A TALE OF TWO TAILS: PREFERRED AND ANTI-PREFERRED NATURAL STIMULI IN VISUAL CORTEX

Anonymous authors
Paper under double-blind review

ABSTRACT

A fundamental quest in neuroscience is to find the preferred stimulus of a sensory neuron. This search lays the foundation for understanding how selectivity emerges in the primate visual stream—from simple edge-detecting neurons to highly-selective face neurons—as well as for the architectures and activation functions of deep neural networks. The prevailing notion is that a visual neuron primarily responds to a single preferred visual feature, like an oriented edge or object identity, resulting in a "one-tailed" distribution of responses to natural images. However, surprisingly, we instead find "two-tailed" response distributions of primate visual cortical neurons, suggesting that these neurons have both preferred *and* anti-preferred stimuli. We verified the existence of anti-preferred stimuli by recording responses from macaque V4 to model-optimized stimuli. We find that these anti-preferred stimuli are important for shaping a neuron's tuning, as only a small number of preferred and anti-preferred images are needed to predict a neuron's responses to natural images. Moreover, in a psychophysics task, humans rely on anti-preferred images to interpret and predict V4 stimulus tuning; this was not the case for internal units from a deep neural network. Interestingly, we find that the features of preferred and anti-preferred images to be seemingly unrelated, suggesting that V4 neurons encode a broader range of features—not just those they prefer—which in turn enriches the V4 population's representational basis for flexible downstream readouts. Overall, we establish anti-preferred stimuli as an important encoding property of V4 neurons. Our work embarks on a new quest in neuroscience to search for anti-preferred stimuli along the visual stream as well as to better understand how feature selectivity arises in visual cortex and deep neural networks.

1 INTRODUCTION

Since the first recordings of action potentials from sensory neurons (Hartline, 1938), neuroscientists have searched for the stimulus features that a neuron prefers. Hubel and Wiesel famously identified the stimulus preferences of early visual cortical (V1) neurons as oriented edges (Hubel and Wiesel, 1962). Deeper into visual cortex are neurons with remarkable selectivity, such as "Jennifer Aniston" neurons that only respond to images of the celebrity, regardless of her profile or hairstyle (Quiroga et al., 2005). This has spurred on new machine learning approaches to identify a visual neuron's preferred stimulus—the stimulus that maximizes a neuron's response (Cowley et al., 2017a; Abbasi-Asl et al., 2018; Ponce et al., 2019; Bashivan et al., 2019; Gu et al., 2022; Pierzchlewicz et al., 2024). Moreover, the concept of a preferred stimulus has been at the heart of modeling visual neurons. For example, the linear-nonlinear (LN) model used to describe retinal ganglion cells and simple V1 neurons (Chichilnisky, 2001; Rust et al., 2005) filters the input to detect a single stimulus pattern (e.g., a localized, oriented edge). The presence of the pattern causes the activity to surpass a ReLU-like threshold, while all other stimulus patterns fail to reach this threshold, silencing the output. This results in a "one-tailed" response distribution (Fig. 1**a**, top row). The deeper units in a task-driven DNN—made up of cascading layers of LN models—achieve the sparse selectivity found in higher-order visual cortex. Indeed, the response distributions of DNN units in deeper layers typically have one extreme tail (Fig. 1**a**, middle row) with a few select stimuli evoking large responses. Unexpectedly, when we recorded from real neurons in macaque V4—a higher-order visual area known for encoding texture, shape, color, etc. (Gallant et al., 1996; Pasupathy and Connor, 1999)—we expected to see similar one-tailed response distributions to natural images. Instead, we found response distributions with two distinct tails (Fig. 1**a**,

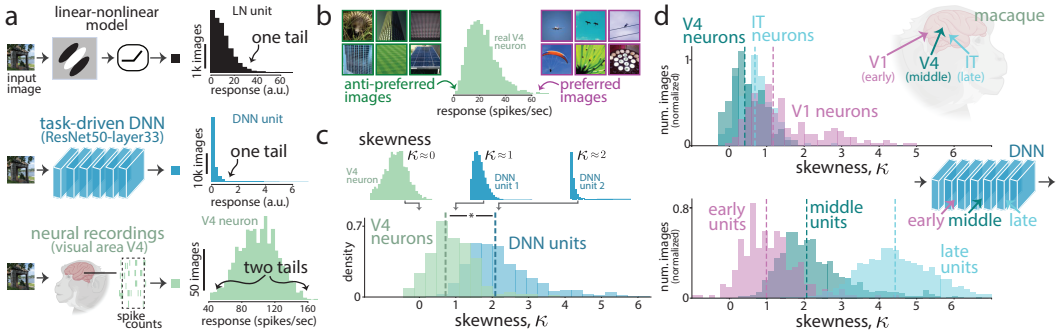


Figure 1: V4 neurons have two-tailed response distributions. a. Response distributions for a Gabor filter (*top*), DNN unit (*middle*), and a real neuron from visual area V4 (*bottom*). **b**. Anti-preferred and preferred images of an example V4 neuron. **c**. Skewness κ of response distributions for V4 neurons and DNN units. **d**. Skewness κ of response distributions for different visual areas in macaque *(top)*, and DNN layers *(bottom)*. Lines: medians.

bottom row, example real V4 neuron). This suggests that, unlike LN models and most DNN units, V4 neurons have preferred (response-maximizing) *and* anti-preferred (response-minimizing) stimuli.

The existence of anti-preferred images for higher-order visual cortical neurons is not obvious. The anti-preferred stimuli have largely been investigated as part of a neuron's tuning for a single stimulus parameter (e.g., a vertical edge drives a V1 neuron's response while a horizontal edge suppresses it). However, little is known about the anti-preferred stimuli of V4 neurons when considering the vast space of natural images varying over many stimulus parameters (Efird et al., 2024). Our prior expectations that the anti-preferred images are mostly featureless and low contrast—a blank, gray screen—were wrong; we find that some anti-preferred visual features are as vivid as those for the preferred images (Fig. 1**b**). This motivated us to systematically investigate the existence of anti-preferred images and their roles in how the visual cortex encodes natural images with the following progression:

1. We first set out to confirm the existence of anti-preferred images by analyzing response distributions of visual cortical neurons from V1, V4, and IT as well as performing our own electrophysiological experiments to validate that anti-preferred images suppress V4 firing rates.

2. If anti-preferred images do exist, we hypothesize a new mapping from DNN features to V4 neurons that takes advantage of pre-ReLU processing. Indeed, we find our new ReLU mapping outperforms other common linear mappings.

3. Further confirming the importance of anti-preferred images in shaping a V4 neuron's tuning, we find that an encoding model must train on responses to anti-preferred images (as well as to preferred images) to best predict responses to natural images. In a similar vein, humans performing a psychophysics task also rely on anti-preferred images to infer a neuron's tuning.

4. How do anti-preferred features contribute to encoding natural images by a V4 population? We find little to no relationship between a neuron's preferred and anti-preferred features, suggesting that anti-preferred images effectively double its capacity for feature selectivity.

5. To encourage further experiments investigating anti-preferred images in visual cortex, we release a tool called *ImageBeagle* that efficiently "hunts" through millions of natural images. We tailored ImageBeagle for closed-loop, real-time experiments.

Our results change our prior conceptions about stimulus encoding in the primate visual cortex: Conceptually, responses are not simply the output of a ReLU with a strong threshold but rather the sum of a baseline offset and a stimulus drive that may enhance or suppress the baseline response, resulting in a two-tailed response distribution. That preferred and anti-preferred features are diverse and independently-distributed across neurons allows the neural population to seemingly dou-

ble its selectivity, providing a rich basis for readout by downstream IT neurons to carry out object recognition and other visual tasks. Our work speaks to neuroscientists studying how feature selectivity arises in the visual cortex as well as to neuroAI researchers building AI models with internal representations that follow the representational principles of the visual cortex.

2 HIGHER-ORDER VISUAL CORTICAL NEURONS IN AREA V4 HAVE TWO-TAILED RESPONSE DISTRIBUTIONS.

To quantify the degree to which V4 responses to natural images have distributions with two tails (a hallmark of the neuron having preferred and anti-preferred images), we computed the skewness κ of response distributions. A distribution with κ close to 0 indicates two tails (Fig. 1**c**, top left panel) while κ close to 2 indicates a one-tailed distribution (Fig. 1**c**, top right panel). As expected, the skewness for ReLU units in a middle layer of the task-driven DNN ResNet50 (He et al., 2016), known to be predictive of V4 responses (Cowley et al., 2023; Yamins and DiCarlo, 2016; Schrimpf et al., 2018a; Zhuang et al., 2021), was close to 2 (Fig. 1**b**, 'DNN units', median $\kappa = 2.06$), indicating that most units in a task-driven DNN have one-tailed response distributions and are selective for one type of visual feature; we confirmed this was true of units from other task-driven DNNs (see Appendix). On the other hand, the response distributions of V4 neurons were better described as two-tailed (Fig. 1**c**, 'V4 neurons', median $\kappa = 0.87$), suggesting that V4 neurons have both preferred and anti-preferred images (See Methods in the Appendix for a description of V4 data collection; V4 responses were repeat-averaged spike counts). Here, we ignore the trivial effects of adaptation (Kohn, 2007)—in which presenting any image for long periods of time would lead to response suppression—by taking spike counts in 100 ms bins after the stimulus onset of a natural image (presented for 100 ms). Thus, V4 neurons appear to dynamically increase their baseline firing rate to encode a newly presented image (Pasupathy and Connor, 1999; Maunsell, 2015), allowing images to both excite and suppress their response from baseline (investigated in the next section). This goes against the conventional notion that a visual neuron responds selectively to certain stimuli by discarding most other stimuli that fail to drive the neuron past its spiking threshold. In other words, V4 responses do not appear to be the output of ReLU-like activation functions.

These findings motivated us to further investigate whether neurons from other areas of visual cortex also exhibit two-tailed response distributions. Using publicly-available datasets for V1 (Cadena et al., 2019) as well as for V4 and IT (Majaj et al., 2015), we re-computed skewness for each area. We found that skewness values from the V4 dataset matched our own data (Fig. 1**d**, 'V4 neurons', median $\kappa = 0.41$). In addition, we found that neurons from V1 and IT also exhibit a two-tailed selectivity (Fig. 1**d**, 'V1 neurons', median $\kappa = 1.17$ and 'IT neurons', median $\kappa = 0.69$). In contrast, the activations from increasingly-deeper layers of ResNet-50 exhibited much larger skewness values. DNN units in an early layer had the lowest skewness (Fig. 1**d**, 'early units', median $\kappa = 0.99$) on par with that observed for V1 neurons. A late layer had the highest skewness value (Fig. 1**d**, 'late units', median $\kappa = 4.43$), revealing a trend of increasing skewness (or one-tailedness) deeper into the network. Taken together, our results indicate a gap between biological and artificial visual systems: Neurons along the visual cortical hierarchy tend to have two-tailed response distributions, whereas DNN units in the deepest layers are most likely to have one-tailed response distributions. In other words, most real neurons encode anti-preferred images, but DNN units (post-ReLU) often encode only preferred features, especially in deeper layers.

