Deep learning-driven characterization of single cell tuning in primate visual area V4 supports topological organization

Konstantin F. Willeke[1,2,7-9,*], **Kelli Restivo**[3-4,*], **Katrin Franke**[7-9], **Arne F. Nix**[1-2], **Santiago A. Cadena**[2], **Tori Shinn**[10], **Cate Nealley**[7-9], **Gabrielle Rodriguez**[7-9], **Saumil Patel**[7-9], **Alexander S. Ecker**[2,5], **Fabian H. Sinz**[1-4,†,✉], **and Andreas S. Tolias**[7-9†,✉]

[1]Institute for Bioinformatics and Medical Informatics, Tübingen University, Tübingen, Germany
[2]Institute of Computer Science and Campus Institute Data Science, University of Göttingen, Germany
[3]Department of Neuroscience, Baylor College of Medicine, Houston, TX, USA
[4]Center for Neuroscience and Artificial Intelligence, Baylor College of Medicine, Houston, TX, USA
[5]Max Planck Institute for Dynamics and Self-Organization, Göttingen, Germnay
[6]Department of Electrical and Computer Engineering, Rice University, Houston, TX, USA
[7]Department of Ophthalmology, Byers Eye Institute, Stanford University School of Medicine, Stanford, CA, US
[8]Stanford Bio-X, Stanford University, Stanford, CA, US
[9]Wu Tsai Neurosciences Institute, Stanford University, Stanford, CA, US
[10]Department of Pediatrics; Allergy & Immunology, Baylor College of Medicine, Houston, Texas, USA
[*,†]Equal contributions

Deciphering the brain's structure-function relationship is key to understanding the neuronal mechanisms underlying perception and cognition. The cortical column, a vertical organization of neurons with similar functions, is a classic example of primate neocortex structure-function organization. While columns have been identified in primary sensory areas using parametric stimuli, their prevalence across higher-level cortex is debated, particularly regarding complex tuning in natural image space. However, a key hurdle in identifying columns is characterizing the complex, nonlinear tuning of neurons to high-dimensional sensory inputs. Building on prior findings of topological organization for features like color and orientation, we investigate functional clustering in macaque visual area V4 in non-parametric natural image space, using large-scale recordings and deep learning–based analysis. We combined linear probe recordings with deep learning methods to systematically characterize the tuning of >1,200 V4 neurons using *in silico* synthesis of most exciting images (MEIs), followed by *in vivo* verification. Single V4 neurons exhibited MEIs containing complex features, including textures and shapes, and even high-level attributes with eye-like appearance. Neurons recorded on the same silicon probe, inserted orthogonal to the cortical surface, often exhibited similarities in their spatial feature selectivity, suggesting a degree of functional organization along the cortical depth. We quantified MEI similarity using human psychophysics and distances in a contrastive learning-derived embedding space. Moreover, the selectivity of the V4 neuronal population showed evidence of clustering into functional groups of shared feature selectivity. These functional groups showed parallels with the feature maps of units in artificial vision systems, suggesting potential shared encoding strategies. These results demonstrate the feasibility and scalability of deep learning–based functional characterization of neuronal selectivity in naturalistic visual contexts, offering a framework for quantitatively mapping cortical organization across multiple levels of the visual hierarchy.

Correspondence: ***sinz@uni-goettingen.de; tolias@stanford.edu***

Introduction

From the intricate layering of neurons with diverse functions in the retina (e.g. Masland, 2001) to the topographic maps of the cerebral cortex (e.g. Penfield & Boldrey, 1937), for decades neuroscientists have been pursing the quest to discover general organizing principles that relate the structure (anatomy) and function (physiology) of the brain. For instance, the concept of functional cortical columns, first described by Mountcastle (1957) in the somatosensory cortex and later in the primary visual cortex (V1) by Hubel & Wiesel (1968), represents one such well-established motif, thought to recur across the primate neocortex (discussed in Horton & Adams, 2005). In this canonical organization, neurons with similar function are vertically organized across cortical layers. Considering that the connections within the cortex are locally dense and span the cortical layers, this configuration enables neurons with similar response functions to synaptically interact, thus facilitating computations to transform information within and across the layers (e.g. Cadwell et al., 2020; Campagnola et al., 2022; Jiang et al., 2015)

Obtaining a comprehensive understanding of the relationship between anatomy and function requires a thorough characterization of neuronal stimulus selectivity or tuning. The selectivity of neurons in monkey and cat V1 for simple visual features, such as orientation, phase, or spatial frequency (Issa et al., 2000; Victor et al., 1994), enables the characterization of the tuning properties of these neurons using well-defined parametric stimuli such as gratings. This has greatly facilitated the identification of general organizing principles of neuronal function in the early visual areas of the cortical hierarchy (e.g. Ohki & Reid, 2014). However, neurons in higher visual areas prefer more complex visual features found in natural scenes, such as shapes, textures, objects, and faces (e.g. Bashivan et al., 2019; Kim et al., 2019; Ponce et al., 2019;

Tang et al., 2020; Tsao et al., 2003), which are not easily parameterized. The immense diversity and high dimensionality of the natural image space make it challenging to systematically characterize more complex visual function and link it to an organizing structure.

Although prior studies have described feature selectivity in area V4 (e.g. Bashivan et al., 2019; Gallant et al., 1993; Kim et al., 2019; Pasupathy & Connor, 2002; Tang et al., 2020; Wang et al., 2023), less is known about how this selectivity is organized across cortical depth or whether it exhibits topographic structure akin to columnar motifs observed in early visual cortex. While orientation, curvature and color domains have been reported to exhibit spatial clustering within V4 (Conway & Tsao, 2009; Ghose & Ts'o, 1997; Tang et al., 2020; Tanigawa et al., 2010; Wang et al., 2024; Zhang et al., 2023), other work has suggested a more patchy or irregular distribution of feature selectivity (Kotake et al., 2009; Namima et al., 2025), raising debate over whether V4 exhibits consistent columnar or topographic organization. It is therefore unclear whether higher-order tuning, especially under natural image stimulation, follows similar spatial principles. Addressing this question requires scalable, data-driven approaches capable of characterizing neuronal tuning within the rich statistics of natural images, without relying on strong assumptions about feature space.

Advancements in deep learning promise to overcome these challenges. Specifically, recent deep learning functional models of the brain can accurately predict responses to arbitrary stimuli (Bashivan et al., 2019; Cadena et al., 2019; Walker et al., 2019), enabling essentially unlimited *in silico* experiments that complement rather than replace in vivo recordings, and allow exploration of high-dimensional visual feature spaces that are experimentally inaccessible. This can be used for a comprehensive characterization of neuronal tuning function, such as identifying the neurons' optimal stimuli (Bashivan et al., 2019; Franke et al., 2022; Höfling et al., 2022; Kindel et al., 2019; Walker et al., 2019), map their invariances (Ding et al., 2023b), characterize contextual modulation (Fu et al., 2023) or characterize how multiple distinct tuning properties or nonlinear contextual effects relate to each other (Ustyuzhaninov et al., 2022). The predictions derived from these *in silico* analyses can then be verified through *in vivo* closed-loop experiments, known as inception loops, which have been successfully applied to single neurons in mice (Ding et al., 2023b; Franke et al., 2022; Fu et al., 2023; Höfling et al., 2022; Walker et al., 2019) and populations of neurons in macaque visual cortex (Bashivan et al., 2019).

In this study, we adapted the inception loop paradigm for macaque electrophysiological single-unit recordings to systematically map stimulus selectivity and analyze the structure-function organization of neurons in visual area V4. We used deep neural networks to build an accurate model of >1,200 recorded V4 neurons, capable of predicting responses to arbitrary images and used it to synthesize the most exciting image (MEI) for individual neurons, which we subsequently verified *in vivo*. We found that neurons recorded on the same silicon probe orthogonal to the cortical surface appeared to have similar spatial features compared to MEIs of neurons recorded in different locations, and verified this impression with human psychophysics and a non-linear embedding space based on image similarity. Furthermore, the MEIs formed isolated clusters in the non-linear embedding space, indicating that V4 neurons separate into distinct functional groups that are selective for specific complex visual features, such as oriented fur patterns, grid-like motifs, curvatures, or even high-level attributes reminiscent of eyes. Interestingly, these functional groups closely resemble the feature maps of early- to mid-level units in deep neural networks trained on image classification (Olah et al., 2020), suggesting that computational principles are shared among biological and artificial visual systems. These findings reveal functional clustering within V4 that is consistent with a topographic organization of feature selectivity and support the notion of a columnar-like organization of natural image selectivity in area V4.

Results

Deep neural network approach captures tuning properties of individual monkey V4 neurons To systematically study the neuronal tuning properties of monkey V4 neurons in the context of natural scenes, we combined large-scale neuronal recordings with deep neural network modeling. To this end, we presented natural images to awake, head-fixed macaque monkeys and monitored the spiking population activity of V4 neurons using acute electrophysiological recordings with 32-channel linear arrays spanning 1,920 µm in depth, covering the majority of the 2 mm cortical depth (Fig. 1a; Denfield et al., 2018). In each recording session, we displayed 9,000–12,075 gray-scale images from the ImageNet database (Deng et al., 2009) organized in a trial structure, where each trial consisted of 15 images, each presented for 120 ms, followed by a gray screen 1,200 ms inter-trial period. During image presentation, the monkey was trained to maintain fixation on a fixation spot offset from the center of the monitor (Fig. 1a). The spot's exact location was selected prior to each recording session such that the neurons' population receptive field (RF), determined by using a sparse random dot stimulus, was centered on the monitor. Post-hoc spike sorting of the neuronal activity recorded across 100 sessions from two monkeys isolated the single-unit visual activity of 1,224 individual V4 neurons (Fig. 1c), resulting in a large dataset of well-isolated single-unit activity in monkey V4.

To predict the responses of the recorded neurons and characterize the neurons' tuning properties, we used a deep convolutional neural network (CNN) model (Fig. 1b). Based on previous work in monkey (Bashivan et al., 2019; Cadena et al., 2019, 2022), we used a pre-trained goal-directed neural network as a non-linear feature space shared across all neurons and fitted only a simple linear-nonlinear neuron-specific readout (Lurz et al., 2020). Specifically, we chose a robust and high-

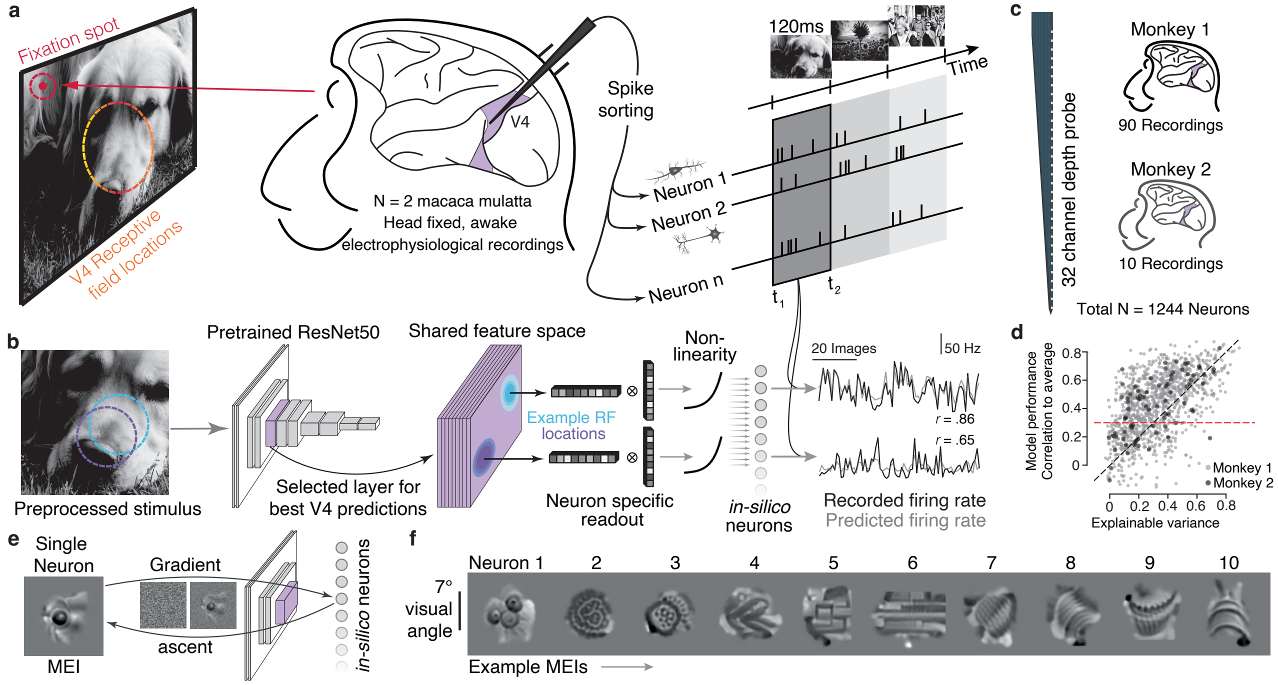


Fig. 1. Deep neural network approach captures tuning properties of individual monkey V4 neurons a, Schematic illustrating experimental setup: Awake, head-fixed macaque monkeys were presented with static natural images after fixating for 300 ms (120 ms presentation time per image, 15 images per trial, 1200 ms inter-trial period), while recording the neuronal activity in V4 using 32-channel probes. Animals were fixating on a fixation spot such that the recorded neurons' population receptive field was centered on the monitor. Post-hoc spike sorting resulted in single-unit activity of individual V4 neurons. **b,** Schematic illustrating model architecture: The pre-processed stimuli (100 x 100 pixels crop) and neuronal responses were used to train a neuron-specific read-out of a ResNet50 pre-trained on an image classification task. Specifically, we selected the ResNet50 layer with the best V4 predictions and computed the neuronal responses by passing the feature activations to a neuron-specific Gaussian readout and a subsequent non-linearity. Traces on the right show average responses (gray) to 75 test images of two example neurons and their corresponding model predictions (black). **c,** Schematic illustrating 32-channels along the probe used for electrophysiological recordings and number of recording sessions per monkey. In total, we recorded the single-unit activity of n=1,244 neurons. **d,** Explainable variance as a measure of response reliability to natural images plotted versus model prediction performance (correlation between prediction and average neural response to repeated presentations) of all cells. Dotted red line indicates a prediction performance of 0.3 used in subsequent analyses (explainable variance mean $\pm$ s.d. $= 0.33 \pm 0.19$, correlation to average mean $\pm$ s.d. $= 0.43 \pm 0.21$) **e,** Schematic illustrating optimization of most exciting images (MEIs). For each *in silico* neuron, we optimized its MEI using gradient ascent over n=100 iterations.The whole gray box (full extent) is 14.82° degrees visual angle in width and height. **f,** MEIs of ten example neurons.

performing ResNet-50 (Salman et al., 2020) as goal-directed neural network, trained on an image classification task. We selected one of its intermediate layers (layer 3.0) as non-linear feature space because it resulted in the best response predictions of the recorded V4 neurons. This yielded a correlation between response predictions and mean neuronal responses across repetitions of 0.43 (Fig. 1d), indicating that the model reliably captured key tuning properties of the V4 population. In the following, we will refer to this measure as model performance.

Treating our CNN model as a functional digital twin of the population of V4 neurons, we synthesized maximally exciting images (MEIs) (Bashivan et al., 2019; Walker et al., 2019) for individual V4 neurons *in silico* (Fig. 1e). To this end, we optimized a contrast-constrained image to produce the highest activation in the model neuron using regularized gradient ascent. The resulting MEI corresponds to the optimal stimulus of a neuron according to the model, depicting the peak of its tuning curve. We found that MEIs strongly differ across neurons, indicating selectivity for distinct stimulus features like texture, curvature and edges (Fig. 1f), which resemble the features found in the MEIs of V4 multi-unit activity (Bashivan et al., 2019). Our MEIs were also consistent with tuning properties of macaque V4, such as shape, curvature, and texture selectivity, previously, identified using parametric stimuli (e.g. Kim et al., 2019; Pasupathy & Connor, 2001; Pasupathy et al., 2020). In contrast to these previous studies, our data-driven approach enables characterization of single-neuron selectivity in a natural image context, without requiring prior assumptions about feature space or pre-selection of stimuli.

