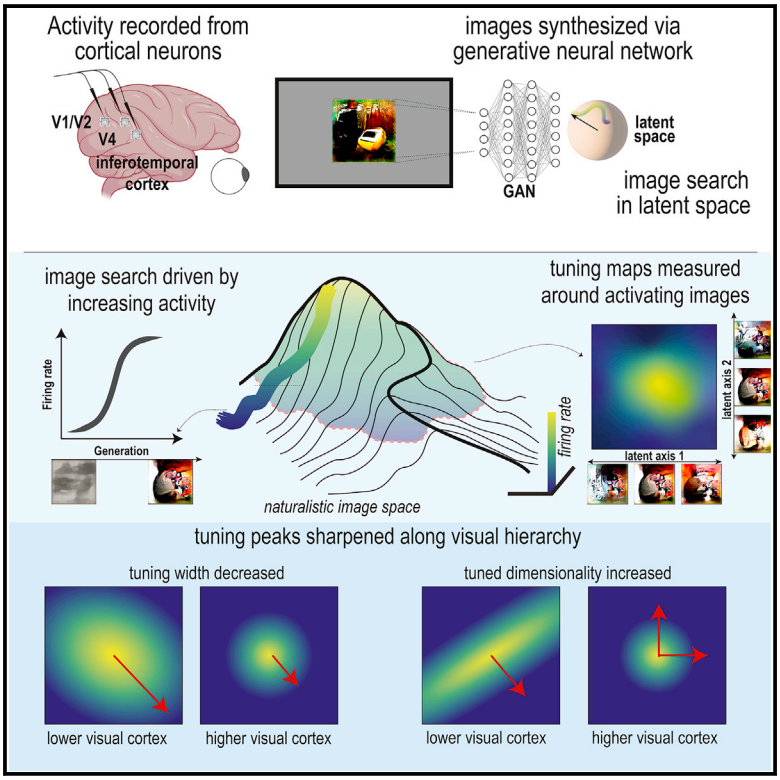
Article

Cell Reports

Tuning landscapes of the ventral stream

Graphical abstract



Highlights

- Images are optimized to strongly activate ventral stream neurons
- Neuronal tuning is measured around optimized images in a naturalistic image manifold
- Neurons show smooth bell-shape tuning, higher and wider than in classic spaces
- Neurons in anterior areas show narrower tuning with higher intrinsic dimensionality

Authors

Binxu Wang, Carlos R. Ponce

Correspondence

carlos_ponce@hms.harvard.edu

In brief

In this study, Wang and Ponce investigate the responses of V1, V4, and IT neurons to naturalistic images on a manifold. They optimize images to drive the neurons, then measure generalized tuning curves or "landscapes" around each response peak, showing a systematic difference in landscape geometry across the visual hierarchy.

Wang & Ponce, 2022, Cell Reports *41*, 111595
November 8, 2022
https://doi.org/10.1016/j.celrep.2022.111595

CellPress

Cell Reports

CellPress
OPEN ACCESS

Article

Tuning landscapes of the ventral stream

Binxu Wang[1,2] and Carlos R. Ponce[2,3,*]
[1]Division of Biology and Biomedical Sciences, Washington University in St. Louis, St. Louis, MO, USA
[2]Department of Neurobiology, Harvard Medical School, Boston, MA, USA
[3]Lead contact
*Correspondence: carlos_ponce@hms.harvard.edu
https://doi.org/10.1016/j.celrep.2022.111595

SUMMARY

A goal in visual neuroscience is to explain how neurons respond to natural scenes. However, neurons are generally tested using simpler stimuli, often because they can be transformed smoothly, allowing the measurement of tuning functions (i.e., response peaks and slopes). Here, we test the idea that all classic tuning curves can be viewed as slices of a higher-dimensional tuning landscape. We use activation-maximizing stimuli ("prototypes") as landmarks in a generative image space and map tuning functions around these peaks. We find that neurons show smooth bell-shaped tuning consistent with radial basis functions, spanning a vast image transformation range, with systematic differences in landscape geometry from V1 to inferotemporal cortex. By modeling these trends, we infer that neurons in the higher visual cortex have higher intrinsic feature dimensionality. Overall, these results suggest that visual neurons are better viewed as signaling distances to prototypes on an image manifold.

INTRODUCTION

A central goal in sensory neuroscience is to explain how neurons respond to natural images. Yet, to approach that understanding, the field has had to rely on simplified stimuli. Simple visual stimuli are easier to transform smoothly, which means that they can be used to generate tuning curves, descriptions of the relationship between a neuron's firing rate and a key "variable of interest" (Seung and Sompolinsky, 1993). The best-known example is the orientation tuning curve in primary visual cortex (V1) (Figure 1A). Here, the variable of interest is orientation, implemented using images of gratings at different slants. If a V1 neuron shows higher activity in response to a particular orientation, with smoothly decreasing activity to gratings with dissimilar orientations, then it is concluded that this space is encoded by the cell (Campbell et al., 1968). This type of tuning function has been a reliable tool in studies of the early visual cortex (V1, V2, V4) (Anzai et al., 2007; Hubel and Livingstone, 1987; Yau et al., 2013) and in dorsal cortical areas, such as the middle temporal (Maunsell and van Essen, 1983). Decades of work have resulted in the accumulation of an open set of variables describing neuronal function. However, there is no rationale about the uniqueness of any of these variables, nor an estimated number of all other potential variables that could be encoded by the visual cortex. How many other explicit variables could there be? How many more to explain a neuron's response to natural images?

Here, we take a broader perspective on this problem, by considering the concept of a tuning landscape, defined as the neuronal response function over the entire image manifold. Since the natural image manifold is a bounded space (topologically compact), if neuronal responses represent a continuous function, they have to exhibit a maximum in that space (as per the extreme value theorem) (Munkres, 2000); when input deviates from the maximum, the function will smoothly decay or stay constant. Given this maximum, we can revisit a framework where such a peak represents a special combination of visual attributes stored by the neuron—a *prototype* (Edelman, 1999; Rosch, 1973) (Figure 1B). In this framework, prototypes serve as landmarks in this representational space, and the neuron's tuning function signals the similarity between any given visual input and its prototype(s) (Poggio and Girosi, 1990a, 1990b). If this mechanism is accurate, then it follows that classic tuning axes (e.g., orientation) are accurate descriptions of neuronal function, but not unique. A more general distance-based function could explain why neurons respond to nearly all kinds of image types, from artificial computer-generated stimuli to randomly selected natural images (Figure 1C). The theoretical foundations for this concept have been developed extensively, partially as solutions to view invariance (e.g., radial basis function networks) (Maruyama et al., 1992; Poggio and Girosi, 1990a, 1990b), and as kernel machines (Anselmi et al., 2015) for visual recognition. While this idea has been tested in some electrophysiological (Logothetis et al., 1995) and imaging studies (Kay and Yeatman, 2017), its full explanatory potential has not been outlined, partially because there has been (1) no clear way to identify the specific landmarks (prototypes) represented by neurons and (2) no easy way to smoothly manipulate complex naturalistic images. In this study, we solve both problems using image generative models. We used a deep generator (Dosovitskiy and Brox, 2016) that maps 4,096-dimensional latent vectors to naturalistic images. This model provided a parametric proxy for natural image space, allowing us to smoothly manipulate natural images and to use search algorithms (optimizers) to

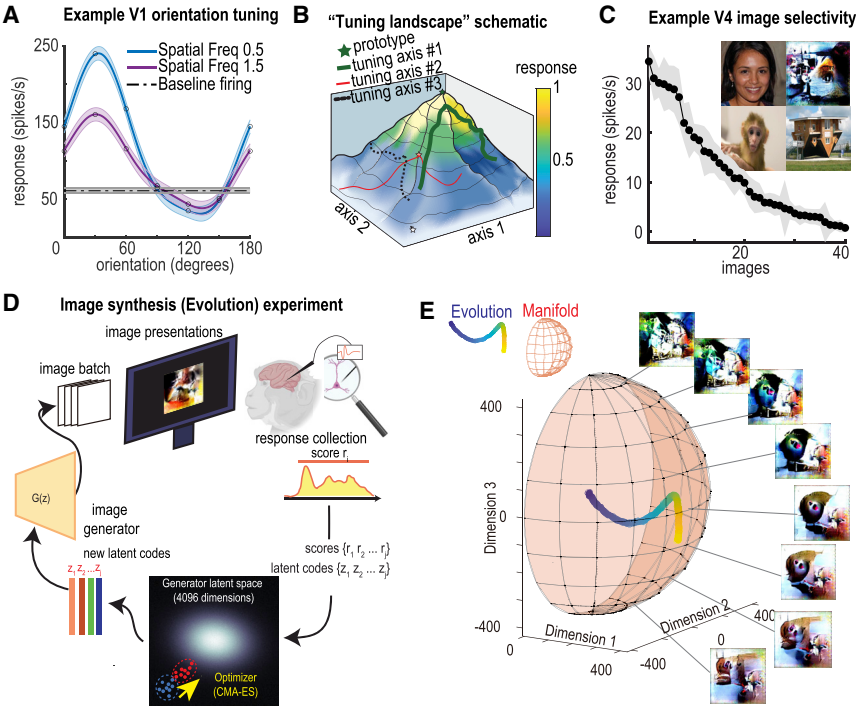


Figure 1. Conceptual approach

(A) Orientation tuning curve (mean ± SEM) for a V1 multiunit site as evoked by Gabor patches.

(B) Schematic of a conceptual "tuning landscape", the neuron's activation function over the space of all visual inputs (2D images); schematic shows two input axes, green star shows location of peak; lines over the landscape illustrate different experimentally defined tuning axes (e.g., orientation, curvature) as represented by images.

(C) Responses (mean ± SEM) of a single neuron in anterior V4 to randomly selected stimuli; inset shows top four images.

(D) Workflow for the Evolution of a preferred image (prototype), combining the image generator (*G*, latent codes as inputs and images as outputs), neuronal responses to each generator image, and an optimization algorithm (CMA-ES) that samples the generator's latent space to maximize neuronal responses.

(E) Main experimental strategy: a 2-fold approach to identify a peak in the image generator space (Evolution) followed by an experiment characterizing neuronal responses as image moves away from preferred location in the generator space (Manifold) (Figure S1).

find tuning peaks in this space. We characterized the tuning landscape of neurons in V1/V2, V4, and inferotemporal cortex (IT) (Figures 1D and 1E), by mapping neuronal responses on a 2D manifold sampled around the peak, then identifying neuronal tuning shape, extent, and smoothness. We also explored the global distribution of peaks by conducting image searches in reduced subspaces. We found three systematic changes in both *in vivo* and *in silico* visual hierarchies: the tuning width decreased along the hierarchy, the search convergence time increased, and there was an increasing gap in responses between reduced subspaces and the full space. We explain these trends using a simple model of radial tuning varying in tuning width and intrinsic tuned dimensionality. Overall, our results suggest that ventral stream neurons can be viewed as operating in a multidimensional distance space, in analogy to hippocampal place cells, where responses signal the proximity between two points in physical space—the neuron's spatial field center at (x_0, y_0) and the input location at (x_1, y_1). In this analogy, the radial distance is the primary functional feature, while the direction of approach to the spatial field center is secondary.