3 EXPERIMENTAL EVIDENCE FOR ANTI-PREFERRED IMAGES IN HIGHER-ORDER VISUAL CORTICAL NEURONS

The existence of anti-preferred images immediately suggests that the way we predict V4 responses—typically a linear mapping between task-driven DNN features and V4 responses to natural images (see Methods) (Yamins and DiCarlo, 2016; Schrimpf et al., 2018b)—is suboptimal. Our first naïve hypothesis was that predicting V4 responses from pre-ReLU activity should outperform post-ReLU activity of the DNN features, as the pre-ReLU activity would have two-tailed response distributions; however, prediction was better for post-ReLU activity (Fig. 2**a**, *i* vs *ii*). The ReLU threshold was near optimal—other thresholds based on quantiles of the activity failed to outperform the original ReLU threshold (Fig. 2**a**, *iii*). Similarly, optimizing the scale and offset of each filter channel's pre-ReLU activity did not boost performance (Fig. 2**a**, *iv* vs. *ii*). Why is the ReLU important for prediction? We reasoned that by combining the post-ReLU activity of one-tailed response distributions across filter channels allows for greater flexibility to "mix and match"

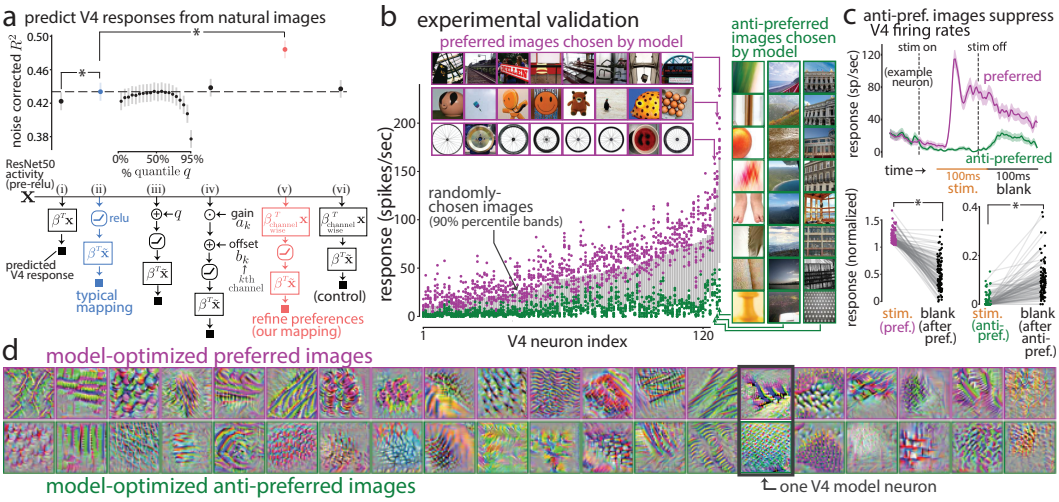


Figure 2: Experimental evidence that V4 neurons have anti-preferred images. a. Predicting V4 responses to randomly-chosen images from a linear mapping of ResNet-50 features. Each dot reflects the median and lines denote standard error. Asterisks denote p<0.001, permutation test. **b**. Experimental validation of preferred and anti-preferred images as predicted by V4 model neurons. Each dot is the repeat-averaged response to one image; gray bands denote 90% percentiles of responses to randomly-chosen images. Insets: Model-chosen images for the 3 neurons with largest baseline responses. **c**. Top: Example PSTH of a V4 neuron (*top*). Bottom: Normalized response to preferred, anti-preferred, and blank images shown right after; each dot denotes the average normalized response across the top 10 images (*bottom*). Asterisks denote p<0.001, permutation test. **d**. Preferred and anti-preferred images synthesized via gradient techniques, one for each V4 model neuron. Traces denote means, shaded areas denote 1 s.e.m.

preferences to estimate a neuron's preferred and anti-preferred features—in contrast, combining two-tailed response distributions of the pre-ReLU activity requires a filter channel to match a neuron in both preferred and anti-preferred features. It follows that we can improve upon this dictionary of one-tailed responses by allowing the linear mapping to form new preferred features before the ReLU step. To do this, we linearly combine filter channels (i.e., a convolution with kernel shape 1×1 and output channels equal to the number of input channels), pass the resulting activity through ReLUs and a final linear mapping. This simple approach significantly improved prediction (Fig. 2**a**, *iv* vs. *ii*); without the ReLUs, performance is no better than before (Fig. 2**a**, *vi*; R^2 for *vi* $>$ R^2 for *i* due to the use of layernorm in *vi*, see Methods). This algorithmic improvement is a direct result of assuming that V4 neurons encode anti-preferred images.

The skewness of response distributions and the improved prediction by mixing preferences before the ReLUs both hint at the existence of anti-preferred images; here, we seek experimental evidence. We build upon recent work that identified highly-predictive DNN models of V4 neurons by training on responses to many natural images (Cowley et al., 2023); these data-driven models predicted the preferred images of real V4 neurons in validation experiments by presenting the model-chosen preferred images in a following recording session (a causal test). We wondered whether the same framework could be used to predict neurons' anti-preferred images. To test this, we recorded V4 responses while the awake, fixating animal (macaque monkey) watched flashes of many images over multiple recording sessions (see Methods), and used the image-response pairs to train a set of data-driven DNN models (which we refer to as V4 model neurons). We identified each V4 model neuron's preferred and anti-preferred images by passing as input 500,000 natural images and keeping the 10 images that either maximized or minimized the model's output response (example chosen images in Fig. 2**b**). We then experimentally validated these predictions by recording V4 responses to these model-chosen preferred and anti-preferred images, along with hundreds of randomly-chosen natural images. After matching V4 neurons to their corresponding model units (see Methods), we found that the predicted preferred images resulted in responses above the 90% density interval of responses to randomly-chosen images (Fig. 2**b**, purple dots above gray lines, quantile of the median response to preferred images for responses to randomly-chosen images

$q = 0.985$, median across neurons), while the predicted anti-preferred images resulted in responses below these density intervals (Fig. 2**b**, magenta dots below gray lines, quantile of median response to anti-preferred images for responses to randomly-chosen images $q = 0.055$, median across neurons). This experimental validation provides clear evidence that V4 neurons have anti-preferred images.

A visual neuroscientist may wonder how the responses to anti-preferred images compare with responses to blank, gray screens—the *de facto* stimulus used between stimulus presentations to bring the neurons' firing rates to rest and presumably the stimulus yielding the smallest responses. To make this comparison, we analyzed V4 responses during the 100 ms stimulus presentation versus responses in the 100 ms immediately following stimulus presentation during which a gray, blank screen was presented (Fig. 2**c**, top, 'stim' and 'blank'; windows lagged to account for synaptic delays). As expected, we found that preferred images strongly drove responses above baseline (Fig. 2**c**, top, 'preferred' and bottom left). However, remarkably, we found that anti-preferred images often suppress a neuron's response below its baseline firing rate (Fig. 2**c**, top, 'anti-preferred' and bottom right). The level of suppression is beyond what we imagined and rules out the possibility that most anti-preferred images are blank within a neuron's receptive field with no discernible features. Indeed, the diversity and specificity of model-optimized anti-preferred images is on par with those of model-optimized preferred images (Fig. 2**d**).

4 Anti-preferred stimuli shape a V4 neuron's tuning.

The existence of anti-preferences alone does not necessarily imply that these images are crucial for stimulus encoding. If the anti-preferred images were indeed an important component shaping a V4 neuron's tuning, we would expect excluding them would result in a a poor estimate. On the other hand, using responses to both preferred and anti-preferred images should result in better estimates than relying on responses to preferred images alone. Thus, we can assess the information content of anti-preferred images by including them or leaving them out when estimating a neuron's tuning. Inspired by this approach, we devised the following data pruning analysis (Paul et al., 2021; Sorscher et al., 2022). We chose the data-driven V4 model neurons to serve as "digital twins" for real V4 neurons, as we required responses to 500,000 images—beyond the limits of current recording experiments. For each set of training images, we considered either preferred, anti-preferred, both, randomly-chosen, or non-preferred images whose responses were closest to the median response (Fig. 3**a**). We found that with a small number of training images (<5k images), training on both preferred and anti-preferred outperformed randomly-chosen images (Fig. 3**b**, 'pref.+anti-pref.' versus 'random'). These results suggest that the two tails of the response distribution alone provide rich information about intermediate responses (DiMattina and Zhang, 2008; Cowley and Pillow, 2020) up to a point—random sampling eventually outperforms other biased training images that are out-of-distribution.

Importantly, training on preferred images alone (Fig. 3**b**, 'pref. only') did not surpass or match the prediction performance of random selection, suggesting preferred images alone are not enough to estimate a neuron's tuning. Likewise, training on anti-preferred images alone (Fig. 3**b**, 'anti-pref. only') was not enough to achieve prediction as good as random selection. Thus, both are needed together to achieve good tuning estimates. This is further exemplified by training solely on non-preferred images (Fig. 3**b**, 'non-pref.'), which led to even worse prediction than training only on preferred or only on anti-preferred images. **Our results together, suggest that if one knows a neuron's preferred *and* anti-preferred images, they can reasonably estimate the rest of the neuron's tuning to other natural images.**

How far can we push a neuron's response to its limits, and how informative are these extreme preferred and anti-preferred images? To answer this, as a first step, we identified the preferred and anti-preferred images from a pool of 1M images and found an increase in prediction performance (Fig. 3**b**, 'pref.+anti-pref. (1M)'). Next, we considered synthesized images optimized via gradient techniques to maximize and minimize a neuron's response (see Methods) (Bashivan et al., 2019; Walker et al., 2019; Gu et al., 2022; Cowley et al., 2023; Willeke et al., 2023). While these synthesized imagse elicit more extreme responses than of natural preferred and anti-preferred images (see Sec. 8), they result in poor prediction performance (Fig. 3**b**, 'synthesized'). This is likely because the synthesized images depart too far from the response range of the test images (i.e., out-of-distribution) as well as suffer from a lack of a diversity, a noted problem with synthesizing images

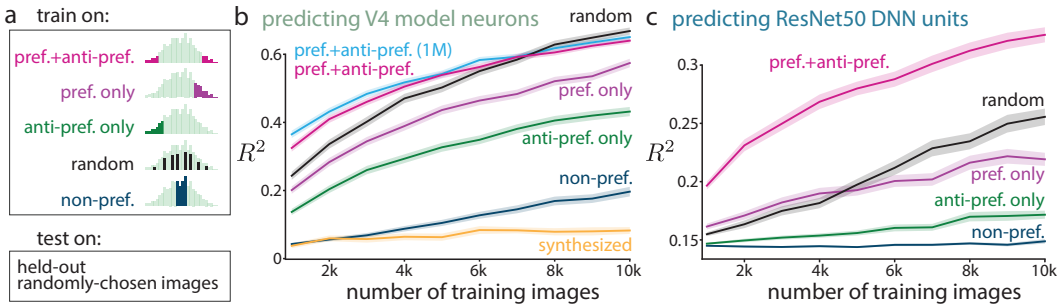


Figure 3: Anti-preferred images contribute to a V4 neuron's tuning. a. Data pruning analysis where we train on one set of images but always compute R^2 on responses to the held-out natural images. Training sets are sampled from the response distribution over 500k images; for example, non-preferred images ('non-pref.') are sampled from images with responses closest to the median response. **b**. We train a DNN (5-layer CNN, see Methods) to predict responses of individual V4 model neurons (219 in total), varying the number of training images. We also consider preferred and anti-preferred images drawn from 1M images ('pref.+anti-pref. (1M)') as well as synthesized images (see Methods). **c**. Same as in **b** except for predicting responses of individual ResNet50 DNN units (219 in total). Traces denote means, shaded areas denote 1 s.e.m.

(Pierzchlewicz et al., 2024; Nguyen et al., 2015). This highlights the usefulness of searching natural images for preferred and anti-preferred images when inferring a neuron's tuning (Borowski et al., 2020; Geirhos et al., 2021).

