# Constrained brain volume in an efficient coding model explains the fraction of excitatory and inhibitory neurons in sensory cortices

Arish Alreja[1], Ilya Nemenman[2], Christopher Rozell[3]

[1] Neuroscience Institute, Center for the Neural Basis of Cognition and Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA | [2] Department of Physics, Department of Biology and Initiative in Theory and Modeling of Living Systems, Emory University, Atlanta, GA 30322, USA | [3] School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA

## Abstract

The number of neurons in mammalian cortex varies by multiple orders of magnitude across different species. In contrast, the ratio of excitatory to inhibitory neurons (E:I ratio) varies in a much smaller range, from 3:1 to 9:1 and remains roughly constant for different sensory areas within a species. Despite this structure being important for understanding the function of neural circuits, the reason for this consistency is not yet understood. While recent models of vision based on the efficient coding hypothesis show that increasing the number of both excitatory and inhibitory cells improves stimulus representation, the two cannot increase simultaneously due to constraints on brain volume. In this work, we implement an efficient coding model of vision under a volume (i.e., total number of neurons) constraint while varying the E:I ratio. We show that the performance of the model is optimal at biologically observed E:I ratios under several metrics. We argue that this happens due to tradeoffs between the computational accuracy and the representation capacity for natural stimuli. Further, we make experimentally testable predictions that 1) the optimal E:I ratio should be higher for species with a higher sparsity in the neural activity and 2) the character of inhibitory synaptic distributions and firing rates should change depending on E:I ratio. Our findings, which are supported by our new preliminary analyses of publicly available data, provide the first quantitative and testable hypothesis based on optimal coding models for the distribution of neural types in the mammalian sensory cortices.

## Introduction

Neural circuits are responsible for a wide variety of tasks, including encoding and processing sensory information. Understanding the design principles as well as the functional computations in such circuits has been a foundational challenge of neuroscience, with potential applications to a wide variety of fields ranging from human health to artificial intelligence. However, the structural complexity and dynamic response properties of these circuits present significant challenges to uncovering their fundamental governing principles. Some of the brain's structural properties are extremely variable across species and individuals [1], while properties such as the structure of cortical microcircuits seem to be reasonably conserved [2–5]. These conserved properties offer hope of revealing general principles of how canonical neural computations are organized.

While multiple experimental [6–13] and computational [14–16] studies have offered insights about inhibitory interneurons at different scales, their precise computational role in sensory information processing remains elusive. The relative abundance of excitatory and inhibitory neurons in primary sensory areas appears to be one of the better conserved structural properties of cortical microcircuits, and this conserved circuit structure should be an important clue for determining neural circuit function. For example, despite wide variations over several orders of magnitude in the total number of neurons across species and sensory cortical areas, morphological studies indicate that the ratios of excitatory to inhibitory neurons (E:I ratio) stay within a relatively narrower nominal range of 3:1–9:1 (i.e., inhibitory interneurons are 10% – 25% of the neural population) across species and are relatively consistent across sensory cortical areas within species (Table 1) [17–27] even when other cortical areas show variations (e.g., motor cortex [28] or medial prefrontal cortex [29]).

| | E:I Ratio/Area | | | # of Neurons/Area | | |
|---|---|---|---|---|---|---|
| Species | A1 | V1 | S1 | A1 | V1 | S1 |
| Rodents | 5.3-7.3:1[18] | 5.7-9:1[19–21] | 5.7-7.7:1[20, 30] | $10^5$[31] | $10^5$[31] | $10^5$[31] |
| Cat | 3:1[22] | 4:1[23, 24] | 2.4-3.2:1[25] | Not reported | $10^7$ [32] | Not reported |
| Primate | Not reported | 4-4.3:1 [26, 27] | 3.1-3.9:1 [27] | $10^6$[33] | $10^7$ [33] | $10^6$[33] |

Table 1: E:I ratios and # of Neurons in primary auditory (A1), visual (V1) and somatosensory (S1) cortices for different species from morphological studies.

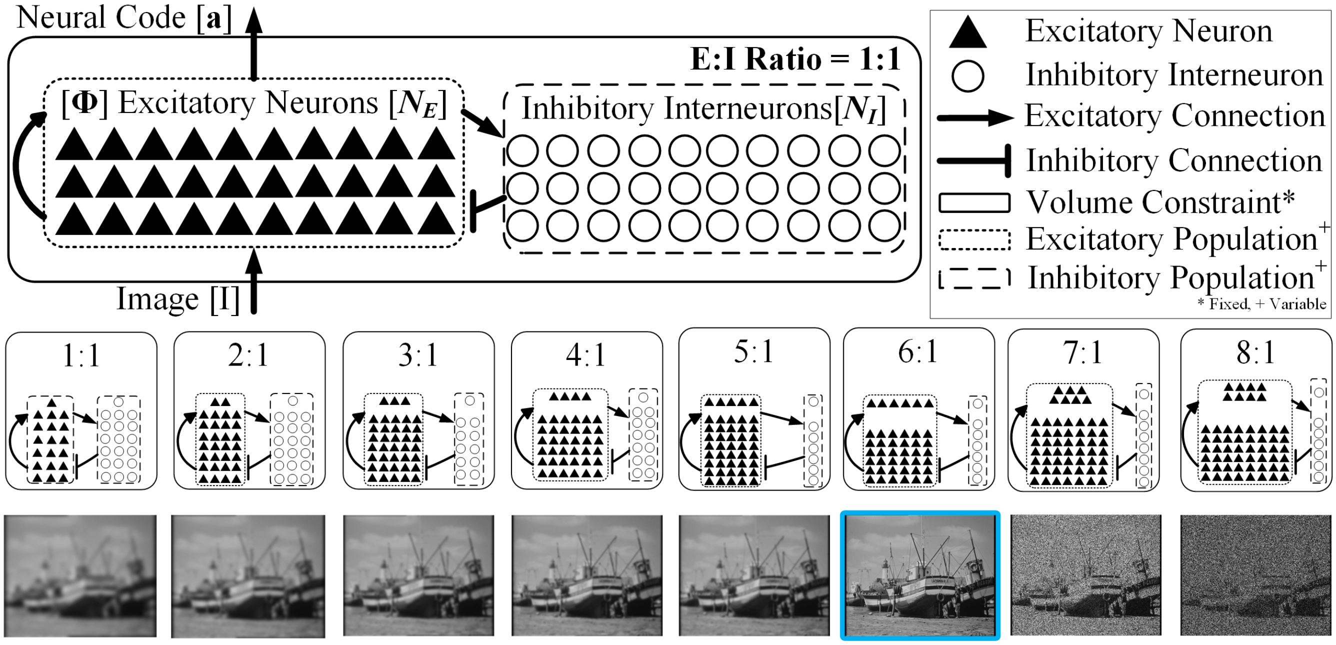

Figure 1: **Optimal E:I ratio for coding fidelity: (top row)** A sparse coding model is placed under a volume constraint by restricting the total number of neurons to $N$. Excitatory neurons receive recurrent as well as feed forward (stimulus) input and are responsible for coding the stimulus. Inhibitory interneurons are driven by recurrent excitatory inputs, and enable accurate computation of the neural encoding to induce sparsity in the excitatory neurons. **(middle row)** We vary the relative size of the excitatory ($N_\mathrm{E}$) and inhibitory ($N_\mathrm{I}$) subpopulations and evaluate the model at different E:I ratios under the volume constraint, $N = N_\mathrm{E} + N_\mathrm{I}$. **(bottom row)** We show that coding fidelity is optimal (boxed image at 6:1) at a unique, biologically plausible E:I ratio for the fixed volume. We evaluate models coding $16 \times 16 = 256$ pixel natural image patches [34] with $N = 1200$ ($\approx 5\times$ overcomplete representation).

This relative constancy of the E:I ratio must be understood within the context of sensory computations. Inhibitory interneurons in sensory cortical microcircuits have connectivity patterns contained within local circuits [2, 4], leading to inhibitory cells being generally viewed as performing a modulatory role in computation while excitatory cells code the sensory information directly [5]. For a given sensory cortical area, there are potential computational benefits to increasing the size of both the excitatory and the inhibitory subpopulations. For example, more excitatory cells may provide higher fidelity stimulus encoding, while more inhibitory cells may enable more complexity or accuracy in the computations being performed. However, volume is a critical constrained resource for cortical structures [35], and increasing one of these subpopulations in a fixed volume necessitates decreasing the other. We propose that the narrow variability of the E:I ratio can be explained as an optimal trade-off in the fidelity of the sensory representation contained in the excitatory subpopulation vs. the fidelity of the information processing mediated by the inhibitory subpopulation. Understanding this trade-off may play a critical role in determining the principles underlying the structure and function of cortical circuits.

Specifically, we propose to understand this trade-off in the context of efficient coding models [34, 36–38] under a volume constraint. In this initial study, the volume constraint is defined as the total number of neurons and does not explicitly model either volume differences by cell type or non-somatic elements such as axons and dendrites (though those extensions could be added in the future). We implement an efficient coding model known as sparse coding [34, 39], which uses recurrent circuit computations to encode a stimulus in the excitatory cell activities (denoted $a_j$) using as few excitatory neurons as possible (i.e., having high population sparsity). In detail, the sparse coding model proposes encoding a stimulus (e.g., an image) $I$ in terms of the sum of the activity $a_j$ of excitatory neurons with receptive fields $\phi_j$, by minimizing a cost function that balances representation error (i.e., fidelity) with the sparsity of the neural population activity:

$$\text{Cost} = \underbrace{||I - \textstyle\sum_j \phi_j \, a_j||_2^2}_{\text{Representation Error (MSE)}} + \lambda \underbrace{\textstyle\sum_j |a_j|}_{\text{Sparsity}} . \qquad (1)$$

Note that the population sparsity constraint only includes excitatory cells and does not include the activity of the inhibitory cells necessary to enact the required computation (i.e., solve the optimization program). Sparse coding models have been shown to account for many observed response properties of the visual cortex [34, 39, 40] and can be implemented in biophysically plausible recurrent circuits [41, 42] with a desired sparsity level and a given E:I ratio [14, 15] (optimally approximating the ideal circuit implementation). See *Methods* for details. Recent work has shown also that increasing the population of excitatory [15, 43] and inhibitory [15] cell types in sparse coding models can improve stimulus representation in models where the size of neural populations is unrestricted.

Here we show that, for a fixed neural population size (representing a volume constraint), there exists an optimal E:I ratio where the stimulus representation, the sparseness of the sensory representation, and the metabolic efficiency of the entire network are all optimized in the model. This model-optimal E:I ratio is consistent with observed biophysical ranges and it varies based on the sparsity level of the encoding, potentially accounting for species specific variations within the observed biophysical ranges. Furthermore, higher optimal E:I ratios (at higher sparsity levels) produce inhibitory synaptic distributions that are more specific while approximately preserving the total inhibitory influence in the circuit (to retain balanced levels of excitation and inhibition). These results constitute specific and testable theoretical predictions requiring comparative neurophysiology and neuroanatomy experiments for full validation. We also perform novel analyses of experimental recordings of neural populations in area V1 for multiple species (mice, cats and monkeys), constituting the first steps in comparative analyses of population sparsity in large-scale electrophysiology recordings. The results of this analysis are consistent with the model prediction of a correlation between E:I ratio and population sparsity level. Taken together, these results suggest that a combination of optimal coding models with physical constraints (e.g., volume) may provide a potential normative explanation for conserved structures observed in sensory cortical microcircuits across species.