RESULTS

Neurons show bell-shaped tuning around peaks in the generator space

To measure the shape and extent of neuronal tuning in the generator space, our experimental approach comprised two parts, an Evolution and a Manifold experiment (Figures 1D and 1E). The Evolution experiment identified images that maximized neuronal responses, and the Manifold experiment sampled the neighborhood around such images. This allowed us to measure a local tuning landscape and to characterize the width and smoothness of tuning around the observed peak.

Evolution experiments

This methodology was used previously (Rose et al., 2021). In brief, after selection of a neuronal site (single or multiunit) in V1/V2, V4, or IT (receptive fields in Figure S1A), we presented shapeless, texture generator images, and recorded the firing rates evoked by each image. These single-trial responses were used as scores for the input vectors behind each generated image; then the optimizer (CMA-ES algorithm) (Loshchilov, 2017) took the scores and vectors and proposed new input vectors likely to maximize firing rates in the next generation. These input vectors were fed back into the generator to create stimuli for the next round, closing the loop. Each cycle lasted for 0.5–1 min; each experiment ran for ~30 min, until the neuronal site's mean response stabilized into a local maximum—having "evolved" a preferred stimulus (Figure 2B). We refer to these GAN-derived preferred stimuli as prototypes, as inspired by prototype and similarity theory, which suggests that complicated concepts may be summarized by templates (Edelman, 1999; Leopold et al., 2001). Prototypes are pictorial illustrations of the specific combinations of colors, shapes, and textures encoded at each given site; they are often complex (Tanaka, 2003), and, while consistent in shape and color over time, upon repeated recordings they can vary in size, position, and other nuisance variables. Each experiment also included the presentation of reference images used to track neuronal isolation stability; these included photographs, Gabor patches, or other stimuli, depending on the site's preferences, as established in separate experiments (Rose et al., 2021).

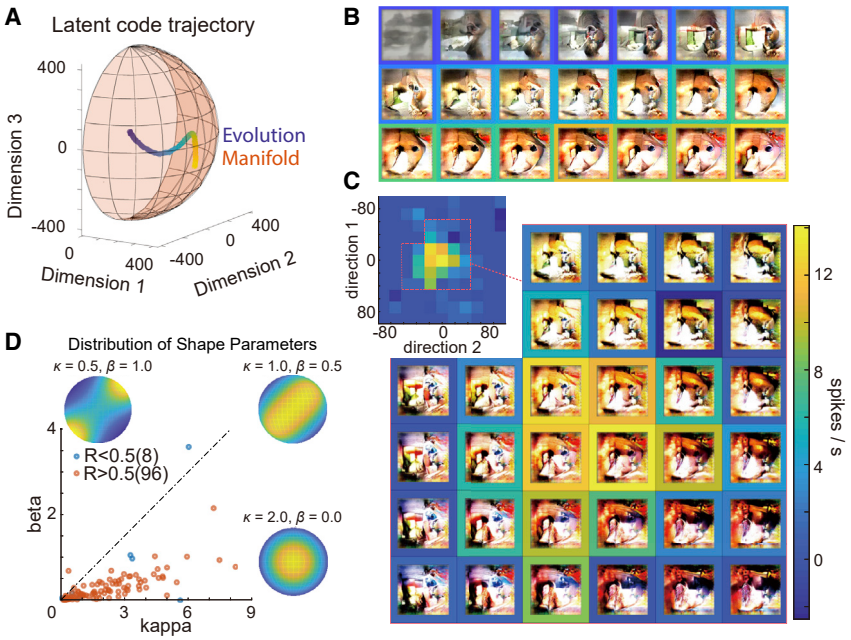


Figure 2. Characterizing tuning landscapes

(A) Schematic illustrating the two-part design for defining the tuning landscape: Evolution and Manifold experiments. The blue-yellow curve illustrates the trajectory of latent vectors during an Evolution experiment, as projected onto its top-three principal-component space, with color representing generation number (blue, first; yellow, last). This trajectory was measured in an Evolution experiment driven by a single neuron in posterior IT. The orange hemisphere illustrates the sampling space for Manifold experiments, based on the hemisphere in the top-three principal-component space with the longitude and latitude grid (18° intervals). This hemisphere explored the manifold around the approximate endpoint of the Evolution trajectory.

(B) Change in generated images throughout the Evolution experiment for the same single unit. Images illustrate mean latent code (every fourth generation, from left to right, top to bottom). These latent codes correspond to the curved trajectory in (A). Color around each image represents the mean activation for that generation.

(C) Example tuning map for the same single unit. Inset shows the image corresponding to each position of the map.

(D) Distribution of the shape parameters κ, β of the Kent function fits; maps well-fit by Kent are colored red; the dashed line $\kappa = 2\beta$ is the boundary in parameter space between unimodal and bimodal Kent functions (Figures S1, S2, S3, S4, S7, and S8).

Manifold experiments

The location of the prototype in the generator space was an anchor point to sample images along a manifold. To define this sampling manifold, we performed a principal-component analysis on the trajectory of latent codes during evolution. The first principal component corresponded roughly to the direction taken during prototype synthesis (Figures S2A–S2C); we then created a 2D sphere centered at the origin in the subspace spanned by the first three components. The radius of this sphere was defined by the norm of the latent code in the last generation of the Evolution experiment (Figure 2A; STAR Methods). The choice of exploring on the sphere was motivated by the geometry of the generator latent space (Wang and Ponce, 2022) and by literature on sampling GANs (Kilcher et al., 2017; Wang and Ponce, 2021; White, 2016). Images sampled from a grid around the prototype (termed the manifold image space) were then displayed in a rapid-serial-visual presentation design, along with images from other stimulus sets used to define neuronal tuning (Enroth-Cugell and Robson, 1966; Lin et al., 2014; Pasupathy and Connor, 1999; Russakovsky et al., 2015), such as Gabor patches and curved 2D contours. We refer to the responses over the manifold image space as the tuning map, defined as a function over the 2D hemisphere. In the next three sections, we report results pooling single and multiunits (SUs and MUs, collectively named *sites*); differences between SUs and MUs are examined in Figure 4.

Most Evolution experiments (79.1% of 91) resulted in a significant increase in the firing rate of the site under study (per two-sample t tests of firing rate in initial two generations versus last two generations, with criterion $p < 0.001$). Having identified tuning peaks in the generator space, we then examined the shape of tuning function in the Manifold experiments. We found that neuronal sites decreased their responses smoothly when moving away from the prototype in any direction (Figures 2C and S1B). This kind of tuning was reminiscent of tuning curves of V1 simple cells responding in the space of oriented gratings. Sites were significantly modulated by the manifold image set (104/110 of experiments, $p < 0.001$, one-way ANOVA, 101 images per test, F-stat range [1.56,68.64]). For more relationships between the success of the Evolution experiments and other properties of tuning maps, see Figure S2.

Above, we noted that a common aspect of tuning curves is their smoothness. Theoretical work suggests that tuning smoothness can serve as a key inductive bias that enables downstream readout neurons to learn from few examples (Bordelon et al., 2021). We hypothesized that smoothness is important because the neuronal code is noisy, and so a neuronal population comprising smooth tuning maps would allow information decoding with smaller errors. To characterize the smoothness of these manifold tuning maps without committing to a specific function type, we measured the Dirichlet energy (DE) of the observed maps and compared this value with the energy of maps where the image-response relationships were shuffled (see STAR Methods and Figure S1E). This energy metric integrates the squared norm of the gradient vectors over the hemisphere, thus quantifying the relative smoothness of the tuning function: a constant function will have zero DE, while a rugged tuning function will result in a high DE. We found that the observed tuning maps were remarkably smooth, showing a significantly smaller DE than the upper ceiling defined by shuffled data (observed versus image-identity shuffled maps, two-sample t test, $p < 0.001$ in 96/110 experiments, T-stat ranged from [−186.6,−8.3], Cohen's d' ranged from [−8.32,−0.37] for the smooth channels).

Next, we set out to quantify the tuning shape around the prototype. By visual inspection, we noticed that most tuning functions appeared bell-shaped. To quantify this, we used the Kent function to fit the 2D tuning maps. The Kent function is defined on a sphere and is analogous to a 2D Gaussian in Euclidean space. This function comprises parameters for (1) activity baseline, (2) tuning amplitude, (3) two parameters for peak location, and (4) three shape parameters. We found that Kent functions generally fit the tuning maps well, with a median R^2 of 0.72 (N = 110) (Figure 2D; for examples of raw and fitted tuning maps, see Figure S3). For reference, the R^2 noise ceiling obtained by bootstrapping single-trial responses had a median 0.876 (N = 110) (STAR Methods). To contextualize this result, we also repeated the Evolution-Manifold experiments using convolutional neural networks (CNNs), which are noise-free, and measured how well the Kent function could fit these maps. The tuning map fits from the VGG-16 network had an R^2 value of 0.918 ± 0.050 (mean ± SD, median 0.922, N = 427 units sampled across nine layers). The tuning maps of V1 driving sites were less-well fit by the Kent function than V4 and IT driving sites: R^2 values for V1 were 0.665 ± 0.027 (N = 31), for V4 0.787 ± 0.029 (N = 21), and IT 0.787 ± 0.020 (N = 50).

Having established that the Kent function could serve as a good model, we then used its shape parameters β,κ to quantify tuning, as the ratio of these parameters indicates the isotropy versus anisotropy of the peak on the manifold sphere. The parameter $\beta = 0$ indicates a perfectly isotropic peak, $\kappa/\beta > 2$ indicates a unimodal function, and $\kappa/\beta < 2$ indicates a bimodal function (see Figures 2D and S1D). In experiments with good Kent fits ($R^2 > 0.5$, $n = 96$), the 95% confidence interval of κ/β was $[7.02, 9.51]$, significantly larger than 2 (one-sample t test, $t = 18.37, p = 5.0 \times 10^{-33}$). This showed that most tuning functions were indeed unimodal, bell shaped, and relatively isotropic within the measured domain.

In summary, these analyses showed that, when neurons were presented with smooth and complex deformations from their preferred stimulus, they also showed a smooth reduction in firing rate; this reduction was similar no matter in what direction the transformation occurred. This is the signature of a radial basis function, where the response change depends largely on the distance from the preferred stimulus, and less so in the direction taken by the deformation. This type of response change is likely to be characteristic of neurons in natural conditions, where changes in the retinal image can be smooth (as in viewpoint rotations) while varying across multiple visual features.

Relating neuronal tuning in generator space to other image spaces

Neurons showed radial tuning around their preferred stimulus, and we interpreted this tuning as a likely default behavior of neurons when presented with generic but smooth image changes. However, our overarching motivation was to define how neurons respond across even larger image domains, involving images with different low-level statistics, such as photographs and computer-generated artificial stimuli. While the generator produces an astronomical number of variations, it has limits. It produces images with a certain textural style and is less photorealistic than newer generators (Brock et al., 2018). Here, we advanced our goal of defining tuning function shapes by including images created both within and outside the generator space. Specifically, we collected responses of given sites to Manifold images, to stimuli from classic image spaces, such as Gabor patches (parametrized by orientation and spatial frequency), to bounded 2D contours (parameterized by orientation and curvature) (Pasupathy and Connor, 2002), and to ImageNet photographs. Together with our generator images, this image set spanned a range from simple/artificial to complex/natural images. While we reasoned that it would be easy to measure neuronal responses to these images, the key challenge was to find a common axis that would fit images with such diverse statistics. To solve this problem, we developed a radial tuning curve analysis (Figure 3A, left), aimed to construct 1D tuning curves using image-response pairs with stimuli created from different spaces, and to define the shape and extent of this tuning function. We first considered each space separately and fit the neuronal responses as a function of image distance to the most activating image in that space (responses were smoothed using Gaussian process regression). Across these spaces, two features of each tuning curve were comparable: (1) peak amplitude and (2) radial tuning width (Figure 3A, right). The peak amplitude was defined as the maximum response among images in each space, signaling the effectiveness of the image space. For the radial tuning width, we smoothed and integrated the area under this radial tuning curve. This estimated the tuning width of the peak, i.e., how far away images could deviate from an optimum while still allowing the neuron to stay responsive. Our image distance was a perceptual similarity measure (LPIPS) based on CNNs trained to match human perceptual judgments (Zhang et al., 2018). We performed this analysis for each driving site and compared the peak amplitude values and radial tuning widths across each space.