For comparison, we performed the same analysis on randomly-chosen ResNet50 DNN units and found a different picture: Preferred images alone outperformed random selection (Fig. 3**c**, 'pref. only' versus 'random'). Furthermore, although anti-preferred images alone were not informative (Fig. 3**c**, 'anti-pref. only' close to 'non-pref.'), preferred and anti-preferred images together outperformed random selection (Fig. 3**c**, 'pref.+anti-pref.' versus 'random'). This surprised us, as it suggests anti-preferred images do convey information for DNN units; on closer inspection, we found some DNN units to have response distributions with skewness similar to V4 neurons (Fig. 1**c**), and a unit's skewness negatively correlates with the extent to which preferred and anti-preferred images boost prediction performance (see Appendix). We also note that this boost appears magnified for the DNN units versus V4 model neurons, but we point out that R^2 overall is lower for the DNN units as they require more training data. This is likely because these units have complicated response functions as well as our finding that DNN units and V4 model neurons with one-tailed response distributions are harder to predict (see Appendix). Our data pruning analyses highlight important stimulus encoding differences between DNN units and V4 neurons, which we further explore in the next section.

5 Humans rely on anti-preferred images to determine a V4 neuron's tuning.

Access to both preferred and anti-preferred images was most informative for estimating V4 tuning (Fig. 3). This was likely because the preferred and anti-preferred images contained easy-to-identify visual features that resulted in learned filters to extract these features. We wondered to what extent the visual features of the preferred and anti-preferred images were readily accessible and interpretable by humans. To test this, we ran a simple psychophysics experiment in which human subjects chose one of two images that they thought would lead to a larger model response (Fig. 4**a**, see Methods); this task was inspired by recent work in explainable AI (Borowski et al., 2020; Zimmermann et al., 2024). Prior to the task, we gave subjects one of four possible sets of reference images: both preferred and anti-preferred images, preferred images only, anti-preferred images only, and no prior images. Subjects improved their performance via feedback of the correct image after each trial. We tested these four conditions for 10 different V4 model neurons and 10 different DNN units (80 tasks total); we found real V4 responses to be too noisy to predict accurately (see Methods).

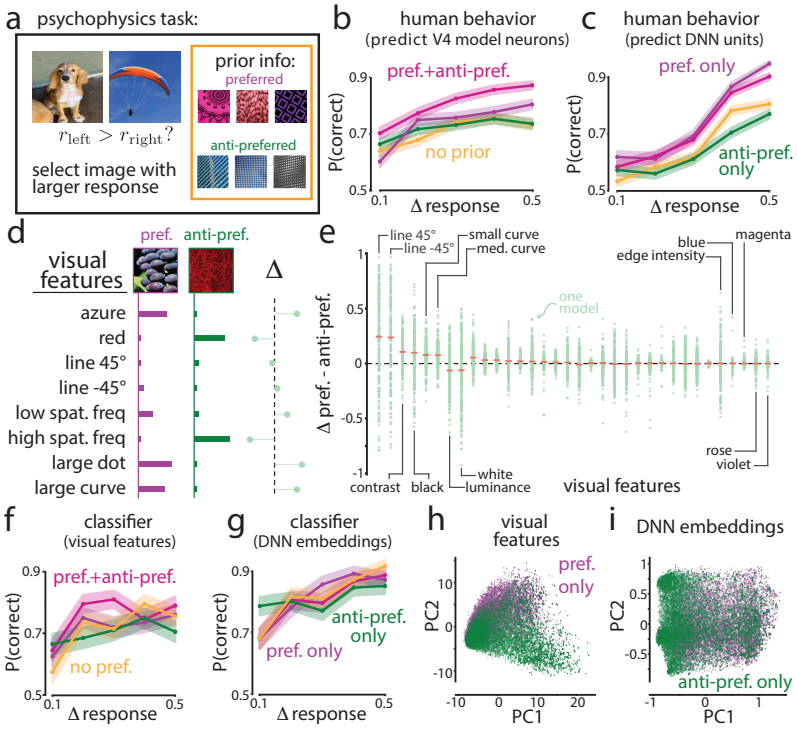


Figure 4: The visual features of anti-preferred images. a. Psychophysics task. **b-c**. Human performance predicting V4 model neurons (**b**) and ResNet50 DNN units (**c**). Lines: subject mean, shade: 1 s.e.m. **d**. Visual features differ between pref. and anti-pref. images. **e**. Differences in mean features. Lines: medians, dots: one V4 model neuron. **f-g**. Task performance in **a** for a classifier on visual features (**f**) or on DNN embeddings (**g**). **h-i**. Projection of preferred and anti-preferred features using interpretable embeddings (**h**) and DNN embeddings (**i**). Dot: one model.

The impressive performance of human subjects (Fig. 4**b**, 80.5% accuracy) suggests that the preferred and anti-preferred images have distinguishable and interpretable visual features. When given prior access to both preferred and anti-preferred images, subjects outperformed other types of prior information while predicting responses of V4 model neurons (Fig. 4**b**, 'pref.+anti-pref.'); performance for preferred-only and anti-preferred-only was comparable to no prior images (Fig. 4**b**, 'no prior' trace). This suggests that subjects relied on both preferred and anti-preferred images to infer a V4 model neuron's tuning. Interestingly, when predicting responses of ResNet50 DNN units, subjects performed similarly when given for reference either solely preferred images or both preferred and anti-preferred images together (Fig. 4**c**, 'pref.+anti-pref.' versus 'pref. only'), whereas performance dropped for anti-preferred images (Fig. 4**c**, 'anti-pref. only'). Thus, the anti-preferred images of DNN units have visual features that are not interpretable by humans, and unlike V4 neurons, a DNN unit's tuning can be mostly explained by its preferred images. This poses an important difference between V4 neurons and current task-driven DNN models.

6 The visual features of preferred and anti-preferred images

What visual features does a human use to distinguish between preferred and anti-preferred images? To get at this question, we considered a visual feature bank of 34 interpretable image statistics (each an index between 0 and 1, see Methods) that included contrast, luminance, edge intensity, orientation, color, among others (Fig. 4**d**). For each V4 model neuron, we computed the difference of each visual feature between preferred and anti-preferred images (mean over 100 images each). We found large differences for individual models (Fig. 4**e**, dots far from dashed line), suggesting these visual features were able to differentiate between preferred and anti-preferred images for a given V4 model neuron. We used these features as input into a classifier to perform the same psychophysics task performed by the human subjects (Fig. 4**f**); we re-trained the classifier each trial given the feedback about the correct image for that trial (classifier was difference-of-means, see Methods). While the classifier had good performance (Fig. 4**b**, 75.7% accuracy), it failed to take advantage of the prior information, unlike humans (Fig. 4**f**, 'pref.+anti-pref.' trace not noticeably higher than other traces), suggesting humans rely on other visual features to make their choices. To further explore this, we also tested a classifier that used a large number of DNN embeddings from ResNet50 and found performance that matched that of humans (Fig. 4**g**, 80.6% accuracy),

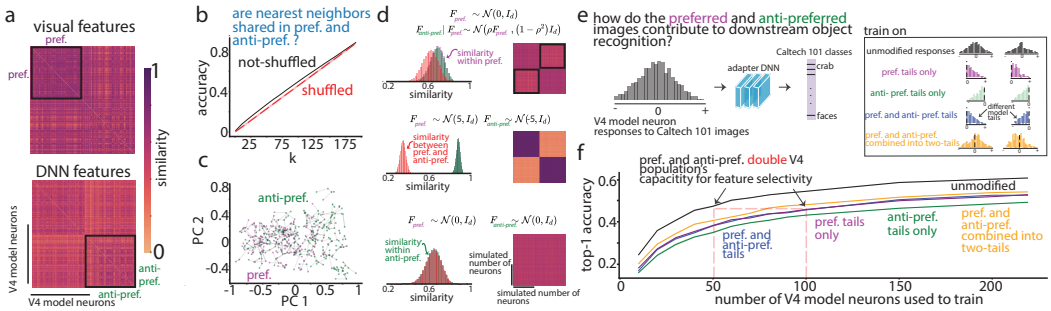


Figure 5: Anti-preferred images double V4 population's capacity for feature selectivity. a. Similarity matrix for 100 preferred and 100 anti-preferred images of 219 V4 model neurons using either visual *(top)* or DNN *(bottom)* features. **b**. Nearest-neighbor overlap as a function of k, for not-shuffled and shuffled matrices. **c**. Nearest-neighbors constructed with access to both preferred and anti-preferred features. **d**. Simulated example of 3 different distributions for preferred and anti-preferred features and the corresponding similarity matrices. **e**. Object recognition analysis using two-tailed responses. We train on one set of responses and compute the top-1 accuracy. **f**. Top-1 accuracy as a function of number of V4 model neuron responses considered, for each set of responses. Traces denote means, shaded areas denote 1 s.e.m.

but this classifier still failed to use prior information (Fig. 4**g**, traces overlapping). Thus, although these features can differentiate between preferred and anti-preferred images, open questions remain about which visual features humans use and how they incorporate prior information into their choices.

Do anti-preferred images share visual features across V4 neurons? Despite the ability of the visual features to distinguish between preferred and anti-preferred images for individual models, we found few visual features that largely differed from 0 across V4 model neurons (Fig. 4**e**, orange lines not far from black dashed line). This suggests that there is little to no relationship between the two image types across neurons. This was also evident from the lack of structure in the low dimensional projections of preferred and anti-preferred images both with interpretable and DNN embeddings (Fig. 4**h**-**i**). We further analyze this relationship in the next section.

7 Anti-preferred images double V4 population's capacity for feature selectivity

Inspired by these findings, we investigated the relationship between preferred and anti-preferred images. We computed the 34 visual features of top 100 preferred and 100 anti-preferred images for each of the 219 V4 model neurons. We combined these features together, to compute the similarity matrix in which the diagonal blocks correspond to within group similarity and off-diagonal corresponds to between group similarity (Fig.5a, top). Despite the separability of these features for some V4 model neurons (Fig.4d), the similarity matrix was mixed with no clear clustering in anti-preferred and preferred population. To rule out the possibility that these features might not be enough to pick up on the abstract features that the true V4 neurons could use, we computed the similarity matrix using 1024 dimensional DNN features (Fig.5a, bottom). Despite the dimensionality of the DNN embeddings, it yielded similar results to that of interpretable visual features. To utilize the similarity information between these images better, we used the preferred distances to compute the nearest neighbors of each V4 model neuron, and repeated the same for anti-preferred images. Are the similarity relationships of V4 neurons preserved in preferred and anti-preferred spaces? To asses this, we compared each V4 model neuron's k-nearest neighbors in the preferred matrix (Fig.5a, top, black square) with its corresponding neighbors in the anti-preferred matrix (Fig.5a, bottom, black square), and quantified the accuracy by the number of overlapping neighbors for each k. We found that, as k increased, the fraction of accuracy also increased (Fig.5b, black trace). To determine whether the observed relationship exceeded chance, we constructed a null distribution by shuffling neuron identities and recomputed the accuracy again (Fig.5b, red trace). This closely aligned with our results, indicating that there is no detectable correspondence in the neighborhood structure between preferred and anti-preferred images. This rules out the pos-

sibility that, if one knew the preferred images of a neuron, they would be able to directly identify its anti-preferred images. Moreover, when given access to the full similarity matrix to construct the nearest neighbors, we find that the nearest neighbors of preferred images are not always the other preferred images, likewise the nearest neighbors of anti-preferred images are not always the other anti-preferred images (Fig.5c). This shows that there is not a common feature within preferred or anti-preferred images that can separate these two groups. The existence of anti-preferred images raises an important question about how these relate to preferred images. For example, these images could be correlated with each other where knowing one could be informative of our knowledge of other (Fig.5d, top), or they could be coming from completely independent distributions with no overlap (Fig.5d, middle). If these were the case, we would have expected our nearest neighbor analyses to share the same neighbors across preferred and anti-preferred images, and show clear clustering. However, instead we find that preferred and anti-preferred images often overlap (Fig.5d, bottom), i.e. they are sampled from the same distribution, thus preferred and anti-preferred images of a neuron share little to no relationship. This might be advantageous for the V4 population: Randomly assigning preferred and anti-preferred visual features to each neuron seemingly doubles the population's capacity for feature selectivity. To test this, we decode the responses from V4 model neurons to perform an object recognition task (Caltech 101 (Li et al., 2022)) (Fig.5e). We find that to achieve the same accuracy, a population of V4 models with one-tailed responses requires double the number of neurons versus the population with original two-tailed responses (Fig.5f, black vs blue trace). This suggests that anti-preferred features provide another entire set of features for downstream neurons to use, thus effectively doubling the capacity for feature selectivity for the same number of neurons.