Closed-loop paradigm verifies model-derived optimal stimuli of single V4 neurons To demonstrate the model's accuracy and empirically validate that the synthesized MEIs effectively activated the recorded neurons, we developed a closed-loop paradigm for acute electrophysiological recordings of single neurons (Fig. 2a). Specifically, after fitting the readout for single-unit responses recorded in a "generation" session where natural images were shown, we selected the best six units based on model prediction for *in vivo* verification, generated their MEIs and presented them back to the animal on the same day while recording from the same neurons in a "verification" session. Single units were matched across the generation and verification session using spike waveform similarity and functional consistency of responses to the same nat-

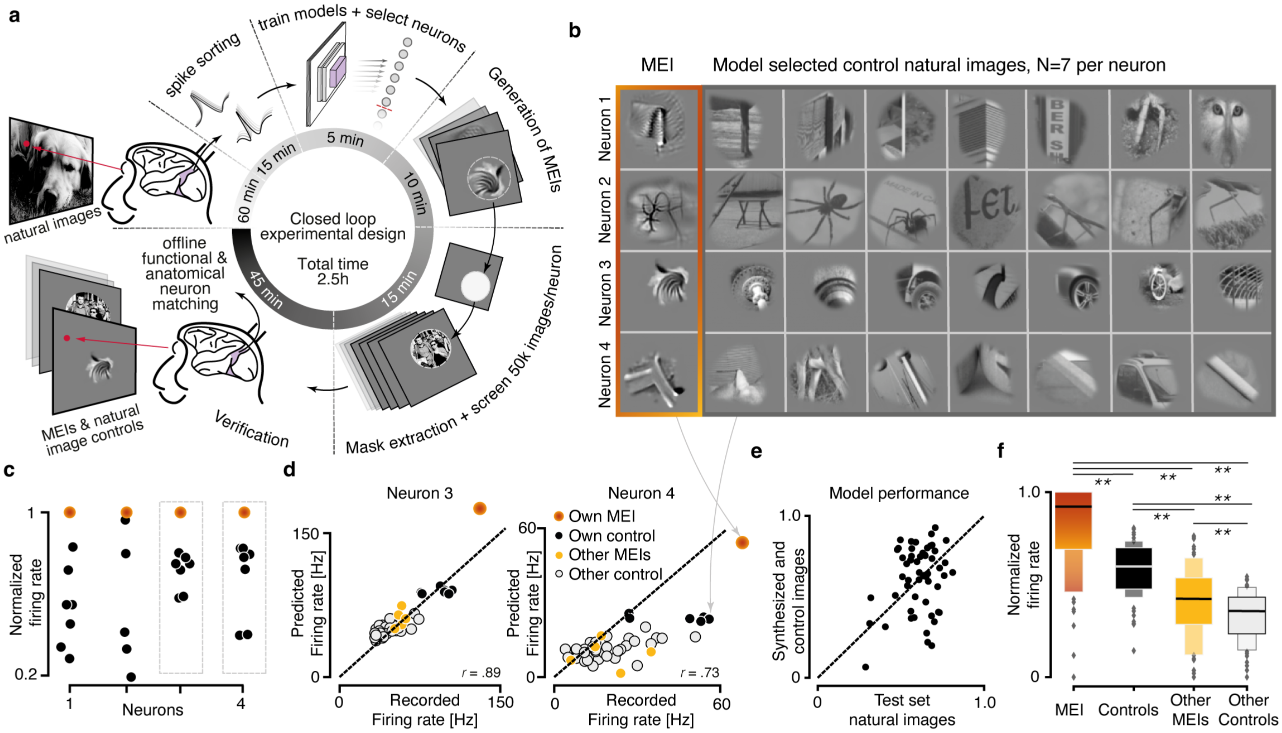


Fig. 2. Closed-loop paradigm verifies model-derived optimal stimuli of single V4 neurons a, Schematic illustrating closed-loop experimental paradigm for acute recordings in monkey V4. In brief, after recording and spike sorting of the "generation session", we train a model, select neurons for experimental confirmation, generate MEIs and identify the most exciting natural image control stimuli, and present both MEIs and controls back to the animal while recording from the same neurons in the "verification session". Functional and waveform matching of units across recordings is performed offline. **b,** MEI and the seven most exciting natural image crops, selected from 5,000 natural images, for four example neurons. Natural images were matched in size, location and contrast to the MEI. **c,** Peak-normalized recorded responses of the neurons in (b) to their MEI (orange) and control images (black; mean across n=20 repeats). **d,** Recorded versus predicted neuronal activity of two example neurons to their MEI and control stimuli, as well as to MEIs and control stimuli of other neurons. **e,** Scatter plot of model performance on the test set of natural images and the closed-loop stimuli (as shown in d, but for all neurons). Correlation to average: mean ± s.d. $= 0.61 \pm 0.11$; Synthesized and selected stimuli: mean ± s.d. 0.61 ± 0.20, $n = 55$ neurons. A paired t-test showed no significant difference $p = .61$. **f,** Distribution of peak-normalized mean responses to each neuron's MEI and control stimuli, as well as MEIs and control stimuli of other neurons for all closed-loop neurons ($n = 55$ neurons, $n = 24$ sessions, $n = 1$ monkey). P-values for a paired t-test are: MEI-Control, 3.22e-08; MEI-OtherMEIs, 2.57e-14; MEI-OtherControls, 3.06e-19; Control-OtherMEIs, 2.86e-07; Control-OtherControls, 1.62e-19; OtherMEIs-OtherControls, 2.99e-05. P-values were corrected for multiple comparisons with Bonferroni correction.

ural images (Suppl. Fig. 1). As a control stimulus for each selected unit, we presented the seven most exciting natural image crops identified by the model by screening 5,000 natural images not used during model training. Each crop was matched to the size, position, and contrast of the MEI of a particular neuron. Control stimuli qualitatively resembled the MEIs (Fig. 2b), confirming that model-synthesized features reflect elements commonly found in natural scenes.

Overall, the model predicted V4 neuronal responses to both full-field natural images and synthesized MEIs with high reliability. Despite the high structural similarity of MEI and control images, the MEI consistently elicited higher neuronal responses than the control images, as well as MEIs and control images of other neurons (Fig. 2c,d,f), indicating neuron-specific selectivity captured by the model. In addition, the model accurately predicted the neurons' average responses to their own MEIs, to control stimuli, and to the MEIs and control stimuli of other neurons of the same session (two example neurons in Fig. 2d). The absolute scale of the firing rate predictions did not always perfectly match the recorded firing rate, likely due to slow drifts in overall firing rates of some neurons (e. g. Fig. 2d, right). Nevertheless, across neurons, the model trained on the generation session generalized robustly to the verification session, with no significant difference in prediction performance between full-field natural images and synthesized MEIs ($\rho = 0.61$, Fig. 2e). Moreover, the amplitude of neuronal responses to control stimuli and MEIs of other neurons only slightly differed, suggesting that both sets of stimuli share similar low-level visual statistics despite the neuron-specific optimization of MEIs (Fig. 2f).

Topographic organization of model-derived feature selectivity in macaque V4 Studying how visual selectivity is organized in a particular brain area has revealed key principles of vision, including the pinwheel of orientation columns in primary visual cortex (Bonhoeffer & Grinvald, 1991). In monkey V4, neurons are tuned to more complex visual features like shape and texture (Kim et al., 2019; Pasupathy & Connor, 2001; Pasupathy et al., 2020; Srinath et al., 2020), but it remains unclear whether V4 tuning properties are organized in a columnar manner (Ghose & Ts'o, 1997; Hatanaka et al., 2022; Namima et al., 2025; Tang et al., 2020). We next asked whether our data-driven approach could reveal topographic patterns in feature selectivity, as inferred from model-derived optimal stimuli.

We noticed that MEIs from individual sessions tended to exhibit higher mutual visual similarity compared to MEIs from other sessions (Fig. 3a,b). While the range of preferred stimuli we found spanned a large variety from oriented and comb-like patterns, to grid-like motifs and high-level attributes, the perceived variability within many sessions was much smaller than across sessions. For example, most neurons in an example session preferred oriented and comb-like patterns (Fig. 3a), while neurons from other example sessions preferred curved edges (session 2 in Fig. 3b) and grid-like patterns (session 3 in Fig. 3b).

To quantify the perceptual similarity of MEIs within a session, we performed a simple psychophysics experiment, where human observers were presented with two sets of MEIs: (1) a selection of nine neurons recorded within one session and (2) a set of nine neurons randomly sampled across sessions. The two sets were presented side-by-side (as shown in Fig. 3c, but without the colored frames), with each set being shown on the left or right at random. The observers then had to report in a two-alternative forced-choice paradigm which set of MEIs looked perceptually more similar. On average, the observers classified MEIs of the same session as being more consistent than a random set of MEIs from different sessions for 73% of the sessions (Fig. 3d), indicating that neurons recorded along the same electrode penetration tend to share similar model-derived feature preferences. Because neurons recorded within a session are arranged roughly along the vertical axis of cortex, these results suggest a local topographic organization of feature selectivity in macaque V4. This pattern is consistent with columnar-like functional clustering but does not constitute direct anatomical evidence for columns.

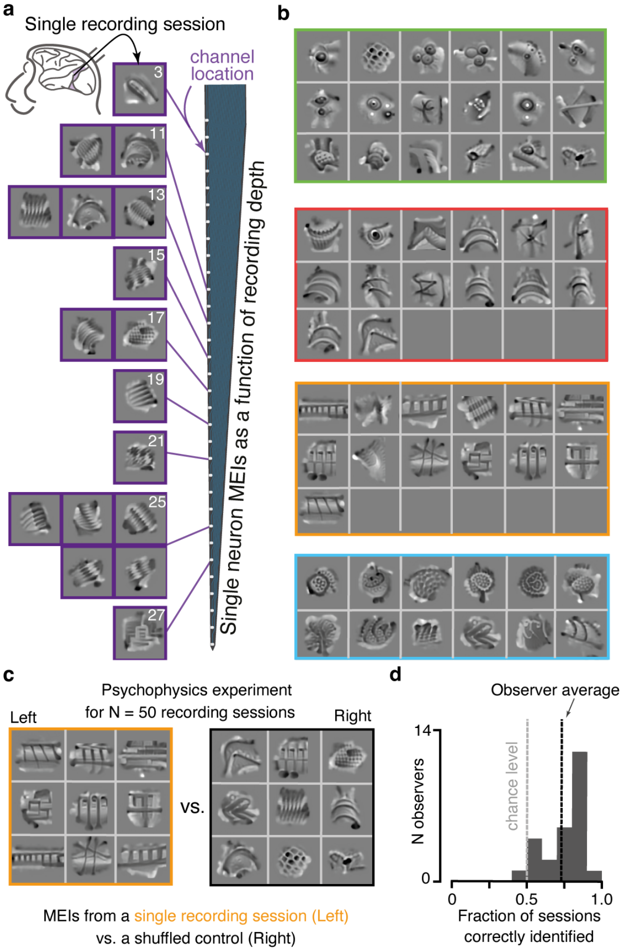


Fig. 3. Topographic organization of model-derived feature selectivity in macaque V4 a, MEIs of 17 neurons recorded in a single experimental session, arranged according to each neuron's channel location along the recording probe. Numbers indicate channel, with higher channel numbers meaning greater recording depth. **b,** MEIs of varying numbers of neurons for four different sessions (indicated by different colors). **c,** Schematic illustrating paradigm of simple psychophysics experiment. In one trial, subjects were presented with MEIs of 9 neurons recorded within one session (left) or randomly sampled from all neurons except the target session (right), and reported the location (left or right) of the set of MEIs that looked more consistent (i.e. shared the same image features). The experiment included n=50 trials/sessions. **d,** Distribution of fraction of sessions correctly identified across n=25 observers, with change level and observer average indicated by dotted lines. Mean across subjects, $= 0.73$; subject-variability in s.d., $= 0.13$; session-variability in s.d., $= 0.21$.

To further quantify whether tuning properties of V4 neurons are functionally organized in a topographic manner, we performed nonlinear dimensionality reduction to embed the MEIs in a two-dimensional space based on model-derived image feature similarity rather than recording session identity.. In contrast to V1 neurons, whose tuning properties can be compared along clearly defined axes such as orientation or spatial frequency, the complex MEI structure of V4 neurons makes it challenging to quantify the tuning similarity between neurons. To resolve this problem, we used an unsupervised deep learning technique that learns a two-dimensional image-embedding based on mutual similarity of images (Böhm et al., 2023). The model is trained using a contrastive objective: embeddings of augmented versions of the same image are attracted to each other, while embeddings of distinct images are repelled (Fig. 4a). The data augmentations used to generate image variants determine which transformations the model learns to ignore as irrelevant differences. By using random rotations, shifts, and scaling as augmentations, we ensured that MEIs differing only by these geometric transformations would cluster together in the embedding space (Fig. 4a) We chose these augmentations because they generally preserve the identity of many mid- and high-level image features: for instance, an eye remains an eye even after rotating, shifting or scaling. Moreover, we observed that multiple MEIs of the same neuron, generated by starting from different initial noise images during optimization, often exhibited variations in some or all of these dimensions (Suppl. Fig. 2). Since the training of the model is exclusively based on image identity, it does not provide any information about which recording session a particular MEI originated from. Therefore, any grouping or clustering of MEIs by session in the learned embedding space reflects only their visual similarity and not any explicit experimental labels

The resulting embedding space placed neurons with similar MEIs close to each other (neurons 1 and 3 in Fig. 4b)

