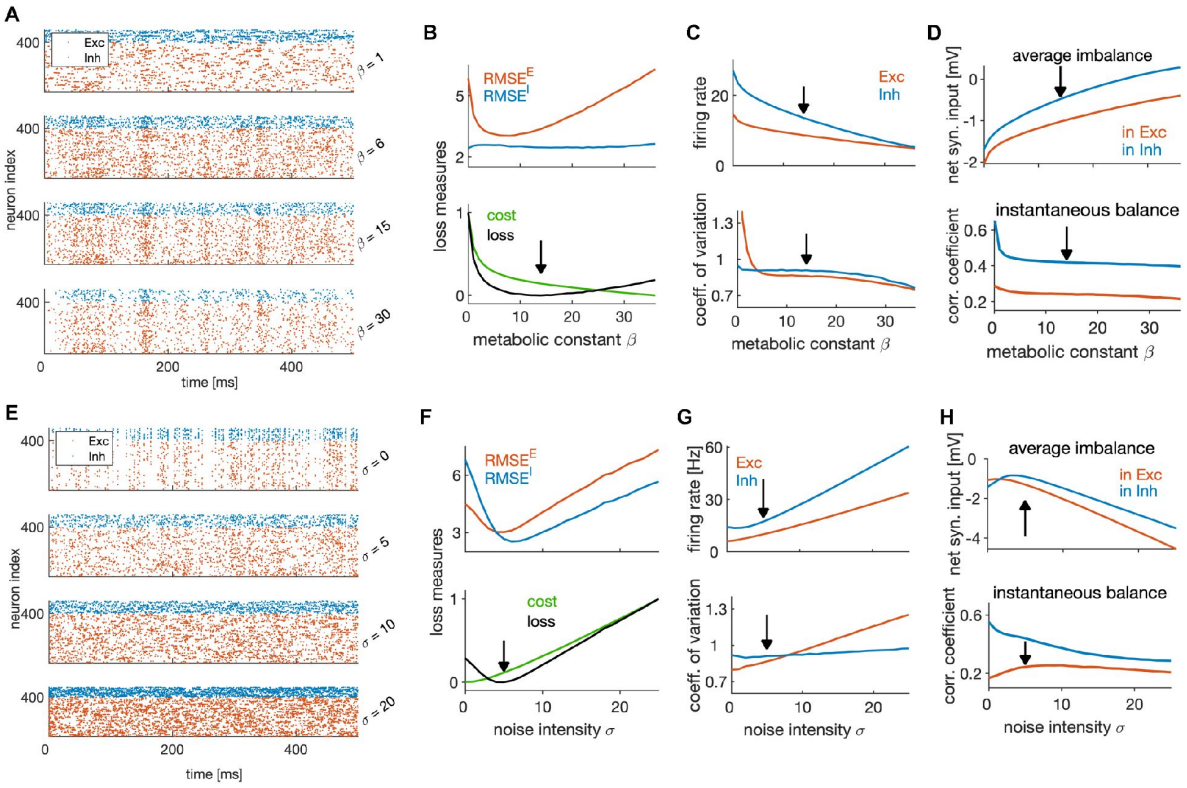


Figure 5.

State-dependent coding and dynamics are controlled by non-specific currents.

(A) Spike trains of the efficient E-I network in one simulation trial, with different values of the metabolic constant β. The network received identical stimulus across trials.
(B) Top: RMSE of E (red) and I (blue) estimates as a function of the metabolic constant. Bottom: Normalized average metabolic cost and average loss as a function of the metabolic constant. Black arrow indicates the minimum loss and therefore the optimal metabolic constant.
(C) Average firing rate (top) and the coefficient of variation of the spiking activity (bottom), as a function of the metabolic constant. Black arrow marks the metabolic constant leading to optimal network efficiency in **B**.
(D) Average imbalance (top) and instantaneous balance (bottom) balance as a function of the metabolic constant.
(E) Same as in **A**, but for different values of the noise intensity σ.
(F) Same as in **B**, as a function of the noise intensity. The noise is a Gaussian random process, independent over time and across neurons.
(G) Same as **C**, as a function of the noise intensity.
(H) Top: Same as in **D**, as a function of the noise intensity. For plots in **B-D** and **F-H**, we computed and averaged results over 100 simulation trials with 1 second of simulation time. For other parameters, see **Table 1**.

adjustment in the firing rates. These results are consistent with the observation in neuronal cultures of a linear change in the rate of postsynaptic events but unchanged postsynaptic current in either E and I neurons for variations in the E-I neuron number ratio[73].

The ratio of the number of E to I neurons had a significant influence on coding efficiency. We found a unique minimum of the encoding error of each cell type, while the metabolic cost increased linearly with the ratio of the number of E and I neurons (**Fig. 6D**). We found the optimal ratio of E to I neuron numbers to be in range observed experimentally in cortical circuits (**Fig. 6D**, bottom, black arrow, $N^E : N^I = 3.75 : 1$;[74]). Due to the linear increase of the cost with the ratio of the number of E and I neurons (**Fig. 6D**, bottom, green), strong weighting of the error predicted higher ratios (**Fig. 6E**, bottom). Also the encoding error (RMSE) alone, without considering the metabolic cost, predicted optimal ratio of the number of E to I neurons within a plausible physiological range, $N^E : N^I = [3.75 : 1, 5.25 : 1]$, with stronger weightings of the encoding error by I neurons predicting higher ratios (**Fig. 6E**, top).

Next, we investigated the impact of the strength of excitatory and inhibitory synaptic efficacy (EPSPs and IPSPs). In our model, the mean synaptic efficacy is fully determined by the distribution of tuning parameters (see Methods). As evident from the expression for the population readouts (**Eq. 2**), the amplitude of tuning parameters (which are also decoding weights) determines the amplitude of jumps of the population readout caused by spikes (**Fig. 6F**). The stronger the amplitude of these weights, the larger is the average impact of spikes on the population signals.

We parametrized the distribution of decoding weights as a normal distributions centered at zero, but allowed the standard deviation (SD) of distributions relative to E and I neurons (σ_w^E and σ_w^I) to vary across E and I cell type. With such parametrization, we were able to analytically evaluate the mean E-I, I-I and I-E synaptic efficacy (see Methods). We found that in the optimally efficient network, the mean E-I and I-E synaptic efficacy is exactly balanced.

We next searched for the optimal ratio of the mean I-I to E-I efficacy as the parameter setting that maximizes network efficiency. Network efficiency was maximized when such ratio was about 3 to 1 (**Fig. 6G**). Our results predict the maximum E-I and I-E synaptic efficacy, averaged across neuronal pairs, of 0.75 mV, and the maximal I-I efficacy of 2.25 mV, values that are consistent with empirical measurements in the primary sensory cortex[75,52,53].

Similarly to the ratio of E-I neuron numbers, a change in the ratio of mean E-I to I-E synaptic efficacy was compensated for by a change in firing rates, with stronger I-I synapses leading to a decrease in the firing rate of I neurons (**Fig. 6H**). Conversely, weakening the E-I and I-E synapses resulted in an increase in the firing rate in E neurons (**Supplementary Fig. S5**). This is easily understood by considering that weakening the E-I and I-E synapses activates less strongly the lateral inhibition in E neurons (**Fig. 2**) and thus leads to an increase in the firing rate of E neurons. We also found that single neuron variability remained almost unchanged when varying the ratio of mean I-I to E-I efficacy (**Fig. 6H**, bottom) and the optimal ratio corresponded with previously found optimal levels of average and instantaneous balance of synaptic inputs (**Fig. 6I**). The instantaneous E-I balance monotonically decreased with increasing ratio of I-I to E-I efficacy (**Fig. 6I**, bottom, **Supplementary Fig. S5**).

In summary, our analysis suggests that optimal coding efficiency is achieved with four times more E neurons than I neurons and with mean I-I synaptic efficacy about 3 times stronger than the E-I and I-E efficacy. The optimal network has less I than E neurons, but the impact of spikes of I neurons on the population readout is stronger, also suggesting that spikes of I neurons convey more information.

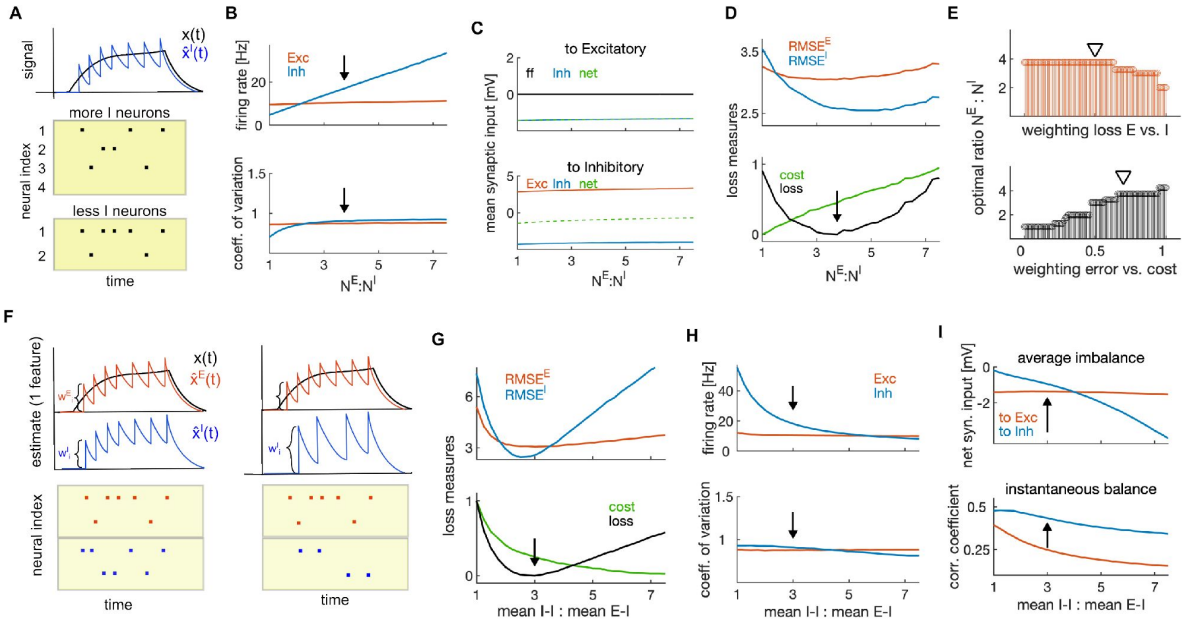


Figure 6.

Optimal ratios of E-I neuron numbers and of mean I-I to E-I efficacy.

(A) Schematic of the effect of changing the number of I neurons on firing rates of I neurons. As encoding of the stimulus is distributed among more I neurons, the number of spikes per I neuron decreases.
(B) Average firing rate as a function of the ratio of the number of E to I neurons. Black arrow marks the optimal ratio.
(C) Average net synaptic currents in E neurons (top) and in I neurons (bottom).
(D) Top: Encoding error (RMSE) of the E (red) and I (blue) estimates, as a function of the ratio of E-I neuron numbers. Bottom: Same as on top, showing the cost and the average loss. Black arrow shows the minimum of the loss, indicating the optimal parameter.
(E) Top: Optimal ratio of the number of E to I neurons as a function of the weighting of the average loss of E and I cell type (using the weighting of the error and cost of 0.7 and 0.3, respectively). Bottom: Same as on top, measured as a function of the weighting of the error and the cost when computing the loss. (The weighting of the losses of E and I neurons is 0.5.) Black triangles mark weightings that we typically used.
(F) Schematic of the readout of the spiking activity of an E neuron (red) and an I neuron (blue) with equal amplitude of decoding weight (left) and with stronger decoding weight in the I neuron (right). Stronger decoding weight in the I neuron results in a stronger effect of spikes of the I neuron on the readout, leading to less spikes by the I neuron.
(G) Same as in (D), as a function of the ratio of mean I-I to E-I efficacy.
(H) Same as in **B**, as a function of the ratio of mean I-I to E-I efficacy.
(I) Average imbalance (top) and instantaneous balance (bottom) balance, as a function of the ratio of mean I-I to E-I efficacy.
For other parameters, see **Table 1**.

Dependence of efficient coding and neural dynamics on the timescales and dimensionality of the stimulus

We finally investigated how the network's behavior depends on the timescales and dimensionality of the input stimulus features. We manipulated the stimulus timescales by changing the time constant of the Ornstein-Uhlenbeck (O-U) process. The network efficiently encoded stimulus features when their time constants varied between 1 and 200 ms, with stable encoding error, metabolic cost (**Fig. 7A**) and neural dynamics (**Supplementary Fig. S6**).

Finally, we tested how the network's behavior changed when we varied the number of stimulus features M processed by the network. The encoding error of E ($RMSE^E$) and I neurons ($RMSE^I$) had a minimum at 3 and 4 stimulus features, respectively (**Fig.7B**, top), while the metabolic cost increased monotonically with the number of features (**Fig.7B**, bottom). The number of features that optimized network efficiency (the average loss) ranged between M = [1, 4]. With strong weighting of the error ($g_L \geq 0.89$), the optimal number of features was M = 4, and with strong weighting of the cost, ($g_L < 0.27$), the optimal number of features was M = 1. It is intriguing that the optimal encoding performance, when assuming the weighting for the error is stronger than for the cost, is achieved not for a single stimulus feature, but for 3 or 4 independent features. Increasing the number of features beyond the optimal number resulted in a monotonic increase in firing rates for both cell types and in a contrasting effect on average and instantaneous balance, as it increased the average E-I balance and weakened the instantaneous balance (**Supplementary Fig. S6**).

In sum, we found the optimal network efficiency in presence of several (3 or 4) stimulus features, and a surprising ability of the network to accurately encode stimuli on a wide range of timescales.

Advantages of E-I versus one cell type model architecture for coding efficiency and robustness to parameter variations

Neurons in the brain are either excitatory or inhibitory. To understand how differentiating E and I neurons benefits efficient coding, we compared the properties of our efficient E-I network with an efficient network with a single cell type (1CT). The 1CT model is a simplification of the E-I model (see Supplementary Text 1) and has been derived and analyzed in previous studies[29,28,36,33,44,42]. We compared the average encoding error (RMSE), the average metabolic cost (MC), and the average loss (see Supplementary Text 2) of the E-I model against the one cell type (1CT) model. Compared to the 1CT model, the E-I model exhibited a higher encoding error and metabolic cost in the E population, but a lower encoding error and metabolic cost in the I population (**Fig. 7C**). The average loss of the E-I model was significantly smaller than that of the 1CT model when using the typical weighting of the error and the cost of g_L = 0.7 (**Fig. 7D**), as well as for the vast majority of other weightings ($g_L \leq 0.95$; **Supplementary Fig. S1**).

We further compared the 1CT and E-I models in terms of the robustness of firing rates to changes in the metabolic constant. Consistently with previous studies[36,35], firing rates in the 1CT model were highly sensitive to variations in the metabolic constant (**Fig. 7E**, note the logarithmic scale on the y-axis), with a superexponential growth of the firing rate with the inverse of the metabolic constant in regimes with metabolic cost lower than optimal. This is in contrast to the E-I model, whose firing rates exhibited lower sensitivity to the metabolic constant, and never exceeded physiological limits (**Fig. 5C**). Because our E-I model does not incorporate a saturating input-output function as in[34] that would constrain the range of firing rates, the ability of the E-I model to maintain firing rates within biologically plausible limits emerges as a highly desirable dynamic property.

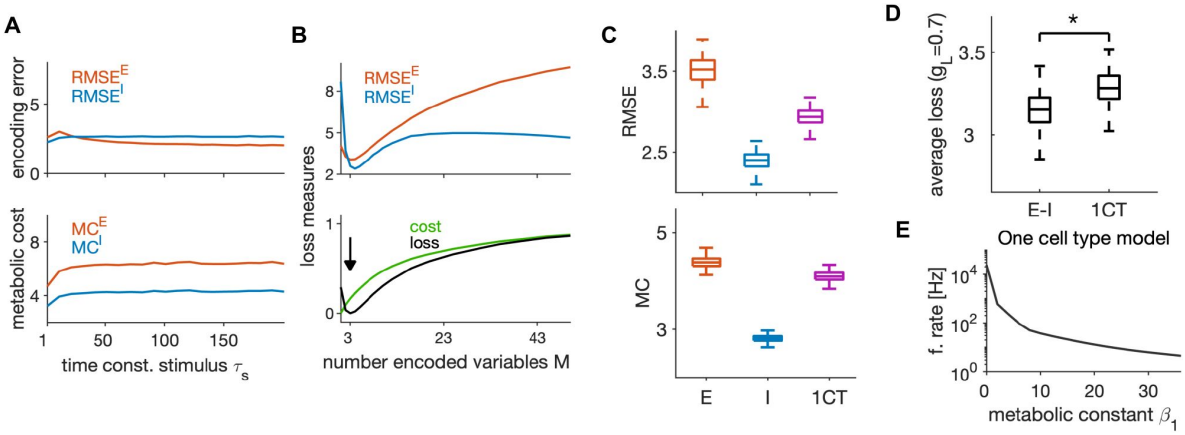


Figure 7.

Dependence of efficient coding and neural dynamics on stimulus parameters and advantages of E-I versus one cell type model architecture.

(A) Top: Root mean squared error (RMSE) of E estimates (red) and I estimates (blue), as a function of the time constant of stimulus features. Bottom: Same as on top, showing the metabolic cost (MC) of E and I cell type. The time constant τ_s is the same for all stimulus features.
(B) Top: Same as in **A** top, measured as a function of the number of stimulus features *M*. Bottom: Normalized cost and the average loss as a function of the number of input features. Black arrow marks the minimum loss and the optimal parameter *M*.
(C) Root mean squared error (top) and metabolic cost (bottom) in E and I populations in the E-I model and in the 1CT model. The distribution is across simulation trials.
(D) Average loss in the E-I and 1CT models with weighting g_L = 0.7 for the error (and 0.3 for the cost).
(E) Firing rate in the 1CT model as a function of the metabolic constant. For other parameters of the E-I model see **Table 1**, and for the 1CT model see **Supplementary Table S1**.

In summary, we found that the optimal E-I model is more efficient than the 1CT model. Beyond the performance of optimal models, the E-I model is advantageous with respect to the 1CT model also because it does not enter into states of physiologically unrealistic firing rates.

Discussion

We analyzed comprehensively the structural, dynamical and coding properties that emerge in networks of spiking neurons that implement optimally the principle of efficient coding. We demonstrated that efficient recurrent E-I networks form highly accurate and unbiased representations of stimulus features with biologically plausible parameters, biologically plausible neural dynamics, instantaneous E-I balance and like-to-like lateral inhibition. The network can implement efficient coding with stimulus features varying over a wide range of timescales and when encoding even multiple such features. Here we discussed the implications of these findings.