8 SEARCHING THROUGH MILLIONS OF IMAGES WITH IMAGEBEAGLE

Our results establish that both preferred and anti-preferred stimuli shape the tuning of V4 neurons. Because identifying these stimuli depends on ranking a large number of images by response, we wondered how many natural images were needed for this search. We chose a V4 model neuron and computed its responses to preferred and anti-preferred images out of K images randomly subsampled from 30 million natural images (K varied from 10k to 30M images). We found that the number of searched images needed to achieve a linear increase in response exponentially scales (Fig. 6**a**, left panel). The identified images followed this trend: 10k to 100k candidate images were not enough for robust identification (Fig. 6**a**, '10k' and '100k'); only when we searched through 1 million candidate images did we find preferred and anti-preferred images that resembled those of 30 million images (Fig. 6**a**, '1M' versus '30M'). A complementary approach is to synthesize images via gradient techniques (Bashivan et al., 2019; Ponce et al., 2019; Walker et al., 2019) that often identifies images that yield the largest and smallest responses (r_{max}=6.5, r_{min}=-3.5) but can be difficult to interpret versus natural objects (Borowski et al., 2020) and are highly stereotyped (Fig. 6**a**, 'synthesized'). This approach also requires technical expertise and dedicated hardware that few neuroscience labs have available. This motivated us to design a simple tool for visual neuroscientists to efficiently search through millions of natural images to optimize a desired objective (e.g., minimizing a neuron's response).

We developed a new tool, called *ImageBeagle*, that efficiently searches through millions of natural images to "hunt" for a desired stimulus in a short amount of time. The key intuition is that we traverse through the natural image manifold by visiting each image's nearest neighbors, moving to the neighbor with the largest objective value (i.e., a discrete version of gradient ascent, Fig. 6**b**). We collected 30M images from diverse image datasets and computed 1k nearest neighbors for each image, where similarity was defined as the Euclidean distance between DNN features (see Methods). ImageBeagle alternates between a global search determined by a coreset over images (Bachem et al., 2018) and a local search that evaluates an image's nearest neighbors and moves to the one with the largest objective value (see Fig. 4 for example nearest neighbors). We tested the performance of ImageBeagle versus random search and found impressive speed-ups: ImageBeagle often identified preferred and anti-preferred images with resulting responses close to the 30M-optimum after only 10k evaluated images (Fig. 6**c**, orange traces), substantially outperforming random selection (Fig. 6**c**, black traces). ImageBeagle may also be used to connect a neuron's preferred and anti-preferred images together along a smoothly-varying tuning curve—such an interpretable tuning curve has been difficult to identify because of the nearly infinite paths possible between two images (Pasupathy and Connor, 2001; Gallant et al., 1996; David et al., 2006). Because we constrain ImageBeagle to traverse smoothly along the image manifold via nearest neigh-

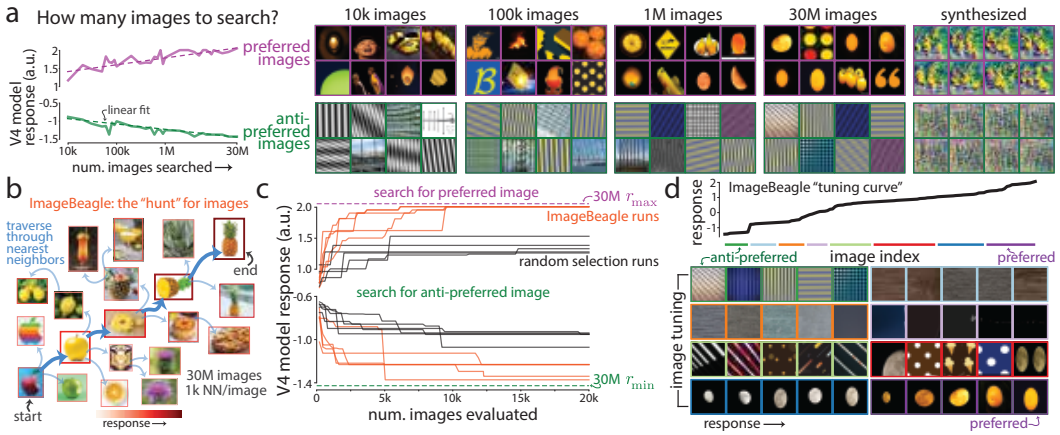


Figure 6: ImageBeagle searches the natural image manifold to efficiently find preferred and anti-preferred stimuli. a. A V4 model neuron's responses to preferred and anti-preferred images after searching through a subsample of K images. Dashed-lines: linear fits; x-axis is log-scale. Right: Top preferred and anti-preferred images for each K as well synthesized images. **b**. ImageBeagle navigates the natural image manifold via a nearest neighbor graph. **c**. ImageBeagle runs (orange traces) versus random selection (black traces) searching for the preferred (*top*) and anti-preferred (*bottom*) image. Dashed lines: optimal out of 30M images. **d**. ImageBeagle tuning curve for a V4 model neuron.

bors, ImageBeagle returns an interpretable sequence of natural images for the chosen V4 model neuron (Fig. 6**d**).

We suspect ImageBeagle will be of practical value to visual neuroscientists interested in optimizing neurons' responses (Cowley et al., 2017b; Walker et al., 2019; Ponce et al., 2019; Bashivan et al., 2019), performing closed-loop experiments with active learning (Benda et al., 2007; Park et al., 2011; Cowley and Pillow, 2020), and estimating tuning curves that smoothly vary in stimulus space (Wang and Ponce, 2024). Unlike most model-optimized stimuli, ImageBeagle does not require technical expertise, lowering the barrier for adoption by many experimental labs.

9 Discussion

Our work establishes the importance of anti-preferred images for stimulus tuning in visual cortex, especially visual area V4. We systematically investigate the properties of anti-preferred images through experimental validation, modeling, data pruning analyses, and human psychophysics. The existence of anti-preferred images is not obvious: Task-driven DNN units, commonly used to model V4 neurons, often do not exhibit anti-preferences due to their ReLU thresholding. This suggests that a V4 neuron's response less resembles the output of a ReLU and more resembles a filter with a dynamic range. Interestingly, we find that V4 responses are better predicted by linear combinations of ReLU DNN units versus pre-ReLU DNN units (Fig. 2**a**), suggesting a V4 neuron may form its two-tail selectivity in part by combining excitatory and inhibitory pre-synaptic input from neurons with one-tailed response distributions (i.e., preferring a single visual feature). Moreover, our results suggest that only characterizing a neuron by its preferred feature misses critical aspects of the neuron's tuning. How two-tailed response distributions and anti-preferred features relate to efficient and sparse coding in the brain (Olshausen et al., 1996; Rozell et al., 2008) remains an open question; two-tailed response distributions may require more energy for spikes but require fewer neurons to encode rich feature selectivity. Overall, our finding of anti-preferred images in V4 marks the beginning of a quest to identify the role anti-preferences play in other biological and artificial visual systems, and improve DNNs inspired by neuroscience principles.

REPRODUCIBILITY STATEMENT

Our V4 data and code will be publicly available upon publication at this link [removed for anonymity]. ImageBeagle will be available upon publication at this link [removed for anonymity]. IRB approval was obtained for all experiments, and the details will be disclosed [currently removed for anonymity] upon camera-ready version.

REFERENCES

Haldan Keffer Hartline. The response of single optic nerve fibers of the vertebrate eye to illumination of the retina. *American Journal of Physiology-Legacy Content*, 121(2):400–415, 1938.

David H Hubel and Torsten N Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1):106, 1962.

R Quian Quiroga, Leila Reddy, Gabriel Kreiman, Christof Koch, and Itzhak Fried. Invariant visual representation by single neurons in the human brain. *Nature*, 435(7045):1102–1107, 2005.

Benjamin R Cowley, Ryan C Williamson, Katerina Acar, Matthew A Smith, and Byron M Yu. Adaptive stimulus selection for optimizing neural population responses. In *Advances in neural information processing systems*, pages 1396–1406, 2017a.

Reza Abbasi-Asl, Yuansi Chen, Adam Bloniarz, Michael Oliver, Ben DB Willmore, Jack L Gallant, and Bin Yu. The deeptune framework for modeling and characterizing neurons in visual cortex area V4. *bioRxiv*, page 465534, 2018.

Carlos R Ponce, Will Xiao, Peter F Schade, Till S Hartmann, Gabriel Kreiman, and Margaret S Livingstone. Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences. *Cell*, 177(4):999–1009, 2019.

Pouya Bashivan, Kohitij Kar, and James J DiCarlo. Neural population control via deep image synthesis. *Science*, 364(6439):eaav9436, 2019.

Zijin Gu, Keith Wakefield Jamison, Meenakshi Khosla, Emily J Allen, Yihan Wu, Ghislain St-Yves, Thomas Naselaris, Kendrick Kay, Mert R Sabuncu, and Amy Kuceyeski. Neurogen: activation optimized image synthesis for discovery neuroscience. *NeuroImage*, 247:118812, 2022.

Pawel Pierzchlewicz, Konstantin Willeke, Arne Nix, Pavithra Elumalai, Kelli Restivo, Tori Shinn, Cate Nealley, Gabrielle Rodriguez, Saumil Patel, Katrin Franke, et al. Energy guided diffusion for generating neurally exciting images. *Advances in Neural Information Processing Systems*, 36, 2024.

EJ Chichilnisky. A simple white noise analysis of neuronal lightresponses. *Network: computation in neural systems*, 12(2):199, 2001.

Nicole C Rust, Odelia Schwartz, J Anthony Movshon, and Eero P Simoncelli. Spatiotemporal elements of macaque v1 receptive fields. *Neuron*, 46(6):945–956, 2005.

Jack L Gallant, Charles E Connor, Subrata Rakshit, James W Lewis, and David C Van Essen. Neural responses to polar, hyperbolic, and cartesian gratings in area V4 of the macaque monkey. *Journal of neurophysiology*, 76(4):2718–2739, 1996.

Anitha Pasupathy and Charles E Connor. Responses to contour features in macaque area V4. *Journal of neurophysiology*, 82(5):2490–2502, 1999.

Cory Efird, Alex Murphy, Joel Zylberberg, and Alona Fyshe. Finding shared decodable concepts and their negations in the brain. *arXiv preprint arXiv:2405.17663*, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Benjamin R Cowley, Patricia L Stan, Jonathan W Pillow, and Matthew A Smith. Compact deep neural network models of visual cortex. *bioRxiv*, 2023.

Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365, 2016.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, page 407007, 2018a.

Chengxu Zhuang, Siming Yan, Aran Nayebi, Martin Schrimpf, Michael C Frank, James J DiCarlo, and Daniel LK Yamins. Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3):e2014196118, 2021.

Adam Kohn. Visual adaptation: physiology, mechanisms, and functional benefits. *Journal of Neurophysiology*, 97(5):3155–3164, 2007.

John HR Maunsell. Neuronal mechanisms of visual attention. *Annual Review of Vision Science*, 1:373–391, 2015.

Santiago A Cadena, George H Denfield, Edgar Y Walker, Leon A Gatys, Andreas S Tolias, Matthias Bethge, and Alexander S Ecker. Deep convolutional models improve predictions of macaque v1 responses to natural images. *PLoS Computational Biology*, 15(4):e1006897, 2019.