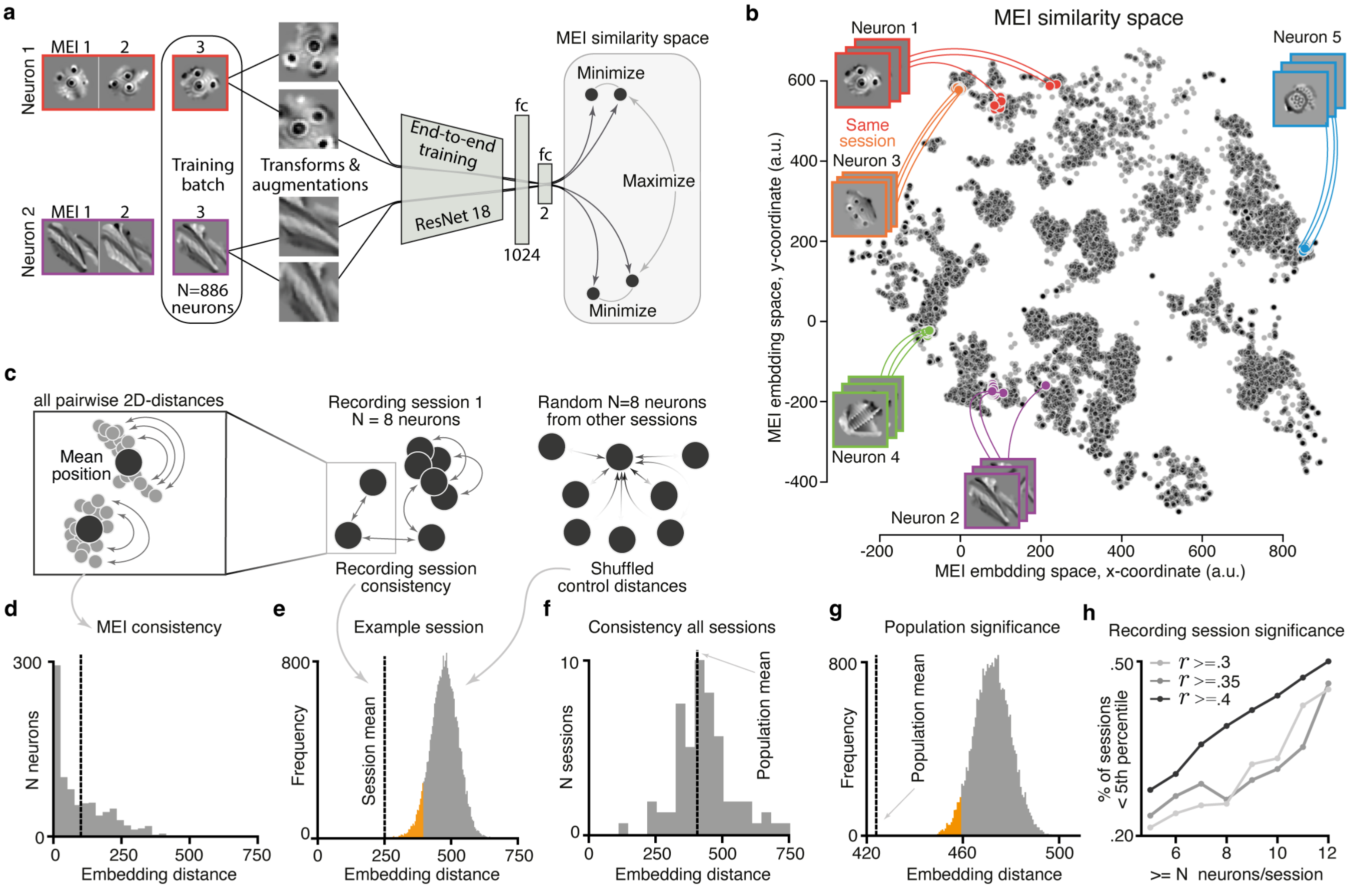


Fig. 4. Contrastive clustering of MEIs confirms topographic organization of V4 visual tuning selectivity a, Schematic illustrating contrastive learning approach to quantify MEI similarity. Per neuron, we optimize n=50 MEIs initialized with different random seeds, then select highly activating MEIs, and use one MEI per neuron (n=889) as a training batch. Each MEI is transformed and augmented twice and the model's objective then is to minimize the distance in a 2D MEI similarity space between different transforms of the same MEI, while maximizing the distance to MEI transforms of other neurons. **b,** Position of all highly activating MEIs (n=19688) of n=889 neurons in a 2D MEI similarity space, with MEIs of five example neurons indicated in different colors. Dots of the same color indicate MEIs optimized from different random seeds of the same neuron. **c,** Schematic illustrating analyses performed on the 2D MEI similarity space. We computed the pairwise 2D distances across all MEIs of one neuron to estimate MEI consistency (left), and all pairwise distances across MEIs of the same recording session to estimate recording session consistency (right). For the latter, we used the distances across a random selection of neurons from other sessions as control. **d,** Distribution of embedding-distances across MEIs of the same neuron. Vertical dotted line indicates mean of the distribution. **e,** Mean distance across neurons from one example session (vertical line), with a null distribution generated by bootstrapping distances across the same number of neurons randomly sampled from all other sessions. Orange shading indicates values <5% percentile. Note that the null distribution depends on how many neurons were recorded in each session, as it estimates the standard error of the mean for each session. **f,** Histogram of session means like in e), but for all sessions. Grand mean across all sessions is indicated by the vertical line. Mean = 423.97 ± 105.20 s.d. **g,** Mean within-session distance across all sessions from f) along with the mean null distribution across sessions in gray. The population mean significantly deviates from the null distribution ($p < 4 \times 10^{-5}$; 25,000 bootstrap samples). Orange shading indicates values <5% percentile. **h,** Percentage of sessions with the within-session distance <5% percentile of the null distribution for different numbers of neurons per session (x-axis) and different model predictions thresholds (shades of gray). The percentiles obtained from the embedding space (including all neurons above a prediction threshold of 0.3) were significantly correlated with the observer agreement (percent correct) of the psychophysics experiment ($\rho = -0.33$, p=.019, n=50 sessions).

and neurons with different MEI features far away (neurons 1 and 2). Similarly, multiple MEIs of the same neuron, generated by starting from different initial noise images during optimization, were placed nearby in the embedding space as well (groups of the same color in Fig. 4b and Fig. 4c,d). These observations suggest that the model indeed learned to embed MEIs based on image similarity. We next quantified whether neurons recorded within one session share tuning properties as suggested by the observed MEI similarity within sessions (cf. Fig. 3). To this end, we computed the mean pairwise distance in the embedding space across neurons from one session and compared it to a null distribution of distances obtained by computing the mean pairwise distance across randomly picked neurons from different sessions (Fig. 4c,e). While the within-session distance varied across sessions (Fig. 4f), on a population level, it was significantly smaller than the across-session distance (Fig. 4g). The percentage of sessions that showed a significantly smaller distance in MEI similarity than the null distribution increased with higher numbers of neurons recorded per session and with higher prediction performance of the model (Fig. 4h). For more than 12 neurons per session and with the highest performance threshold (correlation to average >0.4), half of the sessions displayed a significantly smaller within- than across-session distance, indicative for high similarity of MEIs. Importantly, the within-session distances estimated based on the embedding space significantly correlated with the observer agreement (percent of observers who correctly classified a specific session) from the psychophysics results ($\rho = -.33$, p=.019), suggesting that MEI distance in the embedding space is informative about MEI perceptual similarity. We additionally confirmed that MEIs of neurons recorded within one session are more similar to each other than to MEIs of neurons recorded in other sessions using an independent similarity metric, namely the representational similarity of MEIs in neuronal response space, which closely mimics perceptual similarity (Kriegeskorte, 2008) (Suppl. Fig. 3). Taken together, these analyses demonstrate consistent functional clustering of tuning properties among V4 neurons recorded along the same cortical penetration. This pattern is consistent with a topographic or columnar-like organization of feature selectivity across cortical depth, as inferred from model-derived representations.

V4 neurons cluster into distinct functional groups that resemble feature maps of artificial vision systems At the level of retinal ganglion cells, neurons cluster into specific functional groups or output channels (Baden et al., 2016; Goetz et al., 2022). Whether a similar functional clustering persists for cortical neurons is still an open question. For instance, in V1 it is not clear whether the distinction between simple and complex cells represents two ends of a continuous spectrum or two discrete categories (Mechler & Ringach, 2002). Previous results in mouse primary visual cortex suggest that neurons cluster according to function, but not in an entirely discrete manner (Ustyuzhaninov et al., 2019). Motivated by the structure of the MEI embedding space that exhibited isolated “islands” of MEIs (cf. Fig. 4b), we asked whether tuning properties of V4 neurons fell into similar functional groups, characterized by their preferred stimulus. To address this question, we applied hierarchical clustering (DBSCAN; McInnes et al., 2017) to the two-dimensional MEI embeddings, yielding 17 distinct functional groups (Fig. 5). MEIs within each group exhibited high perceptual similarity and shared stimulus preferences—for instance, neurons in group 11 all responded preferentially to patterns resembling eyes. Conversely, neurons in different groups showed markedly different feature selectivity, particularly for groups positioned distantly in embedding space. For example, while group 11 MEIs contained eye-resembling patterns, group 3 and group 8 MEIs displayed grid-like and comb-like textures, respectively

To ensure that the group structure in the embedding space is not an artifact of the contrastive neighborhood embedding method, we additionally used an independent method to compare within-group to across-group similarities. To that end, we computed the representational similarity of the MEIs using the predicted neuronal responses from our model. Specifically, we centered the receptive fields of all neurons in the model and compared the similarity of two MEIs via the cosine of the predicted population response vectors. Importantly, this similarity metric is unrelated to the embedding space used for clustering and, therefore, provides an independent verification of the identified functional groups. We found that MEIs of neurons assigned to the same group were significantly closer to each other in the neuronal response space than MEIs of neurons assigned to different groups (Suppl. Fig. 3): The within-group similarity in neuronal response space was significantly higher than the across-group similarity for all of the 17 functional groups.

Interestingly, the MEIs of the identified V4 clusters closely resembled the feature visualizations of single units in modern deep neural networks trained on image recognition tasks. For example, a similar preference for specific complex features can be found in the early layers, specifically layer `mixed3a`, of the InceptionV1 deep network (Olah et al., 2020; Szegedy et al., 2015). This alignment between V4 and deep network features is in line with previous results that found boundary selective units in the AlexNet deep network (Krizhevsky et al., 2012; Pospisil et al., 2018), similar to boundary neurons in monkey V4 (Pasupathy & Connor, 2002). Olah et al. (2020) manually grouped the feature visualizations into different categories like “Oriented fur,” “Eyes/circles/loops” and “Divots/boundaries” (Fig. 5b). These categories exhibit perceptual similarities to the functional groups of V4 neurons identified through hierarchical clustering, suggesting potential convergence in encoding strategies between primate and artificial vision systems. The correspondence between V4 neuronal selectivity and deep neural network features provides a basis for generating hypotheses about visual tuning properties in primate V4 beyond spatial patterns. For example, given that the artificial vision systems were trained on color images, their feature visualizations also characterize color tuning and may inform predictions about color boundary encoding in monkey V4 functional groups (Fig. 5c), amenable to verification through *in vivo* experiments. These comparisons highlight how AI models and neuroscience can inform each other, with artificial representations guiding hypotheses while neural data constrain and refine computational models.

Discussion

Our work provides evidence for a topographic and columnar-like organization of tuning to spatial patterns in visual area V4, derived from detailed single-neuron tuning properties. This organization was revealed using an iterative image synthesis approach based on a deep neural network model of neuronal activity. This model, serving as a digital twin of V4, enabled characterization of spatial pattern tuning without imposing prior parametric assumptions about feature selectivity. By combining this approach with a learned non-linear image-similarity embedding space and human psychophysics, we found that MEIs from neurons recorded along the same cortical penetration tended to be more similar to each other than those from randomly selected neurons, and that V4 neurons formed functionally coherent clusters in the embedding space. Together, these findings indicate that despite the complex-

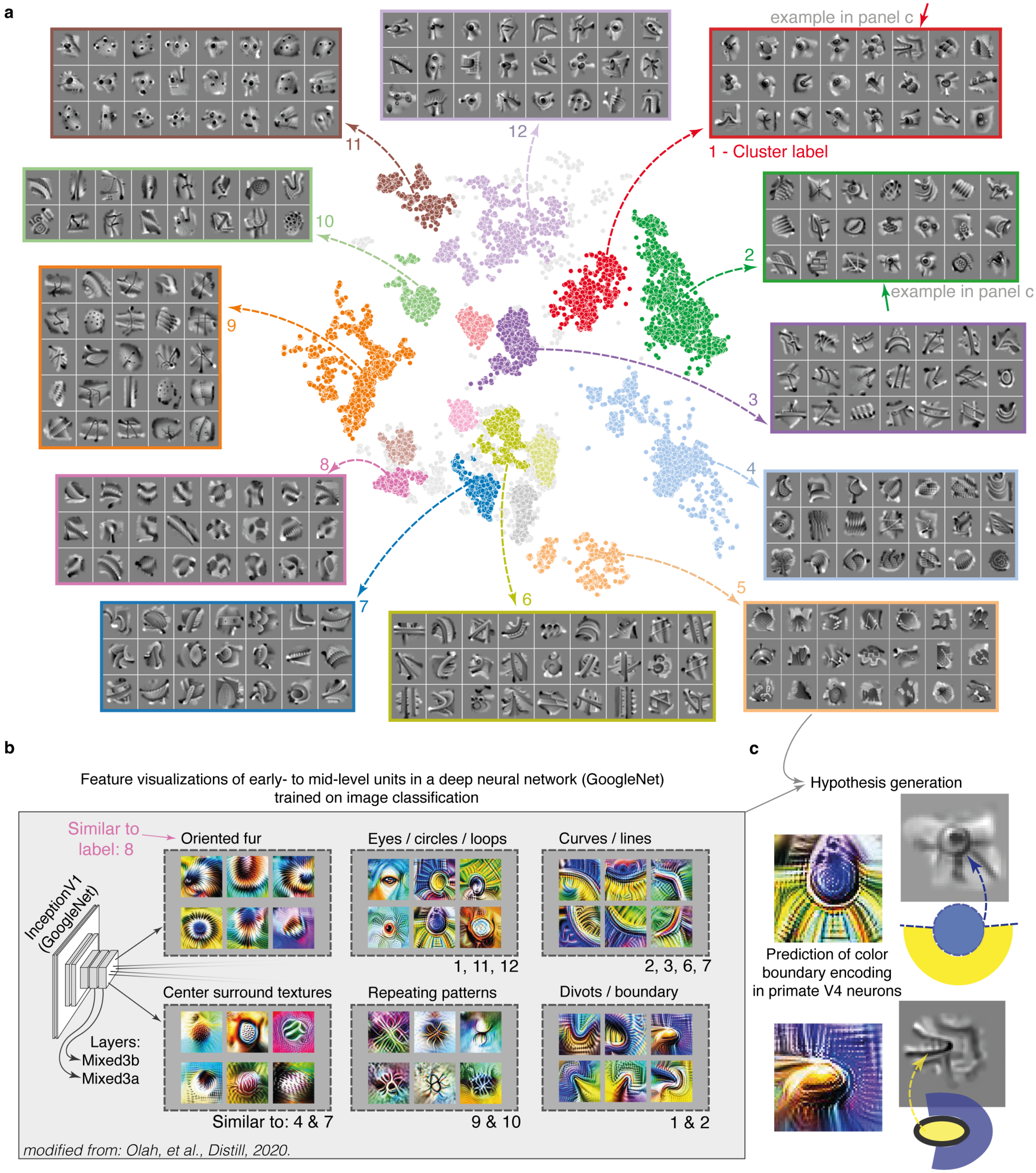


Fig. 5. V4 neurons cluster into distinct response modes that resemble feature maps of artificial vision systems a, Position of all highly activating MEIs (n=19,688) of n=889 neurons in the 2D MEI similarity space, color coded based on cluster assignment obtained from the hierarchical clustering algorithm HDBSCAN. For n=12 clusters, we show a random selection of MEIs of different neurons assigned to this cluster. For examples of the other clusters, see Suppl. Fig. 4 and for independent verification of the clusters, see Suppl. Fig. 3. Light gray dots indicate MEIs that could not be assigned to any of the clusters with high probability. **b,** Feature visualizations of early- to mid-level units in the deep neural network InceptionV1 trained in an image classification task (Olah et al., 2020). Units are grouped into distinct categories based on (Olah et al., 2020), with clusters from (a) resembling these categories indicated below. **c,** Example units of the neural network trained on image classification compared with example MEIs exhibiting similar spatial patterns. The resemblance between the two can be used to generate hypotheses, such as to predict color boundary encoding in primate V4 neurons, that can be subsequently tested experimentally.

ity of V4 stimulus preferences, feature selectivity is locally clustered in a manner consistent with columnar-like or topographic organization.

Are our MEIs consistent with previous findings? The idea to use neuronal encoding models to synthesize optimal stimuli for the brain is well established (Lehky et al., 1992) and has already successfully been used in mouse primary visual cortex (Franke et al., 2022; Walker et al., 2019), mouse retina (Höfling et al., 2022), and also macaque V4 (Bashivan et al., 2019). Additionally, there exist deep learning-based methods for synthesizing optimal stimuli which do not rely on neuronal encoding predictive models. Instead, these methods utilize a genetic algorithm to search through the space of a generative deep neural network of images **?**. Many of our MEIs contain features, such as different types of textures, that qualitatively resembled those found by Bashivan et al. (2019). However, some of our MEIs showed shape-like features, including curved strokes, corners, and even higher-level attributes such as individual eye-like stimuli (cluster 7 in Fig. 5a), which have been less emphasized in previous work.

One distinction between our study and that of Bashivan et al. (2019) is that we used silicon probes, enabling us to separate spikes from individual neurons, while they employed chronically implanted Utah arrays and optimized stimuli for single 'sites' (multi-units) that likely comprise a mix of multiple neurons. Consequently, it remains unclear from their study whether the complexity of the MEIs stems from the mixing of spikes from multiple neurons or if single neurons already prefer such intricate spatial patterns as displayed by the MEIs. Additionally, MEIs for multi-unit activity could average out specific features like shapes. By isolating and verifying MEIs at single-cell resolution, our results demonstrate that the diversity and complexity of optimal stimuli are already present at the level of individual neurons in macaque V4.