We found the neuron-optimized manifold space had the highest peak amplitude (Figures 3B and 3C): the peak firing rate was 228.5 ± 16.8 spikes/s (mean Z score activation 2.532 ± 0.094, combining single and multiunits) compared with 144.6 ± 12.7 spikes/s (Z score 0.630 ± 0.074) for the best natural images and 181.2 ± 16.2 spikes/s (Z score 1.074 ± 0.112) and 160.9 ± 15.7 spikes/s (Z score 0.592 ± 0.086) for the best curved-object images and Gabor patches (Figure 3C, paired t test, Z-scored amplitude, versus natural reference $t_{90} = 17.4$, $p = 1.5 \times 10^{-30}$; versus curvature images $t_{71} = 10.2$, $p = 1.5 \times 10^{-15}$; versus Gabor patches $t_{69} = 16.3$, $p = 2.8 \times 10^{-25}$, Z score specifics, STAR Methods). This confirms that optimized prototypes were more activating for neurons than manually selected ones (Ponce et al., 2019), even in the context of great V1/V4 artificial stimuli, such as curvature patches (Figure S5, for comparisons focusing on V1 and V4—for the V1 population, the peaks in the manifold space were higher than those for Gabor patches; for the V4 population, the peaks in manifold space were more similar to those in curvature space).

Next, we computed the area under the interpolated radial tuning curves (AUC), above the baseline firing rate. We normalized the AUC by dividing the peak activation of each tuning curve, to compare the tuning width while controlling for peak height (Figure 3B; STAR Methods). We found that the median normalized AUC for the manifold space was 0.370 ± 0.009 (Figure 3D),

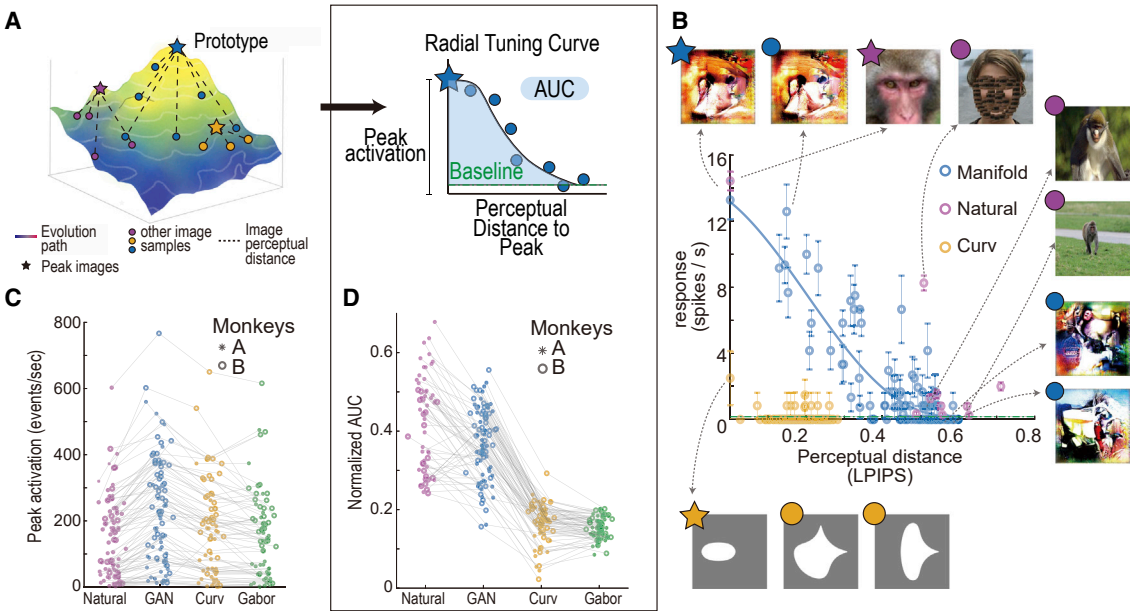


Figure 3. Characterizing tuning landscapes with radial tuning curves

(A) Schematic of radial tuning curve analysis: in each image space, we identified the most activating image and measured the perceptual distance of all other images (in that space) to it. After plotting the activation as a function of image distance, we computed two statistics for each space: peak activation (C) and normalized area under the curve (normAUC, D).

(B) Example of a radial tuning curve (same IT neuron as in Figure 2), quantifying responses to images across different spaces (generator images, blue; photographs, purple; curvature images, yellow). The scatter plot shows the mean ± SEM response to individual images. The solid blue line shows a fit through the Manifold image responses, which were sampled most continuously in the space.

(C) Comparison of peak activations across image spaces, for the same driving sites recorded in both animals. Each site is denoted by a point; the gray lines connect the values across image spaces for the same site.

(D) Tuning width estimated by the normalized AUC of the radial tuning curve, also compared across image spaces (Figure S5).

compared with 0.164 ± 0.005 for the curvature space (paired t test, $p = 3.9 \times 10^{-42}$, Student $t_{80} = 27.16$), 0.153 ± 0.003 for Gabor shapes ($p = 1.4 \times 10^{-40}$, $t_{78} = 26.34$), 0.424 ± 0.016 for natural images ($p = 2.0 \times 10^{-8}$, $t_{57} = -6.52$); when including sessions where V1 sites drove the Evolution and Gabors were used as evolution references, the average was 0.334 ± 0.016 ($p = 8.4 \times 10^{-2}$, $t_{90} = 1.75$, n.s.). We should emphasize that, even when we focused on neurons sampled from V1 and V4 and tested with their classic stimuli spaces (Gabor patches and curvature images), our results held: neurons had a larger absolute and normalized AUC in the manifold space than in classic stimuli spaces (normalized AUC, for the V4 population, Manifold versus curvature, $t_{19} = 10.22$, $p = 3.7 \times 10^{-9}$; for the V1 population, Manifold versus Gabor, $t_{34} = 8.24$, $p = 1.3 \times 10^{-9}$, Figure S5). This shows that tuning curves measured in simpler image spaces can underestimate both the dynamic response range of the neuron and the extent (width) of its tuning. This can be viewed as evidence that classic tuning curves are local sections of a higher-dimensional tuning landscape.

What does the radial tuning width mean? To contextualize these results, we characterized the geometry of these image spaces using LPIPS. First, we measured the diversity of each space as the average distance between two random samples. We calculated the distance between 1,000 random samples from the 50,000 ImageNet validation set and found that the mean pairwise distance was 0.56 ± 0.07 (mean ± SD, [5, 95] percentile [0.46, 0.67]). In comparison, when we randomly sampled 500 Manifold images, they spanned a mean difference of 0.47 ± 0.04 ([5, 95] percentile [0.41, 0.53])—curvature and Gabor patch spaces spanned a mean distance of 0.17 ± 0.05 ([5, 95] percentile [0.08, 0.24]) and 0.16 ± 0.04 ([5, 95] percentile [0.08, 0.22]). Thus, generator images covered a range of shape diversity closer to that in the real-world images. Next, we examined the sample density in these image space using nearest-neighbor distances. For natural images, we computed the distance to a nearest neighbor (within the 50K image set) for 1,000 images from the ImageNet validation set: the mean distance was 0.364 ± 0.051 (mean ± SD), ([5, 95] percentile [0.271, 0.440]). In contrast, in the Manifold experiments, the mean distance between nearest samples was 0.087 ± 0.035 ([5, 95] percentile [0.029, 0.136]); for curvature and Gabor-patch spaces this distance was 0.031 ± 0.011 ([5, 95] percentile [0.022, 0.045]) and 0.081 ± 0.006 ([5, 95] percentile [0.068, 0.087]). For the radial tuning analysis, this difference in sampling density made the tuning width estimate in natural image space less accurate than the more densely sampled spaces, such as the generator, curvature, and Gabor spaces.

We have shown that the tuning width estimates in the generator space were smaller compared with those measured in the space of real-world images. Why is this? First, our image generator is an imperfect approximation of the natural image space,

namely, it is not diverse and as realistic as the natural world itself; a more photorealistic GAN might reveal tuning widths closer to natural images. However, tuning width values estimated from natural images were less reliable, due to the larger sampling gap in photographs. This sampling gap was difficult to overcome by increasing the sample size of natural images—even in a 50,000-image set, only 121 out of 1,000 images had any neighbors within 0.3 perceptual distance, already a coarse step. Finally, the gaps between natural image samples were sometimes larger than the tuning width in generator space (Figure 3B). It is possible that natural image samples marked different tuning peaks on the tuning landscape, further inflating the estimation of radial tuning width. This large sampling gap is likely one reason that smooth tuning curves are seldom reported for IT cortex neurons.

In summary, we defined the shape and extent of neuronal tuning across image spaces, including that of the generator, photographs, and classic parametric stimuli, such as Gabor functions. We used a perceptual similarity metric to compare the neuronal responses to this diverse image set. We found that neurons showed a smooth decay consistent with that found in the generator space. The tuning functions measured around the neuronal prototype were larger in amplitude and wider in extent than those measured over classic non-optimized stimuli. This confirmed that manually selected image sets may not always overlap with the preferred domain of the neuron and could underestimate its full dynamic range and input domain. We interpreted this tuning width as a feature that makes these neurons well suited for processing naturalistic images, i.e., allowing them to remain informative over large swaths in image space. Furthermore, we hypothesize this expansive tuning function is the reason we could use search algorithms (e.g., CMA-ES) to find peaks on the tuning landscapes in the first place (see "inferring the geometry of the tuning landscape by modelling"): as we quantified previously (Wang and Ponce, 2022), the mean distance among each generation of images usually started around 0.4 and gradually decayed to ~0.2. This step size of image change seems to be well suited to "climb" the slope of neuronal tuning peaks.

Areal differences of tuning landscapes

We have shown that neurons in the ventral stream have smooth radial tuning functions spanning a large space of image transformations. Next, we investigated how these functions differed across the visual hierarchy, focusing on neuronal sites in V1/V2, V4, and IT. Neurons in these cortical areas are usually studied using different stimuli, limiting comparisons. Here, the generator space served as a common space for all areas. We examined the local and global properties of the tuning landscapes using the Evolution-Manifold design—finding a site's local maximum and then exploring the landscape by inducing deformations to the favored image(s). Furthermore, we indirectly compared the global characteristics of the tuning landscape by analyzing the Evolution trajectories and by using a new set of reduced-dimension Evolution experiments.