By a systematic study of the model, we determined the model parameters that optimize network efficiency. Strikingly, the optimal parameters (including the ratio between the number of E and I neurons, the ratio of I-I to E-I synaptic efficacy and parameters of non-specific currents) were consistent with parameters measured empirically in cortical circuits, and generated plausible spiking dynamics. This result lends credibility to the hypothesis that cortical networks might be designed for efficient coding and may operate close to optimal efficiency, as well as provides a solid intuition about what specific parameter ranges (e.g. higher numbers of E and than I neurons) may be good for. Efficient networks still exhibited realistic dynamics and reasonably efficient coding in the presence of moderate deviations from the optimal parameters, suggesting that the optimal operational point of such networks is relatively robust. We also found that optimally efficient analytical solution derives generalized LIF (gLIF) equations for neuron models[37]. While gLIF[67,40] and LIF[63,64] models are reasonably biologically plausible and are widely used to model and study spiking neural network dynamics, it was unclear how their parameters affect network-level information coding. Our study provides a principled way to determine uniquely the parameter values of gLIF networks that are optimal for efficient information encoding. Studying the dynamics of gLIF networks with such optimal parameters thus provides a direct link between optimal coding and neural dynamics. Moreover, our formalism provides a framework for the optimization of neural parameters that can in principles be used not only for neural network models that study brain function but also for the design of artificial neuromorphic circuits that perform information coding computations[76,77].

Unlike in previous randomly-connected recurrent networks of LIF and gLIF spiking neurons,[63,64] in our efficient-coding solution, a highly structured E-I, I-I and I-E synaptic connectivity emerges as an optimal structural solution to support efficient coding. Our model generates a number of insights about the role of structured connectivity in efficient information processing. A first insight is that I neurons develop stimulus feature selectivity because of the structured recurrent connectivity. This is in line with recent reports of stimulus feature selectivity of inhibitory neurons, including in primary visual cortex[78,79,80]. A second insight is that a network with structured connectivity shows stronger average and instantaneous E-I balance, as well as significantly lower variance in membrane potentials compared to an equivalent network with the same connections organized randomly. This implies that the connectivity structure determines the operating regime of the network. In particular, a network structured as in our efficient coding solution operates in a dynamical regime that is more stimulus-driven, compared to an unstructured network that is more fluctuation driven. A third insight is that the structured network exhibits a several-fold lower encoding error compared to unstructured networks and achieves this precision with lower firing rates. Network with structured recurrent connectivity creates more precise representations with less spikes and is therefore significantly more efficient compared to unstructured networks. Our analysis of the effective connectivity created by the efficient connectivity structure shows that this structure sharpens stimulus representations,

reduces redundancy and increases metabolic efficiency by implementing feature-specific competition, that is a negative effective connectivity between E neurons with similar stimulus tuning, as proposed by recent theories[30] and experiments [61,62] of computations in visual cortex.

Our perturbation experiments on single E neurons predict a negative like-to-like effective connectivity between E neurons with similar tuning, as found experimentally in the mouse primary visual cortex with 2-photon optogenetic perturbations of E neurons[61,62]. This suggests that the effective connectivity found in mouse visual cortex could reflect efficient coding in visual cortex. Comparing effective connectivity in models and experiments is also useful for ruling in and out different theories of how efficient coding may be implemented in primary visual cortex. Earlier theories[4,11] found evidence for efficient coding in visual cortex and proposed that such efficient computations relied only on feedforward connectivity; thus they predicted null effective connectivity between visual neurons and were ruled out by the empirical effective connectivity measures[61,62]. Our model, instead, implements efficient coding with recurrent interactions, suggesting a mechanism that is compatible with these empirical measures. Importantly, we made predictions for further optogenetics experiments that could better constraints models of visual cortical efficient coding. Previous studies[61] optogenetically stimulated E neurons but did not determine whether the recorded neurons where excitatory or inhibitory. Our model predicts that stimulation of E neurons would increase firing in similarly tuned I neurons and decrease firing in similarly tuned E neurons. Our analysis confirms earlier model predictions[81] that like-to-like connectivity between E and I neurons is necessary for lateral inhibition and competition between E neurons. Beyond like-to-like connectivity, our model predicts an optimally efficient connectivity where synaptic strength positively correlates with pair-wise tuning similarity, a connectivity pattern that was recently observed experimentally [82].

Our study determines how structured E-I connectivity affects the dynamics of E-I balancing and how this relates to information coding. Previous work[32] proposed that the E-I balance in efficient spiking networks operates on a finer time scale than in classical balanced E-I networks with random connectivity[64]. However, a theory to determine the exact levels of instantaneous E-I balance that is optimal for coding was lacking. Consistent with the general idea put forth in[32,31,48], we here showed that moderate levels of E-I balance are optimal for coding, and that too strong levels of instantaneous E-I balance are detrimental to coding efficiency. Our results predict that like-to-like structured E-I-E connectivity is necessary for optimal levels of temporal E-I balance. Finally, the E-I-E structured connectivity that we derived supports optimal levels of instantaneous E-I balance and causes desynchronization of the spiking output. Such intrinsically generated desynchronization is a desirable network property that in previously proposed models could only be achieved by the less plausible addition of strong noise to each neuron[31,35].

We found that our efficient network, optimizing the representation of a leaky integration of stimulus features, does not require recurrent E-E connections. Supporting this prediction, recurrent E-E connections were reported to be sparse in primary visual cortex[83]), and the majority of E-E synapses in the visual cortex were suggested to be long-range[84]. However, future studies could address the role of recurrent excitatory synapses, that were shown to emerge in efficient coding networks implementing computations beyond leaky integration such as linear mixing of features[37]. Efficient networks with E-E connectivity show neural dynamics that goes well beyond the canonical case analyzed here and can potentially describe persistent network dynamics[44]. Such networks would also allow to address whether biologically plausible efficient networks exhibit criticality, as suggested by[85]. Finally, we note that efficient encoding might be the primary normative objective in sensory areas, while areas supporting high-level cognitive tasks such as decision-making might include other computational objectives such as efficient transmission of information downstream to generate reliable behavioral outputs[86,87,88,25].

Acknowledgements

V.K. and T.S. thank Tatiana Engel for her contribution to the discussion of results and for her comments on an earlier version of the manuscript. This project was supported by funding from Technische Universität Berlin ("Equal Opportunity Program" to VK), by Internal Research Funding of Technische Universität Berlin (to TS), by NIH Brain Initiative (grants U19 NS107464, R01 NS109961, R01 NS108410 to SP), and the Simons Foundation for Autism Research Initiative (SFARI; grant 982347 to SP).

Code availability

The complete computer code for reproducing the results is available as a Github repository [will be shared upon acceptance].

Methods

Overview of the current approach and of differences with previous approaches

In the following, we present a detailed derivation of the E-I spiking network implementing the efficient coding principle. The analytical derivation is based on previous works on efficient coding with spikes[28,36], and in particular on our recent work[37]. While these previous works analytically derived feedforward and recurrent transmembrane currents in leaky integrate-and fire neuron models, these models did not contain any synaptic current unrelated to feedforward and recurrent processing. Non-specific synaptic current was suggested to be important for an accurate description of coding and dynamics in cortical networks[71]. In the model derivation that follows, we also derived non-specific external current from efficiency objectives.

As we mapped the efficient coding objective on biologically plausible neural implementations, we found that such implementations (with plausible biophysical parameters) requires a transmembrane current that is independent of feedforward and recurrent processing. We interpreted this current as non-specific external current (shortly, non-specific current), collating the ensemble of synaptic projections from other brain areas that are not directly involved in processing of feedforward stimulus features[70], as well as synaptic inputs from the local network from neurons that are not tuned to feedforward stimulus features[69]. The mechanistic effect of the non-specific current is to regulate the distance to firing threshold, a role that is close to the notion of "background" synaptic activity in cortical neurons[71].

Moreover, previous models on efficient coding did not thoroughly consider physical units of variables that were interpreted as biophysical quantities (such as membrane potentials, firing thresholds, etc.). As these biophysical variables were derived from computational variables (such as target signals and population readouts), it remained unclear how biophysical variables might acquire their physical units. Here, we assigned physical units to the computational variables and thus naturally endowed the model with physical units. The network developed here allows for a better compatibility of efficient spiking models with neurobiology compared to previous works on efficient coding with spikes. With this model, we aim to describe neural dynamics and computation in early sensory cortices such as the primary visual cortex in rodents, even though many principles of the model developed here could be relevant throughout the brain.

Introducing variables of the model

We consider two types of neurons, excitatory neurons E and inhibitory neurons I. We denote as N^E and N^I the number of E-cells and I-cells, respectively. The spike train of neuron i of type $y \in \{E, I\}$, $i = 1, 2, ..., N^y$, is defined as a sum of Dirac delta functions,

$$f_i^y(t) = \sum_\alpha \delta(t - t_i^{y,\alpha}), \tag{6}$$

where $t_i^{y,\alpha}$ is the time of the α-th spike of that neuron, defined as a time point at which the membrane potential of neuron i crosses the firing threshold.

We define the readout of the spiking activity of neuron i of type y (in the following, "single neuron readout") as a leaky integration of its spike train,

$$\frac{dr_i^y(t)}{dt} = -\lambda_r^y r_i^y(t) + f_i^y(t), \qquad y \in \{E, I\}, \tag{7}$$

with λ_r denoting the inverse time constant. This way, the quantity $\tilde{r}_i^y(t) = \lambda_r^y r_i^y(t)$ represents an estimate of the instantaneous firing rate of neuron i.

We denote as $s_k(t)$, $k = 1, 2, ..., M$ the set of M dynamical features of the external stimulus (in the following, stimulus features) which are transmitted to the network through a feedforward sensory pathway. The stimulus features have the unit of the square root of millivolt, $(mV)^{\frac{1}{2}}$. The k-th dimension of the target signal is then obtained through a leaky integration of the feedforward variable, $s_k(t)$[29], with inverse time constant λ, as

$$\frac{dx_k(t)}{dt} = -\lambda x_k(t) + s_k(t). \tag{8}$$

Furthermore, we define a linear population readout of the spiking activity of E and I neurons

$$\frac{d\hat{x}_k^y(t)}{dt} = -\lambda \hat{x}_k^y(t) + \sum_{i=1}^{N^y} w_{ki}^y f_i^y(t), \qquad y \in \{E, I\}, \tag{9}$$

with w_{ki}^y in units of $(mV)^{\frac{1}{2}}$. Here, each neuron i of type y is associated with a vector $\boldsymbol{w}_i^y := [w_{1i}^y, \ldots, w_{Mi}^y]^\top$ of M tuning parameters representing the readout weight of neuron i with respect to the M population readouts in Eq. 9. These readout weights can be combined in the $M \times N^y$ matrix $W^y = [w_{ki}^y]$. The rows of this matrix define the patterns of readout weights $\tilde{\boldsymbol{w}}_k^y := [w_{k1}^y, \ldots, w_{kN^y}^y]^\top$ for each signal dimension $k = 1, ..., M$.

Loss functions

We assume that the activity of a population $y \in \{E, I\}$ is set so as to minimize a time-dependent encoding error and a time-dependent metabolic cost:

$$\mathrm{L}^y(t) = \epsilon^y(t) + \beta^y \kappa^y(t), \tag{10}$$

with $\beta^y > 0$ in units of mV the Lagrange multiplier which controls the weight of the metabolic cost relative to the encoding error. The time-dependent encoding error is defined as the squared distance between the targets and their estimates, and the role of estimates is assigned to the

population readouts $\hat{x}_k^y(t)$. In E neurons, the targets are defined as the target signals $x_k(t)$, and their estimators are the population readouts of the spiking activity of E neurons, $\hat{x}_k^E(t)$. In I neurons, the targets are defined as the population readouts of E neurons $\hat{x}_k^E(t)$ and their estimators are the population readouts of I neurons $\hat{x}_k^I(t)$. Furthermore, the time-dependent metabolic cost is proportional to the squared estimate of the instantaneous firing rate, summed across neurons from the same population. Following these assumptions, we define the variables of loss functions in Eq. 10 as

$$\begin{aligned} \epsilon^E(t) &= \sum_{k=1}^{M} \left[x_k(t) - \hat{x}_k^E(t)\right]^2, \qquad \kappa^E(t) = \sum_{i=1}^{N^E} [r_i^E(t)]^2, \\ \epsilon^I(t) &= \sum_{k=1}^{M} \left[\hat{x}_k^E(t) - \hat{x}_k^I(t)\right]^2, \qquad \kappa^I(t) = \sum_{i=1}^{N^I} [r_i^I(t)]^2. \end{aligned} \tag{11}$$

We use a quadratic metabolic cost because it promotes the distribution of spiking across neurons[28]. In particular, the loss function of I neurons, $\mathrm{L}^I(t)$ implies the relevance of the approximation: $\hat{x}_k^E(t) \approx \hat{x}_k^I(t)$ (see ϵ^I in the Eq. 11), which will be used in what follows.

When shall a neuron spike?

We minimize the loss function by positing that neuron i of type $y \in \{E, I\}$ emits a spike as soon as its spike decreases the loss function of its population y in the immediate future[37]. We also define t^- and t^+ as the left- and right-sided limits of a spike time $t = t_i^{y,\alpha}$, respectively. Thus, at the spike time, the following jump condition must hold:

$$\mathrm{L}^y\left(t^+\right) \leq \mathrm{L}^y\left(t^-\right) + \xi_i^y(t^-), \qquad y \in \{E, I\}, \tag{12}$$

with ξ_i^y in units of mV. Here, the arguments t^- and t^+ denote the left- and right-sided limits of the respected functions at time t. Furthermore, we added a noise term on the right-hand side of the Eq. (12) in order to consider the stochastic nature of spike generation in biological networks[50]. A convenient choice for the noise $\xi_i^y(t)$ is the Ornstein-Uhlenbeck process obeying

$$\dot{\xi}_i^y(t) = -\lambda \xi_i^y(t) + \sqrt{2\lambda}\sigma_\xi^y \eta_i^y(t), \tag{13}$$

where η_i^y is a Gaussian white noise with auto-covariance function $\langle \eta_i(t)\eta_j(t')\rangle = \delta_{ij}\delta(t - t')$. The process $\xi_i^y(t)$ has zero mean and auto-covariance function $\langle \xi_i(t)\xi_j(t')\rangle = (\sigma_\xi^y)^2\delta_{ij}e^{-\lambda|t-t'|}$, with $(\sigma_\xi^y)^2$ the variance of the noise.

By applying the condition for spiking in Eq. (12) using $y = E$ and $y = I$, respectively, we get

$$\begin{aligned} &\left[\sum_{k=1}^{M}\left(x_k(t^+) - \hat{x}_k^E(t^+)\right)^2 + \beta^E \sum_{j=1}^{N^E}(r_j^E(t^+))^2\right] - \left[\sum_{k=1}^{M}\left(x_k(t^-) - \hat{x}_k^E(t^-)\right)^2 + \beta^E \sum_{j=1}^{N^E}(r_j^E(t^-))^2\right] \leq \xi_i^E(t^-), \\ &\left[\sum_{k=1}^{M}\left(\hat{x}_k^E(t^+) - \hat{x}_k^I(t^+)\right)^2 + \beta^I \sum_{j=1}^{N^I}(r_j^I(t^+))^2\right] - \left[\sum_{k=1}^{M}\left(\hat{x}_k^E(t^-) - \hat{x}_k^I(t^-)\right)^2 + \beta^I \sum_{j=1}^{N^I}(r_j^I(t^-))^2\right] \leq \xi_i^I(t^-). \end{aligned} \tag{14}$$

According to the definitions in Eqs. (7) and (9), if neuron *i* fires a spike at time $t = t_i^{y,\alpha}$, it causes a jump of its own filtered spike train (but not of other neurons $j \neq i$), as well as of the population readout of the population it belongs to. Therefore, when neuron *i* fires a spike, we have for a given neuron *j* and a given population readout *k*:

$$r_j^y(t^+) = r_j^y(t^-) + \delta_{ij}, \tag{15a}$$

$$\hat{x}_k^y(t^+) = \hat{x}_k^y(t^-) + w_{ki}^y. \tag{15b}$$

By inserting Eq. (15a)-(15b) in Eq. (12), we find that neuron *i* of type *y* should fire a spike if the following condition holds:

$$\begin{aligned}
\sum_{k=1}^{M} \{w_{ki}^E \left(x_k(t) - \hat{x}_k^E(t)\right)\} - \beta^E r_i^E(t) &\geq \frac{1}{2}\left(\sum_{k=1}^{M}(w_{ki}^E)^2 + \beta^E - \xi_i^E(t)\right), \\
\sum_{k=1}^{M} \{w_{ki}^I \left(\hat{x}_k^E(t) - \hat{x}_k^I(t)\right)\} - \beta^I r_i^I(t) &\geq \frac{1}{2}\left(\sum_{k=1}^{M}(w_{ki}^I)^2 + \beta^I - \xi_i^I(t)\right).
\end{aligned} \tag{16a}$$

These equations tell us when the neuron *i* of type *E* and *I*, respectively, emits a spike, and are similar to the ones derived in previous works[37,28]. In addition to what has been found in these previous works, we here also find that each term on the left- and right-hand side in the Eq 16a has the physical units of millivolts.

We note that the expression derived from the minimization of the loss function of E neurons in the top row of Eq. (16a) is independent of the activity of I neurons, and would thus lead to the E population being unconnected with the I population. In order to derive a recurrently connected E-I network, the activity of E neurons must depend on the activity of I neurons. We impose this property by using the approximation of estimates that holds under the assumption of efficient coding in I neurons (see ϵ^I in the Eq. 11), $\hat{x}_k^I(t) \approx \hat{x}_k^E(t)\ \forall k = 1 \ldots, M$. This yields the following conditions:

$$\begin{aligned}
\sum_{k=1}^{M} \{w_{ki}^E \left(x_k(t) - \hat{x}_k^I(t)\right)\} - \beta^E r_i^E(t) &\geq \frac{1}{2}\left(\sum_{k=1}^{M}(w_{ki}^E)^2 + \beta^E - \xi_i^E(t)\right), \\
\sum_{k=1}^{M} \{w_{ki}^I \left(\hat{x}_k^E(t) - \hat{x}_k^I(t)\right)\} - \beta^I r_i^I(t) &\geq \frac{1}{2}\left(\sum_{k=1}^{M}(w_{ki}^I)^2 + \beta^I - \xi_i^I(t)\right).
\end{aligned} \tag{16b}$$

We now define new variables $u_i^y(t)$ and θ_i^y as proportional to the left- and the right-hand side of these expressions,

$$\begin{aligned} u_i^E(t) &:= \sum_{k=1}^{M} \{w_{ki}^E \left(x_k(t) - \hat{x}_k^I(t)\right)\} - \beta^E r_i^E(t), \\ \theta_i^E &:= \frac{1}{2}\left(\sum_{k=1}^{M} (w_{ki}^E)^2 + \beta^E - \xi_i^E(t)\right), \\ u_i^I(t) &:= \sum_{k=1}^{M} \{w_{ki}^I \left(\hat{x}_k^E(t) - \hat{x}_k^I(t)\right)\} - \beta^I r_i^I(t), \\ \theta_i^I &:= \frac{1}{2}\left(\sum_{k=1}^{M} (w_{ki}^I)^2 + \beta^I - \xi_i^I(t)\right). \end{aligned} \tag{17}$$

The variables $u_i^y(t)$ and θ_i^y are interpreted as the membrane potential and the firing threshold of neuron i of cell type $y \in \{E, I\}$.