Our MEIs are also consistent with previously described tuning properties of V4 neurons. V4 is part of the ventral pathway which plays a major role in object and shape recognition (Felleman & Van Essen, 1991; Mishkin et al., 1983). Previous studies have shown that V4 cells are selective for complex shapes (Kobatake & Tanaka, 1994), and be tuned to convex and concave shapes of object boundaries at specific locations in the visual field (Gallant et al., 1993; Pasupathy & Connor, 2001). In addition to shape, V4 neurons are also known to be selective to texture (Kobatake & Tanaka, 1994). Kim et al. (2019) found that tuning of single V4 neurons can be placed along a continuum from strong tuning for boundary curvature of shapes to strong tuning for perceptual dimensions of texture. Consistent with these findings, our MEIs captured a wide range of feature preferences. On the one hand, we find MEIs that clearly exhibit curvature elements (e.g. cluster 7 in Fig. 5a) or 'eye'-like elements (cluster 1). On the other hand, many MEIs have a texture component such as 'fur' (cluster 8), dots (cluster 11), or grid-like elements (cluster 6). Interestingly, generating multiple MEIs from different starting points – known as Diverse Exciting Images (DEIs; Cadena et al., 2018; Ding et al., 2023b) – resulted in multiple MEIs that had similar shape and texture features (Suppl. Fig. 2), indicating that single cells in V4 are neither texture- nor shape-invariant. Together, our findings extend previous work by providing a single-cell, data-driven view of selectivity in V4 under naturalistic conditions.

Topological organization of MEIs in V4 One advantage of our recordings is that we can record simultaneously across cortical layers using silicon probes. This enabled us to characterize the vertical organization of the tuning selectivity to complex spatial patterns in area V4 and to assess whether tuning similarity is consistent with columnar or topographic organization. The presence of columns in V4 has been a matter of debate and controversy, predominantly studied in the domain of color (Kotake et al., 2009) or orientation (Ghose & Ts'o, 1997) because stimuli for these domains are more accessible to low dimensional parametrization. For instance, in the color domain, some studies has reported columnar organization (Conway & Tsao, 2009; Kotake et al., 2009; Tanigawa et al., 2010), while another found no evidence (Tanaka et al., 1986). Beyond that, other studies using natural images or parametric stimuli for curvature found local clustering of similarly tuned neurons (Hatanaka et al., 2022; Tang et al., 2020; Tanigawa et al., 2010; Wang et al., 2024). However, since these studies focused mainly on neurons from superficial layers, it remains unknown whether such topographic clustering extends across cortical depth.

Moreover, it is well-established that V4 neurons exhibit selectivity for complex spatial features (Oliver et al., 2024; Wang et al., 2024) that extend beyond those characterized by parametric stimuli defined by orientation, or parameterized curvature (e.g. Tang et al., 2020). Consequently, in the absence of a detailed identification of the non-parametric optimal stimuli, assessing the hypothesis of functional columns proves to be challenging. For example, if neurons are tuned to similar grid-like textures but with different orientations, using grating stimuli (Ghose & Ts'o, 1997) or predetermined parameterized curvature stimuli (Tang et al., 2020) will obscure the true underlying organization. Our deep learning based image synthesis method avoids these challenges by identifying most exciting stimuli in the high-dimensional pixel space.

To compare MEIs regardless of their spatial complexity, we used human psychophysics and deep learning techniques to assess the similarity between them, specifically employing contrastive learning to create a non-linear embedding based on similarity among MEIs (Böhm et al., 2023). We discovered that neurons across layers recorded in orthogonal penetrations to the cortical surface had MEIs that were perceptually more similar than those from randomly sampled neurons. However, the strength of this clustering effect varied across recording sessions, and was not equally evident in all cases This variability in effect size could have been caused by several factors. First, despite best efforts, our electrode penetrations may not have been perfectly orthogonal to the cortical surface, because the electrodes were aligned relative to the recording chamber. The recording chamber was implanted such that its center was orthogonal to the surface of the cortex. Since the cortex is curved, penetrations further away from the center may not have been perfectly orthogonal to the cortical surface. The brain may also have moved or have been slightly compressed during insertion, potentially resulting in slightly angled penetrations. Second, if there is a topographical organization in V4, it may feature both homogeneous zones and regions with more mixed selectivity, analogous to pinwheels in orientation maps in V1. Thus, we expect a certain fraction of penetrations to be close to such heterogeneous zones. Because extracellular electrodes record the activity of neurons in a roughly cylindrical region around the electrode, we expect a fraction of penetrations to exhibit mixed tuning even if the penetrations were perfectly vertical. Estimating what fraction of penetrations should exhibit consistent tuning is difficult, because the size of the columns and the relationships between them are not yet understood. we estimate that approximately half of the penetrations show a statistically significant bias toward certain feature selectivities, consistent with a topographic organization. While our findings point toward a columnar-like organization of functional selectivity in V4, this interpretation contrasts with recent Neuropixels recordings that revealed sparse, non-columnar clusters for shape and texture tuning (Namima et al., 2025). One possibility is that such discrepancies arise from differences in stimulus complexity, suggesting that columnar organization in V4 may depend on the specific feature space analyzed.

To more accurately delineate and map the topological organization of spatial form tuning in V4, future studies need to combine recording techniques like two-photon functional imaging with deep learning and inception loop approaches. To this end, a parallel study Wang et al. (2024) used wide-field and two-photon calcium imaging across the surface of V4 and identified a topographical map of natural stimulus preference. Their 2D map contained distinct (clustered) functional domains preferring a variety of natural image features, ranging from surface-related features such as color and texture to shape-related features such as edge, curvature, and facial features — reflecting the MEI features we identified. Although Wang and colleagues did not study the columnar organization of V4, our combined results provide evidence for the hypothesis that the functional cortical columns we find are meticulously arranged into a cortical map.

Similarity to tuning in deep networks The MEIs for V4 neurons visually resemble MEIs of units in deep artificial neural networks trained on image recognition tasks. While there are differences between biological and artificial vision (reviewed in Sinz et al., 2019), deep networks trained on large-scale vision tasks are the closest human engineered system to biological visual system we know. Several previous works have found similarities between the primate visual system and deep network representations of visual stimuli (Güçlü & van Gerven, 2015; Khaligh-Razavi & Kriegeskorte, 2014; Yamins & DiCarlo, 2016; Yamins et al., 2014) and this similarity has recently also been shown in mice Bakhtiari et al. (2021); Nayebi et al. (2021). On a single neuron level, previous work has pointed out similarities between tuning in early vision and selectivities of single units in deep networks (Krizhevsky et al., 2012; Olah et al., 2020; Zeiler & Fergus, 2014) and a recent investigation has shown that single units in deep networks can exhibit similar object boundary tuning as V4 neurons (Pospisil et al., 2018). The striking similarity between our single cell MEIs of V4 neurons and single units in the InceptionV1 architecture (Olah et al., 2020) provide an even stronger case for similarities in tuning in primate early vision and deep networks. What's more, by using this similarity we can derive predictions about the color selectivity of V4 neurons despite having shown only grayscale images in the experiment and during model training (Fig. 5b).

Future directions Our research highlights the power of applying deep learning to comprehensively characterize neuronal representations and to elucidate the relationships between brain structure and function. The concept of cortical columns offers an appealing framework for understanding cortical computation, as it decomposes complex processing into smaller, repeated building blocks. Building on prior evidence for columnar and clustered organization in sensory cortex, our findings support the hypothesis that groups of functionally similar neurons in the mid-level visual area V4 exhibit local clustering across cortical depth. Because neuronal connections within a cortical area tend to be locally dense Gilbert & Wiesel (1989), neighboring neurons within a column—both within and across layers—are much more likely to be interconnected. The purpose of a columnar architecture, therefore, is not to impose strict spatial boundaries between neurons, but rather to enhance synaptic opportunities among neurons that share similar response properties. Such local connectivity facilitates circuit computations that may underlie the emergence of new types of invariances or the construction of increasingly complex feature selectivity.

In this study, we focused on the most activating stimuli and on tuning similarities among neurons recorded along the same probe. However, recent work has extended this approach to include both the most and least activating stimuli, revealing that many visual cortical neurons encode information bidirectionally—as a contrast between two distinct feature types—rather than purely through excitation **?**. This bidirectional framework suggests that structured excitation and suppression jointly define functional microcircuits, raising the intriguing possibility that such dual-feature coding may also manifest within columnar organizations in higher visual areas.

A comprehensive understanding of local cortical organization will therefore require characterizing full tuning functions—including invariances and the diversity of encoded features among nearby neurons. Future work combining large-scale functional recordings with synaptic-resolution connectomics Bock et al. (2011); Consortium et al. (2021); Ding et al. (2023a); Reid (2012) will be essential for clarifying how local circuits implement these computations.

ACKNOWLEDGEMENTS The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Konstantin Willeke and Arne Nix. The authors would also like to thank Jan Niklas Böhm, Philipp Berens, and Dmitry Kobak for their guidance with using their recently developed t-simCNE model and Greg Horwitz for feedback on the manuscript. The authors also thank Edgar Y. Walker, George Denfield, Christoph Blessing, Mohammad Bashiri, Konstantin-Klemens Lurz, Max Burg, Shanqian Ma, Robert Petrovic, Elena Offenberg and Paul Fahey for technical support and helpful discussions. The research was funded by the Carl-Zeiss-Stiftung (KW, FHS), the Cyber Valley Research Fund (AN, FHS). FHS is further supported by the German Federal Ministry of Education and Research (BMBF) via the Collaborative Research in Computational Neuroscience (CRCNS) (FKZ 01GQ2107), as well as the Collaborative Research Center (SFB 1233, Robust Vision) and the Cluster of Excellence "Machine Learning – New Perspectives for Science" (EXC 2064/1, project number 390727645). ASE received funding for this project from the European Research Council (ERC) under the European Union's Horizon Europe research and innovation programme (Grant agreement No. 101041669) and Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), project ID 432680300 (SFB 1456, project B05). The work was also supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/ Interior Business Center (DoI/IBC) contract numbers D16PC00003, D16PC00004, and D16PC0005. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. We also acknowledge support from the National Institute of Mental Health and National Institute of Neurological Disorders And Stroke under Award Number U19MH114830 and National Eye Institute award numbers R01 EY026927 and Core Grant for Vision Research T32-EY-002520-37 as well as the National Science Foundation Collaborative Research in Computational Neuroscience IIS-2113173. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/IBC, or the U.S. Government.

Reference

Baden, T., Berens, P., Franke, K., Román Rosón, M., Bethge, M., & Euler, T. (2016). The functional diversity of retinal ganglion cells in the mouse. *Nature*, *529*(7586), 345–350.

Bakhtiari, S., Mineault, P., Lillicrap, T., Pack, C., & Richards, B. (2021). The functional specialization of visual cortex emerges from training parallel pathways with self-supervised predictive learning. *Advances in Neural Information Processing Systems*, *34*, 25164–25178.

Bashivan, P., Kar, K., & DiCarlo, J. J. (2019). Neural population control via deep image synthesis. *Science*, *364*(6439).

Bock, D. D., Lee, W.-C. A., Kerlin, A. M., Andermann, M. L., Hood, G., Wetzel, A. W., Yurgenson, S., Soucy, E. R., Kim, H. S., & Reid, R. C. (2011). Network anatomy and in vivo physiology of visual cortical neurons. *Nature*, *471*(7337), 177–182.

Böhm, N., Berens, P., & Kobak, D. (2023). Unsupervised visualization of image datasets using contrastive learning. In *The Eleventh International Conference on Learning Representations*.
URL https://openreview.net/forum?id=nI2HmVA0hvt

Bonhoeffer, T., & Grinvald, A. (1991). Iso-orientation domains in cat visual cortex are arranged in pinwheel-like patterns. *Nature*, *353*(6343), 429–431.

Cadena, S. A., Denfield, G. H., Walker, E. Y., Gatys, L. A., Tolias, A. S., Bethge, M., & Ecker, A. S. (2019). Deep convolutional models improve predictions of macaque v1 responses to natural images. *PLoS computational biology*, *15*(4), e1006897.

Cadena, S. A., Weis, M. A., Gatys, L. A., Bethge, M., & Ecker, A. S. (2018). Diverse feature visualizations reveal invariances in early layers of deep neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, (pp. 217–232).

Cadena, S. A., Willeke, K. F., Restivo, K., Denfield, G., Sinz, F. H., Bethge, M., Tolias, A. S., & Ecker, A. S. (2022). Diverse task-driven modeling of macaque V4 reveals functional specialization towards semantic tasks. *bioRxiv*, (p. 2022.05.18.492503).

Cadwell, C. R., Scala, F., Fahey, P. G., Kobak, D., Mulherkar, S., Sinz, F. H., Papadopoulos, S., Tan, Z. H., Johnsson, P., Hartmanis, L., et al. (2020). Cell type composition and circuit organization of clonally related excitatory neurons in the juvenile mouse neocortex. *Elife*, *9*, e52951.

Calabrese, A., & Paninski, L. (2011). Kalman filter mixture model for spike sorting of non-stationary data. *Journal of neuroscience methods*, *196*(1), 159–169.

Campagnola, L., Seeman, S. C., Chartrand, T., Kim, L., Hoggarth, A., Gamlin, C., Ito, S., Trinh, J., Davoudian, P., Radaelli, C., et al. (2022). Local connectivity and synaptic dynamics in mouse and human neocortex. *Science*, *375*(6585), eabj5861.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations.
URL https://arxiv.org/abs/2002.05709

Clevert, D.-A., Unterthiner, T., & Hochreiter, S. (2015). Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*.

Consortium, M., Bae, J. A., Baptiste, M., Bishop, C. A., Bodor, A. L., Brittain, D., Buchanan, J., Bumbarger, D. J., Castro, M. A., Celii, B., et al. (2021). Functional connectomics spanning multiple areas of mouse visual cortex. *BioRxiv*, (pp. 2021–07).

Conway, B. R., & Tsao, D. Y. (2009). Color-tuned neurons are spatially clustered according to color preference within alert macaque posterior inferior temporal cortex. *Proceedings of the National Academy of Sciences*, *106*(42), 18034–18039.
URL https://www.pnas.org/doi/abs/10.1073/pnas.0810943106

Denfield, G. H., Ecker, A. S., Shinn, T. J., Bethge, M., & Tolias, A. S. (2018). Attentional fluctuations induce shared variability in macaque primary visual cortex. *Nature communications*, *9*(1), 1–14.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, (pp. 248–255). Ieee.

Ding, Z., Fahey, P. G., Papadopoulos, S., Wang, E., Celii, B., Papadopoulos, C., Kunin, A., Chang, A., Fu, J., Ding, Z., et al. (2023a). Functional connectomics reveals general wiring rule in mouse visual cortex. *bioRxiv*, (pp. 2023–03).

Ding, Z., Tran, D. T., Ponder, K., Cobos, E., Ding, Z., Fahey, P. G., Wang, E., Muhammad, T., Fu, J., Cadena, S. A., Papadopoulos, S., Patel, S., Franke, K., Reimer, J., Sinz, F. H., Ecker, A. S., Pitkow, X., & Tolias, A. S. (2023b). Bipartite invariance in mouse primary visual cortex.

Ecker, A. S., Berens, P., Cotton, R. J., Subramaniyan, M., Denfield, G. H., Cadwell, C. R., Smirnakis, S. M., Bethge, M., & Tolias, A. S. (2014). State dependence of noise correlations in macaque primary visual cortex. *Neuron*, *82*(1), 235–248.

Engstrom, L., Ilyas, A., Salman, H., Santurkar, S., & Tsipras, D. (2019a). Robustness (python library).
URL https://github.com/MadryLab/robustness

Engstrom, L., Ilyas, A., Santurkar, S., Tsipras, D., Tran, B., & Madry, A. (2019b). Adversarial robustness as a prior for learned representations.
URL https://arxiv.org/abs/1906.00945

Feather, J., Leclerc, G., Mądry, A., & McDermott, J. H. (2022). Model metamers illuminate divergences between biological and artificial neural networks.
URL https://doi.org/10.1101/2022.05.19.492678

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex (New York, NY: 1991)*, *1*(1), 1–47.

Franke, K., Willeke, K. F., Ponder, K., Galdamez, M., Zhou, N., Muhammad, T., Patel, S., Froudarakis, E., Reimer, J., Sinz, F. H., et al. (2022). State-dependent pupil dilation rapidly shifts visual feature selectivity. *Nature*, *610*(7930), 128–134.

Fu, J., Shrinivasan, S., Ponder, K., Muhammad, T., Ding, Z., Wang, E., Ding, Z., Tran, D. T., Fahey, P. G., Papadopoulos, S., Patel, S., Reimer, J., Ecker, A. S., Pitkow, X., Haefner, R. M., Sinz, F. H., Franke, K., & Tolias, A. S. (2023). Pattern completion and disruption characterize contextual modulation in mouse visual cortex.