Local properties

Because neurons become increasingly sensitive to specific visual patterns over the cortical hierarchy (Kobatake and Tanaka, 1994; Rust and DiCarlo, 2010), one prediction is that tuning width in the generator space might vary across areas (Figure 4A). To quantify tuning width across areas, we used the κ parameter of the Kent function, which characterizes tuning sharpness: the higher the κ value, the narrower the tuning width. In tuning maps with reasonable fits ($R^2 > 0.5$, $N = 96$), we found that sites from higher visual areas showed a larger κ value than those from lower areas—V1 sites showed a mean (±SD) κ of 0.72 ± 0.13; V4 sites, 1.82 ± 0.23, and IT sites, 3.19 ± 0.26 (Figures 4B and 4C(i), one-way ANOVA, $F_{2,93} = 26.7$, $p = 7 \times 10^{-10}$; Spearman correlation between κ and area level was $0.661 (df = 96, p = 2.4 \times 10^{-13})$). We confirmed this progression using two non-parametric statistics: we compared (1) the normalized AUC for the radial tuning curves in each area (smaller values indicate narrower tuning) and (2) the normalized volume under the surface (VUS) for each tuning map (smaller means narrower, STAR Methods). Both of these values showed a similar trend: more anterior visual neurons showed sharper tuning peak values than posterior neurons (Spearman correlation of AUC values with ordinate hierarchical position [V1/V2, V4, and IT], $\rho = -0.56$ $(df = 103, p = 5 \times 10^{-10})$; for VUS $\rho = -0.57$ $(df = 103, p = 3 \times 10^{-10})$. We also compared the tuning width values for single versus multiunit signals in the same channels. Single units had narrower tuning, consistent with the view that multiunits represent the local aggregated signal from single units (Figure 4C(ii), for more examples see Figure S7). As noted above, the tuning maps of V1 sites were less-well fit by the Kent function than those of V4 and IT sites: mean R^2 for V1 was 0.67 ± 0.03 $(N = 31)$, for V4, 0.79 ± 0.03 $(N = 21)$, and for IT, 0.79 ± 0.02 $(N = 50)$. This trend was in line with the noise-ceiling R^2 of the three microelectrode array populations, measured by randomly splitting trials into two sets and computing the R^2: for V1, the noise ceiling for R^2 was 0.74 ± 0.03 $(N = 31)$, for V4, 0.89 ± 0.02 $(N = 23)$, and for IT, 0.88 ± 0.02 $(N = 50)$. Overall, this suggests that the tuning maps in V1 were more affected by fluctuation, less bell-shaped, and thus harder to fit by the Kent function.

As a control, we asked if there were any significant differences in the perceptual similarity of images in the V1, V4, and IT manifolds. We were surprised to find a small but significant difference in image diversity (quantified by the mean LPIPS image distance between two images on the hemisphere) between manifolds created by IT and V1 (IT, 0.421 ± 0.003, $n = 33$; V1, 0.410 ± 0.003, $n = 36$, two-sample t test $t_{67} = 3.083$, $p = 0.003$). This overall modulation of image diversity per visual area was only marginally significant (one-way ANOVA across the three areas $F = 4.669$, $p = 0.012$). This effect size was small: for example, neuronal tuning in perceptual metric space ranged from 0 to 0.60 (Figure 3), and the mean difference in perceptual similarity between V1 and IT was ~0.01. We hypothesize that this difference in image diversity was a byproduct of the Evolution and Manifold design: searching for features of different complexity (i.e., for V1 versus for IT) may lead to trajectories in different subspaces; since images were manipulated along the second and third principal-component subspaces of the search trajectory, the manifold exploration in different subspaces could result in small but consistent differences in terms of image diversity, reproducible per *in silico* Evolution and Manifold experiments. As another control, we also conducted

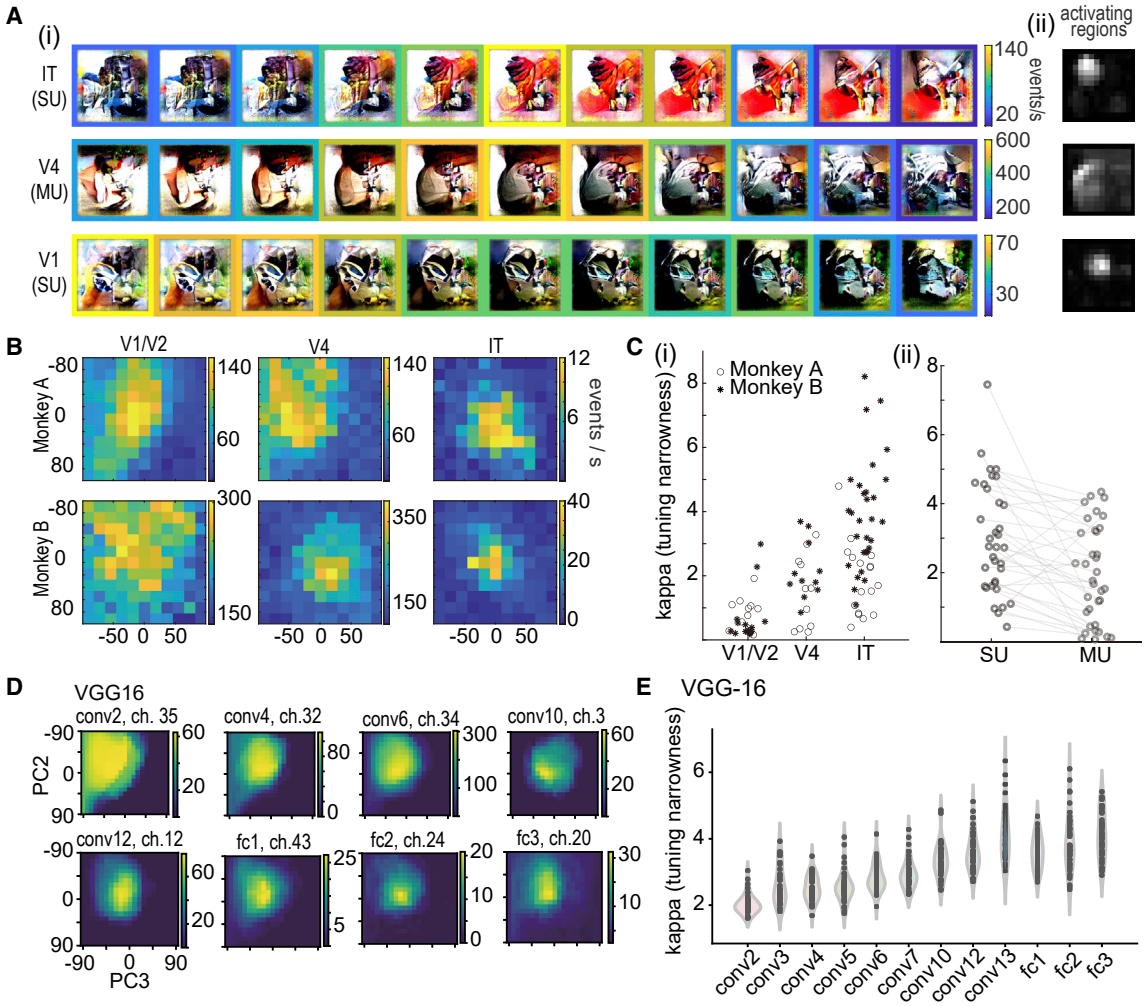


Figure 4. Comparison of tuning landscapes along the hierarchy

(A) (i) Examples of tuning axes sampled from manifolds in IT, V4, and V1 (rows); images show deformations from each site's prototype, framecolor denotes neuronal response. (ii) Activating regions within the synthetic image, identified via correlation-based feature localization (see STAR Methods).

(B) Example 2D tuning maps of sites at the V1/V2 border, in V4, and in PIT (columns), for both animals (rows).

(C) (i) Population tuning width values as quantified by the κ coefficient (higher values → narrower tuning) across visual areas and animals. (ii) Tuning sharpness values κ for single and multiunits.

(D) Example tuning maps of units in VGG-16. Maps ordered by layer depth.

(E) Tuning sharpness value κ as a function of layer in VGG-16. Each point represents one hidden-unit tuning map (Figures S3, S4, and S6).

Evolution-Manifold experiments in V1 sites using textures that were 1° wide (versus 3° wide as in the experiments above) and found that the results above continued to hold (Figure S4). We also investigated the shape of tuning landscapes far away from the peak, by examining the concurrent responses of sites that were not driving the evolution: those tuning maps were more ramp and slope shaped (Figure S8).

We also asked if this pattern of increasing tuning sharpness was specific to visual cortex or if it could emerge in artificial visual hierarchies. We performed the same Evolution/Manifold experiments *in silico*, driven by the activation of randomly selected units from pre-trained CNNs. We found the same pattern of results: tuning peaks over the manifold became sharper as a function of layer depth. For example, in the VGG16 network (Simonyan and Zisserman, 2014), κ values correlated positively with layer depth (linear regression slope, 0.189 ± 0.008, $p = 1.9 \times 10^{-91}$, $N = 568$ units in 12 layers, Figure 4E). We found the same result using AlexNet, ResNet-101, ResNet50, ResNet50-Robust, DenseNet, and CorNet-S (p values ranging from 3×10^{-22} to 2×10^{-124}, N ranging from 600 to 1,200, STAR Methods, Figures S6A–S6F).

So far, we have explored a local property of the tuning landscape, specifically that visual hierarchies comprise units with progressively sharper tuning peaks in this naturalistic space. One straightforward interpretation of this result relates to sparse coding (Rolls and Tovee, 1995), as higher-order cortical units respond to more specific combinations of visual attributes, such as motifs present in faces (Desimone et al., 1984; Tsao

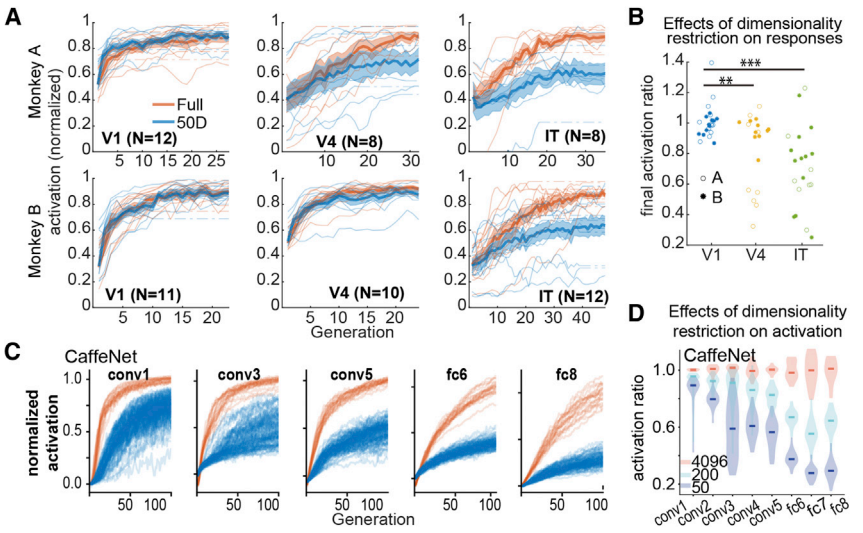


Figure 5. Comparison of required search dimensionality along the hierarchy

(A) Activity as a function of generation during Evolution for neurons in V1/V2, V4, and IT using the full input space (4,096D, orange) or a reduced space (50D, blue), for each animal (top, bottom). Activity normalized by maximal response within each session, with mean ± SEM computed across sessions, and smoothed via moving average (N = 3 generations). Evolution experiments that converged with fewer generations were extrapolated with the same activation value as the last generation.