Dynamic equations for the membrane potentials

In this section we develop the exact dynamic equations of the membrane potentials $\dot{u}_i^y(t)$ for $y \in \{E, I\}$ according to the efficient coding assumption. It is practical to use the vector notation and rewrite variables in Eq. (17) as

$$\begin{aligned} u_i^E(t) &= \left(\boldsymbol{w}_i^E\right)^\top \left(\boldsymbol{x}(t) - \hat{\boldsymbol{x}}_I(t)\right) - \beta^E r_i^E(t), \\ u_i^I(t) &= \left(\boldsymbol{w}_i^I\right)^\top \left(\hat{\boldsymbol{x}}_E(t) - \hat{\boldsymbol{x}}_I(t)\right)\} - \beta^I r_i^I(t), \\ \theta_i^y &= \frac{1}{2}\left(\|\boldsymbol{w}_i^y\|_2^2 + \beta^y\right) - \frac{1}{2}\xi_i^y(t), \end{aligned} \tag{18}$$

with $\|\boldsymbol{w}_i^y\|_2^2 := \sum_{k=1}^{M} (w_{ki}^y)^2$ the squared length of the tuning vector of neuron i of type y. We also rewrite Eq. (8)-(9) in vector notation as

$$\begin{aligned} \dot{\boldsymbol{x}}(t) &= -\lambda \boldsymbol{x}(t) + \boldsymbol{s}(t), \\ \dot{\hat{\boldsymbol{x}}}_E(t) &= -\lambda \hat{\boldsymbol{x}}_E(t) + W_E \boldsymbol{f}_E(t), \\ \dot{\hat{\boldsymbol{x}}}_I(t) &= -\lambda \hat{\boldsymbol{x}}_I(t) + W_I \boldsymbol{f}_I(t), \end{aligned} \tag{19}$$

with $\boldsymbol{x}(t) := [x_1(t), ..., x_M(t)]^\top$ the vector of M target signals, $\hat{\boldsymbol{x}}_y(t) := [\hat{x}_1^y(t), \ldots, \hat{x}_M^y(t)]^\top$ the vector of estimates of cell type y, and $\boldsymbol{f}_y(t) := [f_1^y(t), \ldots, f_{N^y}^y(t)]^\top$ the vector of spike trains for N^y cell type $y \in \{E, I\}$.

In the case of E neurons, the time-derivative of the membrane potential $\dot{u}_i^E(t)$ in Eq. (18), is obtained as

$$\dot{u}_i^E(t) = \left(\boldsymbol{w}_i^E\right)^\top \left(\dot{\boldsymbol{x}}(t) - \dot{\hat{\boldsymbol{x}}}_I(t)\right) - \beta^E \dot{r}_i^E(t). \tag{20}$$

By inserting the dynamic equations of the target signal $\dot{\boldsymbol{x}}(t)$, its estimate $\dot{\hat{\boldsymbol{x}}}_I(t)$ (Eq. 19) and of the single neuron readout $\dot{r}_i^E(t)$ (Eq. 7 in the case $y = E$), we get

$$\begin{aligned}\dot{u}_i^E(t) &= \left(\boldsymbol{w}_i^E\right)^\top \left[-\lambda \boldsymbol{x}(t) + \boldsymbol{s}(t) + \lambda \hat{\boldsymbol{x}}_I(t) - W_I \boldsymbol{f}_I(t)\right] - \beta^E \left[-\lambda_r^E r_i^E(t) + f_i^E(t)\right], \\ &= -\lambda \left[\left(\boldsymbol{w}_i^E\right)^\top \left(\boldsymbol{x}(t) - \hat{\boldsymbol{x}}_I(t)\right) - \beta^E r_i^E(t)\right] + \left(\boldsymbol{w}_i^E\right)^\top \boldsymbol{s}(t) - \left(\boldsymbol{w}_i^E\right)^\top W_I \boldsymbol{f}_I(t) \\ &\quad - \beta^E(\lambda - \lambda_r^E) r_i^E(t) - \beta^E f_i^E(t), \\ &= -\lambda u_i^E(t) + \left(\boldsymbol{w}_i^E\right)^\top \boldsymbol{s}(t) - \left(\boldsymbol{w}_i^E\right)^\top W_I \boldsymbol{f}_I(t) - \beta^E(\lambda - \lambda_r^E) r_i^E(t) - \beta^E f_i^E(t), \end{aligned} \tag{21}$$

where in the last line we used the definition of $u_i^E(t)$ from the Eq. (18).

In the case of I neurons, the time derivative of the membrane potential $\dot{u}_i^I(t)$ in Eq. (18) is

$$\dot{u}_i^I(t) = \left(\boldsymbol{w}_i^I\right)^\top \left(\dot{\hat{\boldsymbol{x}}}_E(t) - \dot{\hat{\boldsymbol{x}}}_I(t)\right) - \beta^I \dot{r}_i^I(t). \tag{22}$$

By inserting the dynamic equations of the population readouts of E neurons $\dot{\hat{\boldsymbol{x}}}_E(t)$ and of the I neurons $\dot{\hat{\boldsymbol{x}}}_I(t)$ (Eq. 19) and of the single neuron readout $\dot{r}_i^I(t)$ (Eq. 7 in the case $y = I$), we get

$$\begin{aligned}\dot{u}_i^I(t) &= \left(\boldsymbol{w}_i^I\right)^\top \left[-\lambda \hat{\boldsymbol{x}}_E(t) + W_E \boldsymbol{f}_E(t) + \lambda \hat{\boldsymbol{x}}_I(t) - W_I \boldsymbol{f}_I(t)\right] - \beta^I \left[-\lambda_r^I r_i^I(t) + f_i^I(t)\right], \\ &= -\lambda \left[\left(\boldsymbol{w}_i^I\right)^\top \left(\hat{\boldsymbol{x}}_E(t) - \hat{\boldsymbol{x}}_I(t)\right) - \beta^I r_i^I(t)\right] + \left(\boldsymbol{w}_i^I\right)^\top W_E \boldsymbol{f}_E(t) - \left(\boldsymbol{w}_i^I\right)^\top W_I \boldsymbol{f}_I(t) \\ &\quad - \beta^I(\lambda - \lambda_r^I) r_i^I(t) - \beta^I f_i^E I(t), \\ &= -\lambda u_i^I(t) + \left(\boldsymbol{w}_i^I\right)^\top W_E \boldsymbol{f}_E(t) - \left(\boldsymbol{w}_i^E\right)^\top W_I \boldsymbol{f}_I(t) - \beta^I(\lambda - \lambda_r^I) r_i^I(t) - \beta^I f_i^I(t). \end{aligned} \tag{23}$$

where in the last line we used the definition of $u_i^I(t)$ from Eq. (18).

Leaky integrate-and-fire neurons

The terms on the right-hand-side in Eqs. (21) and (23) can be interpreted as transmembrane currents. The last term in these equations, $-\beta^y f_i^y(t), y \in \{E, I\}$, can be interpreted as a current instantaneously resetting the membrane potential upon reaching the firing threshold[28]. Indeed, when the membrane potential reaches the threshold, it triggers a spike and causes a jump of the membrane potential by an amount $-\beta^y$; this realizes resetting of the membrane potential which is equivalent to the resetting rule of integrate-and-fire neurons[54,56]. Thus, by taking

into account the resetting mechanism and defining the time constants of population and single neuron readout $\tau := \lambda^{-1}$ and $\tau_r^y := (\lambda_r^y)^{-1}$, we can rewrite Eqs. (21) and (23) as a leaky integrate-and-fire neuron model,

$$\begin{aligned}
\dot{u}_i^E(t) &= -\frac{1}{\tau}u_i^E(t) + \left(\boldsymbol{w}_i^E\right)^\top \boldsymbol{s}(t) - \sum_{j=1}^{N^I}\left(\boldsymbol{w}_i^E\right)^\top \boldsymbol{w}_j^I f_j^I(t) - \beta^E\left(\frac{1}{\tau} - \frac{1}{\tau_r^E}\right) r_i^E(t), \\
\dot{u}_i^I(t) &= -\frac{1}{\tau}u_i^I(t) + \sum_{j=1}^{N^E}\left(\boldsymbol{w}_i^I\right)^\top \boldsymbol{w}_j^E f_j^E(t) - \sum_{\substack{j=1\\ i\neq j}}^{N^I}\left(\boldsymbol{w}_i^I\right)^\top \boldsymbol{w}_j^I f_j^I(t) - \beta^I\left(\frac{1}{\tau} - \frac{1}{\tau_r^I}\right) r_i^I(t), \\
&\text{if } u_i^y(t^-) \geq \theta_i^y \rightarrow u_i^y(t^+) = u_i^{y,\text{reset}}, \\
&\theta_i^y = \frac{1}{2}\left(\|\boldsymbol{w}_i^y\|_2^2 + \beta^y - \xi_i^y(t)\right), \\
&u_i^{E,\text{reset}} = \theta_i^E - \beta^E, \\
&u_i^{I,\text{reset}} = \theta_i^I - \beta^I - \|\boldsymbol{w}_i^I\|_2^2.
\end{aligned} \tag{24}$$

In the Eq. 24 we wrote explicitly the terms $(\boldsymbol{w}_i^y)^\top W_x \boldsymbol{f}_x(t) = \sum_{j=1}^{N^x}(\boldsymbol{w}_i^y)^\top \boldsymbol{w}_j^x f_j^x(t)$, which correspond to the synaptic projections of N^X presynaptic neurons of type *x* to the postsynaptic neuron *i* of type *y*, with the quantity $(\boldsymbol{w}_i^y)^\top \boldsymbol{w}_j^x$ denoting the synaptic weight. We note that, in the case of I neurons, the element with *j* = *i* describes an autapse, i.e., a projection of a neuron with itself; this term is equal to $-\left(\boldsymbol{w}_i^I\right)^\top \boldsymbol{w}_i^I f_i^I(t) = -\|\boldsymbol{w}_i^I\|_2^2 f_i^I(t)$, and thus contributes to the resetting of the neuron *i*.

Imposing Dale's principle on synaptic connectivity

We now examine the synaptic terms in Eq. (24). As a first remark, we see that synaptic weights depend on tuning parameters w_{ki}^y. For the sake of generality we drew tuning parameters w_{ki}^y from a normal distribution with vanishing mean, which yielded both positive and negative values of w_{ki}^y. This has the desirable consequence that a spike of a neuron with a positive tuning parameter in signal dimension *k*, $w_{ki}^y > 0$ pulls the estimate, $\hat{x}_k^y(t)$, up, while a spike of a neuron with $w_{kj}^y < 0$ pulls the estimate down, allowing population readouts to track both positive and negative fluctuations of the target signal on a fast time scale.

Another consequence of synaptic connectivity in the Eq. (24) is that the synaptic weight between a presynaptic neuron *j* of type *x* and a postsynaptic neuron *i* of type *y* is symmetric and depends on the similarity of tuning vectors of the presynaptic and the postsynaptic neuron: $(\boldsymbol{w}_i^y)^\top \boldsymbol{w}_j^x = \sum_{k=1}^M w_{ki}^y w_{kj}^x$. The sign of this scalar product is positive between neurons with similar tuning and negative between neurons with different tuning (and zero when the two tuning vectors are orthogonal). Thus, for a presynaptic neuron *j* of type *x*, the synaptic weights of its outgoing connections can be both positive and negative, because some of its postsynaptic neurons have similar tuning to the neuron *j* while others have different tuning. This is inconsistent with Dale's principle[89], which postulates that a particular neuron can only have one type of effect on postsynaptic neurons (excitatory or inhibitory), but never both. To impose this constraint in our model, we set synaptic weights between neurons with different tuning $(\boldsymbol{w}_i^y)^\top \boldsymbol{w}_j^x < 0)$ to zero. To this end, we define the rectified connectivity matrices,

$$J_{ij}^{yx} = \left[\sum_{k=1}^M w_{ki}^y w_{kj}^x\right]_+, \tag{25}$$

with $x, y \in \{E, I\}$ and $[a]_+ \equiv \max(0, a)$ a rectified linear function. This manipulation is also plausible from a biological point of view, because in the cortex, the connection probability of neurons with very different (e.g. opposite) tuning is typically close to 0[51]. Since the elements of the matrix J^{yx} are all non-negative, it is the sign in front of the synaptic term in the Eq. (24) that determines the sign of the synaptic current between neurons i and j. The synaptic current is excitatory if the sign is positive, and inhibitory if the sign is negative.

It is also interesting to note that rectification affects the rank of connectivity matrices. Without rectification, the product in Eq. (25) yields a connectivity matrix with rank smaller or equal to the number of input features to the network, M, similarly as in previous works[29,43,44]. Since typically the number of input features is much smaller than the number of neurons, i.e., $M << N^y$, this would give a low-rank connectivity matrix. However, rectification in Eq. (25), necessary to ensure Dale's principle in presence of positive and negative tuning parameters, typically results in a substantial increase of the rank of the connectivity matrix.

Using the synaptic connectivity defined in Eq. (25), we rewrite the network dynamics from Eq. (24) as:

$$\begin{aligned}
\dot{u}_i^E(t) &= -\frac{1}{\tau}u_i^E(t) + \sum_{k=1}^{M} w_{ki}^E s_k(t) - \sum_{j=1}^{N^I} J_{ij}^{EI} f_j^I(t) - \beta^E(\frac{1}{\tau} - \frac{1}{\tau_r^E})r_i^E(t), \\
\dot{u}_i^I(t) &= -\frac{1}{\tau}u_i^I(t) + \sum_{j=1}^{N^E} J_{ij}^{IE} f_j^E(t) - \sum_{\substack{j=1 \\ i\neq j}}^{N^I} J_{ij}^{II} f_j^I(t) - \beta^I(\frac{1}{\tau} - \frac{1}{\tau_r^I})r_i^I(t), \\
&\text{if } u_i^y(t^-) \geq \theta_i^y \rightarrow u_i^y(t^+) = u_i^{y,\text{reset}}, \\
&\theta_i^y = \frac{1}{2}\left(\|\boldsymbol{w}_i^y\|_2^2 + \beta^y - \xi_i^y(t)\right), \\
&u_i^{E,\text{reset}} = \theta_i^E - \beta^E, \\
&u_i^{I,\text{reset}} = \theta_i^I - \beta^I - \|\boldsymbol{w}_i^I\|_2^2.
\end{aligned} \tag{26}$$

These equations express the neural dynamics which minimizes the loss functions (Eq. (10)) in terms of a generalized leaky integrate-and-fire model with E and I cell types, and are consistent with Dale's principle.

In principle, it is possible to use the same strategy as for the E-I network to enforce Dale's principle in model with one cell type (introduced by[28]). To do so, we constrained the recurrent connectivity of the model with a single cell type from[36] by keeping only connections between neurons with similar tuning vectors and setting other connections to 0 (see Supplementary text). This led to a network of only inhibitory neurons, a type of network model which is less relevant for the description of biological networks.

Model with resting potential and an external current

In the model given by the Eq. (26) the resting potential is equal to zero. In order to account for biophysical values of the resting potential and to introduce an implementation of the metabolic constant that is consistent with neurobiology, we add a constant value to the dynamical equations of the membrane potentials $\dot{u}_i^y$, the firing thresholds θ_i^y and the reset potentials $u_i^{y,\text{reset}}$. This does not change the spiking dynamics of the model, as what matters to correctly infer the efficient spiking times of neurons is the distance between the membrane potential and the threshold.

Furthermore, in the same equations, the role of the metabolic constant β^y as a biophysical quantity is questionable. The metabolic constant β^y is an important parameter that weights the metabolic cost over the encoding error in the objective functions (Eq. 10). On the level of

computational objectives, the metabolic constant naturally controls firing rates, as it allows the network to fire more or less spikes to correct for a certain encoding error. A flexible control of the firing rates is a desirable property, as gives the possibility to potentially capture different operating regimes of efficient spiking networks[36]. In the spiking model we developed thus far (Eq. 26), similarly to previous efficient spiking models[36,33], the metabolic constant β^y controls the firing threshold. In neurobiology, however, strong changes to the firing threshold that would reflect metabolic constraints of the network are not plausible. We thus searched for an implementation of the metabolic constant β^y that is consistent with neurobiology.

The condition for threshold crossing of the neuron i can be written by Eq. (26) as

$$u_i^y(t) + V_{\text{rest}}^y + \frac{1}{2}\left(c - \beta^y + \xi_i^y(t)\right) \geq \frac{1}{2}\left(\|\boldsymbol{w}_i^y\|_2^2 + c\right) + V_{\text{rest}}^y, \tag{27}$$

with c an arbitrary constant in units of millivolts. In Eq. (27) we added a constant $c/2$ and a resting potential V_{rest}^y on the left- and right-hand side of the firing rule. Moreover, we shifted the noise and the dependency on the parameter β from the firing threshold to the membrane potential. Thus, we assumed that the firing threshold is independent of the metabolic constant and the noise, and we instead assumed the dependence on the metabolic constant and noise in the membrane potentials.

We now define new variables for $y \in \{E, I\}$:

$$\begin{aligned} V_i^y(t) &:\equiv u_i^y(t) + V_{\text{rest}}^y + \frac{1}{2}\left(c - \beta^y + \xi_i^y(t)\right), \qquad V_{\text{rest}}^y < 0, \\ \vartheta_i^y &:\equiv V_{\text{rest}}^y + \frac{1}{2}\left(\|\boldsymbol{w}_i^y\|_2^2 + c\right), \end{aligned} \tag{28}$$

and rewrite the model in Eq. 26 in these new variables

$$\begin{aligned} \tau \dot{V}_i^E(t) &= -\left(V_i^E(t) - V_{\text{rest}}^E\right) + \tau \sum_{k=1}^{M} w_{ki}^E s_k(t) - \tau \sum_{j=1}^{N^I} J_{ij}^{EI} f_j^I(t) - \beta^E\left(1 - \frac{\tau}{\tau_r^E}\right) r_i^E(t) + \frac{\tau}{2}\left(c - \beta^E\right) + \sqrt{\frac{\tau}{2}}\sigma_\xi^E \eta_i^E(t), \\ \tau \dot{V}_i^I(t) &= -\left(V_i^I(t) - V_{\text{rest}}^I\right) + \tau \sum_{j=1}^{N^E} J_{ij}^{IE} f_j^E(t) - \tau \sum_{\substack{j=1 \\ i \neq j}}^{N^I} J_{ij}^{II} - \beta^I\left(1 - \frac{\tau}{\tau_r^I}\right) r_i^I(t) + \frac{\tau}{2}\left(c - \beta^I\right) + \sqrt{\frac{\tau}{2}}\sigma_\xi^I \eta_i^I(t), \\ &\text{if } V_i^y(t^-) \geq \vartheta_i^y \rightarrow V_i^y(t^+) = V_i^{y,\text{reset}}, \\ &\vartheta_i^y = V_{\text{rest}}^y + \frac{1}{2}\left(\|\boldsymbol{w}_i^y\|_2^2 + c\right), \\ &V_i^{E,\text{reset}} = V_{\text{rest}}^E - \beta^E + \frac{1}{2}\left(c + \|\boldsymbol{w}_i^E\|_2^2\right), \\ &V_i^{I,\text{reset}} = V_{\text{rest}}^I - \beta^I + \frac{1}{2}\left(c - \|\boldsymbol{w}_i^I\|_2^2\right), \end{aligned} \tag{29}$$

where $\eta_i^E(t)$ and $\eta_i^I(t)$ are the independent Gaussian white noise processes defined in Eq. (13) above. We note that all terms on the right-hand side of Eq. (29) have the desired units of mV. The model in Eq. (29) is an efficient E-I spiking network with improved compatibility with neurobiology. We have expressed two new terms in the membrane potentials of E and I neurons, one dependent on the metabolic constant β^y and one on the noise that we assumed in the condition for spiking (see Eq. 12). We will group these two terms to define an external current, a current that is well known in spiking models of neural dynamics[40].