Najib J Majaj, Ha Hong, Ethan A Solomon, and James J DiCarlo. Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance. *Journal of Neuroscience*, 35(39):13402–13418, 2015.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, page 407007, 2018b.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. *Advances in neural information processing systems*, 34:20596–20607, 2021.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. *Advances in Neural Information Processing Systems*, 35: 19523–19536, 2022.

Christopher DiMattina and Kechen Zhang. How optimal stimuli for sensory neurons are constrained by network architecture. *Neural Computation*, 20(3):668–708, 2008.

Benjamin Cowley and Jonathan W Pillow. High-contrast "gaudy" images improve the training of deep neural network models of visual cortex. *Advances in Neural Information Processing Systems*, 33:21591–21603, 2020.

Edgar Y Walker, Fabian H Sinz, Erick Cobos, Taliah Muhammad, Emmanouil Froudarakis, Paul G Fahey, Alexander S Ecker, Jacob Reimer, Xaq Pitkow, and Andreas S Tolias. Inception loops discover what excites neurons most using deep predictive models. *Nature neuroscience*, 22(12):2060–2065, 2019.

Konstantin F Willeke, Kelli Restivo, Katrin Franke, Arne F Nix, Santiago A Cadena, Tori Shinn, Cate Nealley, Gabby Rodriguez, Saumil Patel, Alexander S Ecker, et al. Deep learning-driven characterization of single cell tuning in primate visual area V4 unveils topological organization. *bioRxiv*, pages 2023–05, 2023.

Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436, 2015.

Judy Borowski, Roland S Zimmermann, Judith Schepers, Robert Geirhos, Thomas SA Wallis, Matthias Bethge, and Wieland Brendel. Exemplary natural images explain cnn activations better than state-of-the-art feature visualization. *arXiv preprint arXiv:2010.12606*, 2020.

Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Partial success in closing the gap between human and machine vision. *Advances in Neural Information Processing Systems*, 34:23885–23899, 2021.

Roland S Zimmermann, David A Klindt, and Wieland Brendel. Measuring mechanistic interpretability at scale without humans. In *ICLR 2024 Workshop on Representational Alignment*, 2024.

Fei-Fei Li, Marco Andreeto, Marc'Aurelio Ranzato, and Pietro Perona. Caltech 101, Apr 2022.

Olivier Bachem, Mario Lucic, and Andreas Krause. Scalable k-means clustering via lightweight coresets. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1119–1127, 2018.

Anitha Pasupathy and Charles E Connor. Shape representation in area v4: position-specific tuning for boundary conformation. *Journal of neurophysiology*, 2001.

Stephen V David, Benjamin Y Hayden, and Jack L Gallant. Spectral receptive field properties explain shape selectivity in area V4. *Journal of neurophysiology*, 96(6):3492–3505, 2006.

Benjamin Cowley, Ryan Williamson, Katerina Acar, Matthew A Smith, and Byron M Yu. Adaptive stimulus selection for optimizing neural population responses. In *Advances in Neural Information Processing Systems*, pages 1395–1405, 2017b.

Jan Benda, Tim Gollisch, Christian K Machens, and Andreas VM Herz. From response to stimulus: adaptive sampling in sensory physiology. *Current opinion in neurobiology*, 17(4):430–436, 2007.

Mijung Park, Greg Horwitz, and Jonathan W Pillow. Active learning of neural response functions with gaussian processes. In *Advances in neural information processing systems*, pages 2043–2051, 2011.

Binxu Wang and Carlos R Ponce. Neural dynamics of object manifold alignment in the ventral stream. *bioRxiv*, pages 2024–06, 2024.

Bruno A Olshausen et al. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

Christopher J Rozell, Don H Johnson, Richard G Baraniuk, and Bruno A Olshausen. Sparse coding via thresholding and local competition in neural circuits. *Neural computation*, 20(10):2526–2563, 2008.

Deepa Issar, Ryan C Williamson, Sanjeev B Khanna, and Matthew A Smith. A neural network for online spike classification that improves decoding accuracy. *Journal of neurophysiology*, 123(4):1472–1485, 2020.

Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59 (2):64–73, 2016.

Johannes Mehrer, Courtney J Spoerer, Emer C Jones, Nikolaus Kriegeskorte, and Tim C Kietzmann. An ecologically motivated image dataset for deep learning yields better models of human vision. *Proceedings of the National Academy of Sciences*, 118(8):e2011417118, 2021.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017.

Seong Tae Kim, Farrukh Mushtaq, and Nassir Navab. Confident coreset for active learning in medical image analysis. *arXiv preprint arXiv:2004.02200*, 2020.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

Gregory Griffin, Alex Holub, Pietro Perona, et al. Caltech-256 object category dataset. Technical report, Technical Report 7694, California Institute of Technology Pasadena, 2007.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, page 4. Granada, 2011.

Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2575–2584, 2020.

Zijie J Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. *arXiv preprint arXiv:2210.14896*, 2022.

Santiago A Cadena, Konstantin F Willeke, Kelli Restivo, George Denfield, Fabian H Sinz, Matthias Bethge, Andreas S Tolias, and Alexander S Ecker. Diverse task-driven modeling of macaque v4 reveals functional specialization towards semantic tasks. *PLOS Computational Biology*, 20(5):e1012056, 2024.

François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8697–8710, 2018.

A Appendix

A.1 Methods

In this section, we provide details for our V4 experiments, data pruning analyses, human psychophysics experiment, and ImageBeagle dataset and algorithm.

A.1.1 V4 experimental data

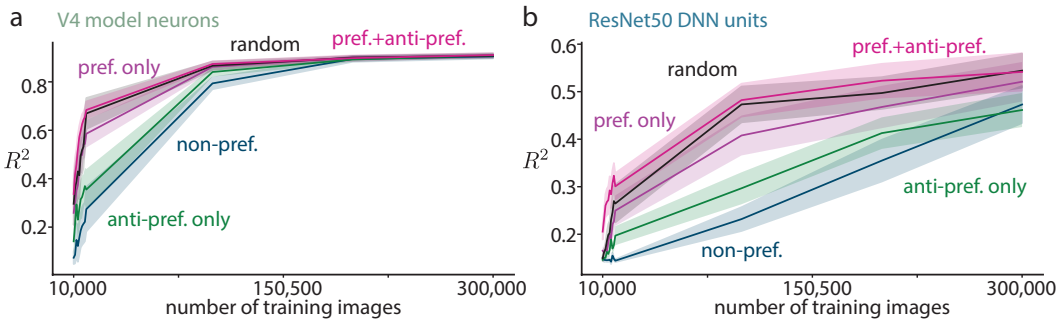
For our neural recordings from macaque V4, we used an experimental setup similar to those of (Bashivan et al., 2019; Cowley et al., 2023). For most of our analyses involving V4 responses, we re-analyzed data from a previous study (Cowley et al., 2023). This includes our results for V4 responses to natural images (Fig. 1) as well as predicting V4 responses using from ResNet-50 embeddings (Fig. 2**a**). We also re-used the study's 'compact models' as our 'V4 model neurons' to synthesize anti-preferred images (Fig. 2**c**), predict their responses using different sets of training images (Fig. 3**b**), include in our human psychophysics task (Fig. 4**b**) and visual feature analyses (Fig. 4**e**, **f**, and **h**), and optimize with ImageBeagle (Fig. 6). However, we needed to run additional experiments to test for the anti-preferred images of V4 neurons (Fig. 2**b**). To do this, we repeated an experimental setup almost identical to previous studies with closed-loop experiments (Bashivan et al., 2019; Ponce et al., 2019; Walker et al., 2019; Cowley et al., 2023). Below, we briefly describe the neural data collection, approved by the IRB [name redacted for anonymity].

Macaque V4 neural data collection: We implanted a 96-electrode array in the left hemisphere of macaque visual area V4, one in each of two macaque monkeys. We extracted spike signals via an automated deep learning pipeline (Issar et al., 2020) that separates spike waveforms from noise on each electrode channel. For each recording session, the awake, head-fixed animal performed thousands of active fixation trials until satiated (typically $\sim$ 2-3 hours). Each trial comprised $\sim$6-8 image flashes ($\sim$100 ms each) interleaved with 100 ms gray blank screens (to prevent adaptation effects between image flashes); image size and location were chosen to cover with the receptive fields of the recorded V4 neurons (8-11 visual degrees in diameter). After maintaining fixation throughout the sequence of images, the central dot disappeared and a target dot appeared 10° away from the central dot; animals received a liquid reward for correctly making a saccade to the target dot. Each recording session had $\sim$1,000 unique images and typically greater than $\sim$5 repeats per image (image repeats shown randomly throughout the session).

Construction of V4 model neurons: We recorded $\sim$10 sessions per animal to train the data-driven model, called the 'ensemble model', with the same architecture and training procedure as in a previous study (Cowley et al., 2023). Briefly, the model first passed the image through ResNet-50 to get the activations of an intermediate layer ('layer 33'). These activations were then passed as input into an ensemble of $\sim$25 small DNNs (each with 4 residual layers). Each ensemble member was trained separately on repeat-averaged responses; at inference, the final predicted response was the average response across the ensemble. We then fixed the ensemble model (with the linear readout weights trained on the last recording session) and searched for preferred and anti-preferred images. To do this, we passed $\sim$500,000 natural images through the ensemble model, and kept the $\sim$10 preferred and $\sim$10 anti-preferred images for each V4 neuron. We then presented these images, along with $\sim$750 randomly-chosen natural images, in the following recording session. Because we could not guarantee that we record from the exact same neurons between sessions (a small number of neurons are lost and added due to small shifts in electrode positions), we had to match up the model neurons (from the ensemble model) to the recorded V4 neurons on the new session. To do this, we computed the predicted R^2 between each model neuron and each V4 neuron for the responses to the randomly-chosen natural images, and kept greedily choosing the pair with the highest R^2 (and removing the chosen model neuron and neuron as future candidates). Then, for each V4 neuron, we take the median response of the $\sim$10 preferred images r_{pref} and $\sim$10 anti-preferred images $r_{\text{anti-pref}}$, and compute the fraction/quantile q to which these median responses are either larger (preferred) or smaller (anti-preferred) to responses to the randomly-chosen images, e.g., $q_{\text{pref}} = \frac{1}{N}\sum_i^N \mathcal{I}(r_{\text{pref}} > r_i)$, where $\mathcal{I}$ is the indicator function and i denotes the ith image randomly-chosen out of N images.

A.1.2 Linear mapping analyses

For our linear mapping analysis (Fig. 2**a**), we used the pre-ReLU activations from ResNet50 ("conv4_block4_add" layer) as our input to all 6 methods, and denote that as "pre-ReLU". Below, we describe each of the methods, *i-vi*, in detail:



Supplementary Figure 1: Extended data pruning plots. a. We train a DNN (5-layer CNN) to predict responses of individual V4 model neurons (10 in total), with larger number of training images than Fig.3. **b**. Same as in **a** except for predicting responses of individual ResNet50 DNN units (10 in total). Traces denote means, shaded areas denote 1 s.e.m.

- Method *i*: We linearly map the pre-ReLU features to V4 responses to predict the V4 responses to held-out images.
- Method *ii*: Similar to *i*, except we now add ReLU activation prior to linear mapping, the classicial approach in neuroscience.
- Method *iii*: Similar to *ii*, except instead of using the regular ReLU thresholding of 0, we vary this based on the different quantiles of the response distribution.
- Method *iv*: Before the linear mapping, we learn a separate gain and offset for each channel, add LayerNorm, and pass the resulting activity through ReLUs and a final linear mapping.
- Method *v*: We linearly combine filter channels via a convolution with kernel shape 1×1 where the output channels equal to the number of input channels, add LayerNorm and pass the resulting activity through ReLUs and a final linear mapping.
- Method *vi*: Same as *v* but we remove the ReLUs before the final linear mapping.