(B) Effects of dimensionality restriction on activation maximization for sites in V1, V4, or IT. The y axis shows the ratio of the final-generation activation values measured after evolving in a 50D space and in a 4,096D space $\frac{r_{50d} - r_{bsl}}{r_{full} - r_{bsl}}$ (see STAR Methods). Each dot shows one neuronal site, for both animals. Asterisks show statistical significance.

(C) Activity as a function of generation during Evolution for hidden units in CaffeNet layers using the full space (orange) or reduced 50D space (blue).

(D) Effects of dimensionality restriction on activation maximization for hidden units in different layers. The y axis shows the ratio of the final-generation activation values measured after evolving in a 50D space (deep blue), a 200D space (light blue), and a 4,096D space (red) (Figure S6).

et al., 2006). If these attribute combinations were less common within the generator space, compared with simpler features, such as colorful curved edges, then it would be easier to reduce the responses of higher-order neurons by "moving away" from the peak by the same distance.

Next, we examined a global property of the tuning landscape: the distribution of prototypes. Evolution experiments showed the existence of highly activating images in the search space and a path following the gradient of the tuning function, from the initial images (textures) to the final images (prototypes). By analyzing these experiments, we could probe the global geometry of the tuning landscape indirectly. The speed for finding a prototype related to the frequency of these highly activating images in the latent space. For example, specific combinations of visual features (e.g., shapes, colors, and textures diagnostic of animal features) should be statistically rarer compared with simpler features (such as oriented edges). If higher-order neurons select for more complex features, it should take longer to find them. This prediction aligns with previous studies of sparse object coding in the ventral stream, for example, those finding that V4 neurons are tuned mostly for acute curvatures (Carlson et al., 2011), in contrast to V1 neurons, which as a population tile a relatively larger portion of orientation space (although V1 does over-represent some orientations; Dragoi et al., 2000). To test this idea in the landscape context, we examined the convergence time of Evolution experiments. We found that the mean firing rate for sites in posterior regions (i.e., V1) converged to a steady state over fewer generations than the firing rate for sites in more anterior regions: the convergence timescale for V1/V2 neurons was 12 ± 1.2 generations (N = 34 experiments), compared with V4, at 21.4 ± 3.3(N = 20), and IT at 28.7 ± 2.5(N = 30) (Figures S6G–S6J). The convergence timescale was quantified by the number of generations needed to reach 0.632 ($1 - 1/e$) of the maximum mean firing rate per generation. This trend confirms a previous report (Rose et al., 2021), also observed *in silico* in AlexNet. This would be expected if higher visual neurons prefer features which were sparsely distributed in the space, and harder to find.

If a neuron's preferred visual features were common in the generative space, the algorithm might still find a path to a peak even if the search was limited to a fraction of the full latent space. To test this, we performed a set of reduced-dimension Evolution experiments, comprising two concurrent Evolutions for each target site: in one track, the search occurred in the full 4,096-dimensional (4,096D) latent space; in the other one, the search algorithm was constrained to a randomly selected 50-dimensional (50D) subspace. Trials from both tracks were interleaved. Then, we measured the ratio R between the activation increase in the 50D evolution and that in the 4,096D evolution. We found that the mean ratio R for V1/V2 sites was 1.009 ± 0.023 $(n = 23)$, for V4 sites, 0.839 ± 0.056 $(n = 19)$, and for IT sites 0.708 ± 0.060 $(n = 20)$ (Figures 5A and 5B); Spearman correlation between hierarchy level and the activation fraction was -0.537 $(P = 6.7 \times 10^{-6})$. We also calculated $\overline{d'}$, the average activation gap (per d-prime d') across all generations and found a growing gap along the hierarchy (Spearman correlation between hierarchy level and mean $\overline{d'}$ was 0.547, p = 1.4×10^{-6}). Thus, early ventral-stream neurons could still guide the evolution of their preferred image in constrained 50D spaces, reaching similar (or sometimes higher) responses compared with the 4,096D space. In contrast, late ventral-stream neurons could not guide the evolution of their preferred image as successfully. This result was also replicated on the artificial visual hierarchy CaffeNet, where we found that limiting search space dimension had a larger impact on units in deeper layers than in shallower layers (Spearman correlation between ordinal layer number and ratio R is -0.893, p = 2.4×10^{-279}; n = 800, 100 experiments per layer for eight layers, Figures 5C and 5D, see STAR Methods).

Together, these results suggest that the features preferred by early cortical neurons were more prevalent in the generator image space than the features preferred by late cortical neurons, likely due to the latter's selectivity for complex features

(Kobatake and Tanaka, 1994); this may also explain why early ventral stream neurons showed broader tuning widths than those of higher-order neurons. We conclude that this trend in tuning specificity is likely to be a general feature of hierarchical networks that learn from natural data.

Inferring the geometry of the tuning landscape by modeling

Above, we reported three systematic changes along the hierarchy of the ventral stream and CNNs: deeper in the hierarchy, (1) the time to convergence in Evolution increased, (2) the tuning width of neurons decreased, and (3) reducing the search dimensionality had a stronger negative effect on activation. Is there a mechanism that accounts for all these effects? We postulated that these trends were manifestations of a systematic difference of tuning landscapes along the ventral stream. To provide more intuition into this, we designed a simple synthetic tuning function to simulate these changes. We modeled the neuron's tuning function in the 4,096D latent space of the generator, with a simple multivariate Gaussian. We then conducted the Evolution, Manifold and reduced-dimension Evolution experiments on this synthetic tuning function while varying two major parameters—the number of tuning dimensions D and the tuning width σ. We quantified these experiments using the above statistics (normalized VUS, convergence timescale, and activation ratio between 50D and full-space evolution) and tested which variations in these two parameters could reproduce the three systematic changes in the ventral pathway.

More formally, the neuronal Gaussian tuning function was parametrized by a center z_0, the bandwidth σ^2, and the Hessian matrix H. Viewed from the peak of that Gaussian z_0, this is a radial basis function, with neural activation decreasing along any direction leading away from the peak with a speed depending on σ and H. The bandwidth σ^2 controlled the general tuning width, while the Hessian H controlled the curvature (or tuning width) along different directions.

$$r(z) = \exp\left[-\frac{1}{2\sigma^2}(z - z_0)^T H (z - z_0)\right]$$

$$\frac{\partial^2 r}{\partial z^2}\Big|_{z_0} = -\frac{1}{\sigma^2} H$$

With no loss of generality, we assumed H is a diagonal matrix, since we can always rotate the coordinates of z to an eigen basis of H, diagonalizing it. In previous work, the number of tuned feature dimensions by an IT neuron was hypothesized to be one variable underlying the trade-off between selectivity and tolerance in IT (Zoccolan et al., 2007). In our model, we changed the number of tuned features by controlling the eigen-spectrum (i.e., diagonal values) of H. We made the simplifying assumption that the Hessian matrix had only two different eigenvalues: 1 and $\varepsilon = 10^{-6} \approx 0$. In the eigen space of $\lambda = 1$, the model neuron had a Gaussian bell-shaped tuning curve with tuning width σ along each dimension. In the eigenspace of $\lambda = \varepsilon$, the neuron had effectively flat tuning—it was agnostic to the feature changes along those dimensions. We let the neuron have D tuning dimensions (with eigenvalue 1) and 4,096D non-tuning dimensions (with eigenvalue ε), then the Hessian matrix is $H = \mathrm{diag}([1, 1, \ldots 1, \varepsilon, \varepsilon, \ldots, \varepsilon])$. In this experiment, we manipulated the number of tuned dimensions D, and the general tuning width σ. We illustrated the effect of D and σ in two dimensions in Figure 6A. The center z_0 of the tuning function was chosen isotropically on the sphere of radius R, by sampling a 4,096D random vector from normal distribution and normalizing its norm $z_0 = R$.

Evolution speed

We hypothesized that the tuning width σ would affect the success rate and convergence time of the Evolution experiments. Here, we found that, given the same number of tuned features D, the larger the tuning width σ, the faster the search converged. At the other extreme, an overly small tuning width led to a failed Evolution for a larger tuned dimension ($D = 20 \sim 160$). When the tuning width σ was fixed, the time required to converge also increased as a function of tuned dimension D. These trends made sense: if the tuning function covers a tiny region in the whole image space and has close to zero value everywhere else, then it will be nearly impossible for the evolutionary algorithm to find a slope and to climb the mountain; and if there are more tuned axes D, it should take longer to optimize each. Thus, the first systematic change of increasing search convergence time along the hierarchy could be captured by moderately decreasing the tuning width σ and increasing the number of tuned features D (note that this model did not incorporate noise as in neuronal recordings, so convergence time comparisons were limited).

Manifold tuning width

We hypothesized that the width of Gaussian tuning function σ would affect the tuning width (normalized VUS) measured in the Manifold experiment. These quantities were not the same: the manifold on which we measured the tuning sharpness was not guaranteed to reside in the tuned subspace of the neuron. Second, the peak found by the Evolution experiment was not guaranteed to be the global maximum of the tuning function. Despite that, we found that the manifold result could be reproduced by the difference in σ and, surprisingly, also by the tuned dimension D (Figure 6D). Intuitively, the larger the tuning width σ, the broader tuned it was at the peak found by the Evolution experiments. Moreover, given the same σ, the smaller the number of tuned dimensions D, the broader tuned it was in a Manifold experiment. Intuitively, along those non-tuned axes, the "neuron" showed an infinitely wide tuning curve. Thus, if the subspace covered by the PC2,3 axes mixed up some tuned and some untuned dimensions, it would lead to a broader tuning curve. Notably, the shape of tuning map in this simplified example (Figure 6B up, middle) resembled the shape of tuning maps for some V1 neurons observed *in vivo*.