Gallant, J. L., Braun, J., & Van Essen, D. C. (1993). Selectivity for polar, hyperbolic, and cartesian gratings in macaque visual cortex. *Science*, *259*(5091), 100–103.

Ghose, G. M., & Ts'o, D. Y. (1997). Form processing modules in primate area V4. *J. Neurophysiol.*, *77*(4), 2191–2196.

Gilbert, C. D., & Wiesel, T. N. (1989). Columnar specificity of intrinsic horizontal and corticocortical connections in cat visual cortex. *Journal of Neuroscience*, *9*(7), 2432–2442.

Goetz, J., Jessen, Z. F., Jacobi, A., Mani, A., Cooler, S., Greer, D., Kadri, S., Segal, J., Shekhar, K., Sanes, J. R., & Schwartz, G. W. (2022). Unified classification of mouse retinal ganglion cells using function, morphology, and gene expression. *Cell Rep.*, *40*(2), 111040.

Güçlü, U., & van Gerven, M. A. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, *35*(27), 10005–10014.

Guo, C., Lee, M. J., Leclerc, G., Dapello, J., Rao, Y., Madry, A., & DiCarlo, J. J. (2022). Adversarially trained neural representations may already be as robust as corresponding biological neural representations.
URL https://arxiv.org/abs/2206.11228

Hansen, L. K., & Salamon, P. (1990). Neural network ensembles. *IEEE transactions on pattern analysis and machine intelligence*, *12*(10), 993–1001.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del R'ıo, J. F., Wiebe, M., Peterson, P., G'erard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., & Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362.
URL https://doi.org/10.1038/s41586-020-2649-2

Hatanaka, G., Inagaki, M., Takeuchi, R. F., Nishimoto, S., Ikezoe, K., & Fujita, I. (2022). Processing of visual statistics of naturalistic videos in macaque visual areas V1 and V4. *Brain Struct. Funct.*, *227*(4), 1385–1403.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, (pp. 770–778).

Höfling, L., Szatko, K. P., Behrens, C., Qiu, Y., Klindt, D. A., Jessen, Z., Schwartz, G. W., Bethge, M., Berens, P., Franke, K., Ecker, A. S., & Euler, T. (2022). A chromatic feature detector in the retina signals visual context changes.

Horton, J. C., & Adams, D. L. (2005). The cortical column: a structure without a function. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, *360*(1456), 837–862.

Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *J. Physiol.*, *195*(1), 215–243.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, (pp. 448–456). PMLR.

Issa, N. P., Trepel, C., & Stryker, M. P. (2000). Spatial frequency maps in cat visual cortex. *Journal of Neuroscience*, *20*(22), 8504–8514.

Jiang, X., Shen, S., Cadwell, C. R., Berens, P., Sinz, F., Ecker, A. S., Patel, S., & Tolias, A. S. (2015). Principles of connectivity among morphologically defined cell types in adult neocortex. *Science*, *350*(6264), aac9462.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, *10*(11), e1003915.

Kim, T., Bair, W., & Pasupathy, A. (2019). Neural coding for shape and texture in macaque area v4. *Journal of Neuroscience*, *39*(24), 4760–4774.

Kindel, W. F., Christensen, E. D., & Zylberberg, J. (2019). Using deep learning to probe the neural code for images in primary visual cortex. *Journal of vision*, *19*(4), 29–29.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., & Willing, C. (2016). Jupyter notebooks – a publishing format for reproducible computational workflows. In F. Loizides, & B. Schmidt (Eds.) *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, (pp. 87 – 90). IOS Press.

Kobatake, E., & Tanaka, K. (1994). Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *J. Neurophysiol.*, *71*(3), 856–867.

Kotake, Y., Morimoto, H., Okazaki, Y., Fujita, I., & Tamura, H. (2009). Organization of color-selective neurons in macaque visual area V4. *J. Neurophysiol.*, *102*(1), 15–27.

Kriegeskorte, N. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*.
URL `https://doi.org/10.3389/neuro.06.004.2008`

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*, 1097–1105.

Lehky, S. R. R. S. R. R., Sejnowski, T. J. J., & Desimone, R. (1992). Predicting responses of nonlinear neurons in monkey striate cortex to complex patterns. *J. Neurosci.*, *12*(9), 3568–3581.

Li, Z., Brendel, W., Walker, E. Y., Cobos, E., Muhammad, T., Reimer, J., Bethge, M., Sinz, F. H., Pitkow, X., & Tolias, A. S. (2019). Learning from brains how to regularize machines.
URL `https://arxiv.org/abs/1911.05072`

Loshchilov, I., & Hutter, F. (2016). Sgdr: Stochastic gradient descent with warm restarts.
URL `https://arxiv.org/abs/1608.03983`

Lurz, K.-K., Bashiri, M., Willeke, K. F., Jagadish, A. K., Wang, E., Walker, E. Y., Cadena, S., Muhammad, T., Cobos, E., Tolias, A., et al. (2020). Generalization in data-driven models of primary visual cortex. *bioRxiv*.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks.
URL `https://arxiv.org/abs/1706.06083`

Masland, R. H. (2001). The fundamental plan of the retina. *Nat. Neurosci.*, *4*(9), 877–886.

McInnes, L., Healy, J., & Astels, S. (2017). hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, *2*(11), 205.
URL `https://doi.org/10.21105/joss.00205`

Mechler, F., & Ringach, D. L. (2002). On the classification of simple and complex cells. *Vision research*, *42*(8), 1017–1033.

Merkel, D. (2014). Docker: lightweight linux containers for consistent development and deployment. *Linux journal*, *2014*(239), 2.

Mishkin, M., Ungerleider, L. G., & Macko, K. A. (1983). Object vision and spatial vision: two cortical pathways. *Trends in neurosciences*, *6*, 414–417.

Mordvintsev, A., Pezzotti, N., Schubert, L., & Olah, C. (2018). Differentiable image parameterizations. *Distill*. Https://distill.pub/2018/differentiable-parameterizations.

Mountcastle, V. B. (1957). Modality and topographic properties of single neurons of cat's somatic sensory cortex. *J. Neurophysiol.*, *20*(4), 408–434.

Namima, T., Kempkes, E., Zamarashkina, P., Owen, N., & Pasupathy, A. (2025). High-density recording reveals sparse clusters (but not columns) for shape and texture encoding in macaque V4. *J. Neurosci.*, *45*(5), e1893232024.

Nayebi, A., Kong, N. C., Zhuang, C., Gardner, J. L., Norcia, A. M., & Yamins, D. L. (2021). Shallow unsupervised models best predict neural responses in mouse visual cortex. *bioRxiv*, (pp. 2021–06).

Ohki, K., & Reid, R. C. (2014). In vivo two-photon calcium imaging in the visual system. *Cold Spring Harbor Protocols*, *2014*(4), pdb–prot081455.

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., & Carter, S. (2020). An overview of early vision in inceptionv1. *Distill*. Https://distill.pub/2020/circuits/early-vision.

Oliver, M., Winter, M., Dupré la Tour, T., Eickenberg, M., & Gallant, J. L. (2024). A biologically-inspired hierarchical convolutional energy model predicts V4 responses to natural videos. *bioRxiv*, (p. 2024.12.16.628781).

Pasupathy, A., & Connor, C. E. (2001). Shape representation in area v4: position-specific tuning for boundary conformation. *Journal of neurophysiology*, *86*(5), 2505–2519.

Pasupathy, A., & Connor, C. E. (2002). Population coding of shape in area V4. *Nat. Neurosci.*, *5*(12), 1332–1338.

Pasupathy, A., Popovkina, D. V., & Kim, T. (2020). Visual functions of primate area V4. *Annu Rev Vis Sci*, *6*, 363–385.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., & Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.) *Advances in Neural Information Processing Systems 32*, (pp. 8024–8035). Curran Associates, Inc.

Penfield, W., & Boldrey, E. (1937). Somatic motor and sensory representation in the cerebral cortex of man as studied by electrical stimulation. *Brain*, *60*, 389–443.

Ponce, C. R., Xiao, W., Schade, P. F., Hartmann, T. S., Kreiman, G., & Livingstone, M. S. (2019). Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences. *Cell*, *177*(4), 999–1009.

Pospisil, D. A., Pasupathy, A., & Bair, W. (2018). 'artiphysiology' reveals v4-like shape tuning in a deep network trained for image classification. *Elife*, *7*.

Reid, R. C. (2012). From functional architecture to functional connectomics. *Neuron*, *75*(2), 209–217.

Safarani, S., Nix, A., Willeke, K., Cadena, S. A., Restivo, K., Denfield, G., Tolias, A. S., & Sinz, F. H. (2021). Towards robust vision by multi-task learning on monkey visual cortex.
URL `https://arxiv.org/abs/2107.14344`

Salman, H., Ilyas, A., Engstrom, L., Kapoor, A., & Madry, A. (2020). Do adversarially robust imagenet models transfer better? In *ArXiv preprint arXiv:2007.08489*.

Shan, K. Q., Lubenov, E. V., & Siapas, A. G. (2017). Model-based spike sorting with a mixture of drifting t-distributions. *Journal of neuroscience methods*, *288*, 82–98.

Sinz, F. H., Pitkow, X., Reimer, J., Bethge, M., & Tolias, A. S. (2019). Engineering a less artificial intelligence. *Neuron*, *103*(6), 967–979.

Srinath, R., Emonds, A., Wang, Q., Lempel, A. A., Dunn-Weiss, E., Connor, C. E., & Nielsen, K. J. (2020). Early emergence of solid shape coding in natural and deep network vision. *Curr. Biol.*.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (pp. 1–9).

Tanaka, M., Weber, H., & Creutzfeldt, O. (1986). Visual properties and spatial distribution of neurones in the visual association area on the prelunate gyrus of the awake monkey. *Experimental Brain Research*, *65*, 11–37.

Tang, R., Song, Q., Li, Y., Zhang, R., Cai, X., & Lu, H. D. (2020). Curvature-processing domains in primate V4. *Elife*, *9*.

Tanigawa, H., Lu, H. D., & Roe, A. W. (2010). Functional organization for color and orientation in macaque V4. *Nat. Neurosci.*, *13*(12), 1542–1548.

Tsao, D. Y., Freiwald, W. A., Knutsen, T. A., Mandeville, J. B., & Tootell, R. B. (2003). Faces and objects in macaque cerebral cortex. *Nature neuroscience*, *6*(9), 989–995.

Ustyuzhaninov, I., Burg, M. F., Cadena, S. A., Fu, J., Muhammad, T., Ponder, K., Froudarakis, E., Ding, Z., Bethge, M., Tolias, A. S., & Ecker, A. S. (2022). Digital twin reveals combinatorial code of non-linear computations in the mouse primary visual cortex. *bioRxiv*.
URL `https://www.biorxiv.org/content/early/2022/02/10/2022.02.10.479884`

Ustyuzhaninov, I., Cadena, S. A., Froudarakis, E., Fahey, P. G., Walker, E. Y., Cobos, E., Reimer, J., Sinz, F. H., Tolias, A. S., Bethge, M., & Ecker, A. S. (2019). Rotation-invariant clustering of neuronal responses in primary visual cortex.

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, *9*(86), 2579–2605.
URL `http://jmlr.org/papers/v9/vandermaaten08a.html`

Van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., Gouillart, E., & Yu, T. (2014). scikit-image: image processing in python. *PeerJ*, *2*, e453.

Victor, J. D., Purpura, K., Katz, E., & Mao, B. (1994). Population encoding of spatial frequency, orientation, and color in macaque v1. *Journal of neurophysiology*, *72*(5), 2151–2166.

Walker, E. Y., Sinz, F. H., Cobos, E., Muhammad, T., Froudarakis, E., Fahey, P. G., Ecker, A. S., Reimer, J., Pitkow, X., & Tolias, A. S. (2019). Inception loops discover what excites neurons most using deep predictive models. *Nat. Neurosci.*.

Wang, T., Lee, T. S., Yao, H., Hong, J., Li, Y., Jiang, H., Andolina, I. M., & Tang, S. (2024). Large-scale calcium imaging reveals a systematic V4 map for encoding natural scenes. *Nat. Commun.*, *15*(1), 6401.

Wang, T., Yao, H., Lee, T. S., Hong, J., Li, Y., Jiang, H., Andolina, I. M., & Tang, S. (2023). A calcium imaging large dataset reveals novel functional organization in macaque v4. *arXiv preprint arXiv:2307.00932*.

Waskom, M., Botvinnik, O., O'Kane, D., Hobson, P., Lukauskas, S., Gemperline, D. C., Augspurger, T., Halchenko, Y., Cole, J. B., Warmenhoven, J., de Ruiter, J., Pye, C., Hoyer, S., Vanderplas, J., Villalba, S., Kunter, G., Quintero, E., Bachant, P., Martin, M., Meyer, K., Miles, A., Ram, Y., Yarkoni, T., Williams, M. L., Evans, C., Fitzgerald, C., Brian, Fonnesbeck, C., Lee, A., & Qalieh, A. (2017). mwaskom/seaborn: v0.8.1 (september 2017).
URL `https://doi.org/10.5281/zenodo.883859`

Willeke, K. F., Fahey, P. G., Bashiri, M., Pede, L., Burg, M. F., Blessing, C., Cadena, S. A., Ding, Z., Lurz, K.-K., Ponder, K., Muhammad, T., Patel, S. S., Ecker, A. S., Tolias, A. S., & Sinz, F. H. (2022). The sensorium competition on predicting large-scale mouse primary visual cortex activity.
URL `https://arxiv.org/abs/2206.08666`

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, *19*(3), 356–365.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, *111*(23), 8619–8624.

Yatsenko, D., Reimer, J., Ecker, A. S., Walker, E. Y., Sinz, F., Berens, P., Hoenselaar, A., Cotton, R. J., Siapas, A. S., & Tolias, A. S. (2015). Datajoint: managing big scientific data using matlab or python. *BioRxiv*, (p. 031658).

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014*, vol. 8689, (pp. 818–833).

Zhang, Y., Schriver, K. E., Hu, J. M., & Roe, A. W. (2023). Spatial frequency representation in V2 and V4 of macaque monkey. *Elife*, *12*(e81794), e81794.

AUTHOR CONTRIBUTIONS

KFW Conceptualization, Methodology, Validation, Software, Formal Analysis, Investigation, Writing - Original Draft, Visualization **KR** Conceptualization, Methodology, Validation, Software, Formal Analysis, Investigation, Data acquisition, Data curation, Writing - Original Draft, Visualization; **KF** Conceptualization, Methodology, Investigation, Experimental and analysis design, Writing - Original Draft, Writing - Review & Editing, Supervision; **AFN** Methodology, Software, Writing - Review & Editing; **SAC** Conceptualization, Methodology, Software, Investigation, Experimental and analysis design; **TS,CN,GR,SP** Data acquisition, Data curation, Methodology; **ASE** Conceptualization, Methodology, Investigation, Experimental and analysis design, Writing - Review & Editing, Supervision; **FHS** Conceptualization, Methodology, Investigation, Experimental and analysis design, Writing - Original Draft, Writing - Review & Editing, Supervision, Funding Acquisition, Project administration; **AST** Conceptualization, Methodology, Investigation, Experimental and analysis design, Writing - Original Draft, Writing - Review & Editing, Supervision, Funding Acquisi-

tion, Project administration;

Materials and Methods

Ethics statement Electrophysiological data were gathered from a pair of healthy male rhesus macaque monkeys (*Macaca mulatta*), aged 17 and 19 years, and weighing 16.4 and 10.5 kg, respectively, at the time of the study. The research adhered to NIH guidelines and received approval from the Institutional Animal Care and Use Committee at Baylor College of Medicine (permit number: AN-4367). The monkeys were individually housed in a spacious room near the training facility, in the company of approximately ten other monkeys, allowing for abundant visual, olfactory, and auditory interactions. They were maintained on a 12-hour light/dark cycle.

The Center for Comparative Medicine at Baylor College of Medicine ensured that the monkeys received regular veterinary check-ups, balanced nutrition, and environmental enrichment. Surgical procedures involving the monkeys were performed using general anesthesia and adhering to standard aseptic techniques. Postoperative pain was managed with analgesics for seven days.