# Results

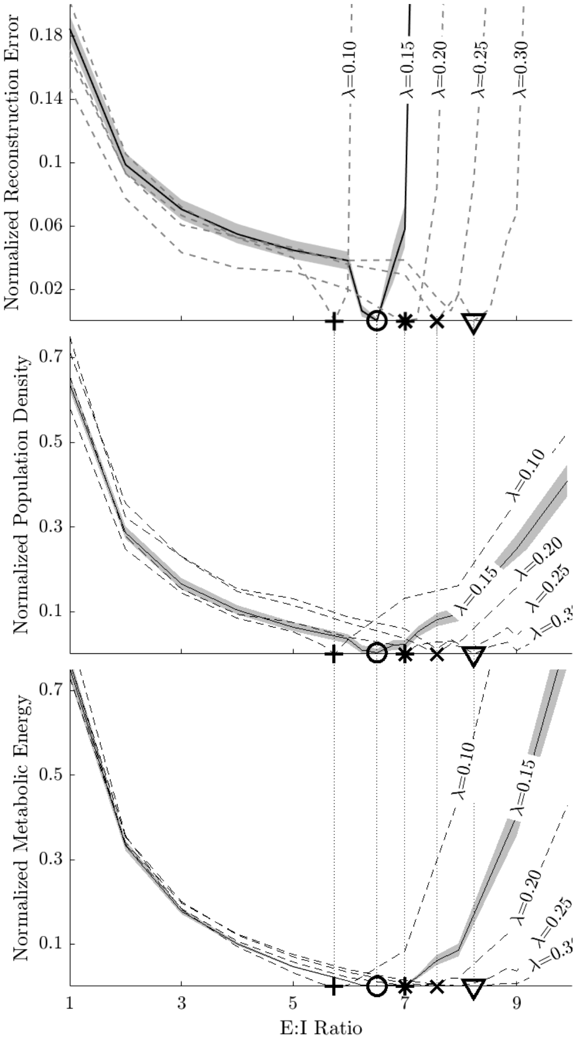

Figure 2: **Optimal E:I ratios for multiple performance measure coincide and increase as sparsity ($\lambda$) increases:** The performance of sparse coding models subject to a volume constraint of $N = 1200$ neurons and under different sparsity constraints ($\lambda \in [0.10, 0.30]$) and using stimuli (100 image patches, 16 x 16 pixels) drawn from a database of 10 natural 512 x 512 pixels images [34]. Performance measures are normalized per Equation 8 and standard error (depicted for $\lambda = 0.15$ with a shaded band) over the natural image database is estimated using a bootstrap procedure (see *Supplemental Methods*). Markers denote the optimal E:I ratio for models at each sparsity constraint for each performance measure. Optimal E:I ratios for different performance measures are essentially identical as illustrated by vertical lines connecting markers across the 3 plots, and increases in model sparsity ($\lambda$) correspond to increases the optimal E:I ratio for each performance measure.**(top)** The coding fidelity for a sparse coding models with different sparsity constraints quantified by the normalized reconstruction error. The coding performance is optimized at an E:I ratio of approximately 6.5:1 (in a biophysically plausible range), with values above (below) that number suffering from lack of diversity in the inhibitory (excitatory) cell population. **(middle)** Population Activity Density (1 - Population Sparsity) for a sparse coding model (see *Methods*) is minimized at nearly the same specific optimal E:I ratio as with coding fidelity. **(bottom)** Lastly, a metabolic energy consumption measure [44] (see *Methods*) reveals minimal metabolic energy consumption at nearly the same specific E:I ratio as with coding fidelity and population density.

We analyze sparse coding models optimized for a variety of E:I ratios (i.e., the ratio of the number of excitatory cells to inhibitory cells while fixing the total number of neurons) and sparsity levels (denoted by model parameter $\lambda$) by unsupervised training using a natural image database [34]. See *Methods* for details. The performance of these models is quantified using stimulus reconstruction error, population sparsity [45], and metabolic energy consumption [44].

For a sparse coding model trained with the sparsity constraint $\lambda$=0.15, we observe that the reconstruction error is minimized at the ratio of $\sim 6.5 : 1$ (Fig. 2 (top)). The reconstruction error is a surrogate measure of the fidelity of the stimulus information preserved in the encoding. As the E:I ratio increases from 1:1, the increase in E cells leads to greater receptive field diversity in the E cell subpopulation [15, 43], allowing for better encoding of the stimulus. This increased representational capacity produces a gradual decline in the reconstruction error. As the E:I ratio increases beyond the optimum, the declining number of inhibitory interneurons results in insufficiently diverse inhibition to accurately solve the desired encoding, leading to a rapid increase in the reconstruction error. Results are independent of the size of the used database (10 images with 512 x 512 pixels each) used for training (See Fig. S2 and *Supplemental Methods*). The tolerance in calculating the optimal reconstruction error was negligible compared to the changes in the error due to varying the E:I ratio.

Efficient coding models seek a parsimonious representation of sensory inputs in the excitatory neural activity in addition to an accurate encoding. To quantify this parsimony, we plot the density of activity of excitatory neurons in the sparse coding model (Fig. 2 (middle)), as measured by population density, an additive inverse of the commonly used modified Treves-Rolls (TR) metric [45] that quantifies population sparsity (see *Methods*). Notably, the population activity density is minimized (i.e., population sparsity is maximized) at approximately the same E:I ratio that optimizes reconstruction fidelity. At low E:I ratios, the stimulus representation is not rich enough to admit a sparse representation of natural scene statistics with available receptive fields of excitatory cells. With high E:I ratios, the available inhibition is insufficient to achieve sparse population activity in the excitatory cells.

A common rationale for the efficient coding hypothesis (including sparse coding models) is that efficient codes may reduce the metabolic cost of the neural activity [46–49]. While decreasing the mean firing rate of excitatory neurons would decrease the metabolic cost of producing action potentials in those cells, it is not clear which network architecture minimizes the total metabolic energy consumption when accounting for the cost of supporting the non-sparse activity of the inhibitory interneurons [9, 50, 51]. We quantify and plot (Fig. 2 (bottom)) the total metabolic energy cost of the network (see *Methods*). Once again, the optimal E:I ratio achieving minimal energy consumption for different sparsity constraints is approximately the same E:I ratio that optimizes reconstruction fidelity and excitatory population sparsity.

When varying the sparsity constraint across a wide range $\lambda = [0.1, 0.3]$, we observe that all three performance measures (reconstruction error, population sparsity, metabolic energy consumption) demonstrate an optimal E:I ratio that is consistent across metrics. This indicates that there is a clear optimal E:I ratio for a given sparsity level that is robust to the choice of optimality criteria. Crucially, we observe that increasing the model sparsity ($\lambda$) leads to a higher optimal E:I ratio in all three metrics (Fig. 2).

While networks optimized for different sparsity levels have different optimal E:I cell type ratios, it is unclear if either the synaptic distribution (a structural measure) or the total amount of inhibitory activity (a functional measure) change as well. To understand potential structural changes, we first examined the structural nature of the inhibitory interactions in the recurrent network at different sparsity levels ($\lambda$) and optimal E:I ratios. We observe that there are systematic changes in the distribution of weights for Inhibitory→Excitatory connections (Fig. 3 (middle row - left)) as $\lambda$ changes. In particular, lower sparsity levels ($\lambda$) corresponding to lower optimal E:I ratios result in inhibitory synapse distributions that have heavier tails and higher kurtosis (Fig. 3 (middle row - right)). Therefore, at lower E:I ratios when there are relatively more inhibitory interneurons in the circuit, the individual interneurons have more targeted projections to deliver inhibition more selectively to shape excitatory activity (Fig. 3 (top row), see also Fig. S3 (left) and (right)).

Functionally, the total amount of inhibitory influence in a circuit is a combination of the spiking activity in the inhibitory interneurons and the total strengths of the synapses from inhibitory to excitatory neurons. We next examined the inhibitory activity in the recurrent network at different sparsity levels ($\lambda$) and optimal E:I ratios. We observe that lower $\lambda$ corresponding to lower optimal E:I ratios result in higher average activity levels per cell (with higher standard deviations) across the relatively larger inhibitory subpopulation (Fig. 3 (bottom row - left)). Despite significant changes in the synaptic structure and firing rates of inhibitory interneurons as $\lambda$ (and the E:I cell type ratio) changes, the total amount of inhibitory influence in the network does not change substantially (Fig. 3 (bottom row - right), S4). Specifically, as $\lambda$ increases, the reduction in inhibitory subpopulation size and firing rates is offset by the broader tuning of the inhibitory synapses so that the balance between total excitation and inhibition in the network remains relatively constant in a stable regime (Fig. S3 (middle)).

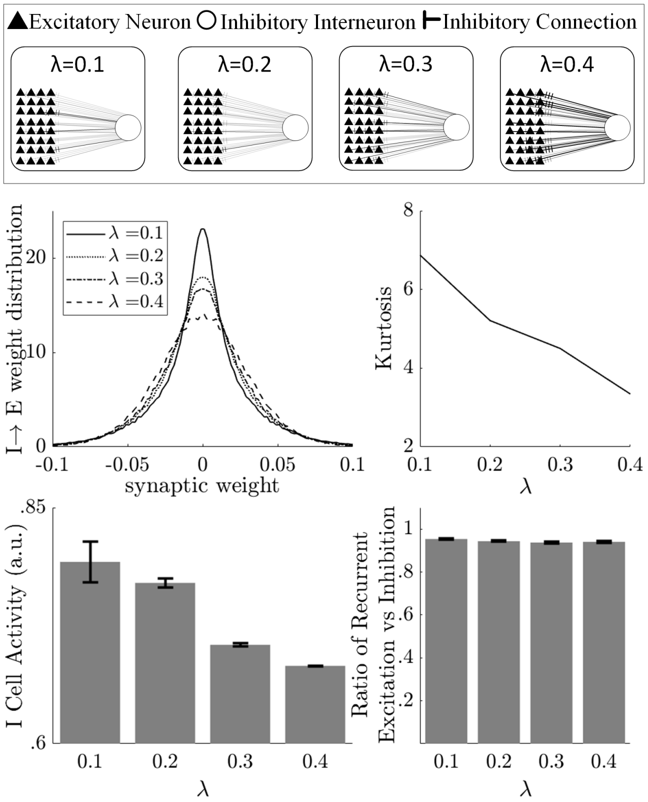

Figure 3: **Structure and function of model inhibition change with sparsity: (top row)** An illustration visualizing the impact of the changing weight distributions on an inhibitory interneuron. **(middle row) (left)** Estimated probability density functions for the inhibitory to excitatory connection weights in the optimal computational models at different sparsity levels reveal an increasing fraction of inhibitory synapses are stronger as sparsity increases. **(middle row) (right)** Estimated kurtosis vs sparsity quantifies the changes visible in the distributions, demonstrating that inhibition is more targeted and less global at lower sparsity levels with smaller E:I ratios. **(bottom row) (left)** With increasing sparsity (corresponding to higher optimal E:I ratios), the inhibitory subpopulation's mean activity level declines and becomes less diverse (exhibiting a lower standard deviation). **(bottom row) (right)** Despite the changes in inhibitory structure and function due to changes in sparsity level (and optimal E:I cell type ratio), the changes to inhibitory synaptic distributions and firing rates counteract each other so that the total inhibitory influence in the network remains constant and the circuit maintains balance between the recurrent excitatory and inhibitory activity.