Efficient generalized leaky integrate-and-fire neuron model

Finally, we rewrite the model from Eq. (29) in a compact form in terms of transmembrane currents, and discuss their biological interpretation. The efficient coding with spikes is realized by the following model for the neuron i of type $y \in \{E, I\}$:

$$\begin{aligned}
\tau \dot{V}_i^E(t) &= -\left(V_i^E(t) - V_{\text{rest}}^E\right) + R_m\left(I_i^{\text{syn},E}(t) - I_i^{\text{ad},E}(t) + I_i^{\text{ext},E}(t)\right), \\
\tau \dot{V}_i^I(t) &= -\left(V_i^I(t) - V_{\text{rest}}^I\right) + R_m\left(I_i^{\text{syn},I}(t) - I_i^{\text{ad},I}(t) + I_i^{\text{ext},I}(t)\right), \\
&\text{if } V_i^y(t^-) \geq \vartheta_i^y \rightarrow V_i^y(t^+) = V_i^{y,\text{reset}}, \\
\vartheta_i^y &= V_{\text{rest}}^y + \frac{1}{2}\left(\|\boldsymbol{w}_i^y\|_2^2 + c\right), \\
V_i^{E,\text{reset}} &= V_{\text{rest}}^E - \beta^E + \frac{1}{2}\left(c + \|\boldsymbol{w}_i^E\|_2^2\right), \\
V_i^{I,\text{reset}} &= V_{\text{rest}}^I - \beta^I + \frac{1}{2}\left(c - \|\boldsymbol{w}_i^I\|_2^2\right),
\end{aligned} \tag{30a}$$

with R_m the current resistance. The leak current,

$$I_i^{\text{leak},y}(t) = -\frac{C_m}{\tau}\left(V_i^y(t) - V_{\text{rest}}^y\right), \qquad y \in \{E, I\}, \tag{30b}$$

with $\tau = R_m C_m$ and C_m the capacitance of the neural membrane[54], arose by assuming the same time constant for the target signals x_k and estimates $\hat{x}_k^E$ and $\hat{x}_k^I$ (see Eq. 19). We see that the passive membrane time constant $\tau = \lambda^{-1}$ can be traced back to the time constant of the population read-out in Eq. (9). The synaptic currents are defined as

$$\begin{aligned}
I_i^{\text{syn},E}(t) &= C_m\left(\sum_{k=1}^{M} w_{ki}^E s_k(t) - \sum_{j=1}^{N^I} J_{ij}^{EI} f_j^I(t)\right), \\
I_i^{\text{syn},I}(t) &= C_m\left(\sum_{j=1}^{N^E} J_{ij}^{IE} f_j^E(t) - \sum_{\substack{j=1 \\ i \neq j}}^{N^I} J_{ij}^{II} f_j^I(t)\right),
\end{aligned} \tag{30c}$$

where we note the presence of a feedforward current to E neurons,

$$\begin{aligned}
I_i^{\text{ff}}(t) &= C_m \sum_{k=1}^{M} w_{ki}^E s_k(t), \\
&= C_m \left(\boldsymbol{w}_i^E\right)^\top \boldsymbol{s}(t),
\end{aligned} \tag{30d}$$

which consist in a linear combination of the stimulus features $\boldsymbol{s}(t)$ weighted by the readout weights $\boldsymbol{w}_i^E$. The stimulus features can be traced back to the definition of the target signals in Eq. (8). This current emerges in E neurons, as a consequence of having the target signals $x_k(t)$ in the loss function of the E population (see Eqs. 10-11). I neurons do not receive the feedforward current because their loss function does not contain the target signal.

The current providing within-neuron feedback triggered by each spike,

$$I_i^{\text{ad},y}(t) = C_m \beta^y \left(\frac{1}{\tau} - \frac{1}{\tau_r^y}\right) r_i^y(t), \tag{30e}$$

was recently recovered[37]. This current has the kinetics of the single neuron readout $r_i^y(t)$ (i.e., low-pass filtered spike train). Its sign depends on the relation between the time constant of the population readout $\tau = \lambda^{-1}$ and single neuron readout $\tau_r^y = (\lambda_r^y)^{-1}$, because the metabolic constant β^y is non-negative by definition (Eq. 10). If the single neuron readout is slower than the population readout, $\tau_r^y > \tau$, within-neuron feedback is negative, and can thus be interpreted as spike-triggered *adaptation*. On the contrary, if the single neuron readout is faster than the population readout, $\tau_r^E < \tau$, the within-neuron feedback is positive and can thus be interpreted as spike-triggered *facilitation*. In a special case where the time constant of the single neuron and population readout are assumed to be equal, within-neuron feedback vanishes.

Finally, we here derived the non-specific external current:

$$I_i^{\mathrm{ext},y}(t) = C_m\left(\frac{c-\beta^y}{2} + \sigma^y \eta_i^y(t)\right), \qquad \sigma^y = \frac{\sigma_\xi^y}{\sqrt{2\tau}} \tag{30f}$$

that captures the ensemble of non-specific synaptic currents received by each single neuron. The non-specific current has a homogeneous mean across all neurons of the same cell type, and a neuron-specific fluctuation. The mean of the non-specific current can be traced back to the weighting of the metabolic cost over the encoding error in model objectives (Eq. 10), while the fluctuation can be traced back to the noise intensity that we assumed in the condition for spiking (Eq. 12). The non-specific external current might arise because of synaptic inputs from other brain areas than the brain area that delivers feedforward projections to the E-I network we consider here, or it might result from synaptic activity of neurons that are part of the local network, but are not tuned to the feedforward input[69].

We also recall the fast and slower time scales of single neuron activity:

$$\begin{aligned} f_i^y(t) &= \sum_\alpha \delta(t - t_i^{y,\alpha}), \\ \frac{dr_i^y(t)}{dt} &= -\frac{1}{\tau_r^y} r_i^y(t) + f_i^y(t), \end{aligned} \tag{30g}$$

and the connectivity matrices

$$J_{ij}^{IE} = \left[(\boldsymbol{w}_i^I)^\top \boldsymbol{w}_j^E\right]_+, \qquad J_{ij}^{II} = \left[(\boldsymbol{w}_i^I)^\top \boldsymbol{w}_j^I\right]_+, i \neq j, \qquad J_{ij}^{EI} = \left[(\boldsymbol{w}_i^E)^\top \boldsymbol{w}_j^I\right]_+. \tag{30h}$$

The structure of synaptic connectivity is fully determined by the similarity of tuning vectors of the presynaptic and the postsynaptic neurons ($\boldsymbol{w}_j^x$ and $\boldsymbol{w}_i^y$, respectively), while the distribution of synaptic connectivity weights is fully determined by the distribution of tuning parameters w_{ki}^y.

Stimulus features

We define stimulus features as a set of $k = 1, \ldots, M$ independent Ornstein-Uhlenbeck processes with vanishing mean, standard deviation σ_s and the correlation time τ_s,

$$\tau_s \frac{ds_k(t)}{dt} = -s_k(t) + \sqrt{2\tau_s}\sigma_s \eta_k(t). \tag{31}$$

If not mentioned otherwise, we use the following parameters: $\sigma_s = 2\ (\mathrm{mV})^{1/2}$ and $\tau_s = 10$ ms. Variables $\eta_k(t)$ are independent Gaussian white noise processes with zero mean and covariance function $\langle \eta_k(t)\eta_l(t')\rangle = \delta_{kl}\delta(t - t')$. These variables should not be confused with the Gaussian white noises $\eta_k^y(t)$ in Eq. (29).

Parametrization of synaptic connectivity

In the efficient E-I model, synaptic weights J_{ij}^{yx} are parametrized by tuning parameters w_{ki}^{y} through Eq. (25). The total number of synapses in the E-I, I-I and I-E connectivity matrices (including silent synapses with zero synaptic weight) is $n_{\text{syn}} = 2N^E N^I + (N^I)^2$, while the number of tuning parameters is $n_w = M(N^E + N^I)$. Because the number of stimulus features M is expected to be much smaller than the number of E or I neurons, the number of tuning parameters n_w is much smaller than the number of synapses n_{syn}.

We can achieve a further substantial decrease in the number of free parameters by using a parametric distribution of tuning parameters w_{ki}^{y}. We have set the tuning parameters following a normal distribution and found that excellent performance can be achieved with random draws of tuning parameters from the normal distribution, thus without searching for a specific set of tuning parameters. This drastically decreased the number of free parameters relative to synaptic weights to only a handful of parameters that determine the distributions of tuning parameters.

Given M features, we sample tuning parameters, $\boldsymbol{w}_i^y = [w_{1,i}, \ldots, w_{M,i}]$, with $i = 1, \ldots, N^y$, $y \in \{E, I\}$, as random points uniformly distributed on a M-dimensional sphere of radius σ_w^y. We obtain this by sampling, for each neuron, a vector of M i.i.d. standard Gaussian random variables, $\boldsymbol{\xi}_i^y = [\xi_{1i}^y, \ldots, \xi_{Mi}^y]^\top$, with $\xi_{ki}^y \sim \mathcal{N}(0,1)$, and normalizing the vector such as to have length equal to σ_w^y [90],

$$\boldsymbol{w}_i^y = \sigma_w^y \frac{\boldsymbol{\xi}_i^y}{\|\boldsymbol{\xi}_i^y\|_2}, \qquad y \in \{E, I\}. \tag{32}$$

This ensures that the length of tuning vectors $\boldsymbol{w}_i^y$ in Eq. (32) is homogeneous across neurons of the same cell type, i.e., $\|w_i^y\|_2 = \sigma_w^y$. Parameters σ_w^E and σ_w^I determine the heterogeneity (spread) of tuning parameters.

By combining Eq. (25) and Eq. (32), we obtain the synaptic weights, J_{ij}^{yx}, as a function of the angle, α_{ij}^{xy}, between the tuning vectors of presynaptic neurons, $\boldsymbol{w}_i^x$, and postsynaptic neurons, $\boldsymbol{w}_j^y$,

$$J_{ij}^{yx} = \sigma_w^y \sigma_w^x \left[\cos \alpha_{ij}^{yx}\right]_+ . \tag{33}$$

In the $M = 3$ dimensional case, we have that the distribution of the angle between two vectors is $p(\alpha_{ij}^{yx}) = \frac{1}{2}\sin(\alpha_{ij}^{yx})$, with $\alpha_{ij}^{yx} \in [0, \pi]$. Thus, the average strength of synaptic weights between the pre- and the postsynaptic population can be calculated as

$$\begin{aligned} \langle J_{ij}^{yx} \rangle &= \frac{1}{2}\sigma_w^y \sigma_w^x \int_0^\pi d\alpha_{ij}^{yx} \sin(\alpha_{ij}^{yx}) \left[\cos(\alpha_{ij}^{yx})\right]_+ \\ &= \frac{1}{4}\sigma_w^y \sigma_w^x. \end{aligned} \tag{34}$$

Thus, the upper bound for the synaptic weight between cell types x and y is simply

$$max\left(J_{ij}^{yx}\right) = \sigma_w^y \sigma_w^x. \tag{35}$$

From the Eq. (34), we have that the mean E-I connectivity is equal to the mean I-E connectivity $\langle J_{ij}^{EI} \rangle = \langle J_{ij}^{IE} \rangle$. As we consider the ratio of the mean connectivity between E-I and I-I neurons, we find that it is given by the following:

$$\frac{\langle J_{ij}^{II} \rangle}{\langle J_{ij}^{EI} \rangle} = \frac{\left(\sigma_w^I\right)^2}{\sigma_w^I \sigma_w^E} = \frac{\sigma_w^I}{\sigma_w^E}. \tag{36}$$

Performance measures

Average encoding error and average metabolic cost

The definition of the time-dependent loss functions (Eq. 10) induces a natural choice for the performance measure: the mean squared error (MSE) between the targets and their estimators for each cell type. In the case of the E population, the time-dependent encoding error is captured by the variable $\epsilon^E(t)$ in the Eq. (11) and in case of I population it is captured by $\epsilon^I(t)$ defined in the same equation. We used the root MSE (RMSE), a standard measure for the performance of an estimator[40]. For the cell type $y \in \{E, I\}$ in trial q, the RMSE is measured as

$$\mathrm{RMSE}^y = \sqrt{\langle \epsilon_q^y(t) \rangle_{t,q}}, \tag{37}$$

with $\langle z_q(t) \rangle_{t,q}$ denoting the time- and trial-average.

Following the definition of the time-dependent metabolic cost in the loss functions (Eq. 10), we measured the average metabolic cost in a trial q for the cell type $y \in \{E, I\}$ as

$$\mathrm{MC}^y = \sqrt{\langle \kappa_q^y(t) \rangle_{t,q}}, \tag{38}$$

with time-dependent metabolic cost $\kappa^y(t)$ as in model's objectives (Eq. 11) and $\langle z_q(t) \rangle_{t,q}$ the time- and trial-average. The square root was taken to have the same scale as for the RMSE (see Eq. 37).

The bias of the estimator

The MSE can be decomposed into the bias and the variance of the estimator. The time-dependent bias of estimates $\hat{x}_k^y(t), y \in \{E, I\}$, were evaluated for each time point over $q = 1, \ldots, Q$ trials. The time-dependent bias in input dimension $k = 1, \ldots, M$ is defined as

$$\begin{aligned} B_k^E(t) &= \frac{1}{Q} \sum_{q=1}^{Q} [\hat{x}_{k,q}^E(t) - x_k(t)], \\ B_k^I(t) &= \frac{1}{Q} \sum_{q=1}^{Q} [\hat{x}_{k,q}^I(t) - \langle \hat{x}_{k,q}^E(t) \rangle_q], \end{aligned} \tag{39a}$$

with $\langle z_q(t) \rangle_{t,q}$ the trial-averaged realization at time t. To have an average measure of the encoding bias, we averaged the bias of estimators over time and over input dimensions:

$$B^y = \frac{1}{TM} \sum_{k=1}^{M} \int_0^T B_k^y(t) dt. \tag{39b}$$

The averaging over time and input dimensions is justified because $s_k(t)$ are independent realizations of the Ornstein-Uhlenbeck process (see Eq.31) with vanishing mean and with the same time constant, and variance across input dimensions.

Criterion for determining optimal model parameters

The equations of the E-I spiking network in Eqs. 30a-30h (Methods), derived from the instantaneous loss functions, give efficient coding solutions valid for any set of parameter values. However, to choose parameters values in simulated data in a principled way, we performed a numerical optimization of the performance function detailed below. Numerical optimization gave the set of optimal parameters listed in **Table 1**. When testing the efficient E-I model with simulations, we used the optimal parameters in **Table 1** and changed only the parameters plotted in the figure axes on a figure-by-figure basis.

To estimate the optimal set of parameters $\theta = \theta^*$, we performed a grid search on each parameter θ_i while keeping all other parameters fixed as specified in **Table 1**. While varying the parameters, we measured a weighted sum of the time- and trial-averaged encoding error and metabolic cost. For each cell type $y \in \{E, I\}$, we computed

$$\mathcal{L}_\theta^y = g_L \sqrt{\langle \epsilon_q^y(t \mid \theta) \rangle_{t,q}} + (1 - g_L) \sqrt{\langle \kappa_q^y(t \mid \theta) \rangle_{t,q}}, \tag{40a}$$

with $\langle z_q(t) \rangle_{t,q}$ the average over time and over trials and with $e^y(t)$ and $\kappa^y(t)$ as in model's objectives (**Eq. 11**).

To optimize the performance measure, we used a value of $g_L = 0.7$. The parameter g_L in the **Eq. (40a)** regulates the relative importance of the average encoding error over the average metabolic cost. Since the performance measure in **Eq. (40a)** is closely related to the average over time and trials of the instantaneous loss function (**Eq. 10**) where the parameter β regulates the relative weight of instantaneous encoding error over the metabolic cost, setting g_L is effectively achieved by setting β.

The optimal parameter set $\theta = \theta^*$ reported in **Table 1** is the parameter set that minimizes the sum of losses across E and I cell type

$$\theta^* = \arg\min_\theta \left(\mathcal{L}_\theta^E + \mathcal{L}_\theta^I \right). \tag{40b}$$

For visualization of the behavior of the average metabolic cost (**Eq. 38**) and average loss (**Eq. 40a**) across a range of a specific parameter θ_i, we summed these measures across the E and I cell type and normalized them across the range of tested parameters.

The exact dynamic and performance of our model depends on the realizations of random variables which describe the the tuning parameters w_{ki}^y, the Gaussian noise in the non-specific currents $\eta_i^y(t)$, and the initial conditions of the membrane potential $V_i^y(t = 0)$, that were randomly drawn from a normal distribution in each simulation trial. To capture the performance of a "typical" network, we iterated the performance measures across trials with different realizations of these random variables, and averaged the performance measures across trials. We used 100 simulation trials for each parameter search.

Functional activity measures

Tuning similarity

The pair-wise tuning similarity was measured as the cosine similarity[91], defined as:

$$\Phi_{ij}^{yx} = \cos\alpha(\boldsymbol{w}_i^y, \boldsymbol{w}_j^x) = \frac{(\boldsymbol{w}_i^y)^\top \boldsymbol{w}_j^x}{||\boldsymbol{w}_i^y||_2 ||\boldsymbol{w}_j^x||_2}, \qquad y \in \{E, I\}, \tag{41}$$

with $||\boldsymbol{w}_i^y||_2 = \sqrt{\sum_{k=1}^{M}(w_{ki}^y)^2}$ the length of the tuning vector in Euclidean space and α the angle between the tuning vectors $\boldsymbol{w}_j^x$ and $\boldsymbol{w}_i^y$.