Effect of reduced dimension

Finally, we hypothesized that the larger the tuned dimensionality D, the larger the detrimental effect of constraining searches to a random 50D subspace. We found we could reproduce this observed effect by changing the number of tuned dimensions D. With a small number of tuned dimensions (e.g., $D = 10$), the two evolutions (4,096D versus 50D) resulted in indistinguishable evolution trajectories, regardless of the tuning width σ (Figures 6B and 6E). This scenario is comparable with what we observed for V1 neurons. However, with a larger D, the difference between 4,096D and 50D evolution emerged. Specifically, with the same

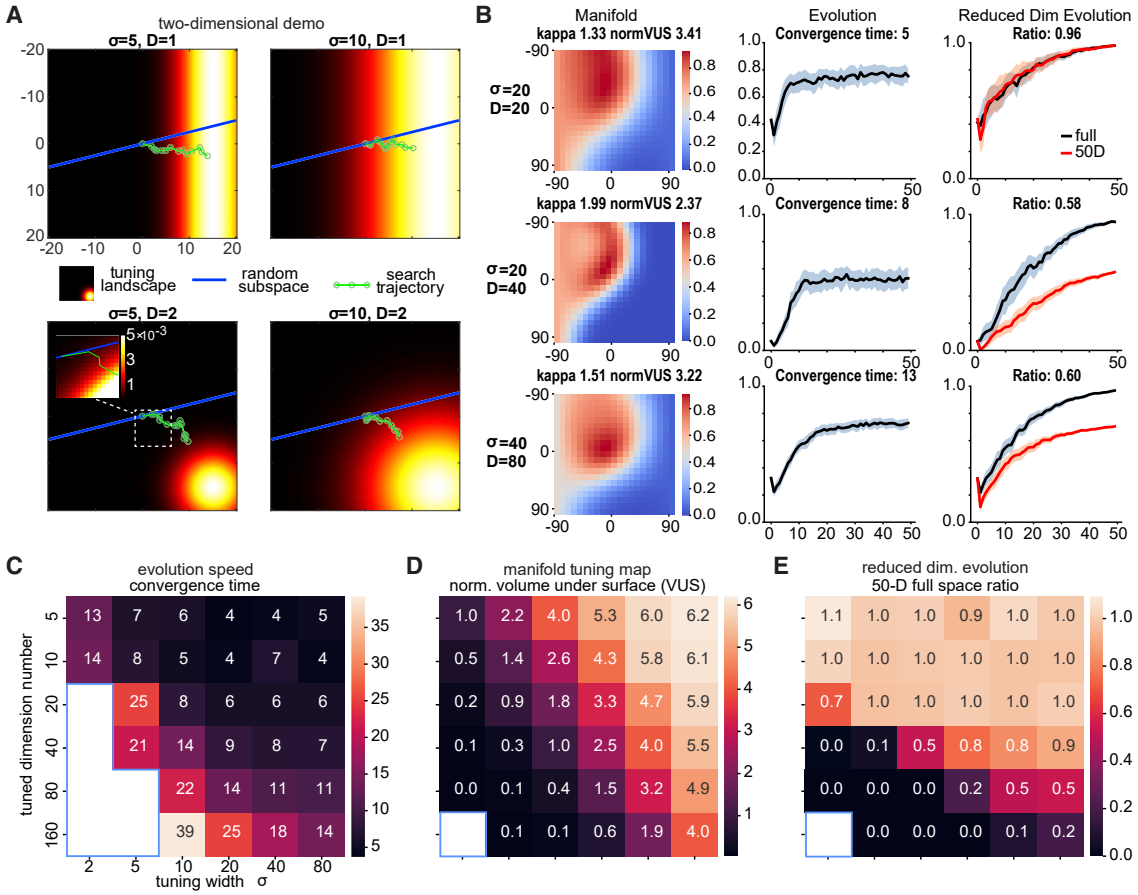


Figure 6. Inferring tuning landscape geometry

(A) 2D demonstration of the main model, showing evolution paths (green) over image space when the activation function has different values for tuning width or dimensionality.

(B) Evolution, Manifold, and reduced-dimensionality experiments in three example conditions ($\sigma = 20, D = 20$; $\sigma = 20, D = 40$; $\sigma = 40, D = 80$). Relevant statistics annotated on top. The shaded area in the Evolution and the Reduced Dim. Evolution panels shows standard deviation of activation values in each generation.

(C–E) The phase space spanned by the tuned dimension D and the tuning width σ. Each panel plots a statistic of the Evolution-Manifold experiment as a function of D and σ, averaged over five repetitions.

(C) Evolution convergence time, in the unit of generation. On this heatmap, the white region defined by blue boundary denotes the conditions where Evolutions failed.

(D) The tuning width of Manifold experiment quantified by normalized volume under surface. The white region outlines conditions where the Manifold tuning map was flat.

(E) The ratio between the final activation of 50D and full-space Evolution search. The white region outlines conditions where both 50D and full-space Evolution failed.

tuning width σ, the performance gap increased with the number of tuned dimensions D; given the same D, the performance gap decreased with the tuning width σ (Figure 6E). Overall, this model provided a simple explanation of the systematic changes we observed *in vivo* and *in silico*. Under the assumptions of this model, there is a likely increase of tuned dimensionality D across stages of the ventral stream, potentially accompanied by a change in tuning width σ. This inference generalizes previous results on the tuned dimensions of IT neurons (Zoccolan et al., 2007).

DISCUSSION

To generalize our understanding of neuronal responses in the natural world, we must characterize tuning using stimuli of appropriate complexity, comprising enough visual attributes to evoke the full response range of given neurons. Driven by this motivation, we have assumed a neurocentric (versus strictly human-interpretable) perspective, asking how neurons respond when the visual world changes smoothly around their preferred stimulus, regardless of the specific transformation. We believe that this brings us one step closer to natural conditions, illustrating how neurons respond to generic smooth transformations before they learn statistical associations between different images—those helpful for invariance (e.g., Li and DiCarlo, 2010). We paired complex-image generators with the classical tuning function, which maps neuronal responses to values of a given parameter space. Tuning functions comprise at least one neuronal response peak and a measure of response decay as

stimuli deviate from this preferred value. This decay is important because, when viewed in the framework of Fisher information, neuronal responses are highly informative about the stimulus regions causing the largest changes in firing rate (Dayan and Abbott, 2005). However, this is complicated in scenarios when neuronal responses reflect information about multiple variables (Benichoux et al., 2017), and in situations of high neuronal variability, when it is the peak that can be most informative (Butts and Goldman, 2006). Consequently, we prioritized identification of response peaks corresponding to the neuron's "preferred" combination of visual attributes, not because we believe that neurons can only be informative when signaling at their maximal response but, because to a downstream decoder, any change in a neuron's response must signify a change from *something*. From views in classical single-neuron studies to those distributed coding frameworks, that *something* is thought to be the visual attributes or parameters that best activate the neuron.

Once we found a location in image space evoking a maximal response, we had to choose how to "move" away from that location. In our data, there seemed to be no special axes for deviating from the peak—all were relatively isotropic, at least within the expected range of neuronal response variability. This is consistent with work applying basis function interpolation as a solution to the problem of categorizing novel views or objects—as described by Edelman, "the shape of the basis function reflects the prior knowledge concerning the change in the output as one moves away from the data point … In the absence of evidence to the contrary, all directions of movement are considered equivalent, making it reasonable to assume that the basis function is radial" (Edelman, 1999). But, since we only measured 2D sections of larger spaces, this may not be true if more directions were probed. Imaginably, there could be special locations linking these peaks, some inducing invariance responses, some inducing sharper activity drop-off. So, ultimately, how does the view of neuronal tuning landscapes relate to invariance? To answer this question, more electrophysiology experiments are required (and underway). Preliminary findings suggest that invariance is not a constant feature of the neuron but depends on the activity evoked by the choice of probe stimuli. While we leave this for future work, these findings highlight the importance of testing the full response dynamics of neurons.

Overall, we conclude that it serves to think of neuronal tuning in multi-attribute feature space the same way as we think about hippocampal place cells acting in a 2D arena, where a neuron's primary responses depend on the proximity to a physical location, less on the direction in which the animal moves to or from it (but see McNaughton et al. [1983] for an effect of entering direction on place field). To push this analogy further, recent works support the view that place cells form a map not just for physical spaces but also for cognitive spaces. Namely, they could encode "locations" in the multidimensional space spanned by task-related variables, such as pitch of auditory cues, accumulated evidence, or even social status of a virtual character (Aronov et al., 2017; Rueckemann and Buffalo, 2017; Tavares et al., 2015). So, the tuning landscape view of visual encoding in the ventral stream connects to the modern view of hippocampus in the cognitive space.

Limitations of the study

Generalization would be broader if more than one generative network were used and if neurons from anterior IT were sampled. If a larger number of natural images could be used as references, the connection between results in the generative network space and natural images would be stronger.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Neuronal recording
  - General animal experiment setup
  - Receptive fields mapping
  - Evolution experiments
  - Manifold experiments
  - Reduced-dimension evolution experiments
  - CNN models of the ventral stream
  - *In silico* evolution-, manifold and reduced-dimension experiments
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Fitting tuning maps with kent function
  - Noise ceiling of explained variance
  - Quantifying tuning width using volume under the surface
  - Quantifying tuning map smoothness by dirichlet energy
  - Quantifying activation increase in *evolution* experiments
  - Comparing tuning across image space via radial tuning curve analysis
  - Convergence speeds of *evolution* experiments
  - Effect of dimensionality restriction on *evolutions*
  - Correlated feature attribution
  - Inclusion criteria for non-driving units
  - Measuring tuning map similarity
  - Naive bayes decoding for population neural activity

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2022.111595.

ACKNOWLEDGMENTS

We thank C. Harvey and C. Pack for comments on the manuscript, J. Freeman and E. Simoncelli for inspiring the title, and T. Holy for the initial motivation and discussion throughout the project. This work was supported by the David and Lucile Packard Foundation (no. 2020-71377), the E. Matilda Ziegler Foundation for the Blind, and the NSF CCF-1231216 Center for Brains, Minds and Machines.

AUTHOR CONTRIBUTIONS

Conceptualization, methodology, software, formal analysis, investigation, data curation, writing – original draft, visualization, B.W.; conceptualization, methodology, validation, investigation, resources, writing – review & editing, visualization, supervision, funding acquisition, C.R.P.

DECLARATION OF INTERESTS

The authors declare no competing interests.

INCLUSION AND DIVERSITY

One or more of the authors of this paper self-identifies as an underrepresented ethnic minority in their field of research or within their geographical location. One or more of the authors of this paper self-identifies as a member of the LGBTQIA+ community. One or more of the authors of this paper received support from a program designed to increase minority representation in their field of research. While citing references scientifically relevant for this work, we also actively worked to promote gender balance in our reference list.

Received: January 31, 2022
Revised: July 20, 2022
Accepted: October 12, 2022
Published: November 8, 2022

REFERENCES

Anselmi, F., Rosasco, L., Tan, C., and Poggio, T. (2015). Deep convolutional networks are hierarchical kernel machines. Preprint at arXiv. https://doi.org/10.48550/arxiv.1508.01084.

Anzai, A., Peng, X., and van Essen, D.C. (2007). Neurons in monkey visual area V2 encode combinations of orientations. Nat. Neurosci. *10*, 1313–1321. https://doi.org/10.1038/nn1975.

Aronov, D., Nevers, R., and Tank, D.W. (2017). Mapping of a non-spatial dimension by the hippocampal–entorhinal circuit. Nature *543*, 719–722, 7647 543. https://doi.org/10.1038/nature21692.

Benichoux, V., Brown, A.D., Anbuhl, K.L., and Tollin, D.J. (2017). Representation of multidimensional stimuli: quantifying the most informative stimulus dimension from neural responses. J. Neurosci. *37*, 7332–7346. https://doi.org/10.1523/JNEUROSCI.0318-17.2017.

Bordelon, B., Paulson, J.A., and Pehlevan, C. (2021). Population codes enable learning from few examples by shaping inductive bias. Preprint at bioRxiv, 2021.03.30.437743. https://doi.org/10.1101/2021.03.30.437743.

Boussaoud, D., Desimone, R., and Ungerleider, L.G. (1991). Visual topography of area TEO in the macaque. J. Comp. Neurol. *306*, 554–575. https://doi.org/10.1002/cne.903060403.

Brock, A., Donahue, J., and Simonyan, K. (2018). Large scale GAN training for high fidelity natural image synthesis. In 7th International Conference on Learning Representations. ICLR 2019 1–35.

Butts, D.A., and Goldman, M.S. (2006). Tuning curves, neuronal variability, and sensory coding. PLoS Biol. *4*, e92–e646. https://doi.org/10.1371/journal.pbio.0040092.