While theoretical modeling often assumes that the sparsity level of an efficient coding model is an unknown parameter that can be fit to data, the analysis above predicts that optimal efficient coding networks should have E:I ratios correlated with population sparsity (Fig. 4). Unfortunately, despite sporadic characterizations of population sparsity reported in the literature (with different data types and analysis methods), we lack a comparative analysis of population sparsity across species. In new analyses of recent publicly available datasets comprised of large-scale V1 electrophysiology recordings, we evaluated the sparsity in population activity in mice [52], monkeys [53, 54] and cats [55] studies featuring natural visual stimuli (movies, images). The similarly low sparsity levels observed in monkeys and cats (both having E:I = 4:1) as well as their contrast with higher sparsity levels in mice (E:I = 5.7-9:1) are consistent with the predictions of the efficient coding model in this study. Specifically, using a hierarchical bootstrap procedure [56] (See *Methods, Supplemental Methods* and Fig. S1) to compare population sparsity for different species, we observed (Fig. 4 (bottom row) that mice have much higher population sparsity (lower density) than monkeys and cats when viewing natural movies ($p_{bootstrap} < 10^{-8}$). Similarly, mice exhibit higher population sparsity than monkeys ($p_{bootstrap} = 0.01966$) in response to natural images.

# Discussion

Using a sparse coding model for early vision and a volume constraint, we showed that the quality and efficiency of stimulus encoding is optimal at E:I ratios consistent with the narrow range observed in biological neuroanatomy. Increasing the E:I ratio improves the representational capacity of the E cell subpopulation through the potential for greater receptive field diversity [15, 43], but at the expense of reducing the ability of the I cells to produce accurate circuit computations to implement the encoding rule. Decreasing the E:I ratio has an opposite effect, increasing the I cells available to improve computational accuracy for the encoding rule at the expense of the representational capacity of the E cell subpopulation, whose receptive field diversity shrinks, diminishing its ability to represent rich sensory statistics.

This model makes several predictions that are testable with comparative electrophysiology experiments. The primary result of this study predicts that the optimal E:I ratio is directly correlated with population sparsity, such that sparser population activity in a species will correspond to a higher E:I cell type ratio (Fig. 4). In secondary results, this model also predicts that species with higher sparsity levels will have inhibitory interneuron subpopulations with both lower average firing rates that are more concentrated around the mean and higher kurtosis of the synaptic distribution than species with lower sparsity levels. These predictions are notable because it is rare for computational theories to make specific and measurable predictions about the relationship between functional and morphological properties of neural systems.

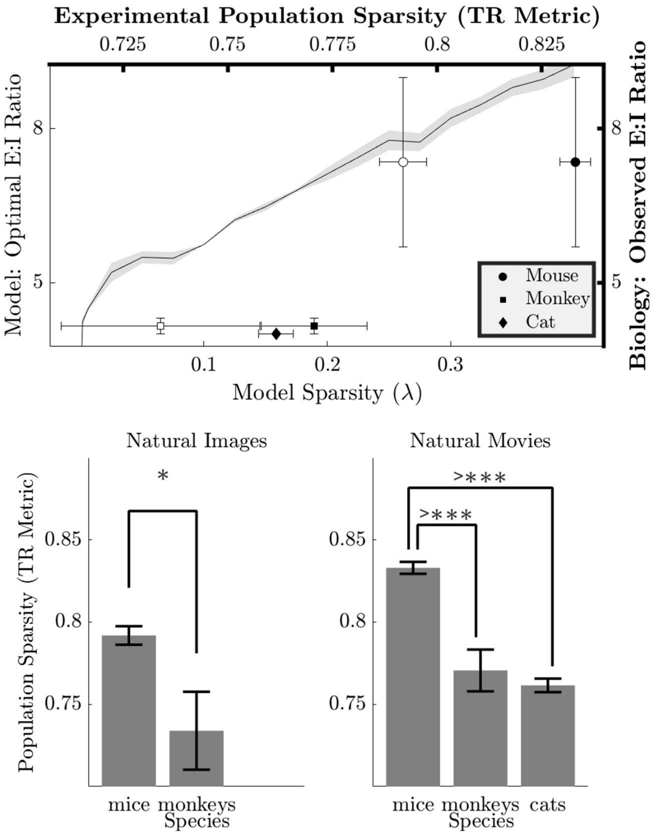

Figure 4: **Model predictions vs experimental data (Top Row): (Left Y and Bottom X axes)** Optimal E:I ratio based on normalized reconstruction error (See Fig. S5 for other performance measures) as a function of model sparsity constraint $\lambda$ is depicted by the solid line (mean) with variability (± standard error) denoted by the shaded band. **(Right Y and Top X axes)** The population sparsity (TR) measure computed for electrophysiology data from experimental studies in mice [52], monkeys [53, 54] and cats [55] is shown (mean (markers) ± standard error (horizontal error bars)) as a function of observed E:I ratio ranges in biology (vertical error bars). Unfilled markers represent natural images and black filled markers represent natural movies. **Interspecies comparisons (Bottom Row)** Statistical significance of hypotheses based on model prediction (i.e., higher E:I ratio in biology corresponds to higher population sparsity) examined via inter-species population sparsity comparisons with all available data using hierarchical boostrapping. **(left)** For natural images, the mice (E:I = 5.7-9:1) exhibit higher population sparsity compared to monkeys (E:I = 4-4.3:1), $p_{bootstrap}$ = 0.01966. **(right)** For natural movies, mice (E:I = 5.7-9:1) exhibit higher population sparsity than both monkeys (E:I = 4-4.3:1) and cats (E:I = 4:1), $p_{boostrap} < 10^{-8}$ for both, which is significant after accounting for multiple comparisons.

The result that networks with a higher level of population sparsity in the excitatory subpopulation are optimized with fewer inhibitory neurons (i.e., higher E:I ratio) may appear counter-intuitive given the apparent need for increased inhibition to achieve higher sparsity. However, a closer look at the specific structure in the inhibitory synaptic distribution (see *Methods* and *Supplementary Information*) provides some insight into this result. Models having higher population sparsity learn to represent natural stimuli differently from models at lower population sparsity. Specifically, in models with higher population sparsity, the smaller inhibitory subpopulation contains cells that have relatively lower firing rates and global synaptic connections, indicating inhibition that is more broadly tuned and less selective than in models with lower population sparsity. This model prediction is consistent with the contrast observed in experimental results from cats (E:I = 4:1) [57] and mice (E:I=5.7-9:1) [58]. In contrast, models at lower population sparsity have inhibitory interneurons with relatively higher firing rates and synapses that are targeted to specific excitatory sub-populations (Fig. 3). We note that while we discuss inhibitory interneurons generally here, we have not attempted to correspond the inhibitory components of the model to a specific genetic subtype of inhibitory interneuron. Future experimental tests of the predictions from this model can and should address the empirical question of which inhibitory interneuron subtypes are the best fit to the inhibitory influences of this model.

To perform a preliminary evaluation of this model prediction with data that is currently available, we analyzed population sparsity in area V1 of mice [52], monkeys [53, 54] and cats [55] using publicly available electrophysiology data sets. We found that the population sparsity trends revealed by this analysis agree with the broad predictions made by the model. Specifically, for a given stimulus type, species with higher E:I ratios demonstrated higher population sparsity levels. To our knowledge this is the first comparative analysis of population sparsity across species, providing valuable insight for future computational and theoretical work beyond the specific predictions of this model.

Despite this apparent agreement between experimental data sets and model predictions, the predicted correlation between optimal E:I ratio and population sparsity is challenging to thoroughly evaluate empirically because the literature currently lacks the necessary reports to provide a substantive comparative analysis of population sparsity between species. The large scale populations recordings necessary to evaluate sparsity have only become possible relatively recently, and comparability of existing studies is often hampered by differences such as recording methodology, experimental conditions (e.g. type and quantity of anaesthesia administered), brain area, number of subjects, number and type of neurons, stimuli, and analysis parameters (e.g. window size substantially influences sparsity measures). The data we analyzed come from experiments whose design was not aimed at facilitating comparisons like those made in this study, and experiments that control for these sources of variability may allow for more robust evaluation of our (and future) model predictions. As an example, population recordings analyzed in this study featured lightly anesthetized mice [52] compared to heavily anesthetized and paralyzed monkeys [53, 54] and cats [55]. Since anesthesia is known to depress neural activity [59–61], we anticipate population sparsity for monkeys/cats is elevated. This bias would make it more difficult to observe the significant differences in sparsity level reported in this study, so it is unlikely to be a major confound in our analysis. However, further studies that explore population level activity in different sensory areas or under different experimental conditions may support/refute whether our model predictions apply more generally.

To illustrate the challenges with making comparative meta-analyses from data that was not collected for that purpose, we note that in addition to the data supporting the model predictions above, we have also encountered a limited number of contrasting exceptions that have known confounds that highlight the subtleties in such comparative analyses. For example, one study [62] captures V1 responses to natural stimuli in ferret and reports population sparsity (TR = 0.42) much lower than cats and monkeys despite a higher E:I ratio of 5:1 [63]. However, this study self-identifies a critical methodological issue that likely resulted in overestimated firing rates due to the use of multi-unit signals instead of isolated single units to compute sparseness, deflating the estimated population sparsity. For another example, [64] captures population sparsity in mouse V1 using spike trains estimated from calcium imaging and reports a lower population sparsity (TR = 0.45-0.55) than a recent calcium imaging study [65] (TR = 0.81), as well results from analysis of electrophysiology data from mice presented in this paper. Closer examination of this inconsistency reveals that [65] features specific targeting of excitatory neurons only while [64] does not employ cell-specific targeting, which can deflate population sparsity estimates due to the elevated firing rates of inhibitory interneurons [9, 50, 51]. The confounding effects present in these two conflicting examples from the literature illustrate a number of important methodological issues to be carefully addressed in future experimental work that aims to perform a conclusive comparative analysis.

The results of this study represent an early step toward understanding the connection between optimal coding rules and the diversity sensory cortical structure in mammals. We expect that additional verifiable predictions will be possible when more relevant biological details are introduced into the models. For example, our analysis does not make distinctions between different kinds of inhibitory interneurons and future work may consider their relative contributions when evaluating the trade-off between computational accuracy and representational capacity. Similarly, modeling thalamic input into inhibitory cells may offer greater insight into the role of inhibition beyond modulating computation performed by the excitatory sub-population.

Finally, we note that the shape of the performance curves (Fig. 2) are asymmetric, with performance degrading very quickly at E:I ratios higher than the optima. While normative models can never ensure they are capturing all constraints that drive evolutionary or developmental goals for a system, this asymmetry indicates that the constraints considered here are more robust to decreasing E:I ratios rather than increasing E:I ratios. This prediction is consistent with the (limited) currently available morphological data (Table 1) that shows the distribution of E:I ratios across species is asymmetric and skewed to smaller values around the mode. Additional morphological studies on animal models not listed in Table 1 may provide additional support or refutation of this prediction. More broadly, we expect that close interplay between computational and experimental studies will further advance our ability to merge functional and physical constraints to better understand the relationship between the information processing in the brain and its structure.