Cross-correlograms of spike timing

The time-dependent coordination of spike timing was measured with the cross-correlogram (CCG) of spike trains, corrected for stimulus-driven coincident spiking. The raw cross-correlogram (CCG) for neuron *i* of cell type *y* and neuron *j* of cell type *x* was measured as follows:

$$C_{ij}^{yx}(\tau) = \frac{1}{Q}\sum_{q=1}^{Q}\int_0^T f_{i,q}^y(t) f_{j,q}^x(t+\tau)\, dt, \tag{42a}$$

with $q = 1, \ldots, Q$ simulation trials with identical stimulus and T the duration of the trial. We subtracted from the raw CCG the CCG of trial-invariant activity. To evaluate the trial-invariant cross-correlogram, we first computed the peri-stimulus time histogram (PSTH) for each neuron as follows:

$$P_i^y(t) = \frac{1}{Q}\sum_{q=1}^{Q} f_{i,q}^y(t). \tag{42b}$$

The trial-invariant CCG was then evaluated as the cross-correlation function of PSTHs between neurons *i* and *j*,

$$S_{ij}^{yx}(\tau) = \int_0^T P_i^y(t) P_j^x(t+\tau) dt. \tag{42c}$$

Finally, the temporal coordination of spike timing was computed by subtracting the correction term from the raw CCG:

$$c_{ij}^{yx}(\tau) = C_{ij}^{yx}(\tau) - S_{ij}^{yx}(\tau). \tag{42d}$$

Average imbalance of synaptic inputs

We considered time and trial-averaged synaptic inputs to each E and I neuron *i* in trial *q*, evaluated as:

$$\begin{aligned} \bar{A}_{i,q}^{\mathrm{net},E} &= \frac{1}{TC_m}\int_0^T I_{i,q}^{\mathrm{syn},E}(t) dt, \\ \bar{A}_{i,q}^{\mathrm{net},I} &= \frac{1}{TC_m}\int_0^T I_{i,q}^{\mathrm{syn},I}(t) dt, \end{aligned} \tag{43}$$

with synaptic currents to E neurons $I_{i,q}^{\mathrm{syn},E}(t)$ and to I neurons $I_{i,q}^{\mathrm{syn},I}(t)$ as in Eq. (30c). Synaptic inputs were measured in units of mV. We reported trial-averages of the net synaptic inputs from the Eq. (43).

Instantaneous balance of synaptic inputs

We measured the instantaneous balance of synaptic inputs as the Pearson correlation of time-dependent synaptic inputs incoming to the neuron *i*. For those synaptic inputs that are defined as weighted delta-spikes (for which the Pearson correlation is not well defined; see Eq. 30c), we convolved spikes with a synaptic filter $F(t) = \exp(-\frac{t}{\tau_{syn}})$,

$$\begin{aligned} A_{i,q}^{IE}(t) &= \sum_{j=1}^{N^E} J_{ij,q}^{IE} \int_0^t f_{j,q}^E(t-s)F(s)ds, \\ A_{i,q}^{II}(t) &= \sum_{\substack{j=1 \\ i\neq j}}^{N^I} J_{ij,q}^{II} \int_0^t f_{j,q}^I(t-s)F(s)ds, \\ A_{i,q}^{EI}(t) &= \sum_{j=1}^{N^I} J_{ij,q}^{EI} \int_0^t f_{j,q}^I(t-s)F(s)ds, \\ A_{i,q}^{\text{ff}}(t) &= C_m^{-1} I_{i,q}^{\text{ff}}(t), \end{aligned} \tag{44}$$

where we used the expression for the feedforward synaptic current from the Eq. (30d). Note that the feedforward synaptic current is already already low-pass filtered (see Eq. 31). Using synaptic inputs from the Eq. 44, we computed the Pearson correlation of synaptic inputs incoming to single E neurons, $\rho_{i,q}^E\left(A_{i,q}^{IE}(t), A_{i,q}^{II}(t)\right)$ for $i = 1, ..., N^E$, and to single I neurons, $\rho_{i,q}^I\left(A_{i,q}^{EI}(t), A_{i,q}^{\text{ff}}(t)\right)$ for $i = 1, ..., N^I$. The coefficients were then averaged across trials.

Perturbation of connectivity

To test the robustness of the model to random perturbations of synaptic weights, we applied a random jitter to optimally efficient recurrent synaptic connectivity weights. The random jitter was proportional to the synaptic weight, $\tilde{J}_{ij}^{yx} = J_{ij}^{yx}(1 + \sigma_J Z_{ij}^{yx})$, where σ_J is the strength of the perturbation and Z_{ij}^{yx} are independent standard normal random variables. All three recurrent connectivity matrices (E-I, I-I and I-E) were randomly perturbed at once.

Computer simulations

We run computer simulations with Matlab R2023b (Mathworks). The membrane equation for each neuron was integrated with Euler integration scheme with the time step of dt = 0.02 ms.

The simulation of the E-I network with 400 E units and 100 I units for an equivalent of 1 second of neural activity lasted approximately 1.65 seconds on a laptop.

Supplementary material

Supplementary text 1: Derivation of the one cell type model

An efficient spiking model network with one cell type (1CT) has been developed previously[28], and properties of the 1CT model where the computation is assumed to be the leaky integration of inputs has been addressed in a number of previous studies[29,43,36,33,42]. Compared to the efficient E-I model, the 1CT model can be seen as a simplification, and can be treated similarly to the E-I model, which is what we demonstrate in this section.

As the name of the model suggests, all neurons in the 1CT model are of the same cell type, and we have $i = 1, \ldots, N$ such neurons. We can then use the definitions in Eqs. (6) - (9) (now without the index y) and a loss function similar to the one in[36], but with only one (quadratic) regularizer

$$\mathrm{L}^{1\mathrm{CT}}(t) = \sum_{k=1}^{M} \left(x_k(t) - \hat{x}_k(t)\right)^2 + \beta_1 \sum_{i=1}^{N} \left[r_i^2(t)\right], \tag{S.1}$$

with $\beta_1 > 0$. The encoding error of the one cell type model minimizes the squared distance between the target signal $x_k(t)$ and the estimate $\hat{x}_k(t)$. As we apply the condition for spiking as for the E-I network (Eq. 12 without the index y) and follow the same steps as for the E-I network, we get

$$\sum_{k=1}^{M} \{w_{ki} \left(x_k(t) - \hat{x}_k(t)\right)\} - \beta_1 r_i(t) > \frac{1}{2}\left(\sum_{k=1}^{M} w_{ki}^2 + \beta_1 - \xi_i(t)\right), \tag{S.2}$$

with $\xi_i(t)$ the noise at the condition for spiking. Same as in the E-I model, we define the noise as an Ornstein-Uhlenbeck process with zero mean, obeying

$$\dot{\xi}_i(t) = -\lambda \xi_i(t) + \sqrt{2\lambda}\sigma_\eta \eta_i(t), \tag{S.3}$$

where η_i is a Gaussian white noise and $\lambda = \tau^{-1}$ is the inverse time constant of the process. We now define proxies of the membrane potential and the firing threshold as

$$\begin{aligned} u_i(t) &:= \sum_{k=1}^{M} \{w_{ki} \left(x_k(t) - \hat{x}_k(t)\right)\} - \beta_1 r_i(t), \\ \theta_i &:= \frac{1}{2}\left(\sum_{k=1}^{M} w_{ki}^2 + \beta_1 - \xi_i(t)\right). \end{aligned} \tag{S.4}$$

Differentiating the proxy of the membrane potential $u_i(t)$ and rewriting the model as an integrate-and-fire neuron, we get

$$\begin{aligned} &\dot{u}_i(t) = -\frac{1}{\tau} u_i(t) + \sum_{k=1}^{M} w_{ki} s_k(t) - \sum_{\substack{j=1 \\ j \neq i}}^{N} \boldsymbol{w}_i^\intercal \boldsymbol{w}_j f_j(t), \\ &\text{if } u_i(t^-) \geq \theta_i \rightarrow u_i(t^+) = u_i^{\text{reset}}, \\ &\theta_i = \frac{1}{2}\left(\|\boldsymbol{w}_i\|_2^2 + \beta_1 - \xi_i(t)\right), \\ &u_i{}^{\text{reset}} = \theta_i - \left(\|\boldsymbol{w}_i\|_2^2 + \beta_1\right). \end{aligned} \tag{S.5}$$

We now proceed in the same way as with the E-I model and define new variables

$$\begin{aligned} V_i(t) &:= u_i(t) + V_{\text{rest}} + \frac{1}{2}\left(c - \beta_1 + \xi_i(t)\right), \qquad V_{\text{rest}} < 0, \\ \vartheta_i &:= V_{\text{rest}} + \frac{1}{2}\left(\|\boldsymbol{w}_i\|_2^2 + c\right). \end{aligned} \tag{S.6}$$

In these new variables, we can rewrite the membrane equation of the 1CT model as follows:

$$\tau\dot{V}_i(t) = -(V_i(t) - V_{\text{rest}}) + \tau\sum_{k=1}^{M} w_{ki}s_k(t) - \tau\sum_{\substack{j=1\\j\neq i}}^{N} \boldsymbol{w}_i^\intercal \boldsymbol{w}_j f_j(t) + \frac{\tau}{2}(c-\beta_1) + \sqrt{\frac{\tau}{2}}\sigma_\xi\eta_i(t). \tag{S.7}$$

Finally, we rewrite the model with a more compact notation of a leaky integrate-and-fire neuron model with transmembrane currents,

$$\begin{aligned} \tau\dot{V}_i(t) &= -\left(V_i(t) - V_{\text{rest}}\right) + R_m\left(I_i^{\text{ff}}(t) + I_i^{\text{syn}}(t) + I_i^{\text{ext}}(t)\right), \\ &\text{if } V_i(t^-) \geq \vartheta_i \rightarrow V_i(t^+) = V_i^{\text{reset}}, \\ \vartheta_i &= V_{\text{rest}} + \frac{1}{2}\left(\|\boldsymbol{w}_i\|_2^2 + c\right), \\ V_i^{\text{reset}} &= V_{\text{rest}} - \beta_1 + \frac{1}{2}\left(c - \|\boldsymbol{w}_i\|_2^2\right), \end{aligned} \tag{S.8a}$$

with currents

$$\begin{aligned} I_i^{\text{ff}}(t) &= C_m\left(\sum_{k=1}^{M} w_{ki}s_k(t)\right), \\ I_i^{\text{syn}}(t) &= C_m\left(\sum_{\substack{j=1\\j\neq i}}^{N} J_{ij}f_j(t)\right), \qquad J_{ij} = -\boldsymbol{w}_i^\intercal\boldsymbol{w}_j, \\ I_i^{\text{ext}}(t) &= C_m\left(\frac{c-\beta_1}{2} + \sigma_1\eta_i(t)\right), \qquad \sigma_1 = \frac{\sigma_\xi}{\sqrt{2\tau}}. \end{aligned} \tag{S.8b}$$

Note that the model with one cell type does not obey Dale's law, since the same neuron sends to its postsynaptic targets excitatory and inhibitory currents, depending on the tuning similarity of the presynaptic and the postsynaptic neuron w_i and w_j (Eq. S.8b). In particular, if the pre- and postsynaptic neurons have similar selectivity $(\boldsymbol{w}_i^\intercal\boldsymbol{w}_j > 0)$, the recurrent interaction is inhibitory, and if the neurons have different selectivity $(\boldsymbol{w}_i^\intercal\boldsymbol{w}_j < 0)$, the interaction is excitatory. Simply put, neurons with similar selectivity inhibit each other while neurons with different selectivity excite each other[36].

Dale's law can be imposed to the 1CT model the same way as in the E-I model. To do so, we removed synaptic interactions between neurons with different selectivity by rectifying the connectivity matrix,

$$\tilde{J}_{ij} = -[\boldsymbol{w}_i^\intercal\boldsymbol{w}_j]_+. \tag{S.9}$$

However, this manipulation results in a network with only inhibitory recurrent synaptic interactions, and thus a network of only inhibitory neurons. Network with only inhibitory interactions is less relevant for the description of recurrently connected biological networks.

Supplementary text 2: Analysis of the one cell type model and comparison with the E-I model

We re-derived the 1CT model as a simplification of the E-I network (Supplementary Text 1, **Supplementary Fig. S1A-B**), with objective function of the same form as L^E and by allowing a single type of neurons sending both excitatory and inhibitory synaptic currents to their post-synaptic targets (**Supplementary Fig. S1C**). Similarly to the E-I model, also the 1CT model exhibits structured connectivity, with synaptic strength depending on the tuning similarity between the presynaptic and the postsynaptic neuron. Pairs of neurons with stronger tuning similarity (dissimilarity) have stronger mutual inhibition (excitation); see **Supplementary Fig. S1D**.

We compared the coding performance of the E-I model with that of a fully connected 1CT model. Both models received the same set of stimulus features and performed the same computation. In the 1CT model, tuning parameters were drawn from the same distribution as used for the E neurons in the E-I model. We used the same membrane time constant τ in both models, while the metabolic constants (β of the E-I model and β_1 of the 1CT model) and the noise intensity (σ of the E-I model and σ_1 of the 1CT model) were chosen such as to optimize the average loss for each model (**Fig. 5B** for E-I model, **Supplementary Fig. S1F-G** for 1CT model). Parameters of the 1CT model are listed in the **Supplementary Table S1**. A qualitative comparison of the E-I and the 1CT model showed that with optimal parameters, both models accurately tracked multiple target signals (**Fig. 1G** and **Supplementary Fig. S1E**).

To compare the performance of the E-I and the 1CT models also quantitatively, we measured the average encoding error (RMSE), metabolic cost (MC) and loss of each model. The RMSE and the MC in the 1CT model were measured as in **Eq. 37** and **38**, while the average loss of each model was evaluated as follows:

$$\begin{aligned}\mathcal{L}^{\text{1CT}} &= g_L\sqrt{\langle\epsilon_q^{\text{1CT}}(t)\rangle_{t,q}} + (1-g_L)\sqrt{\langle\kappa_q^{\text{1CT}}(t)\rangle_{t,q}},\\ \mathcal{L}^{\text{E-I}} &= g_L\frac{\sqrt{\langle\epsilon_q^E(t)\rangle_{t,q}+\langle\epsilon_q^I(t)\rangle}_{t,q}}{2} + (1-g_L)\frac{\sqrt{\langle\kappa_q^E(t)\rangle_{t,q}+\langle\kappa_q^I(t)\rangle}_{t,q}}{2}.\end{aligned} \tag{S.10}$$

Unless mentioned otherwise, we weighted stronger the encoding error compared to the metabolic cost and used g_L = 0.7. Note that our comparison of the losses is conservative, because the metabolic cost is defined as a sum of activities across neurons (**Eq. 38**) and the total number of neurons in the E-I model ($N^E + N^I$) is larger than the number of neurons in the 1CT model ($N^{\text{1CT}} = N^E$).

parameter	notation	value
number of E neurons	N	400
number of the input features	M	3
time constant of the single neuron and population readout	τ	10 ms
noise intensity	σ_1	$1.8\ (\mathrm{mV})^{1/2}$
SD of tuning parameters	σ_w	$1\ (\mathrm{mV})^{1/2}$
metabolic constant	β_1	11.4 mV

Table S1.

Table of default model parameters for the efficient network with one cell type.

The parameters *N, M,* τ and σ_w are chosen identical to the E-I network (see **Table 1** in the main text). Parameters σ_1 and β_1 are determined as values that maximize network efficiency (see section "Performance measures" in the main text).

Supplementary Figures

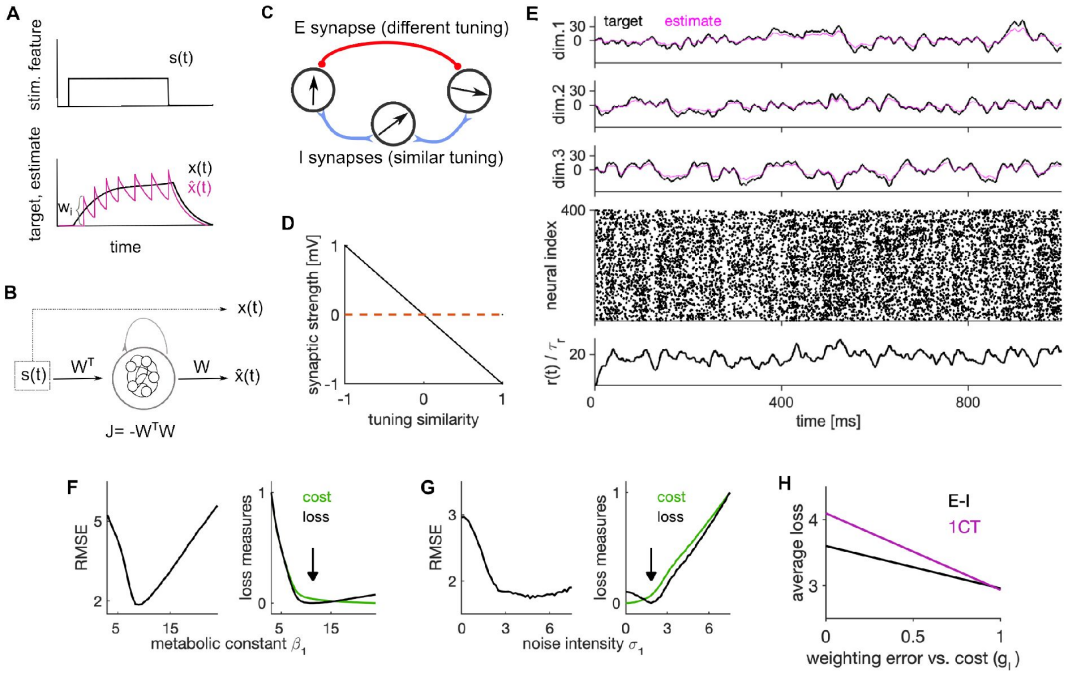


Figure S1.

Efficient spiking model with one cell type.

(A) Schematic of efficient coding with a single spiking neuron with positive weight. The target signal (bottom, black) integrates the input signal (top). The neuron spikes to keep the readout of its activity (magenta) close to the target signal.
(B) Schematic of the efficient 1CT model. Target signal $\boldsymbol{x}(t)$ is computed from stimulus features $\boldsymbol{s}(t)$. The network generates the estimate of the target signal with the population readouts of the spiking activity.
(C) Schematic of excitatory (red) and inhibitory (blue) synaptic interactions in 1CT model. Neurons with similar selectivity inhibit each other (blue), while neurons with different selectivity excite each other (red). The same neuron is sending excitatory and inhibitory synaptic outputs.
(D) Strength of recurrent synapses as a function of the tuning similarity.
(E) Simulation of the network with 1CT. Top three rows show the signal (black), and the estimate (magenta) in each of the 3 input dimensions.
(F) Left: Root mean squared error (RMSE) as a function of the metabolic constant β_1. Right: Normalized metabolic cost (green) and normalized average loss (black) as a function of the metabolic constant β_1. The black arrow denotes the minimum of the loss and thus the optimal parameter β_1.
(G) Same as in **F**, measured as a function of the noise intensity σ_1.
(H) Average loss as a function of the weighting of the encoding error and the metabolic cost, g_L, in the E-I model (black) and in the 1CT model (magenta). For plots F-H, results were computed in 100 simulation trials of duration of 1 second of simulated time. For other parameters, see **Table 1** (E-I model) and **Table S1** (1CT model).