Electrophysiological recordings Non-chronic recordings were conducted using a 32-channel linear silicon probe (NeuroNexus V1x32-Edge-10mm-60-177), with surgical methods and recording protocols previously outlined (Denfield et al., 2018). In summary, custom titanium recording chambers and head posts were implanted under complete anesthesia and sterile conditions. Initially, the bone remained unaltered, and only before recordings were small trephinations (2 mm) made over lateral V4, with eccentricities spanning from 1.7 to 18.3 degrees of visual angle. Recordings took place after a week of each trephination. A Narishige Microdrive (MO-97) and guide tube were used to carefully lower the probes, penetrating the dura.

Data acquisition and spike sorting Electrophysiological data were continuously collected as a broadband signal (0.5Hz–16kHz), digitized at 24 bits. The spike sorting methods employed in this study resemble those used in (Cadena et al., 2019; Denfield et al., 2018), with the code accessible at `https://github.com/aecker/moksm`. The linear array of 32 channels was divided into 14 groups, each containing six neighboring channels (with a stride of two), which were treated as virtual electrodes for spike detection and sorting. Spikes were identified when channel signals exceeded a threshold equal to five times the standard deviation of the noise.

Following spike alignment, the first three principal components of each channel were extracted, resulting in an 18-dimensional feature space utilized for spike sorting. A Kalman filter mixture model was fitted to monitor waveform drift, which is common in non-chronic recordings (Calabrese & Paninski, 2011; Shan et al., 2017). Each cluster's shape was modeled using a multivariate t-distribution (degrees of freedom = 5) with a ridge-regularized covariance matrix. The cluster count was determined based on a penalized average likelihood with a constant cost for each additional cluster (Ecker et al., 2014).

Lastly, a custom graphical user interface was used to manually confirm single-unit isolation by evaluating the units' stability (based on drifts and cell health throughout the session), identifying a refractory period, and inspecting scatter plots of channel principal component pairs.

Visual stimulation and eye tracking Visual stimuli were generated by a specialized graphics workstation and presented on a 16:9 HD widescreen LCD monitor (23.8") with a 100 Hz refresh rate and a resolution of 1920×1080 pixels, positioned at a 100 cm viewing distance (yielding approximately $\sim 63px/^{\circ}$). The monitor underwent gamma correction to ensure a linear luminance response profile. A custom-made, camera-based eye tracking system confirmed that monkeys kept their gaze within roughly $\sim 0.95^{\circ}$ around a $\sim 0.15^{\circ}$-sized red fixation target. Offline evaluations revealed that the monkeys typically fixated with greater accuracy.

Upon maintaining fixation for 300 ms, a visual stimulus was displayed. If the monkeys sustained their gaze throughout the entire trial duration (1.8 s), they were rewarded with a drop of juice at the end of the trial.

Receptive field mapping and stimulus placing At the start of each session, we determined receptive fields in relation to a fixation target using a sparse random dot stimulus. A solitary dot, spanning 1° of the visual field, was displayed on a uniform gray background, with its location and polarity (black or white) randomly changing every 30 ms. Each fixation trial persisted for two seconds. Multi-unit receptive field profiles for each channel were obtained through reverse correlation. The population receptive field location was estimated by fitting a 2D Gaussian to the spike-triggered average across channels at the time lag that optimized the signal-to-noise ratio.

The natural image stimulus occupied the entire screen. The fixation spot was adjusted so that the mean of the population receptive field was as close to the screen's center as feasible. Due to the recording sites' location in both monkeys, this positioning involved placing the fixation spot near the screen's upper border, shifted to the left.

Natural image stimuli We selected a collection of 24,075 images from 964 categories (~ 25 images per category) from ImageNet (Deng et al., 2009), transformed them to grayscale, and cropped the central 420×420 pixels. For images that were smaller than 420×420, a central crop was taken and the resulting image was re-scaled to 420×420. Each image had an 8-bit intensity resolution (values ranging from 0 to 255). From this set, we randomly chose 75 images as our *test-set*. Out of the remaining 24,000 images, we designated 20% as *validation-set* at random, leaving 19,200 images in the *train-set*. Natural images were displayed during the standalone generation recordings of 1,244 units and during the generation phase of closed-loop recordings for 82 units. Specifically, ~ 12k unique *train-set* images were displayed during the standalone generation recordings, and ~ 7.5k unique *train-set* images were displayed during the generation phase of closed-loop recordings. Across sessions, train images were randomly sampled from the *train-set* such that the full set was exhausted before cycling back through the same images. The 75 *test-set* images were displayed in every recording session. Note that selecting images from ImageNet means that the pre-trained convolutional network (see below) has likely seen our natural stimuli (but not the neuronal responses) during training on the classification task.

During standalone generation recording sessions, ~ 1000 successful trials (~ 12k train images and 75 repeated test images) were recorded, whereas 600 successful trials (~ 7.5k train images and 75 repeated test images) were recorded during the generation phase of closed-loop experiments. In both instances, each trial involved continuous fixation for 2.4 seconds, which includes 300 ms of a gray screen (intensity 128) at the beginning and end of the trial, as well as 15 consecutive images displayed for 120 ms each without any gaps. Trials contained either training/validation images that were presented only once or test images, which were repeated during the experiment.

Throughout the recording session, all trails were randomly interleaved, with test images being repeated 40-50 times during standalone generation recordings and 20 times during the generation phase of closed-loop recordings. Training and validation images were sampled without replacement, so each image was effectively displayed once or not at all. Images were upscaled using bicubic interpolation to match the screen width (1920 pixels) while maintaining their aspect ratio. The upper and lower 420-pixel bands were cropped out to cover the entire screen, effectively stimulating both classical and beyond classical receptive fields of V4 neurons. After sorting the neurons, spikes associated with each image presentation were counted within a 70-160 ms time window after stimulus onset.

Image preprocessing for model training. Starting out from an original image size of 420×420 with a resolution of $14\text{px}/^\circ$ we cropped the upper and lower bands to fit the full screen for presentation for a resulting image size of 420×236. We then cropped the images so that only the bottom center 200×200 pixels remained. Then, we downsampled the images to either 80×80 or 100×100 pixels ($5.8\text{px}/^\circ$ or $7\text{px}/^\circ$), for the closed-loop model training and non closed-loop model training, respectively.

Model architecture Our neural predictive model of primate V4 consisted of two main parts: A pretrained *core* that computes nonlinear features of input images, and a *Gaussian readout* (Lurz et al., 2020) that maps these features to the neuronal responses of the single neurons.

As the core of our model, we used a ResNet50 (He et al., 2016) which was adversarially trained on ImageNet (Deng et al., 2009) to have robust visual representations (Salman et al., 2020), which yields improved transfer-learning performance (Engstrom et al., 2019a,b; Madry et al., 2017). Interestingly, it has been previously shown that robust features not only allow for better transfer-learning but they appear to be more similar to biological networks and also improve neural predictivity (Feather et al., 2022; Guo et al., 2022; Li et al., 2019; Safarani et al., 2021). Building on previous work (Cadena et al., 2022), we selected the first residual block of layer 3 of the ResNet, `layer3.0`, to read out from, and found that the adversarially robust training with $\epsilon = 0.1$ yielded the highest predictive performance, compared to all other ResNet models and layers. The corresponding size of the output feature map at layer `layer3.0` was 1024.

The input images $\mathbf{x}$ were forwarded through all layers up to a selected layer, to output a tensor of feature maps. Importantly, the parameters of the pretrained network were always kept fixed. We then applied batch-normalization (Ioffe & Szegedy, 2015). Lastly, we rectified the resulting tensor with a ReLU unit to obtain the final nonlinear feature space $\Phi(\mathbf{x}) \in \mathbb{R}^{w \times h \times c}$ (**w**idth, **h**eight, **c**hannels) shared by all neurons.

To predict the response of a single neuron from the $\Phi(\mathbf{x}) \in \mathbb{R}^{w \times h \times c}$ we use a *Gaussian readout* (Lurz et al., 2020). For each neuron n, this readout learns the coordinates $(x^{(n)}, y^{(n)})$ of the position of the receptive field on the output tensor and extracts a feature vector $\Phi_{x^{(n)},y^{(n)},:} \in \mathbb{R}^c$ at this location from Φ. To this end, the Gaussian readout learns the parameters of a 2D Gaussian distribution $\mathcal{N}(\mu_n, \Sigma_n)$ and samples a location in feature space $\Phi(\mathbf{x})$ during each training step for every neuron n. Σ_n is initialized large enough to ensure that the entire visual field can be covered, and then decreases in size during training to have a more reliable estimate of the mean location μ_n. At inference time (i.e. when evaluating our model), the readout is deterministic and uses the fixed position μ_n. Although this framework allows for rotated and elongated Gaussian functions, we found that for our data, an isotropic formulation of the covariance – parametrized by a single scalar σ_n^2 – was performing equally well as compared to a fully parametrized Gaussian. Taken together, total number of parameters per neuron of the readout were $c+4$ (number of channels, bivariate mean, variance, and bias). This extracted feature vector $\Phi_{x^{(n)},y^{(n)},:}$ is then used in a linear-nonlinear model to predict the neuronal response. To this end, an affine function of the resulting feature vector at the chosen location was computed, followed by a rectifying nonlinearity f, chosen to be an ELU (Clevert et al., 2015) offset by one $(\mathrm{ELU}+1)$ to make responses positive (Eq. 1). The weight vector $\mathbf{w}_n \in \mathbb{R}^c$ was L_1 regularized during training.

$$\hat{r}_n(\mathbf{x}) = f\left(\sum_k \Phi_{x^{(n)},y^{(n)},k}(\mathbf{x}) \cdot w_{n,k} + b_n\right) \quad \textbf{(1)}$$

Model training We trained our model to minimize the summed Poisson loss across N neurons between observed spike counts r and the models' predicted spike counts $\hat{r}$ in addition to the L_1 regularization of the readout parameters.

$$\mathcal{L} = \sum_{i=1}^{N} (\hat{r}_n - r_n \log \hat{r}_n) + \lambda \sum_{n,k} |w_{nk}| \tag{2}$$

We trained the models either on the full dataset of n=100 recording sessions with n=1244 neurons and an image size of 100 by 100 pixels, or on an individual sessions during a closed loop recording with a reduced image size of 80 by 80 pixels in order to save time with model training and stimulus generation. For training on the full dataset, an epoch consisted of the cycling through the whole training set of 19200 images. However, because only ≈9,000 - 13,000 images were shown in each session, if a particular image was not shown in a session, we simply zeroed out the gradients of all neurons in the sessions that did not contain the image in question. When training models on a single closed loop recording, we cycled through all ≈9,000 training images that were shown in that session. In all cases, we used a batch size of 64 and after each batch, we updated the weights using the Adam optimizer (Kingma & Ba, 2014). The initial learning rate was $3 \cdot 10^{-4}$ with a momentum of 0.1.

After each epoch, we computed the Poisson loss on the entire validation set. Similar to the training set, not all images were shown in all sessions, so that we again zeroed out the loss for the sessions in question. We then used early stopping to decide whether to decay the learning rate or stop the training altogether. We scaled the learning rate by a factor of 0.3 once the validation loss did not improve for five consecutive epochs. Before decaying the learning rate, we restored the weights to the best ones based on the poisson loss on the validation set. After four early stopping steps were completed, we stopped the training. On average, this resulted in ≈ 50 training epochs, for a training time of 2 minutes for a closed loop session, and 15 minutes for the entire dataset on a NVIDIA 2080ti GPU.

Ensemble models. Instead of using a single trained model, we used a model ensemble for all of our analyses and for MEI generation. To predict the neuronal responses to individual images, we trained readout weights for each member of an ensemble of five models initialized with different random seeds and used the average prediction across the ensemble for further analyses (Hansen & Salamon, 1990). We always trained ten individual models with a different random seed, which determined the model initialization as well as the drawing of training batches. Then, we selected the five models with the highest performance on the validation set to form a model ensemble. The inputs to the ensemble model were passed to each member, and the resulting predictions were averaged to obtain the final model prediction.

Explainable variance As a measure of response reliability, we estimated the fraction of the stimulus-driven variability as compared to the overall response variability. More specifically, we computed the ratio between each neurons' total variance minus the variance of the observation noise, over the total variance (Eq. 3).To estimate the variance of the observation noise, we averaged the variance of responses across image repeats for all of the 75 repeated full-field natural image test stimuli $\sigma^2_{noise} = \mathrm{E}_j[\mathrm{Var}_t[r_t|x_j]]$ where t corresponds to the repeats and x_j represents a unique image:

$$\mathrm{EV} = \frac{\mathrm{Var}[r] - \sigma^2_{noise}}{\mathrm{Var[r]}}. \tag{3}$$

Model performance measures To measure the predictive performance of our models, we calculated the correlation to average (Cadena et al., 2022; Franke et al., 2022; Willeke et al., 2022) on the held out test set images. Given a neuron's response r_{ij} to image i and repeat j and the model predictions o_i, the correlation is computed between the predicted responses and the average neuronal response $\bar{r}_i$ to the i^{th} test image (averaged across repeated presentations of the same stimulus):

$$\rho_{\text{av}} = \text{corr}(\mathbf{r}_{\text{av}}, \mathbf{o}_{\text{av}}) = \frac{\sum_i (\bar{r}_i - \bar{r})(o_i - \bar{o})}{\sqrt{\sum_i (\bar{r}_i - \bar{r})^2 \sum_i (o_i - \bar{o})^2}}, \tag{4}$$

where $\bar{r}_i = \frac{1}{J} \sum_{j=1}^{J} r_{ij}$ is the average response across J repeats, $\bar{r}$ is the average response across all repeats and images, and $\bar{o}$ is the average prediction across all repeats and images.

Generation of MEIs We used the trained model to synthesize maximally exciting input images (MEIs) for each neuron using regularized gradient ascent (Bashivan et al., 2019; Franke et al., 2022; Walker et al., 2019). Starting out with a randomly initialized Gaussian white noise image given by $\mathbf{x} \in \mathbb{R}^{h \times w}$, with height h and width w, we showed the image to the model and computed the gradients of a single target neuron w.r.t. the image. To avoid high frequency artifacts, after each iteration we applied Gaussian blur with a σ of 1 pixel to smoothen the image. Additionally, we constrained the entire image to have a fixed energy budget, which we implemented as a maximum L2 norm of the standardized image, calculated across all pixel intensities. We chose an L2 norm of 25 for all neurons such that the resulting MEIs had minimal and maximal values similar to those found in our training natural image distribution. If the MEI exceeded the allowed norm budget after any iteration, we divided the MEI by factor f_{norm} with $f_{norm} = \|\text{MEI}\|_2 / b$. Additionally, enforced that the MEI could not contain values outside of the 8-bit pixel range by clipping the MEI outside of the bounds that correspond to 0 or 255 pixel-intensity. We used the stochastic gradient descent (SGD) optimized with learning rate of 10. We ran each optimization for 1000 iterations, without early stopping.

MEIs with transparency masks Furthermore, we employed a novel technique of synthesizing MEIs using a transparency channel based on the idea of Mordvintsev et al. (2018). Given an MEI optimized with a method as described in the section above, it is difficult to distinguish which of the MEI features are important, i.e. strongly activate the neuron, and which ones are not. Through our inbuilt L2 energy constraint, there only is finite contrast that the model is able to distribute across the image. However, this is still uninformative regarding which features of the MEI are most important. Thus, we adopted a transparency method as a differentiable parametrization (Mordvintsev et al., 2018), which jointly optimizes the MEI and a transparency mask, with the objective of making the MEI itself as transparent as possible (i.e. uninformative parts of the MEI), while still retaining the high neuronal activation of the resulting image. More specifically, we optimized an image $\mathbf{x} \in \mathbb{R}^{c \times h \times w}$, with channels c, height h and width w. We set the channels $c = 2$, treated the first channel as the MEI $\mathbf{x}_{mei}$, and the second channel as a transparent mask $\mathbf{x}_{\alpha}$, which we optimized jointly as follows: (1) In each iteration, we drew a random image $\mathbf{x}_{bg}$ from the training set as a background. (2) We clip the second MEI channel, i.e. the transparent mask, between the values 0–1, with 1 meaning that the MEI is fully opaque, and 0 meaning the the MEI is not visible at all. (3) Then, we blend the background and the MEI according to the transparency mask to get the combined image $\mathbf{x}_{combined}$ with

$$\mathbf{x}_{combined} = x_{bg} \times (1 - x_{\alpha}) + x_{mei} \times x_{\alpha} \qquad \textbf{(5)}$$

We then showed the combined image to the model to compute the neuronal response of the target neuron. With no additional constraints, the model did in fact learn to set all values of the transparent mask to 1, so that it is necessary to change the MEI objective function $\mathbf{L}_{old}$ from simply maximizing the neuronal response r to penalize alpha values of zero, as suggested by Mordvintsev et al. (2018):

$$\mathbf{L}_{new} = r \times (1 - mean(x_{\alpha})) \qquad \textbf{(6)}$$

We used the same learning rate and optimizer as for our default gradient descent MEIs. We also applied the same L2 contrast constraint of 25, but only to the MEI image channel x_{mei}. For our closed loop experiments, we ran the initial 16 sessions with the default gradient ascent MEIs, and the final 8 sessions with the transparency method, observing similar success for the *in vivo* verification. Because the transparency method seemed to generate MEIs that were less spread out and had features that were perceptually easier to identify, we used exclusively this method to generate the MEIs for the full dataset. For all the 889 out of 1244 neurons that passed our threshold of correlation to average larger than 0.3, we generated 50 MEIs per neuron for a total of 44,450 MEIs. Each MEI took on average 2 minutes on a NVIDIA 2080ti GPU, resulting in a total compute time of ≈60 GPU weeks.