# Methods

## Sparse coding model of visual computation

Among neural coding models instantiating the efficient coding hypothesis, we concentrate on the sparse coding model [34] that aims to minimize the number of simultaneously active neurons for each stimulus. This model is sufficient to explain the emergence of classical and nonclassical response properties in V1 [34, 42, 66] and is consistent with recent electrophysiological experiments [67–69]. Furthermore, the sparse coding model can be implemented in recurrent network architectures with varying degrees of biophysical plausibility [38, 42, 70–72], including distinct inhibitory interneuron populations [14, 15].

Specifically, in the sparse coding model, a set of neurons encodes an image intensity field $I(x,y)$ through the vector of activities $a = [a_1, a_2, \dots]$ (i.e., firing rates) by minimizing the so called *energy function*:

$$a = \arg\min_a \sum_\mu \left[ \sum_{x,y} \left[ I(x,y) - \sum_i a_i \phi_i(x,y) \right]^2 + \lambda \sum_i |a_i| \right], \tag{2}$$

where the activity of each neuron $a_i$ is associated with a stimulus feature $\phi_i(x,y)$ (similar to a receptive field), and $\mu = 1 \dots M$ sums over all images in a training set. This energy function uses the scalar parameter $\lambda \in [3.78 \times 10^{-4}, 0.4]$ to balance the preservation of stimulus information (measured by the mean-squared reconstruction error in the first term) with the efficiency of the representation (measured by the sum of the activity magnitudes in the second term). We choose the $L_1$ norm for quantifying the efficiency of the representation since it is known to promote sparsity and is (analytically and computationally) tractable. Higher values of $\lambda$ encourage more sparsity and lower values prioritize the fidelity of the stimulus encoding. As has been shown in the past, optimizing the feature set $\phi_i(x, y)$ for this coding rule using a corpus of natural images will produce a set a features that resemble the measured receptive fields in primary visual cortex [34, 42].

## Dynamical System implementation of the sparse coding model

To encode a specified image, we consider a recurrent dynamical circuit model [70] that provably solves the optimization in Eq. [2] [73, 74] (including alternative sparsity penalties [75]) in non-spiking or spiking [71, 76, 77] network architectures. Specifically, the system dynamics for this encoding model are:

$$\begin{aligned} \dot{u}(t) &= \frac{1}{\tau}\left[\Phi^T I - u(t) - Wa(t)\right], \\ a(t) &= T_\lambda(u(t)), \end{aligned} \tag{3}$$

where $I$ is the vectorized version of the stimulus, $\Phi$ is a matrix with a vectorized version of the dictionary element $\phi_i(x, y)$ in the $i^{th}$ column, the vector $u$ contains internal state variables (e.g., membrane potentials), the vector $a$ contains external activations (e.g., spike rates) of excitatory neurons that represent the stimulus, the matrix $W$ governs the connectivity between the neurons (requiring inhibitory interneurons for implementation), and $T_\lambda(\cdot)$ is a pointwise nonlinear activation function (i.e., a soft thresholding function).

When the recurrent influences in the network are governed by $W = G - D = \Phi^T\Phi - D$, where $G$ is a Grammian matrix and $D$ is the diagonal identity matrix, then the network above is guaranteed to converge to the solution of the sparse coding objective function above [70]. In this case, the required connectivities between the excitatory cells (the principal cells encoding the stimulus) must be mediated by a combination of direct excitatory synapses (negative elements of $G$) and a local population of inhibitory interneurons (positive elements of $G$). Deviations from this network structure may result in more efficient implementations (e. g., requiring fewer inhibitory neurons), but will have the consquence of only approximately solving the desired coding objective.

We seek to form a circuit model that approximates the ideal dynamical system above as closely as possible under a fixed size for the inhibitory interneuron population implementing $G$. To reflect the disynaptic connections onto an inhibitory population and back to the excitatory population, consider the factorization of this connectivity matrix using the singular value decomposition (SVD): $G = U\Sigma V^T$. If we consider only the positive entries in this representation as in [14], each column of $V$ contains the synaptic weights of the connections onto a single inhibitory cell, the corresponding element in the diagonal matrix $\Sigma$ represents a dendritic gain term, and the corresponding column of $U$ represents the synaptic weights from that inhibitory cell back onto the population of excitatory principle cells. Following previous work [14], we can use the truncated SVD to find the closest approximation (in terms of the Frobenius norm) to $G$ with a specified rank, which corresponds to specifying the size of the inhibitory population.

Experimental and computational studies have reported that depending upon factors such as location, timing and magnitude, PSPs arriving at the dendritic tree can produce sub-linear, supra-linear and linear gain at the soma [78, 79]. Interpreting $\Sigma$ as a gain term enables us to incorporate the biologically realistic notion of dendritic gain arising from multiple projections from an inhibitory interneuron to an excitatory neuron, into an otherwise abstract circuit model limited to representing a single projection. Under this interpretation, we estimate the activity of inhibitory interneurons as $b = Va$.

## Volume Constraint Implementation

For this study, we represent 16x16 pixel image patches using $N = N_{\mathrm{I}} + N_{\mathrm{E}} = 1200$ total neurons to correspond to a fixed volume constraint (implicitly assuming approximately constant volume per neuron). For each E:I ratio tested, we trained a dictionary using natural images [34] for a dictionary optimized for $N_E$ excitatory cells. After training the dictionary, we implemented the dynamical system described above with the best approximation to the ideal circuit dynamics using $N_I$ inhibitory cells.

In addition to evaluating the model at different E:I ratios, we also trained and evaluated models under different sparsity constraints ($\lambda$). For a given sparsity constraint ($\lambda$) and E:I ratio, we evaluate the network over an image patch database [39] using three different performance measures.

## Performance measures

The first performance measure quantifies the coding fidelity of the model for the reconstruction $\widehat{I} = \sum_i a_i \phi_i$ of an image $I$ encoded by the model. The stimulus reconstruction error is formulated as:

$$\text{Reconstruction Error} = \frac{\|I - \widehat{I}\|_2}{\|I\|_2}. \tag{4}$$

The second performance measure is population sparsity using the modified Treves Rolls (TR) metric [45]. TR scores lie between 0 and 1, with 1 being the highest sparsity. We computed model sparsity using excitatory neuron firing rates ($a_i,\ i = 1......N_E$). Existing literature on experimental evidence for sparse activity in the cortex [50] indicates that typically a small inhibitory interneuron sub-population ($a_i,\ i = N_{E+1}......N$) is far more active than excitatory neurons owing to its role in modulating activity of the entire circuit. Thus sparsity is not expected to be a feature of this sub-population, and these neurons are not included in the TR metric:

$$\text{Population Sparsity(TR)} = \left[\frac{1}{1 - \frac{1}{N_E}}\right]\left[1 - \frac{\left[\sum_{i=1}^{N_E} \frac{a_i}{N_E}\right]^2}{\sum_{i=1}^{N_E} \frac{a_i^2}{N_E}}\right]. \tag{5}$$

We define Population Density (or Population Activity Density) as

$$\text{Population Density} = 1 - \text{Population Sparsity(TR)}. \tag{6}$$

The TR metric is sensitive to bin sizes used to evaluate spike trains and smaller bin sizes lead to higher estimates of sparsity. This consideration does not affect the analysis of model activity ($a$) which is interpreted as a fixed firing rate. However, the inherent variability of spike trains in experimental data means that the choice of bin size does affect population sparsity computation. In this study, a bin size of 100ms is used for natural images, natural movies and spontaneous activity. Analysis for natural images is bound to a 100ms bin size due to a 106ms trial duration constraint in monkey experimental data [53]. A direct comparison between population sparsity of the model and experimental data is not practical given the sensitivity of the TR metric to scaling, since the dynamic ranges for the model coefficients and firing rates of neurons are very different.

The third performance measure is an estimate of the metabolic energy consumption in volume constrained sparse coding models. We compute this measure using metabolic energy consumption models for rodents and primates [44, 80], which are grounded in physiological and anatomical studies. The models estimate the metabolic energy consumption (ATP molecules/gm-minute) for cortical gray matter by aggregating estimates for the granular processes involved in its functioning. The processes include pumping out Na$^+$ entering during signaling, glutamatergic signaling, glutamate recycling, post-synaptic actions of glutamate and pre-synaptic Ca$^{2+}$ fluxes and glial cell activity. While the energy consumption associated with inhibition is thought to be somewhat less than excitation [44], we approximate the energy consumption of spiking activity as being equal in all neuron types due to the relatively smaller prevalance of inhibitory neurons and synapses in the population [44]. We have not included energy consumption due to glial cells due to their relatively small fraction of energy usage [81] and lack of a central role in the current modeling study.

For our study, we compute the metabolic energy consumption of volume constrained sparse coding models using the rodent metabolic energy consumption model, which has two main components. The first component represents the energy expended to maintain resting potentials ($3.42 \times 10^8$ ATP molecules/s-neuron), and the second represents energy spent to sustain action potentials at a given rate ($7.1 \times 10^8$ ATP molecules/neuron-spike × firing rate (Hz)). These estimates are used to compare the performance of a model at different E:I ratios, and they are only weakly affected by whether the rodent or the primate metabolic energy consumption is used:

$$\text{Energy (ATP/s)} = \left(3.42N + \sum_{i=1}^{N_E} 7.1 a_i + \sum_{j=1}^{N_I} 7.1 b_j\right) \times 10^8. \tag{7}$$

## Normalization of Performance Measures

Models with different sparsity constraints ($\lambda$) produce deviations against different baselines for reconstruction error, population sparsity/density and metabolic energy consumption. To compare different models, a common baseline is required. We implement normalization for each of the measures above in the form of a relative increase as a percentage of the minimal value observed across all E:I ratios evaluated for a given model. This normalization is described as

$$\text{Norm. Perf. Measure} = \frac{\text{Perf. Measure} - \min(\text{Perf. Measure})}{\min(\text{Perf. Measure})}. \tag{8}$$

While the normalization makes visualization easier, it does not change the qualitative results.

### Inter-Species Comparisons of Experimental Population Sparsity

We computed the Population Sparsity (TR metric) for electrophysiology data sets for monkeys [53, 54], mice [52] and cats [55] that includes natural images and natural movies as stimuli types. Neural recordings from each study can be viewed as multi-level data sets, with differences in numbers of subjects, trials and neurons across them that can be represented as a hierarchy. For each trial in each data set, we compued a population sparsity value. To test model predictions that higher optimal E:I ratios correspond to greater population sparsity against experimental data from different species, we implemented a hierarchical bootstrap procedure that is more conservative in controlling for Type-I errors with multi-level data sets than traditional paired tests [56].For each species and stimulus type, we run the bootstrap 10,000 times, generating estimates of average population sparsity. We used the resulting distributions to test the hypotheses framed by model predictions. The hierarchical organization for the bootstrap procedure for each species and stimulus type is described in detail in *Supplemental Methods* and Fig. S1.

## Acknowledgements

For cat electrophysiology, neural data were recorded by Tim Blanche in the laboratory of Nicholas Swindale, University of British Columbia, and downloaded from the NSF-funded CRCNS Data Sharing website [55]. For primate electrophysiology, neural data in response to natural images were collected in the laboratory of Adam Kohn at the Albert Einstein College of Medicine and downloaded from the CRCNS web site [53]. Lastly, we appreciate feedback from Bilal Haider, Josh Siegle and Adam Kohn on the early drafts of this manuscript.