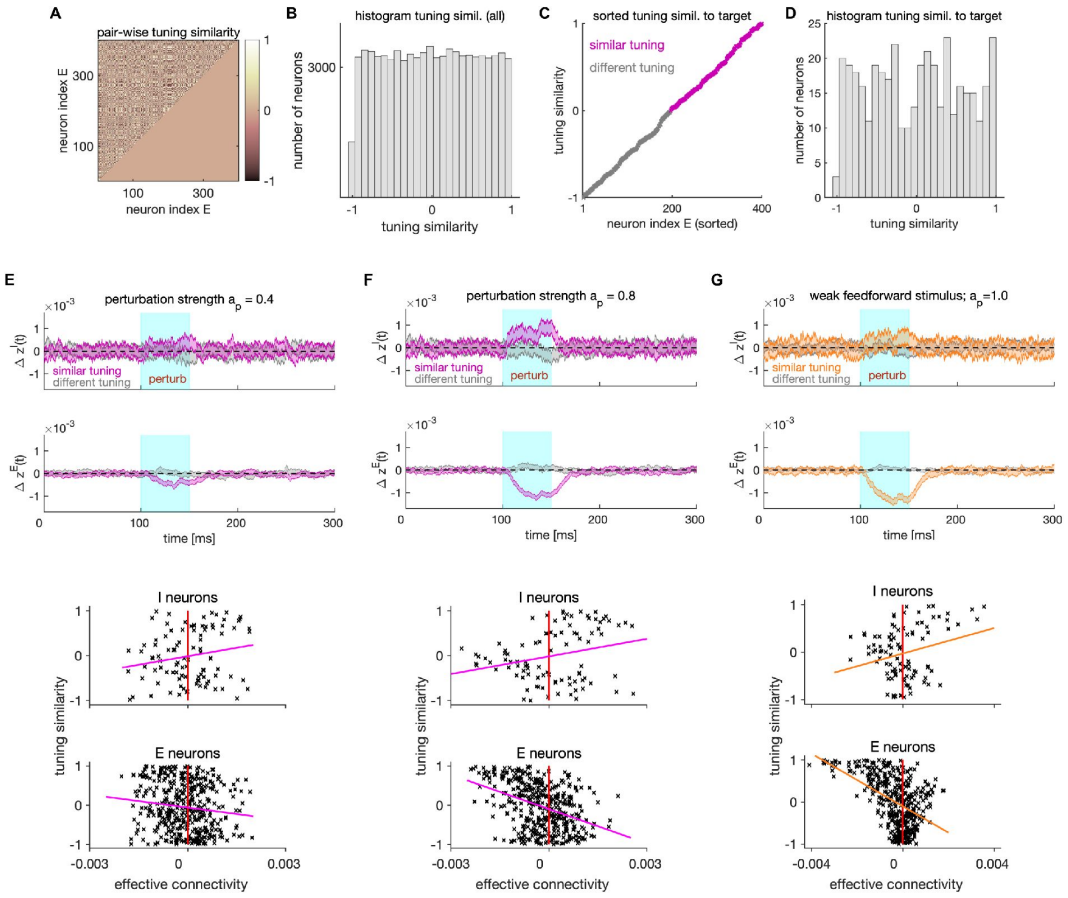


Figure S2.

Tuning similarity and its relation to lateral excitation/inhibition.

(A) Pair-wise tuning similarity for all pairs of E neurons. Tuning similarity between pairs of neurons is measured as the similarity of normalized tuning vectors.
(B) Histogram of tuning similarity across all E-E pairs shown in **A**.
(C) Tuning similarity to a single, randomly selected target neuron. Tuning similarity to a single neuron corresponds to a vector from the tuning similarity matrix in **A**. We sorted the tuning similarity to a single neuron from smallest to biggest value. Neurons with negative similarity are grouped as neurons with different tuning, while neurons with positive tuning similarity are grouped as neurons with similar tuning.
(D) Histogram of tuning similarity of E neurons to the target neuron shown in **C**. With distribution of tuning parameters symmetric around zero as used in our study, any choice of target neuron gives approximately the same number of neurons with similar and different selectivity.
(E) Top: Trial and neuron-averaged deviation of the instantaneous firing rate from the baseline firing rate, for the population of I (top) and E (bottom) neurons with similar tuning (magenta) and different tuning (gray). The baseline firing rates were 6.8 Hz and 12.7 Hz in the E and I cell types, respectively. The stimulation intensity is a_p = 0.4. Figure shows the mean ± standard error of the mean (SEM), with SEM capturing the variance across neurons and across trials. Bottom: Scatter plot of the tuning similarity versus effective connectivity in I (top) and E neurons (bottom). Tuning similarity and effective connectivity are measured with respect to the (same) target neuron. Red line marks zero effective connectivity and magenta line marks the least-squares line.
(F) Same as in **E**, for stimulation intensity of a_p = 0.8.
(G) Same as in **E**, in presence of weak feedforward stimulus, showing the activity of neurons with similar tuning (orange) and different tuning (gray) to the stimulated neuron. We used the stimulation intensity at threshold (a_p = 1.0). The feedforward stimulus was received by all E neurons and it induced, together with the external current, the mean firing rates of 7.3 Hz and 13.5 Hz in E and I neurons, respectively. For model parameters, see **Table 1**. This figure is related to the **Fig. 2** in the main paper.

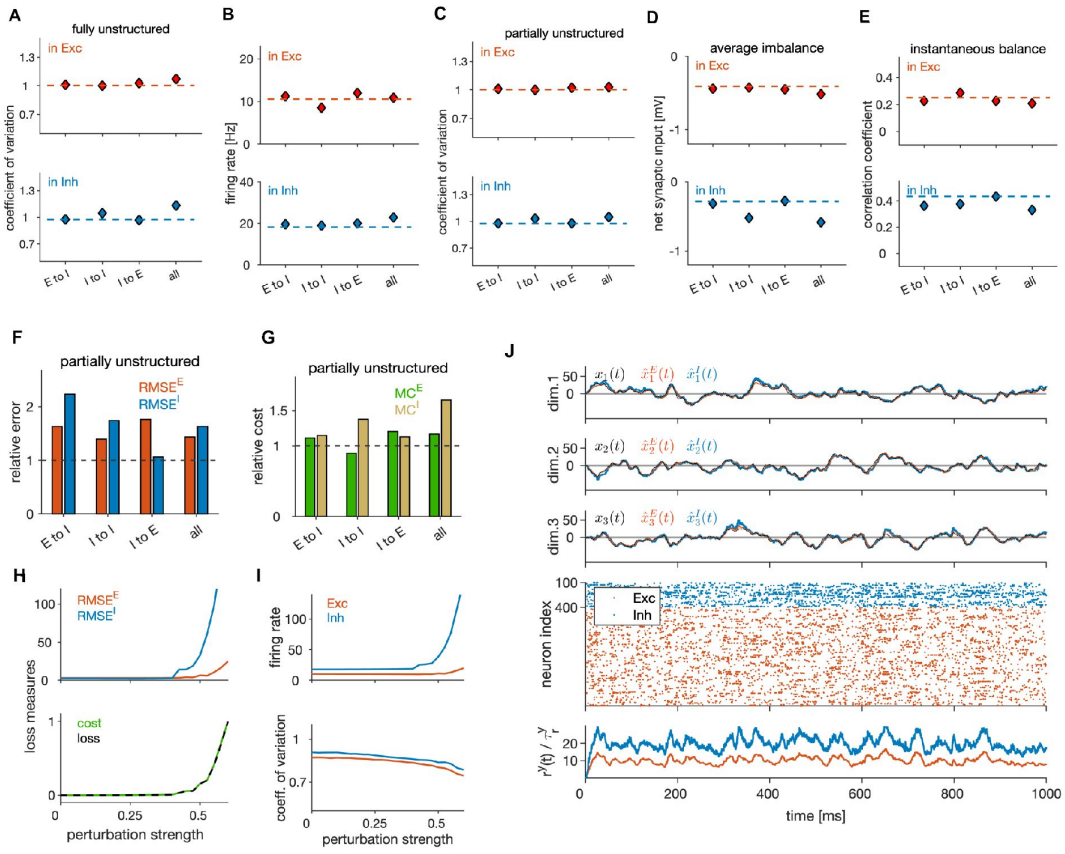


Figure S3.

Effect of complete and partial removal of connectivity structure and of minimal perturbation of synaptic weights.

(A) Average coefficient of variation in networks with fully unstructured connectivity. The dashed line marks the same measure in a structured network.
(B) Mean firing rate in E (top) and I neurons (bottom) in networks with partial removal of connectivity structure in recurrent connectivity. Partial removal of connectivity structure is achieved by limiting the permutation of synaptic connectivity to neuronal pairs with similar tuning.
(C) Same as in **B**, showing the coefficient of variation of spiking activity.
(D) Same as in **B**, showing the average net synaptic current, neural correlate of the average E-I balance.
(E) Same as in **B**, showing the correlation coefficient of synaptic currents, neural correlate of the instantaneous E-I balance.
(F) Encoding error in networks with partially unstructured recurrent connectivity, relative to the encoding error of the structured network (dashed line). From left to right: we perturb synaptic weights in E-I, I-I, I-E and in all three recurrent connectivities at once.
(G) Same as in **F**, showing the metabolic cost on spiking in E and I populations, relative to the metabolic cost in the structured network (dashed line).
(H) The RMSE (top) and the normalized metabolic cost (green) and average loss (black) average firing rate (bottom) in E and I cell type, as a function of the strength of perturbation of the synaptic connectivity.
(I) Average firing rate (top) and the coefficient of variation (bottom) as a function of the strength of random perturbation of all recurrent connectivities.
(J) Target signals, E estimates and I estimates in three input dimensions (three top rows), spike trains (fourth row) and the instantaneous estimate of the firing rate of E and I populations (bottom) in a single simulation trial, with significant perturbation of recurrent connectivity (perturbation strength of 0.5, see Methods). In spite of a relatively strong perturbation, the network shows excellent encoding of the target signal. Other parameters are in **Table 1**. This figure is related to the **Fig. 3** in the main paper.

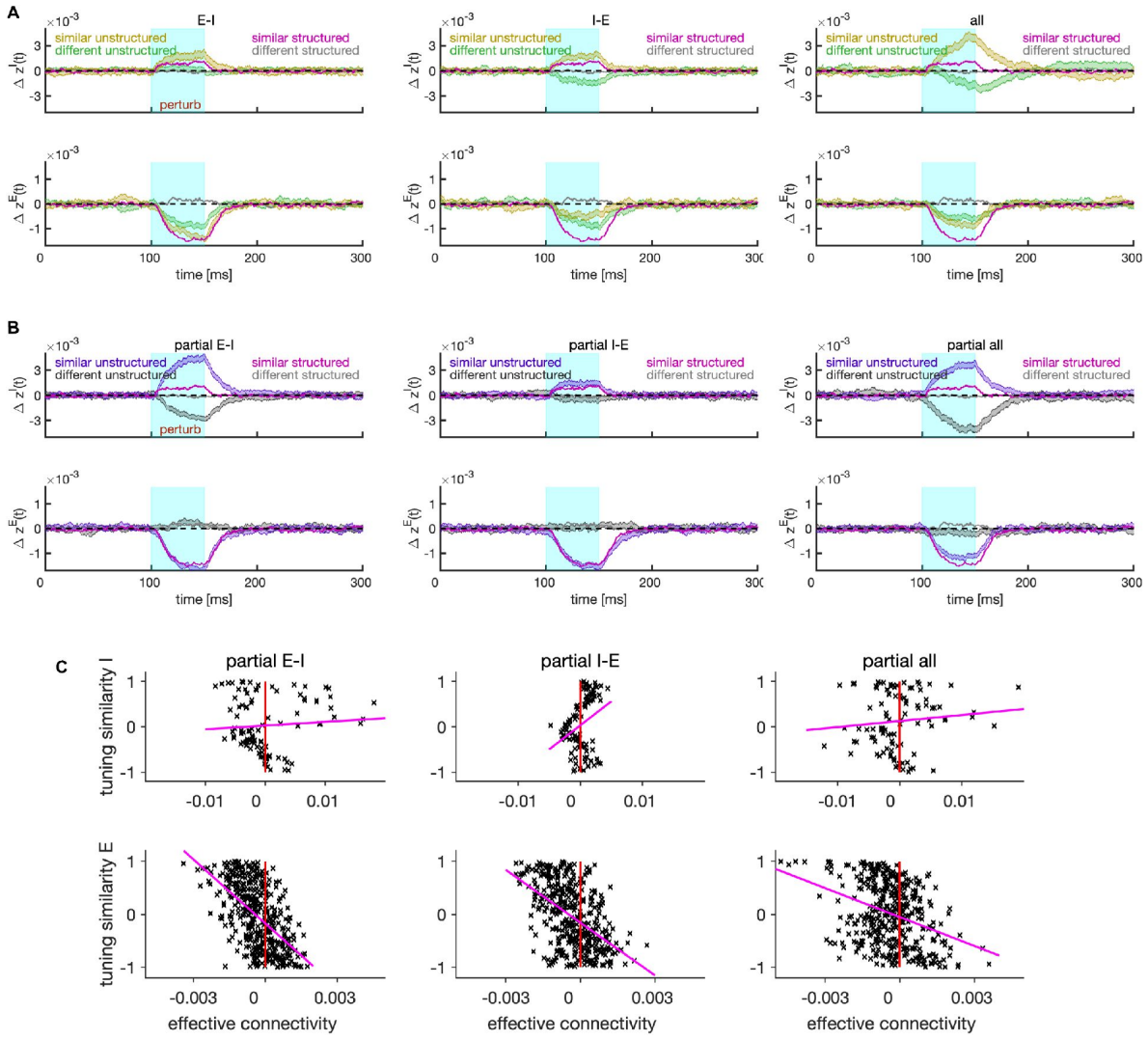


Figure S4.

Lateral excitation/inhibition in models with full and partial removal of connectivity structure.

(A) Average deviation of the instantaneous firing rate from the baseline for the population of I (top) and E (bottom) neurons in networks with fully removed structure in E-I (left), I-E (middle) and in all connectivity matrices (right). We show the mean ± SEM for neurons with similar (ochre) and different (green) tuning to the stimulated neuron. The mean traces of the network with structured connectivity is shown for comparison, magenta and gray for similar and different tuning, respectively.
(B) Same as in **A**, for partial (fine-grained) removal of connectivity structure. Partial removal of connectivity structure is achieved by limiting the permutation of synaptic weights among neurons with similar tuning. Such manipulation maintains the like-like connectivity structure, but removes any structure beyond the like-like.
(C) Scatter plot of tuning similarity versus effective connectivity for networks with partial removal of connectivity structure. In such networks, the specificity of effective connectivity with respect to tuning similarity is largely preserved, in particular in E neurons. For all results, we iterated simulations in 200 trials, where we varied randomly the membrane potential noise and initial conditions of the membrane potentials in each trial, while tuning and synaptic parameters were kept fixed. In all cases, we used stimulation intensity at threshold (a_p = 1.0). For model parameters, see **Table 1**. This figure is related to the **Fig. 3** in the main paper.

Figure S5.

Dependence of coding efficiency and neural dynamics on the ratio of mean I-I to E-I connectivity, computed by changing the mean E-I connectivity.

(A) Top: Encoding error (RMSE) of the E (red) and I (blue) estimates. Bottom: Normalized metabolic cost and average loss.
(B) Average firing rate (top), and average coefficient of variation (bottom) in E and I cell type.
(C) Average imbalance and instantaneous balance of synaptic currents in E and I neurons.
(D) Top: Optimal ratio of mean I-I to E-I connectivity as a function of the weighting of the average loss of E and I cell type. Bottom: Same as on top, as a function of the weighting between the error and the cost. Black triangles mark weightings that are typically used to estimate optimal efficiency. For other parameters, see **Table 1**. This figure is related to the **Fig. 6** in the main paper.

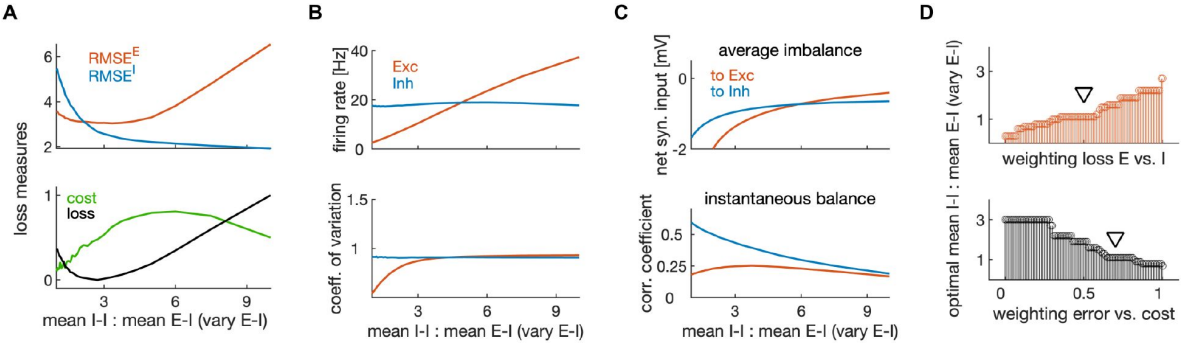
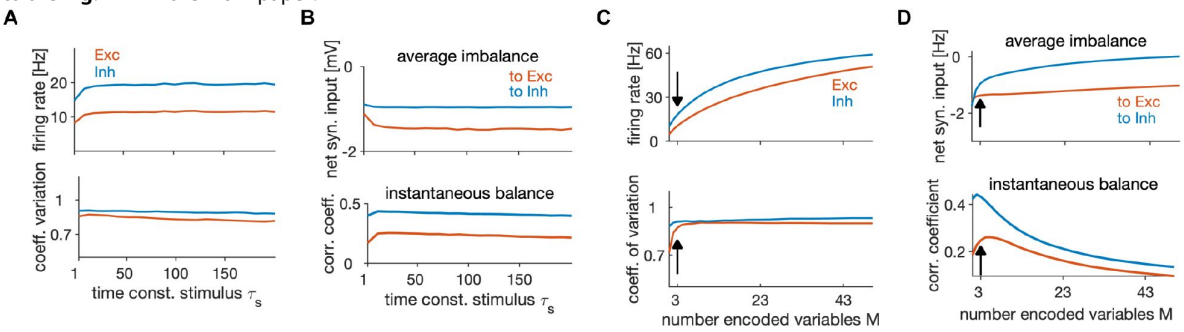


Figure S6.

Effect of stimulus properties on efficient neural coding and dynamics.

(A) Average firing rate (top), and average coefficient of variation (bottom) in E and I cell type, as a function of the time constant of the stimulus τ_s.
(B) Average imbalance (top) and instantaneous balance (bottom) as a function of the time constant of the stimulus τ_s.
(C-D) Same as in **A-B**, as a function of the number of encoded variables. For parameters, see **Table 1**. This figure is related to the **Fig. 7** in the main paper.