Centering of MEIs In this work, we do not further analyze the transparent MEI masks. However, we use the transparent masks to center the MEIs for the similarity quantification. After MEI optimization, the MEI is located at the approximate RF location of each neuron. To center the MEIs, we find the center of mass of the transparent mask, and move the MEI such that the center of mass is at the center. All MEIs shown throughout this manuscript are shown at the center location for ease of comparison.

Psychophysics In our psychophysics experiment,the objective was to assess the perceptual similarity between MEIs originating from either a single recording session as compared to a random selection of neurons from different recording sessions. For this purpose, we devised a two-alternative forced choice task in which we showed two sets of MEIs, arranged in 3x3 grids as the basis for the similarity evaluation task. Participants were instructed to simply report which of the MEIs in the two grids looked more similar to each other. No other information apart from the displayed MEIs was provided.

To obtain the images for the task, we randomly chose 50 out of 52 sessions that had a minimum of 9 well-predicted neurons, using the threshold of correlation to average larger than 0.3. From these selected sessions, the top 9 well-predicted neurons were identified, resulting in a total pool of 450 neurons. For each session, participants were presented with images from the top 9 predicted neurons and a randomly selected set of 9 additional neurons from the pool, sampled without replacement. Specifically, for each neuron, we presented five MEIs, generated by starting from a different initial noise image during optimization, as images and compiled in a GIF. Participants were shown the two example sessions which were not included in the study, together with the correct solution. The subjects had unlimited time to make their choice.

Subject recruitment involved lab members that were familiar with the concept of an MEI from all research groups contributing to this article. In total, we recorded the responses of N=25 observers.

Closed-loop procedure Our closed loop experiments were composed of a generation and verification session. In the generation session, we recorded the electrophysiological responses of V4 neurons to full-field gray-scale natural images for 600 trials, corresponding to 9,000 unique images. After showing of all of these trials, we stopped and promptly restarted the neurophysiological acquisition system to continue recording neuronal activity while we processed the data from the generation session. During this analysis period, we isolated single units and used these units to train several models to predict neuron responses to natural images. We produced an ensemble of the best five models, used this ensemble to select the six highest predicted units and generate ten MEIs per unit. We computed the mask of each MEI by calculating the z-score of each pixel value and setting a threshold ($z = 0.35$) to isolate the area within the mask. We then created a convex hull to close any holes and smoothed the edges with a Gaussian filter ($\sigma = 2$). We used this masking procedure for our natural image control selection procedure.

Specifically, we screened a distinct set of 5,000 natural images, different from our training, validation, and test set images. This set was selected randomly from a database of 100k images, and we screened this same image set for all neurons across all closed-loop experiments. Each image was masked with the ten MEI masks of each unit and re-normalized to the contrast of the MEIs, resulting in 50,000 total screened images for each unit. We selected the seven highest activating masked natural images of each unit to use as controls for the verification recording.

Closed-loop stimulation paradigm During the verification session of the closed-loop experiment, we displayed one MEI and seven controls for each of the six units. These stimuli were centered on the RF and repeated 10 times throughout the experiment. We also showed the same full-field test set natural images 20 times each, to enable functional response matching between the generation and verification units. Each trial contained 15 images, composed of either (i) randomly interleaved generated stimuli (i.e. MEIs and masked controls) or (ii) full-field test set natural images, as described previously. Trials were randomly interleaved. In total, 100 trials of full-field images and 180 trials of generated stimuli were shown.

Closed-loop unit matching To assess the stability of our closed-loop neurons, we developed a procedure for spike waveform matching single units between the generation and verification session recordings. To do this, we performed a *post hoc* spike sorting of the full recording session, which was produced by stitching together the raw data files of the generation and verification recordings. The verification recording includes the analysis period immediately after the generation recording, during which we isolate single units and generate their MEIs. This continuity of recording allows us to more easily track single units throughout the experiment. Furthermore, spike sorting the full session allows us to use a drift correction of the spike sorting model based on a kalman-filter across the entire experiment, which aided the accuracy of our single unit tracking.

We then executed a two-step matching procedure using the spike sorting results of both the generation and full sessions. To determine potentially matched units, we first took the spikes of the generation session units and assigned them to full session units. We then calculated the proportion of spikes that were assigned exclusively to one unit in both the generation and verification sessions. If a unit in the generation session had at least 95% of its spikes assigned to only one single unit in the full session, this was considered a possible match. To confirm true matches, we assigned the full session spikes to the generation session units. If a unit in the full session had at least 95% of its spikes assigned to the potentially matched single unit in the generation session, these units were verified to be a match, and thus certified as stable. We additionally assessed neuron stability by computing the functional consistency of each unit. To compute this measure, we used the test set of 75 full-field natural images that were shown in both generation and verification session and calculated the Pearson correlation to the repeat-averaged responses in both sessions. We set a minimum functional consistency threshold of 0.5. Taken together, we were able to waveform match 82 neurons from 24 closed-loop sessions, with 27 of these failing to meet the functional consistency criterion, resulting in n=55 neurons for the analysis of the closed-loop paradigm.

MEI similarity quantification using contrastive learning We used a recently proposed method of contrastive learning (Böhm et al., 2023) to quantify the perceptual similarity of our neurons' MEIs. The method is based on SimCLR (Chen et al., 2020), a method of self-supervised contrastive learning of visual representations in which images are embedded in a high-dimensional space. In this space, augmentations of the same image are learned to have small distances while simultaneously training the model to increase the distance to all other images. Böhm et al. (2023) adapted this procedure for a two-dimensional space by combining the ideas of contrastive learning with 2-d neighbor embeddings as used in t-SNE (van der Maaten & Hinton, 2008). Their new method, called t-SimCNE, achieves this by changing the similarity function between image embeddings from cosine-similarity used in SimCLR, which would constrain the embeddings to the unit circle, to Euclidean distance $d_{ij} = \|\mathbf{z}_i - \mathbf{z}_j\|$ between the embeddings $\mathbf{z}_i$ and $\mathbf{z}_j$ of two augmentations i and j of the same image. We use this method to train a model end-to-end to embed our MEIs using the loss function as proposed by Böhm et al. (2023):

$$\ell_{t-\mathrm{SimCNE}}(i,j) = -\log \frac{1/\left(1+d_{ij}^2\right)}{\sum_{k\neq i}^{2b} 1/\left(1+d_{ik}^2\right)}. \quad (7)$$

In our adaptation, i and j correspond to two data augmentations of the same MEI, and z denotes the 2-d embedding, which is the model output. For a given batch size b, the resulting batch size is $2b$ because each MEI is augmented twice. As pointed out by the authors of SimCLR (Chen et al., 2020), this approach requires careful selection of data augmentations, as well as choosing the largest batch size possible.

Taken together, our training procedure consists of these steps: We first select all 889 neurons above the model performance threshold of correlation to average larger than 0.3, as described earlier. For each neuron, from the 50 MEIs that we optimized per neuron, we select only the MEIs that elicited at least 90% of the predicted firing rate of the highest-acivating MEI, which resulted in anywhere from 3 to 42 MEIs per neuron, with a total of n=19,688 MEIs used in this and subsequent analyses. This selection was important as there can be failures during optimizing MEIs, resulting in noisy images which the contrastive learning algorithm was very sensitive to. Next, from all the MEIs that we selected, we assembled a batch of size n=889, by randomly drawing one MEI for each neuron.

Subsequently, we preprocessed the MEIs and applied the data augmentations in the following order: (1) centering the MEIs as described above, (2) center cropping to resize the MEIs from 100x100 to 75x75 pixels, (3) random rotation between 0 and 30 degrees, (4) random rescaling between 32x32 and 75x75 pixels, (5) random cropping from the resulting image to a final size of 32x32 pixels. Steps (2) to (4) were applied twice to each MEI to obtain two augmentations per MEI. The resulting training batch thus consisted of a randomly drawn MEI per neuron, augmented twice, for a total batch size of 889*2=1,778 images, with the model's objective being to minimize the distance between each pair, and to maximize the distance from all-to-all images which are not pairs. It is important to point out that with this training scheme, the trained model was given no information about recording sessions or which MEIs belong to which neuron. Once the model was fully trained, we obtained the full 2-d embedding of all MEIs from all neurons by applying transformations (1) and (2), i.e. centering and center cropping the MEIs, as well as rescaling the MEIs to 32x32 pixels, and subsequently showing the resulting images to the model. Finally, We obtained a 2-d location of each neuron by taking the mean location over the embedding locations of all MEIs per neuron.

Contrastive learning: model architecture and training We closely followed the t-simCNE authors (Böhm et al., 2023) in our choice of model architecture and training paradigm. As a model backbone, we employed a randomly initialized ResNet18 (He et al., 2016) with an output size of 512 and a reduced kernel size of the first convolutional layer from 7×7 to 3×3. Following the authors, we also added one hidden ReLU layer with n=1,024 units followed by a linear output layer of n=128 units, which was reduced to n=2 during the final stage of training. The training consisted of three stages: First, the model was trained for 3,000 epochs with the output layer size of 128. In the second stage, the output layer was disregarded and replaced with a linear output layer of n=2, followed by a training of only this layer for 200 epochs while the rest of the model was frozen. Lastly, in the third stage, the whole model was fine-tuned for another 1,000 epochs with a reduction of the learning rate by a factor of 1,000. In each of the three stages, we used the initial learning rate of $0.03 \cdot b/256 \approx 0.1$ with b=889, preceded by linear warm-up for ten epochs, followed by cosine annealing (Loshchilov & Hutter, 2016) with a final learning rate of 0. In each epoch, we accumulated the loss from 10 batches and optimized the model using SGD with a momentum of 0.9.

Contrastive learning: within-neuron MEI distances As mentioned above, it is crucial that our contrastive learning training scheme did not provide the model with any information about either the recording sessions or the neurons' identity of each MEI. Therefore, an important sanity-check for the trained model is to analyze the 2-d embedding distances of all MEIs that we optimized for a single neuron. We performed this analysis by simply calculating all pairwise distances and taking the average of all MEIs per neuron.

Contrastive learning: Session distances We first averaged all pairwise distances across neurons recorded within one session to obtain a within-session distance in the 2-d embedding space. We then compared this within-session distance to a shuffled control. For this control, we used bootstrapping to sample at random the same number of neurons as in a given session, with the constraint that they cannot be from the session in question and that all randomly drawn neurons originate from different sessions. We then computed the mean pairwise distance for these shuffled neurons, and repeated this process 25,000 times for each session to get the null distribution. We then calculated the percentile of the true within-session distance against each sessions' null distribution to obtain a p-value.

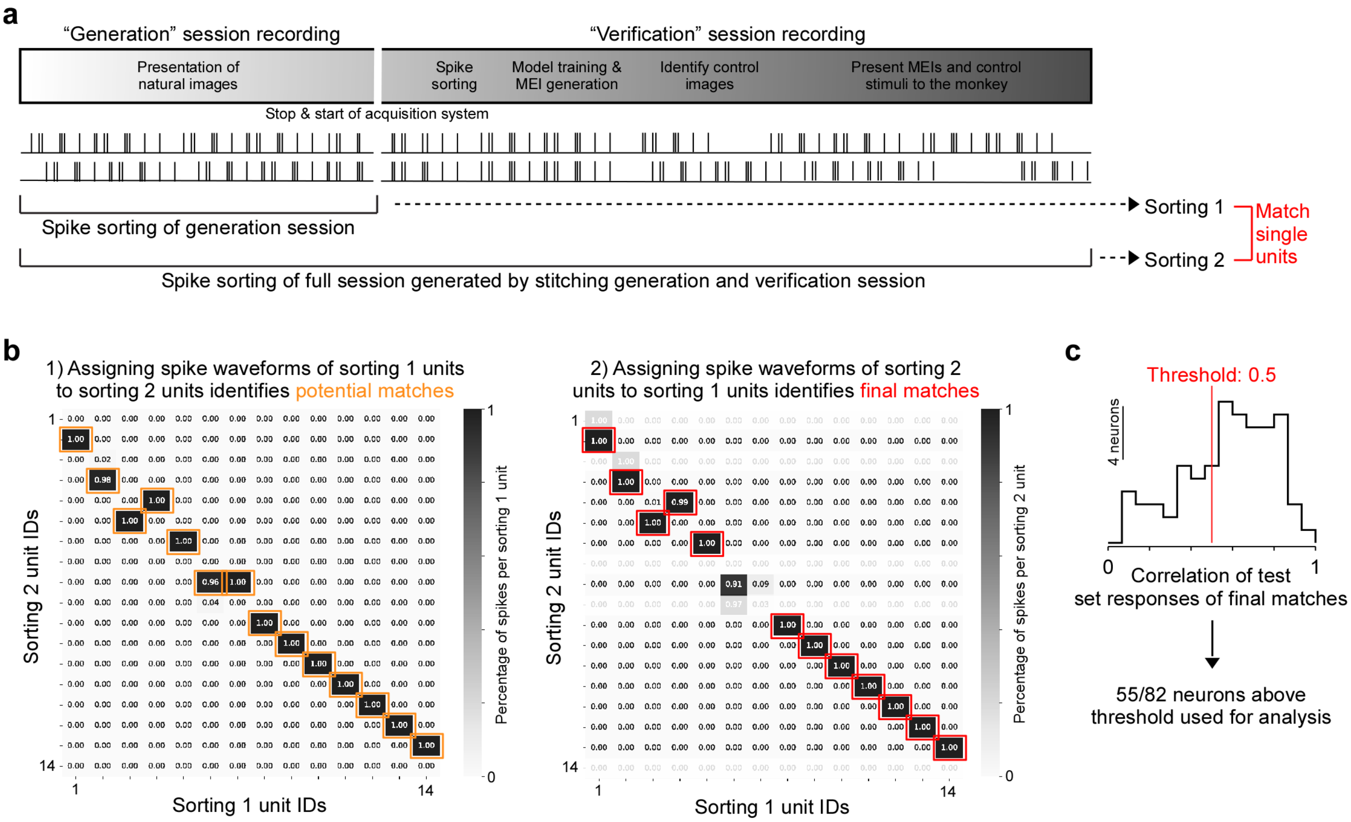
Contrastive learning: HDBSCAN cluster cutting After we obtained the 2-d embedding space of MEIs, we clustered this entire two-dimensional space using hierarchical density-based spatial clustering of applications with noise (HDBSCAN McInnes et al., 2017), the same clustering that was employed by the t-simCNE authors (Böhm et al., 2023). We searched over the parameter grid minimum cluster size $\in \{100, 125, \ldots, 400\} \times$ and minimum samples, which indicates the minimum number of samples to be considered non-noise, $\in \{5, 15, \ldots, 145\}$ with the constraint min-samples $\leq$ min-cluster-size. The resulting clusterings will create one clustered group labelled -1, for MEIs that were unable to be clustered. We selected the parameters min-cluster-size = 200 and min-samples = 10 which resulted in the lowest number of unassigned MEIs. Using this approach, the MEIs of each neuron get assigned a label, with the possibility that not all MEIs of single neuron were assigned to the same cluster. We assigned a cluster ID to a neuron if more than half of its MEIs had the same cluster ID. Furthermore, we excluded neurons for which the more than half of the MEIs were in the unassigned category -1 of the HDBSCAN algorithm. Based on these two criteria, 828 out of 889 neurons were given a valid clustering ID.