For all of the methods above, we ensure that the train, test, and validation sets remain the same for the final comparisons.

A.1.3 Data pruning analyses

For our data pruning analysis, we ran simulations to asses the information content of anti-preferred images by including or leaving them out when we estimated a neuron's tuning. To this end, we used V4 model neurons to serve as surrogate ground truth models of V4 neurons (as recording a real neuron's responses to 500k images is unfeasible, and V4 model neurons closely resemble real V4 neurons, see Supp. Fig.5). We then used these surrogates as "teacher" models to train the "student" models (5-layer CNN architecture, 100 filters per layer) with different curricula (see below). To stay as close to a real neuroscience experiment as possible, we were interested in training with <10k images (Extending the number of training images did not change the results, see Supp. Fig.1. Thus, we trained each model from scratch from 1k to 10k at 1k intervals and reported their R^2. We used the same procedure for ResNet-50 units. In Supp. Fig.1 we extend our pruning plots from Fig.3 to include more training images for 10 models. Below we detail the pre-training and training procedures.

Pre-training details: Prior to training, for every V4 neuron model, we sorted each model's responses to 500k images; we define the top K images as the preferred images and the bottom K images as the anti-preferred images. We sought to quantify the extent to which preferred and anti-preferred images contributed to our estimate of that model's tuning. We designed the following different curricula (corresponding to the traces in Fig. 3), where K refers to the selected number of images and responses on which to train.

- **preferred images only**: We selected the K images that had the highest responses.

- **anti-preferred images only**: We selected the K images that had the lowest responses.
- **preferred and anti-preferred images**: We selected $K/2$ images that had the highest responses and $K/2$ images that had the lowest responses.
- **preferred and anti-preferred images (1 million)**: We considered an entirely different set of images that was double in size to our baseline dataset (1M here versus 500k for the other curricula). All other details were the same as for **preferred and anti-preferred images**.
- **randomly-chosen images**: We randomly selected K images from the pool of 500k images.
- **non-preferred images**: We first found the median response and selected the K images with responses closest to the median response (i.e., $K/2$ images with responses below the median response and $K/2$ images with responses above the median response).
- **synthesized images**: We synthesized $K/2$ images to maximize the model's output response and synthesized $K/2$ images to minimize the model's output response. Synthesized images were optimized with gradient ascent/descent techniques (Bashivan et al., 2019; Walker et al., 2019; Cowley et al., 2023).

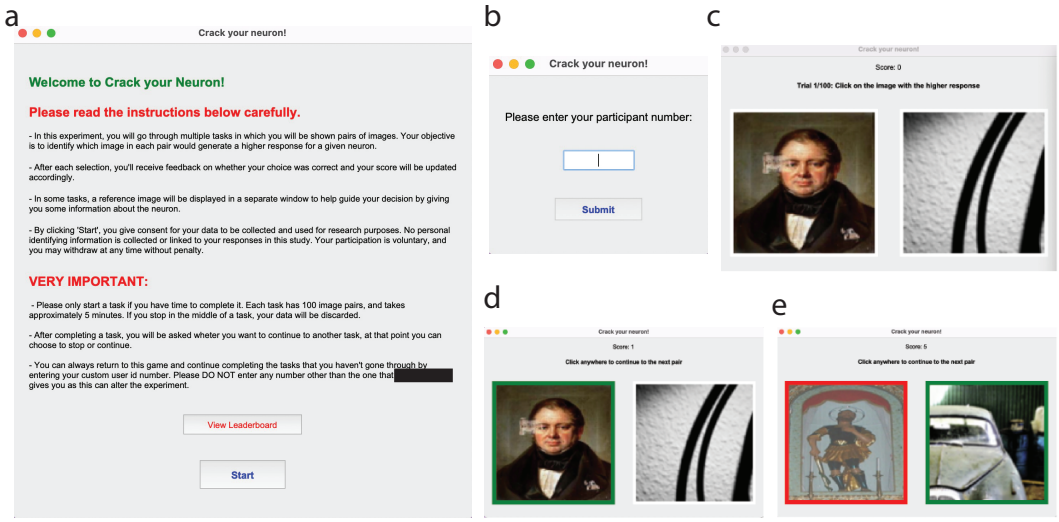
The chosen curricula, except for randomly-chosen images, will likely lead to biases such that the training and test data distributions will not match (i.e., out-of-distribution). To mitigate such biases, we replaced 10% of the images for each curriculum with randomly-selected images (replacing the images with the lowest responses for preferred images and the images with the highest responses for anti-preferred images). We also note that as the number of training images increases, training on randomly-selected images outperforms other curricula, as the training distribution matches the test distribution. Lastly, to make the simulations more similar to real V4 responses, we added Gaussian noise to the V4 model neuron's responses. Instead of predicting the responses one-to-one, we added $(0.2 \times \sigma \times \epsilon)$ to our responses and predicted this value. Here σ was the standard deviation of the responses and $\epsilon \sim \mathcal{N}(0, 1)$. The exact procedure was used for ResNet50 units (Fig. 3**c**); we found these units needed more training data than the V4 model neurons to reach large values of R^2 (Supp. Fig. 1), likely because ResNet50 units computed more complex functions.

Training details: Across models, the training images were sampled from the same pool of 500k images; these 500k images were randomly sampled from ImageBeagle dataset (see Section A.2.5.) comprising 30M images. For testing and validation, we sampled another 20k images from ImageBeagle, different from the 500k images, and used 10k for test and 10k for validation to evaluate the trained models and report the R^2 score. We trained the model with the ADAM optimizer with learning rate $1e-4$, early stopping (based on validation data), and used a batch size of 8 for ResNet-50 and 64 for V4 model neurons. Since ResNet-50 had a more complex architecture than V4 model neurons, it required a smaller batch size than V4 model neurons to achieve a higher R^2.

A.1.4 Human psychophysics experiment

We performed a human psychophysics experiment to test how human subjects rely on preferred or anti-preferred images to guess a neuron's (or model unit's) tuning (Fig.4). The subject pool comprised volunteer scientists with no compensation; IRB approval [identity removed for anonymity] was obtained prior to beginning the experiment.

Task details: The task was as follows (task GUI shown in Supp. Fig. 2). Given a pair of images, the subject is instructed to select the image that would lead to the higher response of a chosen neuron/model. The user's score, the number of times the user picked the correct answer, was displayed above the prompt to show their progress (Supp. Fig.2). Importantly, the subject was given feedback after every decision via a green box around the correct answer and red box around the wrong answer—through this feedback humans learned the task. We include the layout of the general setup in Supp. Fig.2. Each task in our psychophysics experiment consisted of 100 pairs of images. If prior images were provided, they were always 36 images in total (Supp. Fig.3). In order to avoid overlap, we excluded these prior images from the image pairs to avoid duplicates. To ensure that each task had an equal mix of difficult and easy to discriminate image pairs, we took 5 bins of image pairs with increasing response differences (i.e., the larger the response difference between two images, likely the easier the discrimination). The first bin had images that were very close in value (Δ response ~ 0.1), and the last bin had images there were very far apart (Δ response ~ 0.5).

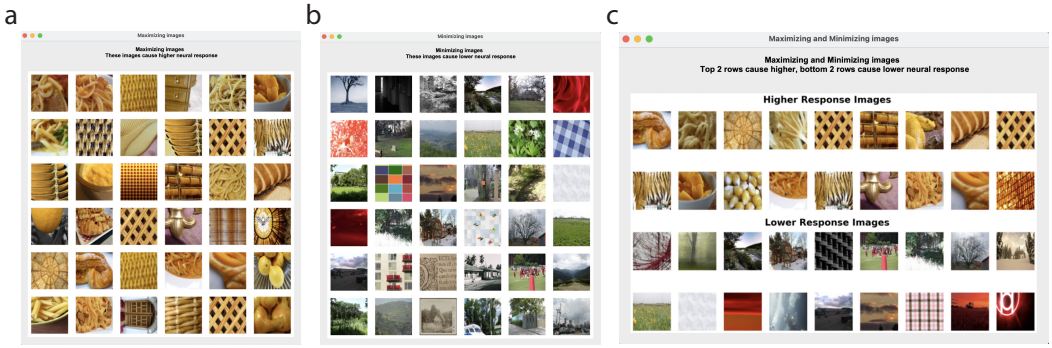


Supplementary Figure 2: GUI of the psychophysics task. a. Welcome screen with pertinent information about the task. Name has been redacted for anonymity. **b**. A user id screen that the user needs to input a specific id given by the instructor. **c**. An example pair from a task. **d**. User is given a positive feedback when the correct answer is picked via a green box. **e**. A negative feedback is given via a red box around the wrong answer, and positive feedback is given via a green box around the correct answer.

Each task had 20 image pairs from each bin (100 total) which were randomly ordered across the task. To create these bins, fo each model we extracted the responses, and calculated all possible response differences and saved the sorted differences in an array. We then used 20th-80th percentiles of this array to compute bin edges to create our 5 bins and filled these bins with non-overlapping images (e.g. if an image appeared in a pair, it can't be used for another pair), until each bin had 20 images.

Task types: For each task, we used one of the 10 units/neurons from a given model. The models could either be V4 model neurons or ResNet-50 units. Hence, creating a total of 40 tasks for each model (4 conditions $\times$ 10 model neurons/units), and 80 tasks total for the entire experiment. For ResNet-50, we used 10 randomly selected units from a mid-layer (layer 33, with 1,024 filters). For V4 model neurons, we used 10 randomly selected units from 219 V4 model neurons. In a pilot dataset, we also attempted the task for real V4 neurons, but found the responses too noisy and likely too few images ($\sim$1,000 images per recording session) for humans to identify meaningful selectivity. In addition, each task consisted of one of 4 conditions describing the prior information provided. Our goal was to investigate how humans use these prior images to guide their decisions. The priors were as follows:

- **no prior:** In this condition, no additional images were shown to the subject. Thus, the subject had to rely heavily on the feedback from the task to guide and improve their decision.
- **preferred prior:** In this condition, we showed the 36 preferred images of the unit. We refer to these images as "maximizing" in the experiment to make it more intuitive for the subjects. Tasks with this condition allow the subject to utilize this prior information by selecting the image from the pair that's most similar to these images.
- **anti-preferred prior:** In this condition, we showed the 36 preferred images of the unit. We refer to these images as "minimizing" in the experiment to make it more intuitive for the subjects. Tasks with this condition allow the subject to utilize this prior information by selecting the image from the pair that is not similar to these images. This condition tends to be more challenging compared to preferred prior because here the subject is given information on the lower response images (anti-preferred), but not the higher.



Supplementary Figure 3: Example preferred and anti-preferred images from the psychophysics task. The user had access to these images throughout the task. a. 36 preferred images. **b**. 36 anti-preferred images. **c**. 18 preferred and 18 anti-preferred images.

Therefore, the user still has to figure out what the maximizing images are through feedback and elimination-based strategies.

- **preferred:** In this condition, we showed the 18 preferred images and 18 anti-preferred images of the model neuron/unit. Tasks with this condition allow the subject to utilize this prior information by selecting the image from the pair that is similar to the preferred images and not similar to anti-preferred images.

A.1.5 IMAGEBEAGLE DATASET AND NEAREST NEIGHBOR GRAPH

ImageBeagle relies on a large bank of millions of diverse images. To collect the image dataset, we scraped images from various popular public datasets and sources. We sampled the images from various sources such as Flickr Creative Commons dataset (Thomee et al., 2016), Ecoset (Mehrer et al., 2021), and Duckduckgo, to name a few, in addition to other sources across the web (see Table 1 for approximate number of images extracted from each source). In addition to these, we also created artificial stimuli that are of interest to neuroscience, e.g. bars, gratings, colorful letters, gaudy images (Cowley and Pillow, 2020), and so on. We stored our 30M images in 1,500 zips, where each zip contains 20k images; we chose zips for easy access and transfer. To make the images consistent across the dataset, each image was resized to a 224x224 RGB PNG file. We make a miniImageBeagle (with 1M images) publicly available to researchers at [url removed for anonymity]. The full ImageBeagle dataset is large (2 TB) and available on request by the authors.