References

1. Abbott L. F., DePasquale B., Memmesheimer R.-M. (2016) **Building functional networks of spiking model neurons** *Nature neuroscience* **19**:350–355

2. Thalmeier D., Uhlmann M., Kappen H. J., Memmesheimer R.-M. (2016) **Learning universal computations with spikes** *PLoS computational biology* **12**

3. Barlow H. B., et al. (1961) **Possible principles underlying the transformation of sensory messages** *Sensory communication* **1**:217–233

4. Olshausen B. A., Field D. J. (1996) **Emergence of simple-cell receptive field properties by learning a sparse code for natural images** *Nature* **381**:607–609

5. Deneve S., Chalk M. (2016) **Efficiency turns the table on neural encoding, decoding and noise** *Current Opinion in Neurobiology* **37**:141–148

6. Simoncelli E. P., Olshausen B. A. (2001) **Natural image statistics and neural representation** *Annual review of neuroscience* **24**:1193–1216

7. Olshausen B. A., Field D. J. (1997) **Sparse coding with an overcomplete basis set: A strategy employed by v1?** *Vision research* **37**:3311–3325

8. Vinje W. E., Gallant J. L. (2000) **Sparse coding and decorrelation in primary visual cortex during natural vision** *Science* **287**:1273–1276

9. Li Z. (2014) **Understanding vision: theory, models, and data**

10. Atick J. J. (1992) **Could information theory provide an ecological theory of sensory processing?** *Network: Computation in neural systems* **3**:213–251

11. Olshausen B. A., Field D. J. (2004) **Sparse coding of sensory inputs** *Current opinion in neurobiology* **14**:481–487

12. Lewicki M. S. (2002) **Efficient coding of natural sounds** *Nature neuroscience* **5**:356–363

13. Koulakov A., Rinberg D. (2011) **Sparse incomplete representations: A potential role of olfactory granule cells** *Neuron* **72**:124–136

14. Bialek W., Rieke F., de Ruyter van Steveninck R. R., Warland D. (1991) **Reading a neural code** *Science* **252**:1854–1857

15. Bialek W., Rieke F. (1992) **Reliability and information transmission in spiking neurons** *Trends in neurosciences* **15**:428–434

16. Panzeri S., Petersen R. S., Schultz S. R., Lebedev M., Diamond M. E. (2001) **The role of spike timing in the coding of stimulus location in rat somatosensory cortex** *Neuron* **29**:769–777

17. Nemenman I., Lewen G. D., Bialek W., de Ruyter van Steveninck R. R. (2008) **Neural coding of natural stimuli: information at sub-millisecond resolution** *PLoS computational biology* **4**

18. Kayser C., Logothetis N. K., Panzeri S. (2010) **Millisecond encoding precision of auditory cortex neurons** *Proceedings of the National Academy of Sciences* **107**:16976–16981

19. Ince R. A., Panzeri S., Kayser C. (2013) **Neural codes formed by small and temporally precise populations in auditory cortex** *Journal of Neuroscience* **33**:18277–18287

20. Panzeri S., Brunel N., Logothetis N. K., Kayser C. (2010) **Sensory neural codes using multiplexed temporal scales** *Trends in neurosciences* **33**:111–120

21. Fairhall A. L., Lewen G. D., Bialek W., de Ruyter van Steveninck R. R. (2001) **Efficiency and ambiguity in an adaptive neural code** *Nature* **412**:787–792

22. Wark B., Fairhall A., Rieke F. (2009) **Timescales of inference in visual adaptation** *Neuron* **61**:750–761

23. Mazzoni A., Panzeri S., Logothetis N. K., Brunel N. (2008) **Encoding of naturalistic stimuli by local field potential spectra in networks of excitatory and inhibitory neurons** *PLoS computational biology* **4**

24. Mlynarski W. F., Hermundstadt A. M. (2021) **Efficient andadaptive sensory codes** *Nature Neuroscience* **24**:998–1009

25. Koren V., Bondanelli G., Panzeri S. (2023) **Computational methods to study information processing in neural circuits** *Computational and Structural Biotechnology Journal* **21**:910–922

26. Lochmann T., Ernst U. A., Deneve S. (2012) **Perceptual inference predicts contextual modulations of sensory responses** *Journal of Neuroscience* **32**:4179–4195

27. Zhu M., Rozell C. J. (2013) **Visual nonclassical receptive field effects emerge from sparse coding in a dynamical system** *PLoS computational biology* **9**

28. Boerlin M., Machens C. K., Denève S. (2013) **Predictive coding of dynamical variables in balanced spiking networks** *PLoS Comput Biol* **9**

29. Bourdoukan R., Barrett D., Deneve S., Machens C. K. (2012) **Learning optimal spike-based representations** *Advances in neural information processing systems* **25**:2285–2293

30. Moreno-Bote R., Drugowitsch J. (2015) **Causal inference and explaining away in a spiking network** *Scientific Reports* **5**

31. Chalk M., Gutkin B., Deneve S. (2016) **Neural oscillations as a signature of efficient coding in the presence of synaptic delays** *Elife* **5**

32. Denève S., Machens C. K. (2016) **Efficient codes and balanced networks** *Nature neuroscience* **19**:375–382

33. Gutierrez G. J., Denève S. (2019) **Population adaptation in efficient balanced networks** *Elife* **8**

34. Kadmon J., Timcheck J., Ganguli S., Larochelle H., Ranzato M., Hadsell R., Balcan M., Lin H (2020) **Predictive coding in balanced neural networks with noise, chaos and delays** *Advances in Neural Information Processing Systems* :16677–16688

35. Buxó C. E. R., Pillow J. W. (2020) **Poisson balanced spiking networks** *PLoS computational biology* **16**

36. Koren V., Denève S. (2017) **Computational account of spontaneous activity as a signature of predictive coding** *PLoS computational biology* **13**

37. Koren V., Panzeri S., Koyejo S., et al. (2022) **Biologically plausible solutions for spiking networks with efficient coding** *Advances in Neural Information Processing Systems* :20607–20620

38. Brette R., Gerstner W. (2005) **Adaptive exponential integrate-and-fire model as an effective description of neuronal activity** *Journal of neurophysiology* **94**:3637–3642

39. Mensi S., et al. (2012) **Parameter extraction and classification of three cortical neuron types reveals two distinct adaptation mechanisms** *Journal of neurophysiology* **107**:1756–1775

40. Gerstner W., Kistler W. M., Naud R., Paninski L. (2014) **Neuronal dynamics: From single neurons to networks and models of cognition**

41. Jolivet R., et al. (2008) **The quantitative single-neuron modeling competition** *Biological cybernetics* **99**

42. Brendel W., Bourdoukan R., Vertechi P., Machens C. K., Denéve S. (2020) **Learning to represent signals spike by spike** *PLoS computational biology* **16**

43. Barrett D. G., Deneve S., Machens C. K. (2016) **Optimal compensation for neuron loss** *Elife* **5**

44. Alemi A., Machens C., Deneve S., Slotine J.-J. (2018) **Learning nonlinear dynamics in efficient, balanced spiking networks using local plasticity rules** *Proceedings of the AAAI Conference on Artificial Intelligence* **32** https://doi.org/10.1609/aaai.v32i1.11320

45. Vogels T. P., Sprekeler H., Zenke F., Clopath C., Gerstner W. (2011) **Inhibitory plasticity balances excitation and inhibition in sensory pathways and memory networks** *Science* **334**:1569–1573

46. Amatrudo J. M., et al. (2012) **Influence of highly distinctive structural properties on the excitability of pyramidal neurons in monkey visual and prefrontal cortices** *Journal of Neuroscience* **32**:13644–13660

47. Rigotti M., et al. (2013) **The importance of mixed selectivity in complex cognitive tasks** *Nature* **497**:585–590

48. Denève S., Alemi A., Bourdoukan R. (2017) **The brain as an efficient and robust adaptive learner** *Neuron* **94**:969–977

49. Tavoni G., Balasubramanian V., Gold J. I. (2019) **What is optimal in optimal inference?** *Current Opinion in Behavioral Sciences* **29**:117–126

50. Faisal A. A., Selen L. P., Wolpert D. M. (2008) **Noise in the nervous system** *Nature reviews neuroscience* **9**:292–303

51. Ko H., et al. (2011) **Functional specificity of local synaptic connections in neocortical networks** *Nature* **473**:87–91

52. Pala A., Petersen C. (2015) **In-vivo measurement of cell-type-specific synaptic connectivity and synaptic transmission in layer 2/3 mouse barrel cortex** *Neuron* **85**:68–75

53. Campagnola L., et al. (2022) **Local connectivity and synaptic dynamics in mouse and human neocortex** *Science* **375**

54. Burkitt A. N. (2006) **A review of the integrate-and-fire neuron model: I. homogeneous synaptic input** *Biological cybernetics* **95**:1–19

55. Schwalger T., Deger M., Gerstner W. (2017) **Towards a theory of cortical columns: From spiking neurons to interacting neural populations of finite size** *PLoS Comput. Biol* **13**

56. Harkin E. F., Béïque J.-C., Naud R. (2021) **A user's guide to generalized integrate-and-fire models** *Computational Modelling of the Brain: Modelling Approaches to Cells, Circuits and Networks* :69–86

57. 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**:301–316

58. Ahmadian Y., Miller K. D. (2021) **What is the dynamical regime of cerebral cortex?** *Neuron* **109**:3373–3391

59. Okun M., Lampl I. (2008) **Instantaneous correlation of excitation and inhibition during ongoing and sensory-evoked activities** *Nature neuroscience* **11**:535–537

60. Xue M., Atallah B. V., Scanziani M. (2014) **Equalizing excitation–inhibition ratios across visual cortical neurons** *Nature* **511**:596–600

61. Chettih S. N., Harvey C. D. (2019) **Single-neuron perturbations reveal feature-specific competition in V1** *Nature* **567**:334–340

62. Oldenburg I. A., et al. (2024) **The logic of recurrent circuits in the primary visual cortex** *Nature Neuroscience* **27**:1–11

63. Brunel N. (2000) **Dynamics of sparsely connected networks of excitatory and inhibitory spiking neurons** *Journal of computational neuroscience* **8**:183–208

64. Renart A., et al. (2010) **The asynchronous state in cortical circuits** *Science* **327**:587–590

65. Abbott L. F., Nelson S. B. (2000) **Synaptic plasticity: taming the beast** *Nature neuroscience* **3**:1178–1183

66. Turrigiano G. G., Nelson S. B. (2004) **Homeostatic plasticity in the developing nervous system** *Nature reviews neuroscience* **5**:97–107

67. Schwalger T., Lindner B. (2013) **Patterns of interval correlations in neural oscillators with adaptation** *Front. Comput. Neurosci* **7**

68. Levy M., Sporns O., MacLean J. N. (2020) **Network analysis of murine cortical dynamics implicates untuned neurons in visual stimulus coding** *Cell Reports* **31**

69. Zylberberg J. (2017) **The role of untuned neurons in sensory information coding** *BioRxiv* **134379**

70. Destexhe A., Paré D. (1999) **Impact of network activity on the integrative properties of neocortical pyramidal neurons in vivo** *Journal of neurophysiology* **81**:1531–1547

71. Destexhe A., Rudolph M., Paré D. (2003) **The high-conductance state of neocortical neurons in vivo** *Nature reviews neuroscience* **4**:739–751

72. Van Vreeswijk C., Sompolinsky H. (1996) **Chaos in neuronal networks with balanced excitatory and inhibitory activity** *Science* **274**:1724–1726

73. Sukenik N., et al. (2021) **Neuronal circuits overcome imbalance in excitation and inhibition by adjusting connection numbers** *Proceedings of the National Academy of Sciences* **118**

74. Markram H., et al. (2004) **Interneurons of the neocortical inhibitory system** *Nature reviews neuroscience* **5**:793–807

75. Cossell L., et al. (2015) **Functional organization of excitatory synaptic strength in primary visual cortex** *Nature* **518**:399–403

76. Roy K., Jaiswal A., Panda P. (2019) **Towards spike-based machine intelligence with neuromorphic computing** *Nature* **575**:607–617

77. Schuman C. D., et al. (2022) **Opportunities for neuromorphic computing algorithms and applications** *Nature Computational Science* **2**:10–19

78. Najafi F., et al. (2020) **Excitatory and inhibitory subnetworks are equally selective during decision-making and emerge simultaneously during learning** *Neuron* **105**:165–179

79. Runyan C. A., et al. (2010) **Response features of parvalbumin-expressing interneurons suggest precise roles for subtypes of inhibition in visual cortex** *Neuron* **67**:847–857

80. Kuan A. T., et al. (2024) **Synaptic wiring motifs in posterior parietal cortex support decision-making** *Nature* **627**:367–373

81. Sadeh S., Clopath C. (2020) **Theory of neuronal perturbome in cortical networks** *Proceedings of the National Academy of Sciences* **117**:26966–26976

82. Znamenskiy P., et al. (2024) **Functional specificity of recurrent inhibition in visual cortex** *Neuron* **112**:991–1000

83. Seeman S. C., et al. (2018) **Sparse recurrent excitatory connectivity in the microcircuit of the adult mouse and human cortex** *Elife* **7**

84. Stepanyants A., Martinez L. M., Ferecskó A. S., Kisvárday Z. F. (2009) **The fractions of short-and long-range connections in the visual cortex** *Proceedings of the National Academy of Sciences* **106**:3555–3560

85. Safavi S., Chalk M., Logothetis N., Levina A. (2023) **Signatures of criticality in efficient coding networks** *bioRxiv*

86. Valente M., et al. (2021) **Correlations enhance the behavioral readout of neural population activity in association cortex** *Nature neuroscience* **24**:975–986

87. Panzeri S., Moroni M., Safaai H., Harvey C. D. (2022) **The structures and functions of correlations in neural population codes** *Nature Reviews Neuroscience* **23**:551–567

88. Manning T. S., et al. (2024) **Transformations of sensory information in the brain suggest changing criteria for optimality** *PLOS Computational Biology* **20**

89. Whittaker V., Osborne N. N. (1983) **What is Dale's principle** *Dale's Principle and Communication Between Neurones* :1–5

90. Muller M. E. (1959) **A note on a method for generating points uniformly on n-dimensional spheres** *Communications of the ACM* **2**:19–20

91. Luo C., et al. (2018) **Cosine normalization: Using cosine similarity instead of dot product in neural networks** *Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks* **27**:382–391

Editors

Reviewing Editor
Richard Naud
University of Ottawa, Ottawa, Canada

Senior Editor
Panayiota Poirazi
FORTH Institute of Molecular Biology and Biotechnology, Heraklion, Greece

Reviewer #1 (Public Review):

Koren et al. derive and analyse a spiking network model optimised to represent external signals using the minimum number of spikes. Unlike most prior work using a similar setup, the network includes separate populations of excitatory and inhibitory neurons. The authors show that the optimised connectivity has a like-to-like structure, leading to the experimentally observed phenomenon of feature competition. They also characterise the impact of various (hyper)parameters, such as adaptation timescale, ratio of excitatory to inhibitory cells, regularisation strength, and background current. These results add useful biological realism to a particular model of efficient coding. However, not all claims seem fully supported by the evidence. Specifically, several biological features, such as the ratio of excitatory to inhibitory neurons, which the authors claim to explain through efficient coding, might be contingent on arbitrary modelling choices. In addition, earlier work has already established the importance of structured connectivity for feature competition. A clearer presentation of modelling choices, limitations, and prior work could improve the manuscript.

Major comments:

(1) Much is made of the 4:1 ratio between excitatory and inhibitory neurons, which the authors claim to explain through efficient coding. I see two issues with this conclusion: (i) The 4:1 ratio is specific to rodents; humans have an approximate 2:1 ratio (see Fang & Xia et al., Science 2022 and references therein); (ii) the optimal ratio in the model depends on a seemingly arbitrary choice of hyperparameters, particularly the weighting of encoding error versus metabolic cost. This second concern applies to several other results, including the strength of inhibitory versus excitatory synapses. While the model can, therefore, be made consistent with biological data, this requires auxiliary assumptions.

(2) A growing body of evidence supports the importance of structured E-I and I-E connectivity for feature selectivity and response to perturbations. For example, this is a major conclusion from the Oldenburg paper (reference 62 in the manuscript), which includes extensive modelling work. Similar conclusions can be found in work from Znamenskiy and colleagues (experiments and spiking network model; bioRxiv 2018, Neuron 2023 (ref. 82)), Sadeh & Clopath (rate network; eLife, 2020), and Mackwood et al. (rate network with plasticity; eLife,

2021). The current manuscript adds to this evidence by showing that (a particular implementation of) efficient coding in spiking networks leads to structured connectivity. The fact that this structured connectivity then explains perturbation responses is, in the light of earlier findings, not new.

(3) The model's limitations are hard to discern, being relegated to the manuscript's last and rather equivocal paragraph. For instance, the lack of recurrent excitation, crucial in neural dynamics and computation, likely influences the results: neuronal time constants must be as large as the target readout (Figure 4), presumably because the network cannot integrate the signal without recurrent excitation. However, this and other results are not presented in tandem with relevant caveats.

(4) On repeated occasions, results from the model are referred to as predictions claimed to match the data. A prediction is a statement about what will happen in the future - but most of the "predictions" from the model are actually findings that broadly match earlier experimental results, making them "postdictions". This distinction is important: compared to postdictions, predictions are a much stronger test because they are falsifiable. This is especially relevant given (my impression) that key parameters of the model were tweaked to match the data.

https://doi.org/10.7554/eLife.99545.1.sa3

Reviewer #2 (Public Review):

Summary:

In this work, the authors present a biologically plausible, efficient E-I spiking network model and study various aspects of the model and its relation to experimental observations. This includes a derivation of the network into two (E-I) populations, the study of single-neuron perturbations and lateral-inhibition, the study of the effects of adaptation and metabolic cost, and considerations of optimal parameters. From this, they conclude that their work puts forth a plausible implementation of efficient coding that matches several experimental findings, including feature-specific inhibition, tight instantaneous balance, a 4 to 1 ratio of excitatory to inhibitory neurons, and a 3 to 1 ratio of I-I to E-I connectivity strength. It thus argues that some of these observations may come as a direct consequence of efficient coding.

Strengths:

While many network implementations of efficient coding have been developed, such normative models are often abstract and lacking sufficient detail to compare directly to experiments. The intention of this work to produce a more plausible and efficient spiking model and compare it with experimental data is important and necessary in order to test these models.

In rigorously deriving the model with real physical units, this work maps efficient spiking networks onto other more classical biophysical spiking neuron models. It also attempts to compare the model to recent single-neuron perturbation experiments, as well as some long-standing puzzles about neural circuits, such as the presence of separate excitatory and inhibitory neurons, the ratio of excitatory to inhibitory neurons, and E/I balance. One of the primary goals of this paper, to determine if these are merely biological constraints or come from some normative efficient coding objective, is also important.

Though several of the observations have been reported and studied before (see below), this work arguably studies them in more depth, which could be useful for comparing more directly to experiments.

Weaknesses:

Though the text of the paper may suggest otherwise, many of the modeling choices and observations found in the paper have been introduced in previous work on efficient spiking models, thereby making this work somewhat repetitive and incremental at times. This includes the derivation of the network into separate excitatory and inhibitory populations, discussion of physical units, comparison of voltage versus spike-timing correlations, and instantaneous E/I balance, all of which can be found in one of the first efficient spiking network papers (Boerlin et al. 2013), as well as in subsequent papers. Metabolic cost and slow adaptation currents were also presented in a previous study (Gutierrez & Deneve 2019). Though it is perfectly fine and reasonable to build upon these previous studies, the language of the text gives them insufficient credit.

Furthermore, the paper makes several claims of optimality that are not convincing enough, as they are only verified by a limited parameter sweep of single parameters at a time, are unintuitive and may be in conflict with previous findings of efficient spiking networks. This includes the following. Coding error (RMSE) has a minimum at intermediate metabolic cost (Figure 5B), despite the fact that intuitively, zero metabolic cost would indicate that the network is solely minimizing coding error and that previous work has suggested that additional costs bias the output. Coding error also appears to have a minimum at intermediate values of the ratio of E to I neurons (effectively the number of I neurons) and the number of encoded variables (Figures 6D, 7B). These both have to do with the redundancy in the network (number of neurons for each encoded variable), and previous work suggests that networks can code for arbitrary numbers of variables provided the redundancy is high enough (e.g., Calaim et al. 2022). Lastly, the performance of the E-I variant of the network is shown to be better than that of a single cell type (1CT: Figure 7C, D). Given that the E-I network is performing a similar computation as to the 1CT model but with more neurons (i.e., instead of an E neuron directly providing lateral inhibition to its neighbor, it goes through an interneuron), this is unintuitive and again not supported by previous work. These may be valid emergent properties of the E-I spiking network derived here, but their presentation and description are not sufficient to determine this.