Representational similarity of MEIs As an independent way to verify the similarity of MEIs, we computed their representational similarity. For this purpose, we used the centered MEIs, and computed the responses of all 889 well-predicted neurons from our neural predictive model to each MEI. We controlled for the different RF positions of each *in silico* neuron by artificially centering the *in silico* neuron's RF to the center by setting the readout location of the Gaussian readout to the center of the core output. We used this approach for each model ensemble member individually and again took the model ensemble average as the predicted neuronal activity. Consequently, for each MEI, we obtained a response vector of length 889, which we then used for pairwise comparisons using cosine similarity. We then compared the average cosine similarity either within clusters or within recording sessions to the average similarity across sessions/clusters. To obtain a p-value for the similarity within vs. across clusters, we computed the similarity matrix 100 times. In each of the 100 runs, per neuron we drew at random one of the neurons MEIs and computed the population response vector for that MEI. We thus ended up with 100 similarity matrices, for which the pairwise comparisons were thus computed based on different MEIs. For each cluster, we then computed the average similarity across the 100 runs, and compared this similarity. We then took (1) the average cosine similarity within each cluster across the 100 runs, and (2) for each of the 100 runs, for each cluster we average the similarity to all other clusters. We report the p-value as percentile of the within-session similarity to the distribution of across-session similarity.

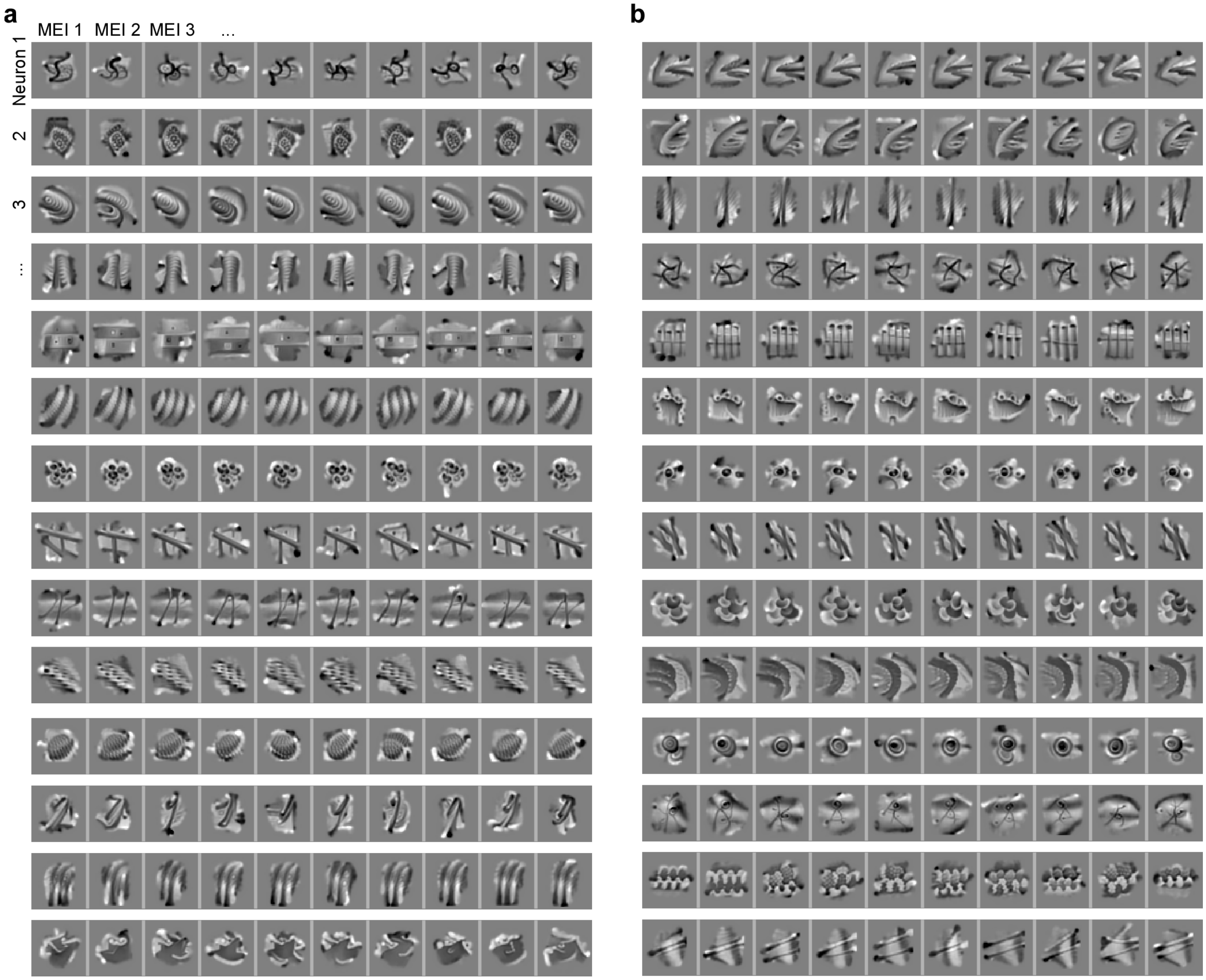
Data and code availability

The analysis code will be publicly available in an online repository latest upon journal publication. Our coding framework uses Pytorch (Paszke et al., 2019), Numpy (Harris et al., 2020), scikit-image (Van der Walt et al., 2014), matplotlib (Hunter, 2007), seaborn (Waskom et al., 2017), DataJoint (Yatsenko et al., 2015), Jupyter (Kluyver et al., 2016), and Docker (Merkel, 2014). We also used the following custom libraries and code: `neuralpredictors` (https://github.com/sinzlab/neuralpredictors) for torch-based custom functions for model implementation, `nnfabrik` (https://github.com/sinzlab/nnfabrik) for automatic model training pipelines using DataJoint, `nnvision` (https://github.com/sinzlab/nnvision) for specific model definitions, analysis, and figures.

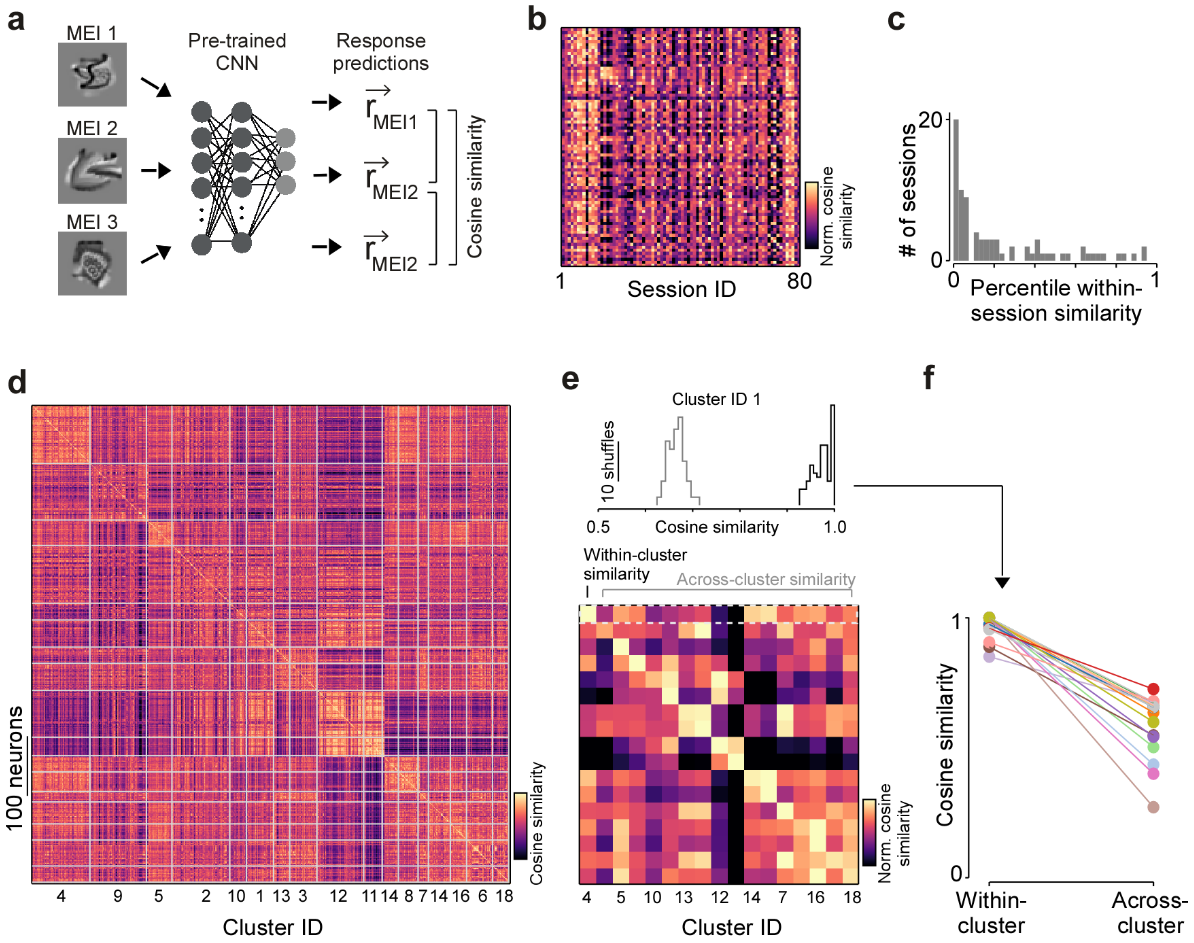
Supplementary Information



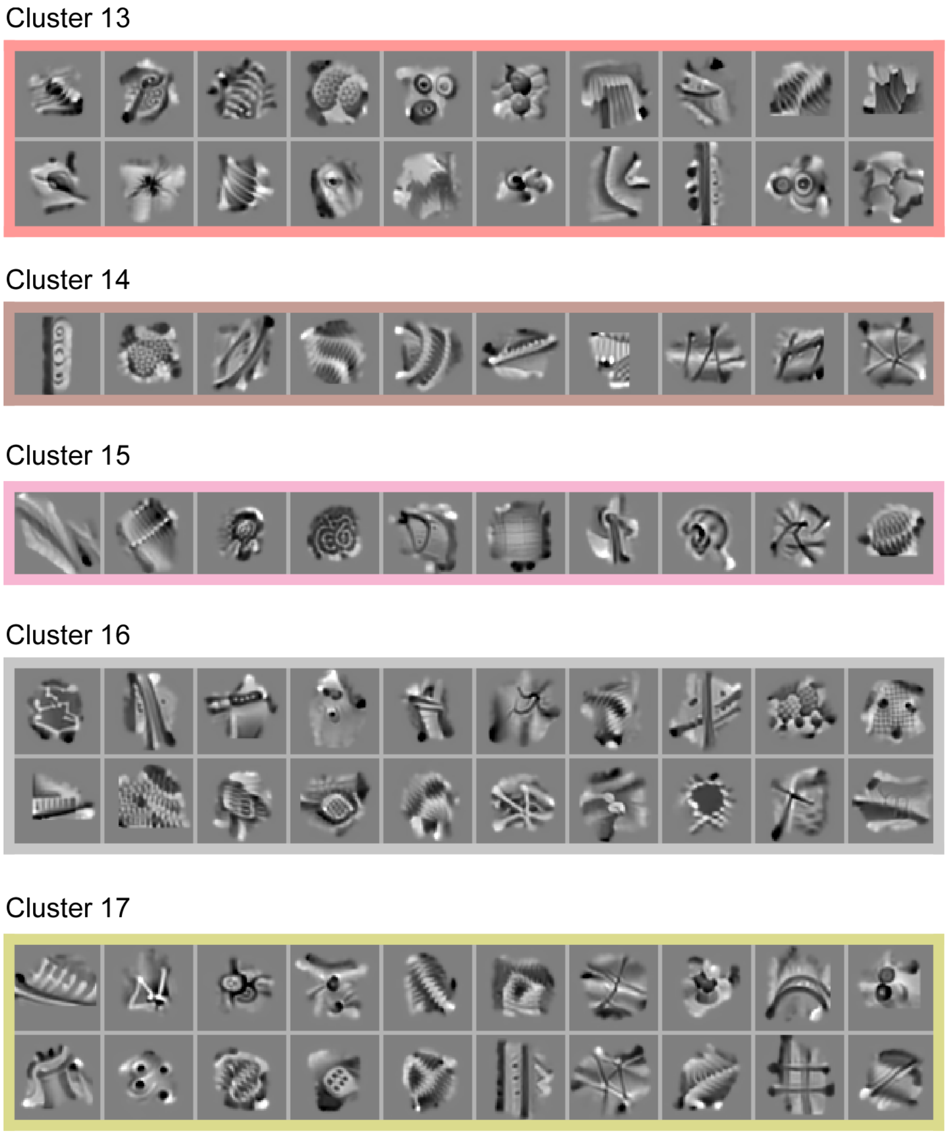
Supplemental Fig. 1. Spike waveform and functional matching of single units across recordings. a, Schematic illustrating spike sorting of the closed-loop experimental paradigm. The "generation session" is spike sorted directly after the recording, resulting in "Sorting 1". This data is then used for model training and optimization of MEIs, which are presented back to the animal during the "verification session". The verification session recording starts immediately after the generation session recording ends, to ensure a continuous monitoring of the recorded units over time. After the experiment, the generation and verification session recordings are concatenated ("full session") and spike sorted, resulting in "Sorting 2". **b,** Unit matching based on spike waveforms across Sorting 1 (generation session) and Sorting 2 (full session) for an example session. The left plot shows the percentage of spikes of the Sorting 1 units assigned to the units of Sorting 2. Units were assigned by passing the principal components of each spike, extracted using the Sorting 1 Gaussian Mixture model (GMM), to the Sorting 2 model. For a potential match (orange), at least 95% of the spikes of a single unit of Sorting 1 had to be assigned to an individual unit of Sorting 2. The right plot shows the percentage of spikes of Sorting 2 units assigned to the units of Sorting 1. For a final match (red), at least 95% of the spikes of a single Sorting 2 unit had to be assigned to the potential match of Sorting 1. **c,** Distribution of correlations of mean test set responses for all final matches. We only included matched units into the analysis, if their functional correlation was at least 0.5.



Supplemental Fig. 2. Diverse model-derived stimuli of individual monkey V4 neurons. a, For a set of 14 example neurons, we show 10 MEIs per neuron, generated from different starting points (random seeds) during the MEI optimization. The different MEIs exhibit the same visual feature but somewhat differ with respect to orientation, scale and position. **b,** Same as (a), for another set of 14 neurons.



Supplemental Fig. 3. Similarity of optimal stimuli in neuronal response space. a, MEIs of three example neurons. Right shows a schematic illustrating how we compared the similarity of MEIs using representational similarity. In brief, each MEI was presented to the trained CNN model that was used to produce the MEIs to obtain a response vector. The response vectors were then compared using cosine similarity. **b,** Mean cosine similarity of MEIs within a single recording session (diagonal) and across recording sessions for n=88 sessions, peak-normalized for each row. **c,** Distribution of percentiles of within-session similarity. For example, a percentile of 0.05 means that the MEI similarity within the session was larger than the MEI similarity to 95% of the other sessions. For n=30/55 sessions, the percentile was <0.05. **d,** Cosine similarity of MEIs of n=889 neurons, sorted based on cluster assignment (cf. Fig. 5). **e,** Mean cosine similarity of MEIs within a cluster (diagonal) and across clusters, peak normalized per row. The matrix depicts the mean across n=100 similarity matrices, each generated based on a different random selection of MEIs per neuron. Top shows distribution of cosine similarity within an example cluster (black) and the mean similarity to all other clusters (gray). **f,** Mean within-cluster and across-cluster similarity for all clusters.



Cluster 13

Cluster 14

Cluster 15

Cluster 16

Cluster 17

Supplemental Fig. 4. Overview of optimal stimuli of V4 response modes. a, MEIs of example neurons for five response modes not shown in Fig. 5.