Nearest neighbor graph: For our 30 million images, we desired each image's 1k nearest neighbors; this allows us to estimate the natural image manifold via local approximations, where the neighbors correspond to possible directions along the manifold. We defined distance as the Euclidean distance between activations from a middle layer of ResNet50 (units come from layer 33 of ResNet50), which are predictive of V4 responses (Schrimpf et al., 2018a; Cowley et al., 2023). We down-sampled the large tensor of activations via a spatial average pool (from $14 \times 14 \times 1,024$ to $3 \times 3 \times 1,024$ with pooling kernel of 4×4 and a stride of 5). We took the top $\sim$1k images with the smallest distances. We confirmed that this similarity metric led to perceptually similar neighbors.

Computing the distance matrix of 30M images was computationally intensive, involving $\sim$500 hours of H100 GPU computation. To make the the distance calculations as efficient as possible for 30M images, instead of calculating the entire distance matrix, we randomly initialized the nearest neighbors and continuously update them by randomly choosing pairs of zip files to compute the distances (keeping track of previously-computed pairs). Thus, the ImageBeagle search algorithm may operate even as nearest neighbors continue to be updated.

A.1.6 IMAGEBEAGLE SEARCH ALGORITHM

Given the nearest neighbors, ImageBeagle consists of 2 steps: Global search and local search. For our global search, we utilize 10 coresets comprising 10k images each (Sener and Savarese, 2017; Bachem et al., 2018; Kim et al., 2020) to ensure that we explore diverse regions of the image manifold. We create an approximate coreset to bias our global search to explore diverse neighbors in our image manifold, thus preventing the algorithm from getting stuck at sub-optimal solutions.

Given the saved nearest neighbors, we randomly initialize the coreset with an image, and add its farthest neighbor to our coreset. We then move on to this image and repeat the process, essentially iteratively traversing the nearest neighbor graph until we filled our coreset with 10k images. Since this coreset is approximate (due to our filing system where we only save 1k neighbors), we randomly initialize 10 coresets. For local search, we use the computed nearest neighbors. We alternate between the global and local searches to explore the image manifold efficiently for a given objective function. ImageBeagle is given a budget of the number of evaluations allowed (i.e., computing the objective value for each image). ImageBeagle stores the objective value for every evaluated image to ensure images are not re-evaluated. In Alg.1 we explain the high level flow of ImageBeagle as well as the required hyperparameters.

Global search: We use the coreset images to get out of a local optimum and explore more areas in the image manifold. Thus, during global search, we take the next L images of the coreset and cycle through coresets whenever we reach M during local search. This allows ImageBeagle access to diverse regions in the image manifold.

Local search: We use the nearest neighbor information of each image during the local search process of ImageBeagle. The local search begins with the image that maximizes our objective function from all previously evaluated images (whose nearest neighbors have not been evaluated). We use this image to do our local search to explore its neighbors. We continue this process until we do not improve our objective function. We repeat this M times, after which we continue to the next global search.

Algorithm 1 ImageBeagle algorithm

Require: 1k nearest neighbors for every image
Require: 10 coresets of 10k images each
Require: **hyperparameters**
  L: number of coreset images searched at each global step
  K: number of nearest neighbors to evaluate per local step
  M: number of local searches
  B: number of images to evaluate (budget)
  $\phi(\mathbf{x}) : \mathcal{R}^{p \times p \times 3} \rightarrow \mathcal{R}$: objective function with input image $\mathbf{x}$
    set: `num_evals` $\leftarrow 0$, $\mathbf{X}_{\text{evaluated}} \leftarrow [\,]$,$\mathrm{X}_{\text{visited}} \leftarrow [\,]$ # empty lists

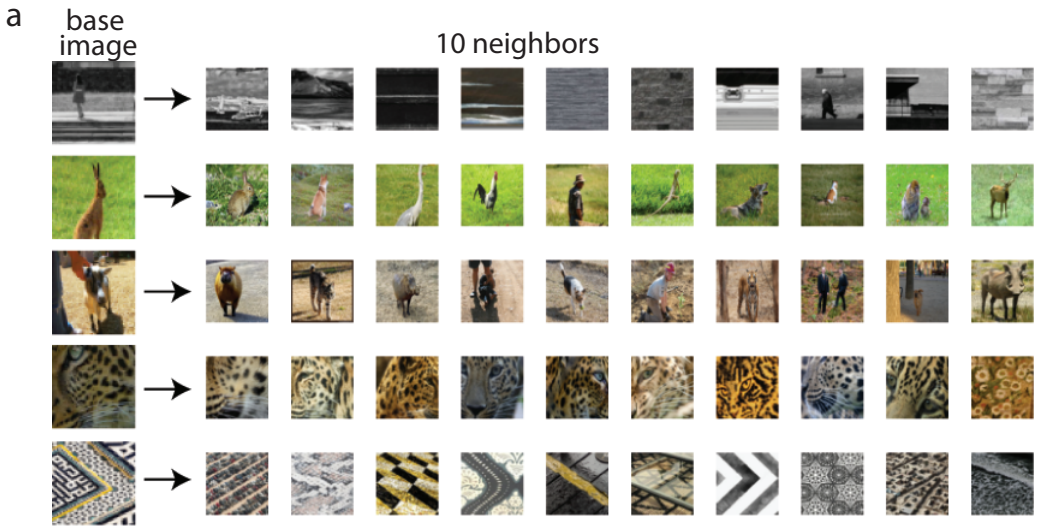
while `len`($\mathbf{X}_{\text{evaluated}}$) < B **do**
  # **global search**
  $\mathbf{X}_{\text{coreset}} \leftarrow$ next L images in coreset
  if coreset empty **then** move to next coreset
  $\mathbf{X}_{\text{evaluated}} \leftarrow [\mathbf{X}_{\text{evaluated}}]$ + $[\mathbf{X}_{\text{coreset}}]$ # combine lists

  # **local search**
  for `isearch` $\leftarrow 1$ to M **do**

    # choose starting image
    $\mathbf{X}_{\neg\text{visited}} \leftarrow \mathbf{X}_{\text{evaluated}} - \mathbf{X}_{\text{visited}}$ # subtract lists
    $\mathbf{x}_{\text{next}} \leftarrow \arg\max(\Phi(\mathbf{X}_{\neg\text{visited}}))$
    $\mathbf{X}_{\text{visited}} \leftarrow [\mathbf{X}_{\text{visited}}] + [\mathbf{x}_{\text{next}}]$
    $\mathbf{X}_{\text{nearest neighbors}} \leftarrow K$ nearest neighbors of $\mathbf{x}_{\text{next}}$
    $\mathbf{X}_{\text{evaluated}} \leftarrow [\mathbf{X}_{\text{evaluated}}] + [\mathbf{X}_{\text{nearest neighbors}}]$ # combine lists
    # walk through nearest neighbors
    while $\max[\Phi(\mathbf{X}_{\text{nearest neighbors}})] > \phi(\mathbf{x}_{\text{next}})$ **do**
      $\mathbf{x}_{\text{next}} \leftarrow \arg\max(\Phi(\mathbf{X}_{\text{nearest neighbors}}))$
      $\mathbf{X}_{\text{visited}} \leftarrow [\mathbf{X}_{\text{visited}}] + [\mathbf{x}_{\text{next}}]$
      $\mathbf{X}_{\text{nearest neighbors}} \leftarrow K$ nearest neighbors of $\mathbf{x}_{\text{next}}$
      $\mathbf{X}_{\text{evaluated}} \leftarrow [\mathbf{X}_{\text{evaluated}}] + [\mathbf{X}_{\text{nearest neighbors}}]$ # combine lists
    end while
  end for
end while
$\mathbf{x}_{\text{optimal}} \leftarrow \arg\max[\Phi(\mathbf{X}_{\text{evaluated}})]$

ImageBeagle	
Source	Approximate Amount
Flickr Creative Commons dataset Thomee et al. (2016)	7 million
Ecoset Mehrer et al. (2021)	1.5 million
CIFAR Krizhevsky et al. (2009)	120,000
CelebA dataset Liu et al. (2015)	202,000
Caltech-256 dataset Griffin et al. (2007)	30,000
Fashion MNIST dataset Xiao et al. (2017)	60,000
SVHN dataset Netzer et al. (2011)	248,000
Google Landmarks dataset Weyand et al. (2020)	4.1 million
DiffusionDB Wang et al. (2022)	5 million
Duckduckgo	155,000
Flickr	2 million
YouTube	1 million
Artificial stimuli	3.1 million
Others	5.5 million

Table 1: Summary of image sources for ImageBeagle

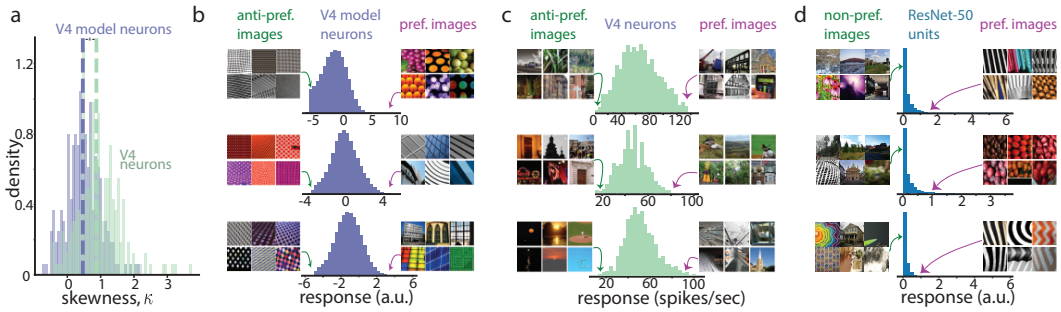


Supplementary Figure 4: Example ImageBeagle neighbors. a. Example 'base' images and their 10 nearest neighbors based on distances of embeddings from our chosen DNN (ResNet50). The base image and its neighbors are perceptually similar in low-level statistics (textures, colors, etc.), allowing ImageBeagle to be useful in identifying preferred and anti-preferred images for neurons in different visual cortical areas (V1, V4, IT, ...) as well as DNN units in different layers.