This work was partially supported by NIH R01EY019965 and NSF CCF-1350954 to C.R and NIH EB022872, NS099375, NS084844 and NSF BCS1822677 to I.N.

# References

[1] M. A. Hofman, "On the evolution and geometry of the brain in mammals," *Progress in Neurobiology*, vol. 32, no. 2, pp. 137–158, 1989.

[2] R. J. Douglas, K. A. Martin, and D. Whitteridge, "A canonical microcircuit for neocortex," *Neural computation*, vol. 1, no. 4, pp. 480–488, 1989.

[3] J. DeFelipe, L. Alonso-Nanclares, and J. I. Arellano, "Microstructure of the neocortex: comparative aspects," *Journal of Neurocytology*, vol. 31, no. 3-5, pp. 299–316, 2002.

[4] K. D. Harris and G. M. Shepherd, "The neocortical circuit: themes and variations," *Nature neuroscience*, vol. 18, no. 2, pp. 170–181, 2015.

[5] K. D. Miller, "Canonical computations of cerebral cortex," *Current Opinion in Neurobiology*, vol. 37, pp. 75–84, 2016.

[6] J. A. Hirsch, L. M. Martinez, C. Pillai, J.-M. Alonso, Q. Wang, and F. T. Sommer, "Functionally distinct inhibitory neurons at the first stage of visual cortical processing," *Nature Neuroscience*, vol. 6, no. 12, pp. 1300–1308, 2003.

[7] S. El-Boustani and M. Sur, "Response-dependent dynamics of cell-specific inhibition in cortical networks in vivo," *Nature Communications*, vol. 5, p. 5689, 2014.

[8] B. V. Atallah, W. Bruns, M. Carandini, and M. Scanziani, "Parvalbumin-expressing interneurons linearly transform cortical responses to visual stimuli," *Neuron*, vol. 73, no. 1, pp. 159–170, 2012.

[9] B. Haider, M. Häusser, and M. Carandini, "Inhibition dominates sensory responses in the awake cortex," *Nature*, vol. 493, no. 7430, pp. 97–100, 2013.

[10] B. Haider and D. A. McCormick, "Rapid neocortical dynamics: cellular and network mechanisms," *Neuron*, vol. 62, no. 2, pp. 171–189, 2009.

[11] B. Haider, D. P. Schulz, M. Häusser, and M. Carandini, "Millisecond coupling of local field potentials to synaptic currents in the awake visual cortex," *Neuron*, vol. 90, no. 1, pp. 35–42, 2016.

[12] H. Adesnik, "Layer-specific excitation/inhibition balances during neuronal synchronization in the visual cortex," *Journal of Physiology*, vol. 596, no. 9, pp. 1639–1657, 2018.

[13] H. Adesnik, W. Bruns, H. Taniguchi, Z. J. Huang, and M. Scanziani, "A neural circuit for spatial summation in visual cortex," *Nature*, vol. 490, no. 7419, pp. 226–231, 2012.

[14] M. Zhu and C. J. Rozell, "Modeling biologically realistic inhibitory interneurons in sensory coding models," *PLoS Computational Biology*, vol. 11, no. 7, p. e1004353, 2015.

[15] P. D. King, J. Zylberberg, and M. R. DeWeese, "Inhibitory interneurons decorrelate excitatory cells to drive sparse code formation in a spiking model of V1," *Journal of Neuroscience*, vol. 33, no. 13, pp. 5475–5485, 2013.

[16] A. Litwin-Kumar, R. Rosenbaum, and B. Doiron, "Inhibitory stabilization and visual coding in cortical circuits with multiple interneuron subtypes," *Journal of Neurophysiology*, vol. 115, no. 3, pp. 1399–1409, 2016.

[17] J. A. Winer and D. T. Larue, "Populations of GABAergic neurons and axons in layer i of rat auditory cortex," *Neuroscience*, vol. 33, no. 3, pp. 499 – 515, 1989.

[18] L. Ouellet and E. de Villers-Sidani, "Trajectory of the main gabaergic interneuron populations from early development to old age in the rat primary auditory cortex," *Frontiers in neuroanatomy*, vol. 8, p. 40, 2014.

[19] V. Braitenberg and A. Schüz, *Cortex: Statistics and Geometry of Neuronal Connectivity*. Springer-Verlag, second ed., 1998.

[20] C. Beaulieu, "Numerical data on neocortical neurons in adult rat, with special reference to the GABA population," *Brain Research*, vol. 609, no. 1–2, pp. 284–292, 1993.

[21] A. Peters and D. A. Kara, "The neuronal composition of area 17 of rat visual cortex. ii. the nonpyramidal cells," *Journal of Comparative Neurology*, vol. 234, no. 2, pp. 242–263, 1985.

[22] J. J. Prieto, B. A. Peterson, and J. A. Winer, "Morphology and spatial distribution of GABAergic neurons in cat primary auditory cortex (AI)," *Journal of Comparative Neurology*, vol. 344, no. 3, pp. 349–382, 1994.

[23] P. L. Gabbott and P. Somogyi, "Quantitative distribution of gaba-immunoreactive neurons in the visual cortex (area 17) of the cat," *Experimental Brain Research*, vol. 61, no. 2, pp. 323–331, 1986.

[24] P. Somogyi, *Synaptic organization of GABAergic neurons and GABAA receptors in the lateral geniculate nucleus and visual cortex*. Houston: Portfolio Publishing, 1989.

[25] J. Li and H. D. Schwark, "Distribution and proportions of GABA-immunoreactive neurons in cat primary somatosensory cortex," *Journal of Comparative Neurology*, vol. 343, no. 3, pp. 353–361, 1994.

[26] C. C. Sherwood, M. A. Raghanti, C. D. Stimpson, C. J. Bonar, A. A. de Sousa, T. M. Preuss, and P. R. Hof, "Scaling of inhibitory interneurons in areas v1 and v2 of anthropoid primates as revealed by calcium-binding protein immunohistochemistry," *Brain, Behavior and Evolution*, vol. 69, no. 3, pp. 176–195, 2007.

[27] S. H. Hendry, H. D. Schwark, E. G. Jones, and J. Yan, "Numbers and proportions of gaba-immunoreactive neurons in different areas of monkey cerebral cortex," *Journal of Neuroscience*, vol. 7, no. 5, pp. 1503–1519, 1987.

[28] N. Tamamaki, Y. Yanagawa, R. Tomioka, J.-I. Miyazaki, K. Obata, and T. Kaneko, "Green fluorescent protein expression and colocalization with calretinin, parvalbumin, and somatostatin in the gad67-gfp knock-in mouse," *Journal of Comparative Neurology*, vol. 467, no. 1, pp. 60–79, 2003.

[29] P. L. Gabbott, B. G. Dickie, R. R. Vaid, A. J. Headlam, and S. J. Bacon, "Local-circuit neurones in the medial prefrontal cortex (areas 25, 32 and 24b) in the rat: morphology and quantitative distribution," *Journal of Comparative Neurology*, vol. 377, no. 4, pp. 465–499, 1997.

[30] H. S. Meyer, D. Schwarz, V. C. Wimmer, A. C. Schmitt, J. N. Kerr, B. Sakmann, and M. Helmstaedter, "Inhibitory interneurons in a cortical column form hot zones of inhibition in layers 2 and 5a," *Proceedings of the National Academy of Sciences*, vol. 108, no. 40, pp. 16807–16812, 2011.

[31] S. Herculano-Houzel, C. R. Watson, and G. Paxinos, "Distribution of neurons in functional areas of the mouse cerebral cortex reveals quantitatively different cortical zones," *Frontiers in Neuroanatomy*, vol. 7, p. 35, 2013.

[32] T. Binzegger, R. J. Douglas, and K. A. Martin, "A quantitative map of the circuit of cat primary visual cortex," *Journal of Neuroscience*, vol. 24, no. 39, pp. 8441–8453, 2004.

[33] C. E. Collins, D. C. Airey, N. A. Young, D. B. Leitch, and J. H. Kaas, "Neuron densities vary across and within cortical areas in primates," *Proceedings of the National Academy of Sciences*, vol. 107, no. 36, pp. 15927–15932, 2010.

[34] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, no. 6583, pp. 607–609, 1996.

[35] L. R. Varshney, P. J. Sjöström, and D. B. Chklovskii, "Optimal information storage in noisy synapses under resource constraints," *Neuron*, vol. 52, no. 3, pp. 409–423, 2006.

[36] H. B. Barlow, "Possible principles underlying the transformations of sensory messages," in *Sensory Communication* (W. A. Rosenblith, ed.), ch. 13, pp. 217–234, MIT Press, 1961.

[37] P. Földiak, "Forming sparse representations by local anti-hebbian learning," *Biological Cybernetics*, vol. 64, no. 2, pp. 165–170, 1990.

[38] J. Zylberberg, J. T. Murphy, and M. R. DeWeese, "A sparse coding model with synaptically local plasticity and spiking neurons can account for the diverse shapes of V1 simple cell receptive fields," *PLoS Computational Biology*, vol. 7, p. e1002250, 10 2011.

[39] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by V1?," *Vision research*, vol. 37, no. 23, pp. 3311–3325, 1997.

[40] M. Zhu and C. J. Rozell, "Visual nonclassical receptive field effects emerge from sparse coding in a dynamical system," *PLoS Computational Biology*, vol. 9, no. 8, p. e1003191, 2013.

[41] C. J. Rozell, D. H. Johnson, R. G. Baraniuk, and B. A. Olshausen, "Sparse coding via thresholding and local competition in neural circuits," *Neural computation*, vol. 20, no. 10, pp. 2526–2563, 2008.

[42] M. Rehn and F. T. Sommer, “A network that uses few active neurones to code visual input predicts the diverse shapes of cortical receptive fields,” *Journal of Computational Neuroscience*, vol. 22, no. 2, pp. 135–146, 2007.

[43] B. A. Olshausen, “Highly overcomplete sparse coding,” in *Human Vision and Electronic Imaging XVIII*, vol. 8651, p. 86510S, International Society for Optics and Photonics, 2013.

[44] D. Attwell and S. B. Laughlin, “An Energy Budget for Signaling in the Grey Matter of the Brain,” *Journal Cerebral Blood Flow and Metabolism*, vol. 21, pp. 1133–1145, Oct. 2001.

[45] W. E. Vinje and J. L. Gallant, “Sparse Coding and Decorrelation in Primary Visual Cortex During Natural Vision,” *Science*, vol. 287, pp. 1273–1276, Feb. 2000.

[46] B. A. Olshausen and D. J. Field, “Sparse coding of sensory inputs,” *Current Opinion in Neurobiology*, vol. 14, no. 4, pp. 481–487, 2004.

[47] J. E. Niven and S. B. Laughlin, “Energy limitation as a selective pressure on the evolution of sensory systems,” *Journal of Experimental Biology*, vol. 211, no. 11, pp. 1792–1804, 2008.

[48] E. B. Baum, J. Moody, and F. Wilczek, “Internal representations for associative memory,” *Biological Cybernetics*, vol. 59, no. 4, pp. 217–228, 1988.

[49] A. S. Charles, H. L. Yap, and C. J. Rozell, “Short term memory capacity in networks via the restricted isometry property,” *Neural Computation*, vol. 26, p. 1198–1235, 2014.