Campbell, F.W., Cleland, B.G., Cooper, G.F., and Enroth-Cugell, C. (1968). The angular selectivity of visual cortical cells to moving gratings. J. Physiol. *198*, 237–250. https://doi.org/10.1113/jphysiol.1968.sp008604.

Carlson, E.T., Rasquinha, R.J., Zhang, K., and Connor, C.E. (2011). A sparse object coding scheme in area V4. Curr. Biol. *21*, 288–293. https://doi.org/10.1016/j.cub.2011.01.013.

Dayan, P., and Abbott, L. (2005). Theoretical Neuroscience, Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems (Computational Neuroscience) (MIT Press).

Desimone, R., Albright, T.D., Gross, C.G., and Bruce, C. (1984). Stimulus-selective properties of inferior temporal neurons in the macaque. J. Neurosci. *4*, 2051–2062. https://doi.org/10.1523/JNEUROSCI.04-08-02051.1984.

Dosovitskiy, A., and Brox, T. (2016). Generating images with perceptual similarity metrics based on deep networks. In Advances in Neural Information Processing Systems (NIPS).

Dragoi, V., Sharma, J., and Sur, M. (2000). Adaptation-induced plasticity of orientation tuning in adult visual cortex. Neuron *28*, 287–298. https://doi.org/10.1016/S0896-6273(00)00103-3.

Edelman, S. (1999). Representation and Recognition in Vision (The MIT Press). https://doi.org/10.7551/mitpress/5890.001.0001.

Enroth-Cugell, C., and Robson, J.G. (1966). The contrast sensitivity of retinal ganglion cells of the cat. J. Physiol. *187*, 517–552.

Gattass, R., Gross, C.G., and Sandell, J.H. (1981). Visual topography of V2 in the macaque. J. Comp. Neurol. *201*, 519–539. https://doi.org/10.1002/cne.902010405.

Gattass, R., Sousa, A.P., and Gross, C.G. (1988). Visuotopic organization and extent of V3 and V4 of the macaque. J. Neurosci. *8*, 1831–1845.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Identity Mappings in Deep Residual Networks.

Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K.Q. (2016). Densely connected convolutional networks. In Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017 2017-Janua, pp. 2261–2269.

Hubel, D.H., and Livingstone, M.S. (1987). Segregation of form, color, and stereopsis in primate area 18. J. Neurosci. *7*, 3378–3415.

Hwang, J., Mitz, A.R., and Murray, E.A. (2019). NIMH MonkeyLogic: behavioral control and data acquisition in MATLAB. J. Neurosci. Methods *323*, 13–21. https://doi.org/10.1016/j.jneumeth.2019.05.002.

Kay, K.N., and Yeatman, J.D. (2017). Bottom-up and top-down computations in word- and face-selective cortex. Elife *6*, e22341. https://doi.org/10.7554/ELIFE.22341.

Kilcher, Y., Lucchi, A., and Hofmann, T. (2017). Semantic interpolation in implicit models. In 6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings. https://doi.org/10.48550/arxiv.1710.11381.

Kobatake, E., and Tanaka, K. (1994). Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. J. Neurophysiol. *71*, 856–867. https://doi.org/10.1152/JN.1994.71.3.856.

Krizhevsky, A., Sutskever, I., and Hinton, G.E. (2012). ImageNet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst., 1097–1105.

Kubilius, J., Schrimpf, M., Nayebi, A., Bear, D., Yamins, D., and DiCarlo, J. (2018). CORnet: modeling the neural mechanisms of core object recognition. Preprint at bioRxiv, 408385. https://doi.org/10.1101/408385.

Lee, D.D., and Seung, H.S. (1999). Learning the parts of objects by non-negative matrix factorization. Nature *401*, 788–791. https://doi.org/10.1038/44565.

Lee, D.D., and Seung, H.S. (2001). Algorithms for Non-negative Matrix Factorization.

Leopold, D.A., O'Toole, A.J., Vetter, T., and Blanz, V. (2001). Prototype-referenced shape encoding revealed by high-level aftereffects. Nat. Neurosci. *4*, 89–94, 1 4. https://doi.org/10.1038/82947.

Li, N., and DiCarlo, J.J. (2010). Unsupervised natural visual experience rapidly reshapes size-invariant object representation in inferior temporal cortex. Neuron *67*, 1062–1075. https://doi.org/10.1016/J.NEURON.2010.08.029.

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C.L. (2014). Microsoft COCO: common objects in context. In Lecture Notes in Computer Science (Springer Verlag), pp. 740–755. https://doi.org/10.1007/978-3-319-10602-1_48.

Logothetis, N.K., Pauls, J., and Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. Curr. Biol. *5*, 552–563.

Loshchilov, I. (2017). LM-CMA: an alternative to L-BFGS for large-scale black box optimization. Evol. Comput. *25*, 143–171. https://doi.org/10.1162/EVCO_a_00168.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. In 6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings.

Maruyama, M., Girosi, F., and Poggio, T. (1992). A Connection between GRBF and MLP.

Maunsell, J.H.R., and van Essen, D.C. (1983). Functional properties of neurons in middle temporal visual area of the macaque monkey. I. Selectivity for stimulus direction, speed, and orientation. https://doi.org/10.1152/jn.1983.49.5.1127 49. https://doi.org/10.1152/JN.1983.49.5.1127. 1127–1147.

McNaughton, B.L., Barnes, C.A., and O'Keefe, J. (1983). The contributions of position, direction, and velocity to single unit activity in the hippocampus of freely-moving rats. Exp. Brain Res. *52*, 41–49. https://doi.org/10.1007/BF00237147.

Munkres, J. (2000). Topology, 2nd Edition (Pearson).

Pasupathy, A., and Connor, C.E. (1999). Responses to contour features in macaque area V4. J. Neurophysiol. *82*, 2490–2502.

Pasupathy, Anitha, and Connor, Charles E. (2002). Population coding of shape in area V4. Nature neuroscience *5*, 1332–1338. https://doi.org/10.1038/972.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury Google, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). PyTorch: an imperative style, high-performance deep learning library. Adv. Neural Inf. Process. Syst.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: machine learning in {P}ython. J. Mach. Learn. Res. *12*, 2825–2830.

Poggio, T., and Girosi, F. (1990a). Networks for approximation and learning. Proc. IEEE *78*, 1481–1497. https://doi.org/10.1109/5.58326.

Poggio, T., and Girosi, F. (1990b). Regularization algorithms for learning that are equivalent to multilayer networks. Science *247*, 978–982, (1979). https://doi.org/10.1126/SCIENCE.247.4945.978.

Ponce, C.R., Xiao, W., Schade, P.F., Hartmann, T.S., Kreiman, G., and Livingstone, M.S. (2019). Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences. Cell *177*, 999–1009.e10. https://doi.org/10.1016/j.cell.2019.04.005.

Portilla, J., and Simoncelli, E.P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. Int. J. Comput. Vis. *40*, 49–70. https://doi.org/10.1023/A:1026553619983.

Rafaely, B. (2019). Sampling the sphere. In Fundamentals of Spherical Array Processing (Springer Topics in Signal Processing. Springer), pp. 59–80. https://doi.org/10.1007/978-3-319-99561-8_3.

Rasmussen, C.E., and Williams, C.K.I. (2006). Gaussian Processes for Machine Learning (MIT Press).

Rish, I., and Rish, I. (2001). An Empirical Study of the Naive Bayes Classifier.

Rolls, E.T., and Tovee, M.J. (1995). Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex. J. Neurophysiol. *73*, 713–726. https://doi.org/10.1152/JN.1995.73.2.713.

Rosch, E.H. (1973). Natural categories. Cogn. Psychol. *4*, 328–350. https://doi.org/10.1016/0010-0285(73)90017-0.

Rose, O., Johnson, J., Wang, B., and Ponce, C.R. (2021). Visual prototypes in the ventral stream are attuned to complexity and gaze behavior. Nat. Commun. *12*, 6723, 1 12. https://doi.org/10.1038/s41467-021-27027-8.

Rueckemann, J.W., and Buffalo, E.A. (2017). Neuroscience:Auditory landscape on the cognitive map. Nature *543*, 631–632, 7647 543. https://doi.org/10.1038/543631a.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. *115*, 211–252. https://doi.org/10.1007/s11263-015-0816-y.

Rust, N.C., and DiCarlo, J.J. (2010). Selectivity and tolerance ("invariance") both increase as visual information propagates from cortical area V4 to IT. J. Neurosci. *30*, 12978–12995. https://doi.org/10.1523/JNEUROSCI.0179-10.2010.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N.J., Rajalingham, R., Issa, E.B., Kar, K., Bashivan, P., Prescott-Roy, J., Schmidt, K., et al. (2018). Brain-score: which artificial neural network for object recognition is most brain-like?. Preprint at bioRxiv, 407007. https://doi.org/10.1101/407007.

Seung, H.S., and Sompolinsky, H. (1993). Simple models for reading neuronal population codes. Proc. Natl. Acad. Sci. USA *90*, 10749–10753, –10753. https://doi.org/10.1073/pnas.90.22.10749.

Simonyan, K., and Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition.

Tanaka, K. (2003). Columns for complex visual object features in the inferotemporal cortex: clustering of cells with similar but slightly different stimulus selectivities. Cereb. Cortex *13*, 90–99. https://doi.org/10.1093/CERCOR/13.1.90.

Tavares, R.M., Mendelsohn, A., Grossman, Y., Williams, C.H., Shapiro, M., Trope, Y., and Schiller, D. (2015). A map for social navigation in the human brain. Neuron *87*, 231–243. https://doi.org/10.1016/J.NEURON.2015.06.011.

Tsao, D.Y., Freiwald, W.A., Tootell, R.B.H., and Livingstone, M.S. (2006). A cortical region consisting entirely of face-selective cells. Science *311*, 670–674. https://doi.org/10.1126/science.1119983.

Wang, B., and Ponce, C.R. (2021). The geometry of deep generative image models and its applications. In International Conference on Learning Representations.

Wang, B., and Ponce, C.R. (2022). High-performance evolutionary algorithms for online neuron control. In Genetic and Evolutionary Computation Conference. https://doi.org/10.1145/3512290.3528725.

White, T. (2016). Sampling Generative Networks. https://doi.org/10.48550/arxiv.1609.04468.

Wikipedia Editors, n.d. Functional correlation - wikipedia [WWW Document]. URL. (accessed 8.28.21). https://en.wikipedia.org/wiki/Functional_correlation#Correlation_as_angle_between_functions.

Yau, J.M., Pasupathy, A., Brincat, S.L., and Connor, C.E. (2013). Curvature processing dynamics in macaque area V4. Cereb. Cortex *23*, 198–209. https://doi.org/10.1093/cercor/bhs004.

Zhang, R., Isola, P., Efros, A.A., Shechtman, E., and Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 586–595.