A.2 Multi-unit activity and anti-preferred images

To record neurons in our experiments, we used a Utah multi-electrode array, which captures the activity of both single- and multi-units. Identifying well-isolated single units by analyzing spike waveforms is possible, but one concern that is hard to fully rule-out is if any unit is truly a single neuron. Therefore, instead, here we argue that multi-unit activity cannot largely explain the existence of anti-preferred images. This is for two reasons:

First, we analyzed a separate dataset of V4 responses to natural images from Cadena et al. (2024) that used NeuroPixel probes (NeuroNexus V1x32-Edge-10mm-60–177) to record neural activity. An advantage of NeuroPixels is that the electrode channels are much closer together ($50\mu m$) than those of the Utah array ($400\mu m$); one can isolate single units based on coincidence timings of spikes between channels. The authors also performed extensive spike sorting to ensure well-



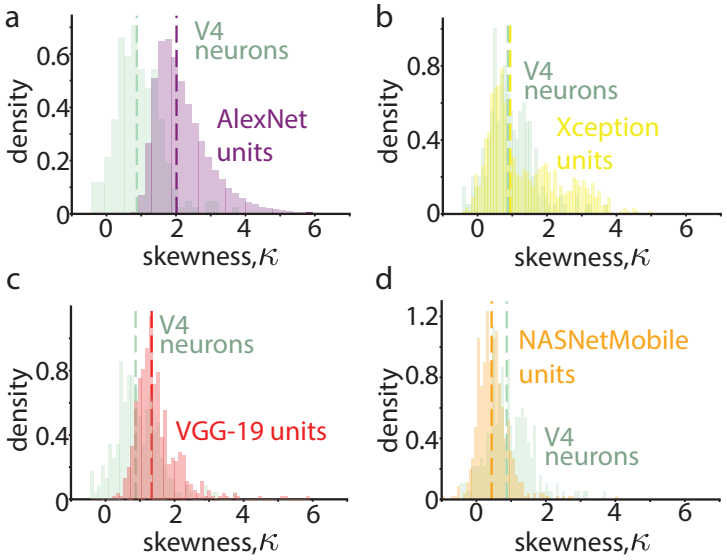
Supplementary Figure 5: Additional examples of preferred and anti-preferred images from 3 ResNet-50 units and V4 neuron models. **a**. Skewness of response distributions for V4 neurons and V4 model neurons. **b**. Anti-preferred and preferred images of 3 V4 model neurons. **c**. Non-preferred and preferred images of 3 ResNet-50 units.

isolated single units. A caveat is that images were presented with only one repeat and in rapid succession without interleaved blank screens—thus, responses are considerably noisier than our analyzed repeat-averaged V4 responses. We kept neurons with a SNR of at least 0.5 (split-repeat analysis). The median skewness of the Cadena dataset was $\kappa = 1.377$(mean firing rate = 8.76 spikes/sec). We found a tight relationship between mean firing rate and skewness: For neurons with firing rates > 10 spikes/sec, the skewness was $\kappa = 1.06$, and for neurons with firing rates > 15 spikes/sec (similar to our analyzed V4 data), the skewness was 0.852, matching closely to our observed $\kappa = 0.87$ (Fig. 1**c**). That neurons with lower firing rates had higher skewness is unsurprising in this dataset due to the Poisson nature of spike counts—unfortunately, each image was repeated only once, making the estimates of the true response distribution difficult. However, we believe that the low skewness of the well-isolated high-firing neurons is interesting, as this need not be the case. That the skewness values match well with our observed ones when controlling for firing rate, further confirms the presence of two-tailed response distributions for V4 neurons.

Second, because multi-unit activity is thought to be additive, it is likely the case that the image that maximizes the response of the multi-unit likely maximizes an individual neuron within the multiple neurons that comprise the multi-unit. This assumption is often made by recent studies that optimize preferred stimuli of V4 and IT neurons (Bashivan et al., 2019; Ponce et al., 2019; Cowley et al., 2023; Pierzchlewicz et al., 2024). With similar logic, assuming each unit has an anti-preferred stimulus, identifying the anti-preferred stimulus of a multi-unit is akin to minimizing the response of one of the individual neurons. However, if the units had one-tailed distributions, adding the responses of enough of these units together would likely yield a Gaussian distribution (based on the central limit theorem). To test how many units would need to be added together, we added DNN units with larger skewness ($\kappa \sim 5$), and found that we needed ~ 400 DNN units to match the skewness of observed V4 neurons ($\kappa = 0.87$)—this is unrealistic for a real multi-unit, which likely comprises only two or three neurons. In addition, averaged over 100 runs, adding responses to 3 DNN units yielded a skewness of $\kappa = 1.87$, much larger than the skewness for V4. We found that the anti-preferred images for the multi-units of 3 DNN units and above did not have discernible shared visual features versus the perceptually-similar anti-preferred images of the real V4 neurons.

A.3 Skewness for task-driven DNN units

In Figure 1, we considered the skewness of one task-driven DNN (ResNet-50). Here, we compare the skewness of four task-driven DNNs to that of V4 neurons (Supp. Fig. 6). Specifically, we wanted to investigate how the highly predictive ReLU layers of popular DNNs organized their responses compared to ResNet-50. To do this, we measured the population skewness from 4 popular DNNs that are known to be predictive of V4 responses, Xception (Chollet, 2017) (728 units, from 'block10_sepconv1_act' layer)), AlexNet (Krizhevsky et al., 2012) (384 units, from 'conv3' layer), VGG-19 (Simonyan and Zisserman, 2014) (512, from 'block5_conv2' layer), and NASNet-Mobile (Zoph et al., 2018) (528 units, from 'activation_104' layer) (Supp. Fig. 6**a**.-**d**.). While units from AlexNet, Xception, and VGG-19 overall appear to be more skewed than those of V4 neurons (median κ = 2.02, median κ = 0.94, median κ = 1.34 respectively), surprisingly for NASNetMobile



Supplementary Figure 6: Skewness κ of V4 neurons against four different DNNs. **a**. Skewness κ of response distributions for V4 neurons from Fig.1**c**. and AlexNet units. **b**. Same as **a** but with Xception units. **c**. Same as **a** but with VGG-19 units. **d**. Same as **a** but with NasNetMobile units. Lines: medians.

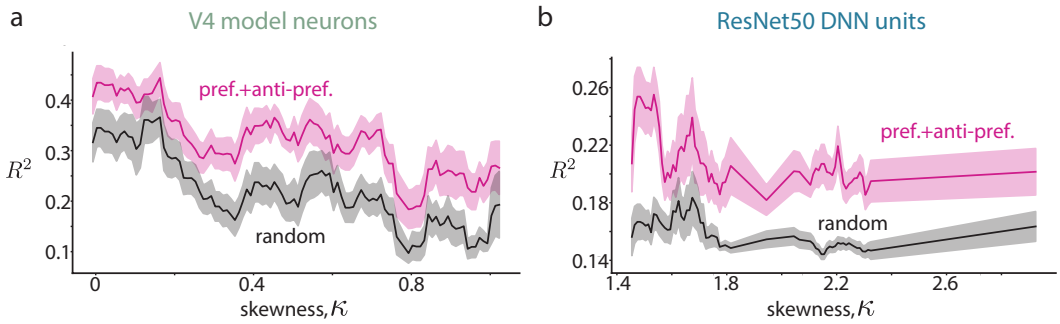
(median $\kappa = 0.44$) we found that to be not the case (Supp. Fig. 6**d**.). Despite the ReLU activation, units in NASNetMobile organize the responses in a way that preserves the two-tailedness of the distribution, thus effectively exhibit a linear behavior. We suspect that this is caused by the high baseline activations where these units rarely operate in the zero-output regime, hence creating two-tails. However, thresholding is still present in the model where a nontrivial fraction of activations still fall below zero in some layers. This finding hints at the importance and effects of architectural designs of DNNs in their selectivity. Compared to ResNet-50, AlexNet, VGG-19, and Xception, NASNetMobile utilizes modular cell structure where each cell combines the outputs from previous layers with addition operations, thus accumulating the activations from multiple layers. Therefore, the accumulation of the activation can increase the baseline activations, hence leading to always 'on' ReLUs. Although the residual skip connections are present in other DNNs such as ResNet-50, we suspect the frequency of these additions in NASNetMobile is what leads them to higher cumulative pre-activations.

A.4 The effects of skewness on R^2

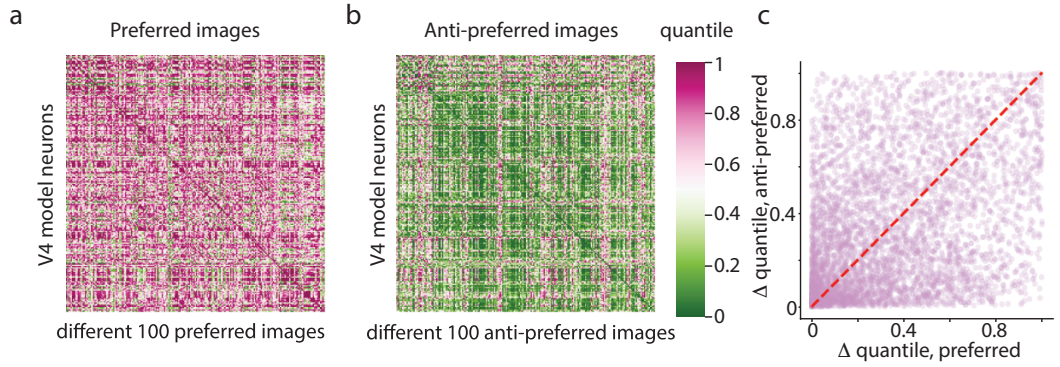
Our finding that preferred and anti-preferred training performed the best for ResNet-50 units was surprising (Fig.3**c**). After all, most ResNet-50 units had $\kappa \sim 2$ and exhibited no structured pattern in their anti-preferred images (Fig.5). To follow-up on this, we investigated the effect of skewness on R^2 for every V4 model neurons and ResNet-50 units from Fig.3. We found that across both models, overall R^2 was lower for more skewed (one-tail) distributions (Supp. Fig. 7, black and pink traces decrease as κ increases). This is in line with the informativeness of two-tails, where if a unit is less skewed, it can utilize more information, hence have higher R^2 compared to if it has one-tail. Therefore, we conclude that more skewed units have less R^2. Moreover, we see that the effect between pref.+anti-pref. and random is larger for more two-tailed distributions (Supp. Fig.7, difference between pink and black traces). This observation is consistent with our findings from Fig.3 where random surpasses the performance of pref.+anti-pref. at 8k training images. Thus, the gap between the traces in Supp. Fig.7 indicates that for less skewed units (two-tails), random can also leverage the structure in preferred and anti-preferred images.

A.5 Little relationship between preferred image similarity and anti-preferred image similarity across models.

To further investigate whether there is a shared structure between preferred and anti-preferred images, for a given V4 model neuron, we fed the sets of 100 preferred images of all V4 model neurons, and recorded the responses (we repeated this process for anti-preferred images as well.). In order to scale the responses proportional to the model's true preferred/anti-preferred, we used quantiles. For each set of preferred/anti-preferred images, we took the median and checked how many of the images out of 500k had responses lower than this, and we normalized this by the total number of images to get the quantile response.



Supplementary Figure 7: The effects of skewness κ on R^2 a. Skewness vs R^2 of each V4 model neurons for 1k images. Skewness vs R^2 of each ResNet-50 units for 1k images. Traces denote to binned average, shaded areas denote to 1 s.e.m.



Supplementary Figure 8: Preferred images are not shared across V4 model neurons. a. Responses of every V4 model neurons (rows) to every other V4 model neurons' 100 preferred images (columns). **b**. Responses of every V4 model neurons to every other V4 model neurons' 100 anti-preferred images. The mixture of pink and green indicates that some anti-preferred images were close to being preferred images of other models. **c**. Differences in quantiles for the preferred images matrix, and their corresponding differences in quantiles in the anti-preferred images matrix. The dashed red line represents the unity line (y=x).

Here, a quantile of 1 indicates that the preferred images of the ith model is also the preferred image of the jth model, and quantile of 0 indicates that the anti-preferred image of one is also the anti-preferred of the other. Moreover, we checked whether the two models with similar preferred images would also have similar anti-preferred images. To this end, for every row in Supp. Fig.8**a**.,we calculated the absolute difference between the model with the highest quantile and the model with the second highest quantile. We did this for the farthest quantiles (highest quantile - lowest quantile) and 20 randomly-chosen quantiles (highest quantile - randomly selected quantiles). We computed the corresponding models' differences from the anti-preferred matrix and compared these Δ q's against each other. Here we find that although some models have similar preferred and anti-preferred images (lower bottom left corner in Supp. Fig.8$_s$), others do not (Supp. Fig.8$_s$, top left corner, bottom right corner). Overall, most images did not fall on the unity line indicating that there is no linear relationship. These results further support our findings from Fig.4 where there was not an apparent structure shared across preferred and anti-preferred images.