[50] A. L. Barth and J. F. Poulet, “Experimental evidence for sparse firing in the neocortex,” *Trends in Neurosciences*, vol. 35, pp. 345–355, June 2012.

[51] A. Hasenstaub, Y. Shu, B. Haider, U. Kraushaar, A. Duque, and D. A. McCormick, “Inhibitory postsynaptic potentials carry synchronized frequency information in active cortical networks,” *Neuron*, vol. 47, no. 3, pp. 423–435, 2005.

[52] J. H. Siegle, X. Jia, S. Durand, S. Gale, C. Bennett, N. Graddis, G. Heller, T. K. Ramirez, H. Choi, J. A. Luviano, P. A. Groblewski, R. Ahmed, A. Arkhipov, A. Bernard, Y. N. Billeh, D. Brown, M. A. Buice, N. Cain, S. Caldejon, L. Casal, A. Cho, M. Chvilicek, T. C. Cox, K. Dai, D. J. Denman, S. E. J. de Vries, R. Dietzman, L. Esposito, C. Farrell, D. Feng, J. Galbraith, M. Garrett, E. C. Gelfand, N. Hancock, J. A. Harris, R. Howard, B. Hu, R. Hytnen, R. Iyer, E. Jessett, K. Johnson, I. Kato, J. Kiggins, S. Lambert, J. Lecoq, P. Ledochowitsch, J. H. Lee, A. Leon, Y. Li, E. Liang, F. Long, K. Mace, J. Melchior, D. Millman, T. Mollenkopf, C. Nayan, L. Ng, K. Ngo, T. Nguyen, P. R. Nicovich, K. North, G. K. Ocker, D. Ollerenshaw, M. Oliver, M. Pachitariu, J. Perkins, M. Reding, D. Reid, M. Robertson, K. Ronellenfitch, S. Seid, C. Slaughterbeck, M. Stoecklin, D. Sullivan, B. Sutton, J. Swapp, C. Thompson, K. Turner, W. Wakeman, J. D. Whitesell, D. Williams, A. Williford, R. Young, H. Zeng, S. Naylor, J. W. Phillips, R. C. Reid, S. Mihalas, S. R. Olsen, and C. Koch, “Data from ”a survey of spiking activity reveals a functional hierarchy of mouse corticothalamic visual areas”.” bioRxiv, 10 2019. https://doi.org/10.1101/805010.

[53] A. Kohn and R. Coen-Cagli, “Data from ”multi-electrode recordings of anesthetized macaque v1 responses to static natural images and gratings.”.” CRCNS.org, 2015. http://dx.doi.org/10.6080/K0SB43P8.

[54] A. Kohn and M. A. Smith, “Data from ”utah array extracellular recordings of spontaneous and visually evoked activity from anesthetized macaque primary visual cortex (v1).”.” CRCNS.org, 2016. http://dx.doi.org/10.6080/K0NC5Z4X.

[55] T. Blanche, “Data from ”multi-neuron recordings in primary visual cortex.”.” CRCNS.org, 2009. http://dx.doi.org/10.6080/K0MW2F2J.

[56] V. Saravanan, G. J. Berman, and S. J. Sober, “Application of the hierarchical bootstrap to multi-level data in neuroscience,” 2019.

[57] L. G. Nowak, M. V. Sanchez-Vives, and D. A. McCormick, “Lack of orientation and direction selectivity in a subgroup of fast-spiking inhibitory interneurons: cellular and synaptic mechanisms and comparison with other electrophysiological cell types,” *Cerebral Cortex*, vol. 18, no. 5, pp. 1058–1078, 2008.

[58] A. M. Kerlin, M. L. Andermann, V. K. Berezovskii, and R. C. Reid, “Broadly tuned response properties of diverse inhibitory neuron subtypes in mouse visual cortex,” *Neuron*, vol. 67, no. 5, pp. 858–871, 2010.

[59] B. Antkowiak and C. Helfrich-Forster, “Effects of Small Concentrations of Volatile Anesthetics on Action Potential Firing of Neocortical Neurons In Vitro ,” *Anesthesiology: The Journal of the American Society of Anesthesiologists*, vol. 88, pp. 1592–1605, 06 1998.

[60] B. Antkowiak, “Different Actions of General Anesthetics on the Firing Patterns of Neocortical Neurons Mediated by the GABAAReceptor ,” *Anesthesiology: The Journal of the American Society of Anesthesiologists*, vol. 91, pp. 500–511, 08 1999.

[61] L. D. Lewis, V. S. Weiner, E. A. Mukamel, J. A. Donoghue, E. N. Eskandar, J. R. Madsen, W. S. Anderson, L. R. Hochberg, S. S. Cash, E. N. Brown, *et al.*, “Rapid fragmentation of neuronal networks at the onset of propofol-induced unconsciousness,” *Proceedings of the National Academy of Sciences*, vol. 109, no. 49, pp. E3377–E3386, 2012.

[62] M. Weliky, J. Fiser, R. H. Hunt, and D. N. Wagner, “Coding of natural scenes in primary visual cortex,” *Neuron*, vol. 37, no. 4, pp. 703–718, 2003.

[63] J. D. Peduzzi, “Genesis of gaba-immunoreactive neurons in the ferret visual cortex,” *Journal of Neuroscience*, vol. 8, no. 3, pp. 920–931, 1988.

[64] E. Froudarakis, P. Berens, A. S. Ecker, R. J. Cotton, F. H. Sinz, D. Yatsenko, P. Saggau, M. Bethge, and A. S. Tolias, “Population code in mouse v1 facilitates readout of natural scenes through increased sparseness,” *Nature neuroscience*, vol. 17, no. 6, p. 851, 2014.

[65] Y. Yu, J. N. Stirman, C. R. Dorsett, and S. L. Smith, “Mesoscale correlation structure with single cell resolution during visual coding.” bioRxiv, 11 2018.

[66] M. Zhu and C. J. Rozell, “Visual nonclassical receptive field effects emerge from sparse coding in a dynamical system,” *PLoS Computational Biology*, vol. 9, p. e1003191, 08 2013.

[67] B. Haider, M. R. Krause, A. Duque, Y. Yu, J. Touryan, J. A. Mazer, and D. A. McCormick, “Synaptic and Network Mechanisms of Sparse and Reliable Visual Cortical Activity during Nonclassical Receptive Field Stimulation,” *Neuron*, vol. 65, no. 1, pp. 107–121, 2010.

[68] W. E. Vinje and J. L. Gallant, “Sparse coding and decorrelation in primary visual cortex during natural vision,” *Science*, vol. 287, no. 5456, pp. 1273–1276, 2000.

[69] J. Wolfe, A. R. Houweling, and M. Brecht, “Sparse and powerful cortical spikes,” *Current Opinion in Neurobiology*, vol. 20, pp. 306–312, 6 2010.

[70] C. J. Rozell, D. H. Johnson, R. G. Baraniuk, and B. A. Olshausen, “Sparse coding via thresholding and local competition in neural circuits,” *Neural Computation*, vol. 20, no. 10, pp. 2526–2563, 2008.

[71] S. Shapero, C. J. Rozell, and P. Hasler, “Configurable hardware integrate and fire neurons for sparse approximation,” *Neural Networks*, vol. 45, no. 0, pp. 134–143, 2013.

[72] T. Hu, A. Genkin, and D. B. Chklovskii, “A network of spiking neurons for computing sparse representations in an energy-efficient way,” *Neural Computation*, vol. 24, pp. 2852–2872, Aug. 2012.

[73] A. Balavoine, J. K. Romberg, and C. J. Rozell, “Convergence and rate analysis of neural networks for sparse approximation,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, pp. 1377–1389, 9 2012.

[74] A. Balavoine, C. J. Rozell, and J. K. Romberg, “Convergence of a neural network for sparse approximation using the nonsmooth łojasiewicz inequality,” in *International Joint Conference in Neural Networks (IJCNN)*, 2013.

[75] A. S. Charles, P. Garrigues, and C. J. Rozell, “A common network architecture efficiently implements a variety of sparsity-based inference problems,” *Neural Computation*, vol. 24, pp. 3317–3339, Sept. 2012.

[76] S. Shapero, A. S. Charles, C. J. Rozell, and P. Hasler, “Low power sparse approximation on reconfigurable analog hardware,” *Emerging and Selected Topics in Circuits and Systems, IEEE Journal on*, vol. 2, pp. 530 –541, 9 2012.

[77] S. Shapero, M. Zhu, J. Hasler, and C. Rozell, “Optimal sparse approximation with integrate and fire neurons,” *International Journal of Neural Systems*, vol. 24, no. 05, p. 1440001, 2014.

[78] J. Schiller, G. Major, H. J. Koester, and Y. Schiller, “Nmda spikes in basal dendrites of cortical pyramidal neurons,” *Nature*, vol. 404, no. 6775, p. 285, 2000.

[79] A. Polsky, B. W. Mel, and J. Schiller, “Computational subunits in thin dendrites of pyramidal cells,” *Nature neuroscience*, vol. 7, no. 6, p. 621, 2004.

[80] P. Lennie, "The cost of cortical computation," *Current Biology*, vol. 13, no. 6, pp. 493–497, 2003.

[81] M. T. Wong-Riley, "Cytochrome oxidase: an endogenous metabolic marker for neuronal activity," *Trends in Neurosciences*, vol. 12, no. 3, pp. 94–101, 1989.

# Supplementary Information: Constrained brain volume in an efficient coding model explains the fraction of excitatory and inhibitory neurons in sensory cortices

Arish Alreja[1], Ilya Nemenman[2], Christopher Rozell[3]

[1] Neuroscience Institute, Center for the Neural Basis of Cognition and Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA | [2] Department of Physics, Department of Biology and Initiative in Theory and Modeling of Living Systems, Emory University, Atlanta, GA 30322, USA | [3] School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA

## Supplementary Methods

### Estimation of statistical errors in the analysis

We evaluate variance and bias in reconstruction error, population sparsity and metabolic energy over the image patch database of 10 natural images from which image patches are sampled for the sparsity constraint $\lambda$=0.15 at each E:I ratio (1:1-10:1).

To estimate the variance, we randomly select 1 out of $N = 10$ images (with replacement) and we select 10 16x16 pixel image patches from this image. Repeating this process 10 times, we gather 100 16x16 pixel image patches. We perform inference in the model and calculate the mean reconstruction error, population sparsity and metabolic energy consumption are computed using the model corresponding to each E:I ratio. This constitutes one run. We collect and aggregate statistics from 100 runs. The standard deviation of the means computed for each of the runs is the standard error for a given performance measure for a given model.

Estimation of the bias is similar, however, instead of choosing patches from $N = 10$ images, we use $N^*$ images, where $N^* = \alpha N$, $\alpha < 1$. Mean reconstruction error, population sparsity and metabolic energy consumption are computed for each E:I ratio. This constitutes a single run. For a given $\alpha$, we perform 100 runs. We repeat this process for each $\alpha \in \{0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$ ($\alpha =$ 1.0 being the variance estimate mentioned above), which amounts to a total of 600 runs. The average (over 100 runs) optimal E:I ratio for $\lambda = 0.15$ at different values of $\alpha$ is then examined to explore if the image patch database size induces any bias in the observed optimal E:I ratio.