Alternatively, the methodology of the model suggests that ad hoc modeling choices may be playing a role. For example, an arbitrary weighting of coding error and metabolic cost of 0.7 to 0.3, respectively, is chosen without mention of how this affects the results. Furthermore, the scaling of synaptic weights appears to be controlled separately for each connection type in the network (Table 1), despite the fact that some of these quantities are likely linked in the optimal network derivation. Finally, the optimal threshold and metabolic constants are an order of magnitude larger than the synaptic weights (Table 1). All of these considerations suggest one of the following two possibilities. One, the model has a substantial number of unconstrained parameters to tune, in which case more parameter sweeps would be necessary to definitively make claims of optimality. Or two, parameters are being decoupled from those constrained by the optimal derivation, and the optima simply corresponds to the values that should come out of the derivation.

https://doi.org/10.7554/eLife.99545.1.sa2

Reviewer #3 (Public Review):

Summary:

In their paper the authors tackle three things at once in a theoretical model: how can spiking neural networks perform efficient coding, how can such networks limit the energy use at the same time, and how can this be done in a more biologically realistic way than previous work?

They start by working from a long-running theory on how networks operating in a precisely balanced state can perform efficient coding. First, they assume split networks of excitatory (E) and inhibitory (I) neurons. The E neurons have the task to represent some lower dimensional input signal, and the I neurons have the task to represent the signal represented by the E neurons. Additionally, the E and I populations should minimize an energy cost represented by the sum of all spikes. All this results in two loss functions for the E and I populations, and the networks are then derived by assuming E and I neurons should only spike if this improves their respective loss. This results in networks of spiking neurons that live in a balanced state, and can accurately represent the network inputs.

They then investigate in-depth different aspects of the resulting networks, such as responses to perturbations, the effect of following Dale's law, spiking statistics, the excitation (E)/inhibition (I) balance, optimal E/I cell ratios, and others. Overall, they expand on previous work by taking a more biological angle on the theory and showing the networks can operate in a biologically realistic regime.

Strengths:

(1) The authors take a much more biological angle on the efficient spiking networks theory than previous work, which is an essential contribution to the field.

(2) They make a very extensive investigation of many aspects of the network in this context, and do so thoroughly.

(3) They put sensible constraints on their networks, while still maintaining the good properties these networks should have.

Weaknesses:

(1) The paper has somewhat overstated the significance of their theoretical contributions, and should make much clearer what aspects of the derivations are novel. Large parts were done in very similar ways in previous papers. Specifically: the split into E and I neurons was also done in Boerlin et al (2008) and in Barrett et al (2016). Defining the networks in terms of realistic units was already done by Boerlin et al (2008). It would also be worth it to discuss Barrett et al (2016) specifically more, as there they also use split E/I networks and perform biologically relevant experiments.

(2) It is not clear from an optimization perspective why the split into E and I neurons and following Dale's law would be beneficial. While the constraints of Dale's law are sensible (splitting the population in E and I neurons, and removing any non-Dalian connection), they are imposed from biology and not from any coding principles. A discussion of how this could be done would be much appreciated, and in the main text, this should be made clear.

(3) Related to the previous point, the claim that the network with split E and I neurons has a lower average loss than a 1 cell-type (1-CT) network seems incorrect to me. Only the E population coding error should be compared to the 1-CT network loss, or the sum of the E and I populations (not their average). In my author recommendations, I go more in-depth on this point.

(4) While the paper is supposed to bring the balanced spiking networks they consider in a more experimentally relevant context, for experimental audiences I don't think it is easy to follow how the model works, and I recommend reworking both the main text and methods to improve on that aspect.

Assessment and context:

Overall, although much of the underlying theory is not necessarily new, the work provides an important addition to the field. The authors succeeded well in their goal of making the networks more biologically realistic, and incorporating aspects of energy efficiency. For computational neuroscientists, this paper is a good example of how to build models that link well to experimental knowledge and constraints, while still being computationally and mathematically tractable. For experimental readers, the model provides a clearer link between efficient coding spiking networks to known experimental constraints and provides a few predictions.

https://doi.org/10.7554/eLife.99545.1.sa1

Author response:

> ***eLife assessment***
>
> *This study offers a useful treatment of how the population of excitatory and inhibitory neurons integrates principles of energy efficiency in their coding strategies. The analysis provides a comprehensive characterisation of the model, highlighting the structured connectivity between excitatory and inhibitory neurons. However, the manuscript provides an incomplete motivation for parameter choices. Furthermore, the work is insufficiently contextualized within the literature, and some of the findings appear overlapping and incremental given previous work.*

We thank the Reviewers and the Reviewing Editor for taking time to provide extremely valuable suggestions and comments, which will help us to substantially improve our paper. In what follows we summarize our current plan to improve the paper taking up on their suggestions.

> ***Public Reviews:***
>
> ***Reviewer #1 (Public Review):***
>
> *Summary: Koren et al. derive and analyse a spiking network model optimised to represent external signals using the minimum number of spikes. Unlike most prior work using a similar setup, the network includes separate populations of excitatory and inhibitory neurons. The authors show that the optimised connectivity has a like-to-like structure, leading to the experimentally observed phenomenon of feature competition. They also characterise the impact of various (hyper)parameters, such as adaptation timescale, ratio of excitatory to inhibitory cells, regularisation strength, and background current. These results add useful biological realism to a particular model of efficient coding. However, not all claims seem fully supported by the evidence. Specifically, several biological features, such as the ratio of excitatory to inhibitory neurons, which the authors claim to explain through efficient coding, might be contingent on arbitrary modelling choices. In addition, earlier work has already established the importance of structured connectivity for feature competition. A clearer presentation of modelling choices, limitations, and prior work could improve the manuscript.*

Thanks for these insights and for this summary of our work.

> *Major comments:*
>
> *(1) Much is made of the 4:1 ratio between excitatory and inhibitory neurons, which the authors claim to explain through efficient coding. I see two issues with this conclusion: (i) The 4:1 ratio is specific to rodents; humans have an approximate 2:1 ratio (see Fang &*

> *Xia et al., Science 2022 and references therein); (ii) the optimal ratio in the model depends on a seemingly arbitrary choice of hyperparameters, particularly the weighting of encoding error versus metabolic cost. This second concern applies to several other results, including the strength of inhibitory versus excitatory synapses. While the model can, therefore, be made consistent with biological data, this requires auxiliary assumptions.*

We will describe better the ratio of numbers of E and I neurons found in real data, as suggested. The first submission already contained an analysis of how this ratio of neuron numbers depends on the weighting of the loss of E and I neurons and on the relative weighting of the encoding error vs the metabolic cost in the loss function (see Fig 6E). We will make sure that these results are suitably expanded and better emphasized in revision. We will also include new analysis of dependence of optimal parameters on the relative weighting of encoding error vs metabolic cost in the loss function when studying other parameters (namely: noise intensity, metabolic constant, ratio of mean I-I to E-I connectivity, time constants of single E and I neurons).

> *(2) A growing body of evidence supports the importance of structured E-I and I-E connectivity for feature selectivity and response to perturbations. For example, this is a major conclusion from the Oldenburg paper (reference 62 in the manuscript), which includes extensive modelling work. Similar conclusions can be found in work from Znamenskiy and colleagues (experiments and spiking network model; bioRxiv 2018, Neuron 2023 (ref. 82)), Sadeh & Clopath (rate network; eLife, 2020), and Mackwood et al. (rate network with plasticity; eLife, 2021). The current manuscript adds to this evidence by showing that (a particular implementation of) efficient coding in spiking networks leads to structured connectivity. The fact that this structured connectivity then explains perturbation responses is, in the light of earlier findings, not new.*

We agree that the main contribution of our manuscript in this respect is to show how efficient coding in spiking networks can lead to structured connectivity similar to those proposed in the above papers. We apologize if this was not clear enough in the previous version. We will make it clearer in revision. We nevertheless think it useful to report the effects of perturbations within this network because the structure derived in our network is not identical to those studied in the above paper, and because these results give information about how lateral inhibition works in this network. Thus, we will keep presenting it in the revised version, although we will de-emphasize and simplify its presentation to give more emphasis to the novelty of the derivation of this connectivity rule from the principles of efficient coding.

> *(3) The model's limitations are hard to discern, being relegated to the manuscript's last and rather equivocal paragraph. For instance, the lack of recurrent excitation, crucial in neural dynamics and computation, likely influences the results: neuronal time constants must be as large as the target readout (Figure 4), presumably because the network cannot integrate the signal without recurrent excitation. However, this and other results are not presented in tandem with relevant caveats.*

We will improve the Limitations paragraph in Discussion, and also anticipate caveats in tandem with results when needed, as suggested.

> *(4) On repeated occasions, results from the model are referred to as predictions claimed to match the data. A prediction is a statement about what will happen in the future - but most of the "predictions" from the model are actually findings that broadly match earlier experimental results, making them "postdictions".*

This distinction is important: compared to postdictions, predictions are a much stronger test because they are falsifiable. This is especially relevant given (my impression) that key parameters of the model were tweaked to match the data.

We will better distinguish between pre- and post-dictions in revision.

> ***Reviewer #2 (Public Review):***
>
> *Summary: In this work, the authors present a biologically plausible, efficient E-I spiking network model and study various aspects of the model and its relation to experimental observations. This includes a derivation of the network into two (E-I) populations, the study of single-neuron perturbations and lateral-inhibition, the study of the effects of adaptation and metabolic cost, and considerations of optimal parameters. From this, they conclude that their work puts forth a plausible implementation of efficient coding that matches several experimental findings, including feature-specific inhibition, tight instantaneous balance, a 4 to 1 ratio of excitatory to inhibitory neurons, and a 3 to 1 ratio of I-I to E-I connectivity strength. It thus argues that some of these observations may come as a direct consequence of efficient coding.*
>
> *Strengths:*
>
> *While many network implementations of efficient coding have been developed, such normative models are often abstract and lacking sufficient detail to compare directly to experiments. The intention of this work to produce a more plausible and efficient spiking model and compare it with experimental data is important and necessary in order to test these models.*
>
> *In rigorously deriving the model with real physical units, this work maps efficient spiking networks onto other more classical biophysical spiking neuron models. It also attempts to compare the model to recent single-neuron perturbation experiments, as well as some long-standing puzzles about neural circuits, such as the presence of separate excitatory and inhibitory neurons, the ratio of excitatory to inhibitory neurons, and E/I balance. One of the primary goals of this paper, to determine if these are merely biological constraints or come from some normative efficient coding objective, is also important.*
>
> *Though several of the observations have been reported and studied before (see below), this work arguably studies them in more depth, which could be useful for comparing more directly to experiments.*

Thanks for these insights and for the kind words of appreciation of the strengths of our work.

> *Weaknesses:*
>
> *Though the text of the paper may suggest otherwise, many of the modeling choices and observations found in the paper have been introduced in previous work on efficient spiking models, thereby making this work somewhat repetitive and incremental at times. This includes the derivation of the network into separate excitatory and inhibitory populations, discussion of physical units, comparison of voltage versus spike-timing correlations, and instantaneous E/I balance, all of which can be found in one of the first efficient spiking network papers (Boerlin et al. 2013), as well as in subsequent papers. Metabolic cost and slow adaptation currents were also presented in a previous study (Gutierrez & Deneve 2019). Though it is perfectly fine and reasonable to build upon these previous studies, the language of the text gives them insufficient credit.*

We will improve the text to make sure that credit to previous studies is more precisely and more clearly given.

> *Furthermore, the paper makes several claims of optimality that are not convincing enough, as they are only verified by a limited parameter sweep of single parameters at a time, are unintuitive and may be in conflict with previous findings of efficient spiking networks. This includes the following. Coding error (RMSE) has a minimum at intermediate metabolic cost (Figure 5B), despite the fact that intuitively, zero metabolic cost would indicate that the network is solely minimizing coding error and that previous work has suggested that additional costs bias the output. Coding error also appears to have a minimum at intermediate values of the ratio of E to I neurons (effectively the number of I neurons) and the number of encoded variables (Figures 6D, 7B). These both have to do with the redundancy in the network (number of neurons for each encoded variable), and previous work suggests that networks can code for arbitrary numbers of variables provided the redundancy is high enough (e.g., Calaim et al. 2022). Lastly, the performance of the E-I variant of the network is shown to be better than that of a single cell type (1CT: Figure 7C, D). Given that the E-I network is performing a similar computation as to the 1CT model but with more neurons (i.e., instead of an E neuron directly providing lateral inhibition to its neighbor, it goes through an interneuron), this is unintuitive and again not supported by previous work. These may be valid emergent properties of the E-I spiking network derived here, but their presentation and description are not sufficient to determine this.*

We are addressing this issue in two ways. First, we will present results of joint sweeps of variations of pairs of parameters whose joint variations are expected to influence optimality in a way that cannot be understood varying one parameter at a time. Namely we plan to vary jointly the noise intensity and the metabolic constant, as well as the ratio of E to I neuron numbers and the ratio of mean I-I to E-I connectivity. Second, we will individuate a reasonable/realistic range of possible variations of each individual parameter and then perform a Monte Carlo search for the optimal point within this range, and compare the so-obtained results with those obtained from the understanding gained from varying one or two parameters at a time. We will also add the suggested citation to Calaim et al. 2022 in regard to the points discussed above.

We will improve the comparison between the Excitatory-Inhibitory and the 1-Cell-Type model (see reply to the suggestions of Referee 3 for more details).

> *Alternatively, the methodology of the model suggests that ad hoc modeling choices may be playing a role. For example, an arbitrary weighting of coding error and metabolic cost of 0.7 to 0.3, respectively, is chosen without mention of how this affects the results. Furthermore, the scaling of synaptic weights appears to be controlled separately for each connection type in the network (Table 1), despite the fact that some of these quantities are likely linked in the optimal network derivation. Finally, the optimal threshold and metabolic constants are an order of magnitude larger than the synaptic weights (Table 1). All of these considerations suggest one of the following two possibilities. One, the model has a substantial number of unconstrained parameters to tune, in which case more parameter sweeps would be necessary to definitively make claims of optimality. Or two, parameters are being decoupled from those constrained by the optimal derivation, and the optima simply corresponds to the values that should come out of the derivation.*

In the previously submitted manuscript we presented both the encoding error and the metabolic cost separately as a function of the parameters, so that readers could get an understanding of how stable optimal parameters would be to the change of the relative weighting of encoding error and metabolic cost. We will improve this work by adding the suggested calculations to provide quantitative measures of the dependence of the optimal network parameters and configurations on this relative weighting.

> ***Reviewer #3 (Public Review):***
>
> *Summary: In their paper the authors tackle three things at once in a theoretical model: how can spiking neural networks perform efficient coding, how can such networks limit the energy use at the same time, and how can this be done in a more biologically realistic way than previous work?*
>
> *They start by working from a long-running theory on how networks operating in a precisely balanced state can perform efficient coding. First, they assume split networks of excitatory (E) and inhibitory (I) neurons. The E neurons have the task to represent some lower dimensional input signal, and the I neurons have the task to represent the signal represented by the E neurons. Additionally, the E and I populations should minimize an energy cost represented by the sum of all spikes. All this results in two loss functions for the E and I populations, and the networks are then derived by assuming E and I neurons should only spike if this improves their respective loss. This results in networks of spiking neurons that live in a balanced state, and can accurately represent the network inputs.*
>
> *They then investigate in-depth different aspects of the resulting networks, such as responses to perturbations, the effect of following Dale's law, spiking statistics, the excitation (E)/inhibition (I) balance, optimal E/I cell ratios, and others. Overall, they expand on previous work by taking a more biological angle on the theory and showing the networks can operate in a biologically realistic regime.*
>
> *Strengths:*
>
> *(1) The authors take a much more biological angle on the efficient spiking networks theory than previous work, which is an essential contribution to the field.*
>
> *(2) They make a very extensive investigation of many aspects of the network in this context, and do so thoroughly.*
>
> *(3) They put sensible constraints on their networks, while still maintaining the good properties these networks should have.*

Thanks for this summary and for these kind words of appreciation of the strengths of our work.

> *Weaknesses:*
>
> *(1) The paper has somewhat overstated the significance of their theoretical contributions, and should make much clearer what aspects of the derivations are novel. Large parts were done in very similar ways in previous papers. Specifically: the split into E and I neurons was also done in Boerlin et al (2008) and in Barrett et al (2016). Defining the networks in terms of realistic units was already done by Boerlin et al (2008). It would also be worth it to discuss Barrett et al (2016) specifically more, as there they also use split E/I networks and perform biologically relevant experiments.*

We will improve the text to make sure that credit to previous studies is more precisely and more clearly given.

> *(2) It is not clear from an optimization perspective why the split into E and I neurons and following Dale's law would be beneficial. While the constraints of Dale's law are sensible (splitting the population in E and I neurons, and removing any non-Dalian connection), they are imposed from biology and not from any coding principles. A discussion of how this could be done would be much appreciated, and in the main text, this should be made clear.*

We indeed removed non-Dalian connections because having only connections respecting Dale's law is a major constraint for biological plausibility. Our logic was to consider efficient coding within the space of networks that satisfy this (and other) biological plausibility constraints. We did not intend to claim that removing the non-Dalian connections was the result of an analytical optimization. However, to get better insights into how Dale's Law constrains or influences the design of efficient networks, we added a comparison of the coding properties of networks that either do or do not satisfy Dale's law. We apologize if this was not sufficiently clear in the previous version and we will clarify this in revision.

> *(3) Related to the previous point, the claim that the network with split E and I neurons has a lower average loss than a 1 cell-type (1-CT) network seems incorrect to me. Only the E population coding error should be compared to the 1-CT network loss, or the sum of the E and I populations (not their average). In my author recommendations, I go more in-depth on this point.*

We will perform the suggested detailed comparisons between the network loss in the 1CT-model and E-I model and then revise or refine conclusions if and as needed, according to the results we will obtain.

> *(4) While the paper is supposed to bring the balanced spiking networks they consider in a more experimentally relevant context, for experimental audiences I don't think it is easy to follow how the model works, and I recommend reworking both the main text and methods to improve on that aspect.*

We will try to make the presentation of the model more accessible to a non-computational audience.

> *Assessment and context: Overall, although much of the underlying theory is not necessarily new, the work provides an important addition to the field. The authors succeeded well in their goal of making the networks more biologically realistic, and incorporating aspects of energy efficiency. For computational neuroscientists, this paper is a good example of how to build models that link well to experimental knowledge and constraints, while still being computationally and mathematically tractable. For experimental readers, the model provides a clearer link between efficient coding spiking networks to known experimental constraints and provides a few predictions.*

Thanks for these kind words. We will make sure that these points emerge more clearly and in a more accessible way from the revised paper.

https://doi.org/10.7554/eLife.99545.1.sa0