### Estimation of the ratio of recurrent excitation vs recurrent inhibition during stimulus representation

To better understand the effects of changes in the size of inhibitory subpopulation (i.e., different E:I ratios), we examine the relationship between recurrent excitation and recurrent inhibition received by active units in response to stimulus image patches. We consider the following decomposition of the low rank approximation ($G$) of the recurrent connectivity matrix utilized in an earlier study [1]

$$\mathrm{G} = \underbrace{U^+\Sigma V^- + U^-\Sigma V^+}_{G_{excite}} + \underbrace{U^+\Sigma V^+ + U^-\Sigma V^-}_{G_{inhib}}. \tag{9}$$

The recurrent excitation and inhibition received by active nodes is computed as

$$\text{Recurrent E} = [(D - G_{excite}) \times a] \circ \mathbb{I}_{a>0}, \tag{10}$$

$$\text{Recurrent I} = [G_{inhib} \times a] \circ \mathbb{I}_{a>0}, \tag{11}$$

where $\mathbb{I}$ is the standard indicator function taking the value 1 if the argument is true and 0 otherwise. We compute a ratio between recurrent excitation and recurrent inhibition received by active neurons as

$$\text{Recurrent Ratio} = \frac{\text{Recurrent E}}{\text{Recurrent I}}. \tag{12}$$

The Recurrent Ratio is computed for each anatomical E:I ratio for models with different sparsity constraints ($\lambda$). Here we evaluate the recurrent E/I balance specifically in response to stimulus, and recognize that the network is more stable when recurrent inhibition is greater than recurrent excitation.

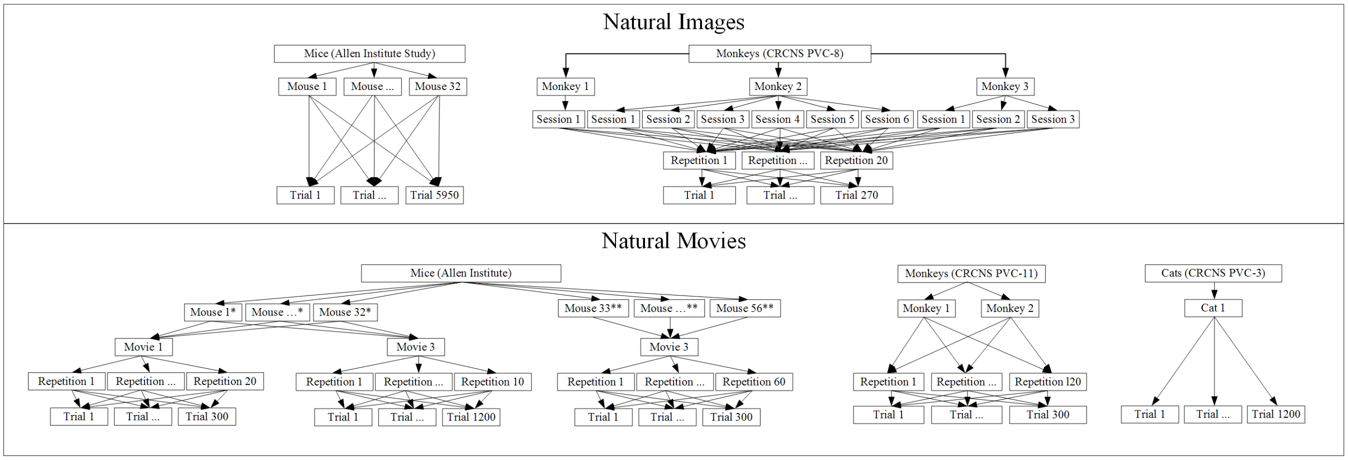

Figure S1: The hierarchy of multi-level experimental data sets used to draw comparisons between population sparsity of different species is represented in each figure for different stimulus types. **(top panel)** For natural images, the available data sets allow us to draw a comparison between 32 mice (E:I = 5.7-9:1) [3] with 44-244 V1 neurons and 3 monkeys (E:I = 4-4.3:1) [4] with 16-76 V1 neurons. The monkey data [4] features multiple recording sessions per subject (1,6,3 sessions for monkeys 1,2,3) and 20 repetitions of all stimuli images in a block structure. **(bottom panel)** For natural movies, the available data sets allow us to draw a comparison between 56 mice [3] with 44-244 V1 neurons, 2 monkeys [5] with 69-104 V1 neurons and 1 cat [6] (E:I = 4:1) with 10 V1 neurons. The mouse subjects come from 2 different experiments (* denotes the Brain Observatory Experiment with 32 subjects, and ** denotes the Functional Connectivity Experiment with 24 subjects) where the key differences include the number of different natural movies shown and how often each movie is repeated.

## Hierarchical bootstrap: Hypothesis testing model predictions against multi-level experimental data

The hierarchical bootstrap procedure is built around sampling with replacement at different levels of a hierarchy in a multi-level dataset at each bootstrap run to estimate averages. The procedure is described in detail in [2].

Our measure of interest in this study is the scalar population sparsity measure (TR metric), which means that the number of recorded neurons don't feature in our hierarchy. As an example specific to its usage in this study, we describe the case of the hierarchical bootstrap for comparing average population sparsity between mice and monkeys for natural image stimuli from multi-level data sets visualized in Figure S1 (top panel), to test the hypothesis/model prediction that mice should exhibit higher population sparsity than monkeys. For computational efficiency, population sparsity is pre-computed for each dataset before the hierarchical bootstrap procedure.

For mice, we first sample subjects (first level of hierarchy) with replacement from the 32 mice with stimulus responses to natural images in the Allen Institute data set [3]. Next, for each sampled mouse, we sample trials (second level of hierarchy) with replacement from the total number of trials 'T' (T=5950 in the example). Finally, we average the population sparsity across all the sampled trials to obtain an average population sparsity for natural images in mice. This process represents a single bootstrap run. A slightly different hierarchy, shown in Figure S1 (top panel) is constructed for monkeys, where the first level represents different recording sessions (with different numbers of neurons) for a single monkey. The same hierarchical bootstrap procedure is repeated. We collect average population sparsity estimates for mice and monkeys for a total of 10,000 bootstrap runs for natural images as well as other stimulus types.

For hypothesis testing related to the example above, we treat the 10,000 average population sparsity values for mice and monkeys as a 2 dimensional distribution. We use this joint distribution to evaluate the model hypothesis/prediction that population sparsity in mice (E:I = 5.7-9:1) should be higher than monkeys (E:I = 4-4.3:1). With mice on the x-axis and monkeys on the y-axis, we compute the volume of the distribution where $x > y$ (i.e. the volume of the distribution below the line $y = x$). If the volume $> 1 - \frac{\alpha}{2}$ then mouse pop sparsity > monkey pop sparsity at level $\alpha$, ($\alpha = 0.05$ in our analysis). In the event of multiple comparisons (e.g. natural movie stimulus), Bonferroni (or other) corrections can be applied the same way as traditional hypothesis testing. The volume of the distribution opposing the hypothesis (above $y = x$ in our example) is the $p$ value for the test. It is referred to as $p_{bootstrap}$ to disambiguate it from $p$ values emanating from traditional hypothesis testing.

# Supplementary Results

## Bias in statistical analysis

We estimated bias in all three performance measures using a bootstrap procedure (See *Supplementary Methods*) which samples from a subset of the natural image database [7] for models trained with a sparsity constraint of $\lambda = 0.15$. Fig. S2 indicates stability of the mean optimal E:I ratio over 100 runs sampling from differently sized subsets (denoted by $\alpha$) of the natural image databases, suggesting negligible bias.

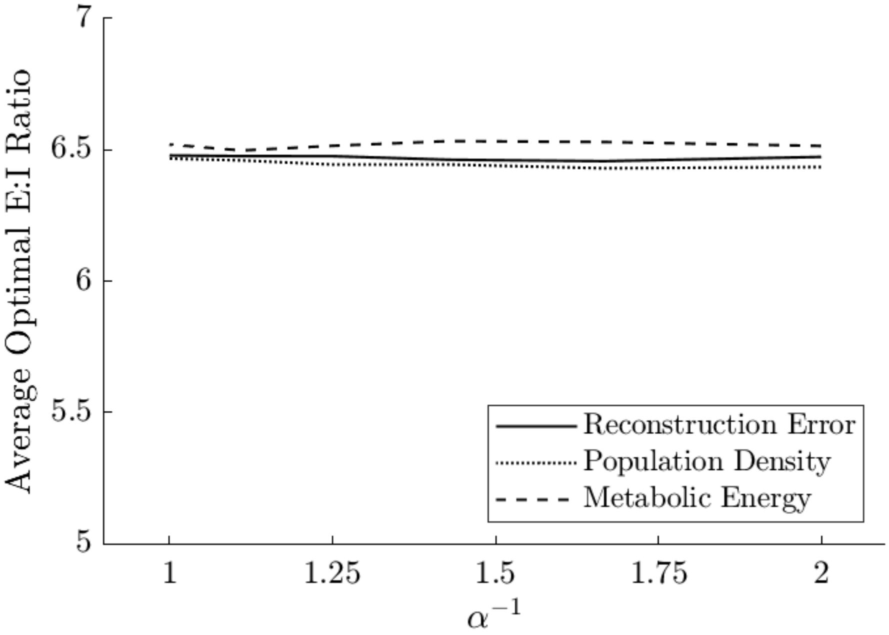

Figure S2: Estimation of the bias involves choosing 100 patches from $N^* = \alpha N$, $\alpha < 1$ of the $N = 10$ natural images. Mean reconstruction error, population sparsity and metabolic energy consumption are computed for each E:I ratio. This constitutes a single run. For a given $\alpha$, we perform 100 runs. This process is repeated for each $\alpha \in \{0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$, which amounts to a total of 600 runs. Average optimal E:I for $\lambda = 0.15$ and different values of $\alpha$ are shown for reconstruction error, population density and metabolic energy consumption. The relative constancy of average optimal E:I ratio over the bias runs for different values of $\alpha$ indicates that any possible bias in estimating the optimal E:I ratio is negligible for the explored sample sizes.

## Structure of Recurrent Inhibition

We examined the static structure of inhibition of the model at the optimal E:I ratio for different sparsity levels. We observed that inhibitory strength, represented by the Singular Values $\Sigma$, interpreted as implementing dendritic gain (see *Methods*) is distributed less evenly across the inhibitory sub-population as sparsity ($\lambda$) increases (Fig. S3(left)), even as the total inhibitory strength/dendritic gain across the inhibitory sub-population remains relatively unchanged across different sparsity levels (Fig. S3(middle)). Next, we use a metric called the stable rank [8] which is defined as

$$\text{Stable Rank} = \frac{\sum_{\mathrm{i=1}}^{\mathrm{N_i}} \sigma_{\mathrm{i}}^2}{\sigma_1^2}, \tag{13}$$

where $\sigma_i$ is the $i^{th}$ singular value of the SVD of the recurrent matrix. The stable rank is relatively robust to smaller singular values. In the context of our interpretation of $\Sigma$ as the dendritic gain, the stable rank can serve as an additional measure of unevenness (lower stable rank implies greater unevenness). The value of the stable rank decreases as sparsity ($\lambda$) increases (Fig. S3(right)) adding support for the preceding result that indicates that unevenness in inhibitory strength of the model at the optimal E:I ratio increases as sparsity increases. Together, these results suggest that while the total amount of inhibition supported by the structure is relatively unchanged for models at optimal E:I ratio at different sparsity levels, it is distributed less evenly in the inhibitory sub-population.

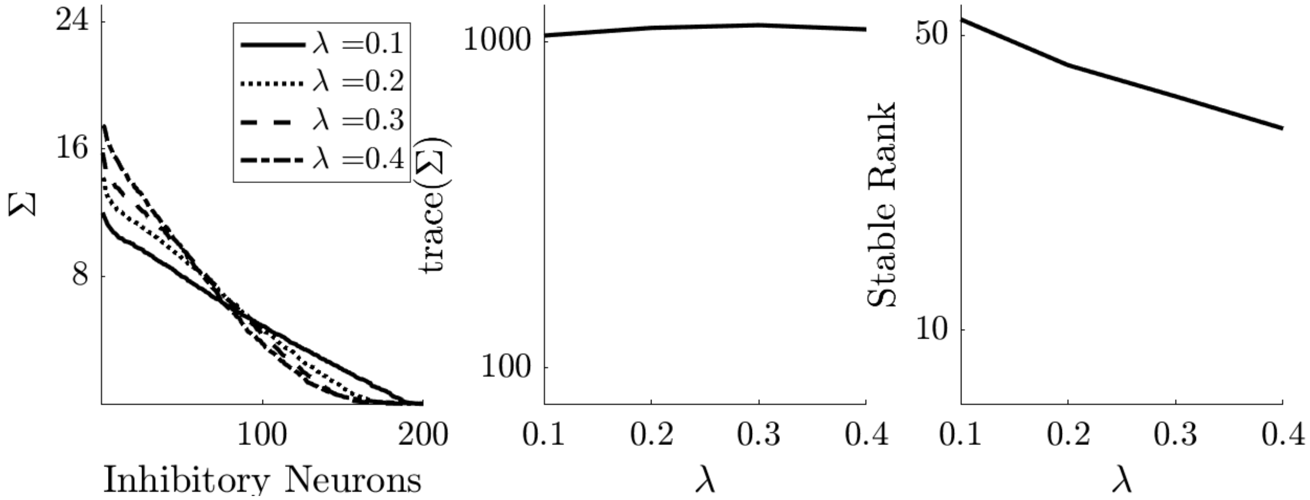

Figure S3: **(left)** Inhibitory strength Σ interpreted as being implemented via dendritic gain (see *Methods*) vs Inhibitory Interneurons (Components) for different sparsity levels ($\lambda$). **(middle)** Total amount of inhibitory strength/dendritic gain across the inhibitory sub-population, relatively unchanged for models at the optimal E:I ratio for different sparsity constraints ($\lambda$). **(right)** The trend of increasing unevenness in inhibitory strength as sparsity ($\lambda$) increases, depicted by the first two plots is also reflected by a decrease in the stable rank measure detailed in [8].

## Inhibitory sub-population activity profiles at different sparsity levels

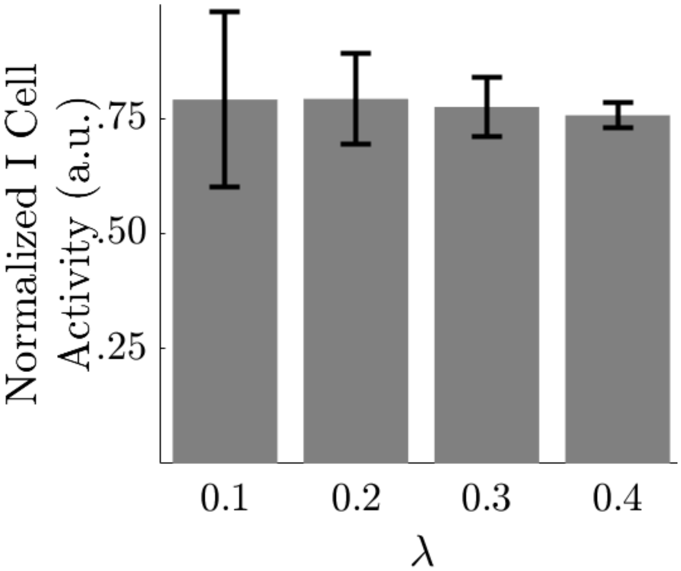

Figure S4: A normalized version of the (bottom row)(left) plot in Fig. 3 shows I cell activity when normalized against the total (E+I cell) activity of the model. The normalized I cell activity is relatively unchanged across optimal models at different sparsity levels, while the diversity of responses (error bars) to different natural image stimuli in the inhibitory sub-population shows (like the un-normalized plot) that I cell responses are more specifically tuned to stimuli at lower sparsity levels/lower optimal E:I ratio and become broadly tuned and less diverse as model sparsity/optimal E:I ratio increases.

## Model performance (all performance measures) vs biology

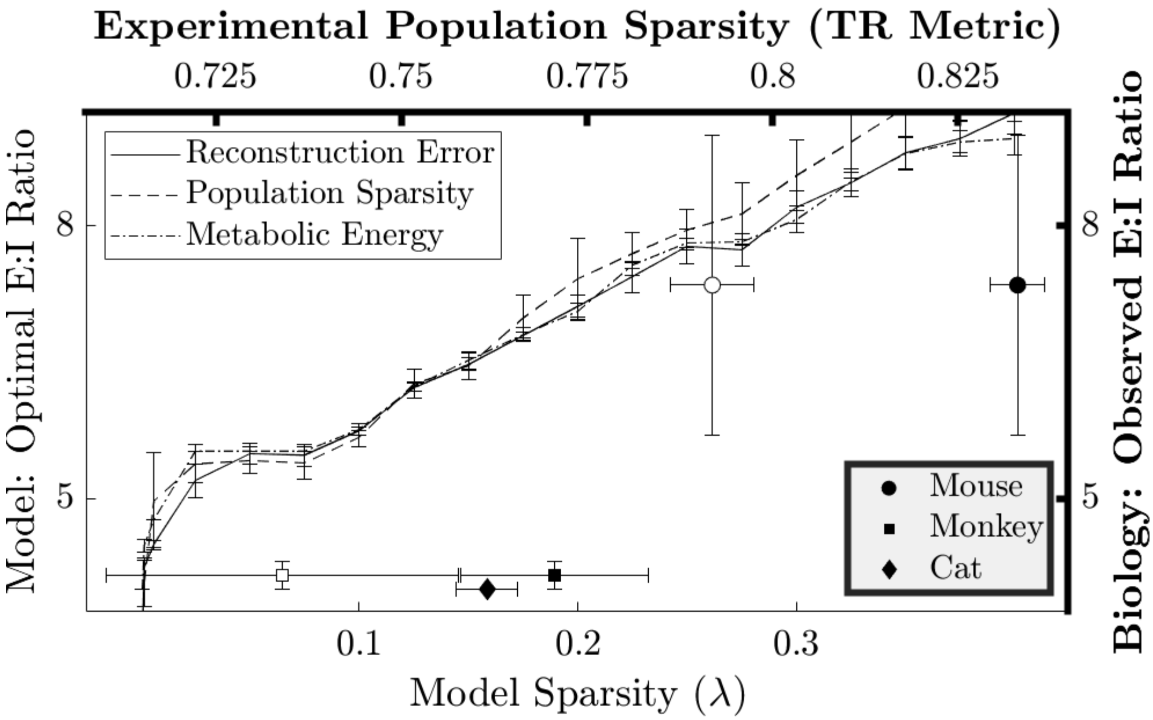

Figure S5: **(Left Y and Bottom X axes)** The Optimal E:I ratio as a function of model sparsity constraints $\lambda$ in computational models according to all three performance measures is captured by the different lines with error bars denoting the standard error for each. A similar (but not identical) trend in the relationship between optimal E:I ratio and model sparsity is revealed for each of the 3 performance measures. **(Right Y and Top X axes)** The Population Sparsity (TR) measure computed for electrophysiology data from experimental studies in mice [3], monkeys [4, 5] and cats [6] is shown as mean (markers) ± standard error(horizontal error bars) w.r.t. observed E:I ratio ranges (vertical error bars) in Biology with unfilled markers representing natural images and black filled markers representing natural movies.

# Supplementary References

[SR1] M. Zhu and C. J. Rozell, “Modeling biologically realistic inhibitory interneurons in sensory coding models,” *PLoS Computational Biology*, vol. 11, no. 7, p. e1004353, 2015.

[SR2] V. Saravanan, G. J. Berman, and S. J. Sober, “Application of the hierarchical bootstrap to multi-level data in neuroscience,” 2019.

[SR3] J. H. Siegle, X. Jia, S. Durand, S. Gale, C. Bennett, N. Graddis, G. Heller, T. K. Ramirez, H. Choi, J. A. Luviano, P. A. Groblewski, R. Ahmed, A. Arkhipov, A. Bernard, Y. N. Billeh, D. Brown, M. A. Buice, N. Cain, S. Caldejon, L. Casal, A. Cho, M. Chvilicek, T. C. Cox, K. Dai, D. J. Denman, S. E. J. de Vries, R. Dietzman, L. Esposito, C. Farrell, D. Feng, J. Galbraith, M. Garrett, E. C. Gelfand, N. Hancock, J. A. Harris, R. Howard, B. Hu, R. Hytnen, R. Iyer, E. Jessett, K. Johnson, I. Kato, J. Kiggins, S. Lambert, J. Lecoq, P. Ledochowitsch, J. H. Lee, A. Leon, Y. Li, E. Liang, F. Long, K. Mace, J. Melchior, D. Millman, T. Mollenkopf, C. Nayan, L. Ng, K. Ngo, T. Nguyen, P. R. Nicovich, K. North, G. K. Ocker, D. Ollerenshaw, M. Oliver, M. Pachitariu, J. Perkins, M. Reding, D. Reid, M. Robertson, K. Ronellenfitch, S. Seid, C. Slaughterbeck, M. Stoecklin, D. Sullivan, B. Sutton, J. Swapp, C. Thompson, K. Turner, W. Wakeman, J. D. Whitesell, D. Williams, A. Williford, R. Young, H. Zeng, S. Naylor, J. W. Phillips, R. C. Reid, S. Mihalas, S. R. Olsen, and C. Koch, “Data from ”a survey of spiking activity reveals a functional hierarchy of mouse corticothalamic visual areas”.” bioRxiv, 10 2019. https://doi.org/10.1101/805010.

[SR4] A. Kohn and R. Coen-Cagli, “Data from ”multi-electrode recordings of anesthetized macaque v1 responses to static natural images and gratings.”.” CRCNS.org, 2015. http://dx.doi.org/10.6080/K0SB43P8.

[SR5] A. Kohn and M. A. Smith, “Data from ”utah array extracellular recordings of spontaneous and visually evoked activity from anesthetized macaque primary visual cortex (v1).”.” CRCNS.org, 2016. http://dx.doi.org/10.6080/K0NC5Z4X.

[SR6] T. Blanche, “Data from ”multi-neuron recordings in primary visual cortex.”.” CRCNS.org, 2009. http://dx.doi.org/10.6080/K0MW2F2J.

[SR7] B. A. Olshausen and D. J. Field, “Emergence of simple-cell receptive field properties by learning a sparse code for natural images,” *Nature*, vol. 381, no. 6583, pp. 607–609, 1996.

[SR8] A. Eftekhari, H. L. Yap, M. B. Wakin, and C. J. Rozell, “Stabilizing embedology: Geometry-preserving delay-coordinate maps,” *Physical Review E*, vol. 97, no. 2, p. 022222, 2018.