Zoccolan, D., Kouh, M., Poggio, T., and DiCarlo, J.J. (2007). Trade-off between object selectivity and tolerance in monkey inferotemporal cortex. J. Neurosci. *27*, 12292–12307. https://doi.org/10.1523/JNEUROSCI.1897-07.2007.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Electrophysiological data (formatted mat file)	This paper	https://doi.org/10.17605/OSF.IO/GPZM5
Image stimuli (natural images, contours and the images synthesized from the generative adversarial network)	This paper; (Pasupathy and Connor, 2002)	https://doi.org/10.17605/OSF.IO/GPZM5 https://depts.washington.edu/shapelab/resources/stimsonly.php
Experimental models: Organisms/Strains		
Rhesus Macaques	Florida facility of PrimGen/PreLabs	Macaca mulatta
Software and algorithms		
MATLAB	Mathworks	https://www.mathworks.com
Python	Python Software Foundation	https://www.python.org/
Pytorch	(Paszke et al., 2019)	https://pytorch.org/
Scikit-learn	(Pedregosa et al., 2011)	https://scikit-learn.org/stable/
Code for analyzing *in vivo* electrophysiological data, calculating statistics and generating figures	This paper	https://github.com/PonceLab/Tuning-Manifold-Charting
Code for conducting *in silico* experiments on neural network and analyzing the results	This paper	https://github.com/PonceLab/Tuning-Manifold-Charting-in-silico

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Carlos R. Ponce (carlos_ponce@hms.harvard.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Formatted neuronal spiking data, natural image stimuli, and the images synthesized from the generative adversarial network have been deposited in Open-Science Framework (OSF) repository and are publicly available as of the date of publication. These data are enough for reproducing results in the publication. The DOI is listed in the key resources table.
- All original code for neurophysiological data analysis has been deposited in a GitHub repository (https://github.com/PonceLab/Tuning-Manifold-Charting), and all original code for conducting and analyzing *in silico* experiments has been deposited in the GitHub repository (https://github.com/PonceLab/Tuning-Manifold-Charting-in-silico). Both are publicly available as of the date of publication. We also made self-contained Jupyter notebooks that allows online exploration of our dataset. DOIs are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this work is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Two male adult rhesus macaques (A and B, *Macaca mulatta*, ages 6–7, 10–11 kg) were used in the study. All procedures conform to the Guide for the Care and Use of Laboratory Animals and were approved by the Institutional Animal Care and Use Committee at Washington University and Harvard Medical School.

METHOD DETAILS

Neuronal recording

Two male adult rhesus macaques (A and B) were implanted with chronic floating microelectrode arrays (Microprobes for Life Sciences, MD) in the right hemisphere: one array was located at the posterior lip of the lunate sulcus (corresponding to the V1/V2 transition), one on the prelunate gyrus (V4) and another anterior to the inferior occipital sulcus (PIT). We refer to these sites as V1/V2, V4 and IT (Boussaoud et al., 1991; Gattass et al., 1981, 1988). The locations of the arrays were chosen based on sulcal landmarks, around local vasculature.

Neurophysiology data was acquired using OmniPlex Neural Recording Data Acquisition Systems (Plexon Inc.), with the PlexControl client to sort electrical events online based on waveform and interspike intervals. Because all *Evolution* experiments were based on a closed loop between neuronal activity and image synthesis, spike sorting was done at the beginning of each experiment. Within each channel, events were estimated as arising from single-units, multi-units or "hash" using a 1-5 scale, where 1 indicated strong confidence on the presence of a single-unit (based on waveform shape, inter-spike interval and separation from the main hash signal) and 5 indicated hash/multiunit activity. We use the term *site* to refer to all signal types; across experiments, sites comprised mostly multiunits/hash and a fraction of single units.

After data collection, spike/event times were discretized into 1-ms bins and convolved with a symmetric Gaussian probability density function with a 2-ms standard deviation.

General animal experiment setup

Experimental sessions were run using MonkeyLogic2 (Hwang et al., 2019), which directed the presentation of visual stimuli on ViewPixx EEG monitors (ViewPixx Technologies). Refresh rate was 120 Hz at a resolution of 1920x1080 pixels. The monkeys were placed 58 cm from the screen. Gaze position was tracked via ISCAN cameras (ISCAN Inc.). Animals fixated 0.25°-diameter circles, with fixation window that permitted eye movements up to 1.0–1.3° from the fixation point during stimulus presentation; they obtained reward if they held their gaze on the target for 2-3 s. Rewards delivered via DARIS Control Module System (Crist Instruments).

Receptive fields mapping

At the start of each experimental session, receptive field locations were estimated as follows. While the animal performed a fixation task, at each moment, a single square test image was presented at a single position for 100-ms. The image could be a photograph or a previously collected generator image and on any given presentation, it could be sized either at 1°-, 2°- or 4°-width. Positions were randomly sampled from grids ranging from [-2°–2°], [-4°–4°] or [-8°–8°] of the central visual field, in steps of 1°, 2°or 4° for the three sizes. After data collection, neuronal responses (events/second) were quantified as a function of stimuli location. A 2-D Gaussian function was fit to this 2-D response grid to estimate the center of the receptive field. Because the test image width was so large and because we did not sample a dense enough position grid, we did not obtain a strict estimate of RF size (particularly for V1/V2 and V4 sites) and our data over-estimates it. The receptive-field center distribution of all neuronal sites from V1/V2, V4 and IT visual areas are shown in Figure S1A. After estimating RF center location, *Evolution* and *Manifold* experiments were carried out with stimuli at the estimated center location, sized to cover the region with most activity. Usually, stimuli were made larger than the estimated RF size. This helped to prevent the site from responding to the high-contrast, salient image edge. Moreover, larger image sizes provided a canvas to engage both the classical RF and its surround during optimization of the neuronal response.

Evolution experiments

After finding the optimal location the evolving textures, the experiment started by presenting 30 synthetic images and 10 reference images, only one presentation per image. Reference images were selected if they were known to evoke high activity from the array site under study, per independent experiments. The initial 30 synthetic images were approximations of Portilla and Simoncelli textures (Portilla and Simoncelli, 2000) recreated in the generator. After all images were presented, their input vectors and the site's spike rate responses (averaged over 50-200 ms after image onset) were provided to the update function of Cholesky CMA-ES algorithm, which then gave 40 new vectors as outputs. These vectors were provided to the generator to create 40 new images, and the cycle began again, i.e., the 40 synthetic images and the same 10 reference images were presented to the subject. Each experiment comprised tens of cycles (*generations*) and were stopped 10-20 generations after firing rate convergence was observed.

Manifold experiments

After evolving a preferred stimulus (a *prototype*), the goal was to measure the tuning landscape around it, by sampling images in a smooth, continuous fashion. By examining the distribution of latent vectors during the *Evolution* experiments and applying principal component analysis, we found a properly scaled PC1 recreated the measured prototype — PC1 vector re-created an image like the evolved image in the final generation, once scaled to the norm and sign of the observed prototype latent vector. Image contrast could be manipulated via positive scaling of the latent vector (illustrated in Figure B.1 A of Wang and Ponce, 2022). Settling on PC1 $\boldsymbol{v}_1$ of the trajectory to represent the prototype, the next two vectors orthogonal to PC1 were used to form a basis of three vectors. A two-dimensional sphere was defined in the subspace spanned by these three vectors, and images were sampled along the azimuth

and elevation angle θ, φ uniformly. Samples were along θ, φ in $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$ with $\frac{\pi}{10}$ interval, chosen to make the transition between neighboring images perceptually smooth. As a result, for each manifold experiment, there were $121 = 11 \times 11$ points sampled on the hemisphere around the prototype.

$$\boldsymbol{z}(\theta, \varphi) = [\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3] \cdot [\cos\theta \cos\varphi, \sin\theta \cos\varphi, \sin\varphi]^T$$

$$\theta \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right], \ \varphi \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$$

Note these points did not form a uniform grid on the sphere. The distortion by mapping a 2d grid onto a spherical surface was greater near the "north pole" and "south pole" of the sphere; for example, the 11 images at either $\varphi = \frac{\pi}{2}$ or $-\frac{\pi}{2}$ were exactly the same, resulting in 101 unique images in each hemisphere. Thus, we correct for this distortion effect in the subsequent computation of tuning functions and statistics of tuning maps.

We performed 46 experiment sessions on monkey A, and 45 experiment sessions on monkey B. As we mentioned above, in 10 of the monkey B experiments, we performed *Manifold* experiments in three subspaces following the *Evolution* experiment, using PC2-3, PC49-50, and two random vectors as the perturbation vectors $\boldsymbol{v}_2, \boldsymbol{v}_3$ in the *Manifold* experiments. Thus, we had $N = 91$ standard experimental pairs with an *Evolution* experiment and a *Manifold* experiment in PC2-3 space. Initial analysis found that the tuning maps in the PC49-50 and random subspaces were comparable to those in PC2-3 space. So, when characterizing the tuning maps, we pooled the additional maps (PC49-50, Random) with the PC2-3 maps, resulting in $N = 111$ tuning maps of neuronal sites driving the *Evolution*.

Reduced-dimension evolution experiments

To probe the global geometry of the tuning landscape, and to investigate how it changed from posterior to anterior visual cortex, we performed *reduced-dimension Evolution* experiments. In each session, after receptive field mapping, we performed paired *Evolution* experiments, using the same neuronal unit as the driving unit in two parallel *Evolution* experiments, using independent optimizers. We called each *Evolution* a thread. The stimuli created by the two *Evolution* threads were presented in an interleaved fashion. One thread was the *reduced-dimension Evolution*, with the optimizer constrained to search in a randomly selected 50-dimensional subspace; the other thread was the full-space *Evolution*, searching in the full 4096-dimensional space. The 50-dimensional subspace was selected independently for each session. We developed a new optimizer that operated on the hypersphere of arbitrary dimension *SphereCMA* (Wang and Ponce, 2022) for these experiments. This ensured that the search codes would have the same norm throughout the *Evolution*. Without this constraint, using the original Cholesky CMA-ES, the vector norm of the codes differed greatly in the two threads, which made it difficult to compare the two threads. The same optimizer was applied to each *Evolution* thread, with the same population size ($N = 31$) but different dimensionalities (4096D vs 50D). The same 10 reference images were used for both threads. We performed 34 sessions on each monkey (A and B), within which 23 sessions were driven by V1 neurons, 20 by V4 neurons, and 25 by IT neurons.

CNN models of the ventral stream

In the past few years, many CNN models have been developed to solve the object recognition problem. Though generally, these models incorporate many working principles of the ventral stream, some of them are better models than others. To find the CNN models that resemble the population representations in the primate ventral stream the most, we consulted the BrainScore (Schrimpf et al., 2018) (http://www.brain-score.org/) leaderboard and selected a few top and classic networks: AlexNet (Krizhevsky et al., 2012) (No. 57), VGG16 (Simonyan and Zisserman, 2014) (No.5), ResNet50 (He et al., 2016) (No.11), ResNet50-Robust (Madry et al., 2017) (No.3), ResNet101 (He et al., 2016) (No.4), DenseNet-169 (Huang et al., 2016) (No.9), CorNet-S (Kubilius et al., 2018) (No. 1). Source and specification of the models are as follows, all implemented in PyTorch (Paszke et al., 2019). ResNet50-Robust weights were obtained from https://github.com/MadryLab/robustness, $\varepsilon = 8/255$ version. CorNet-s model definition and weights from https://github.com/dicarlolab/CORnet. All other models' definitions and weights were obtained from the model zoo in *torchvision* https://pytorch.org/vision/stable/models.html. All seven models were used to analyze tuning sharpness progression along the visual hierarchy (Figures 4D-4E, S6A–S6F). AlexNet, VGG16, ResNet50).

In silico evolution-, manifold and reduced-dimension experiments

To corroborate our findings in the ventral stream hierarchy *in vivo*, we performed parallel experiments for CNN. For each network, we picked the major convolutional and fully connected layers ($n = 8 - 16$). For convolutional layers, we selected the units in the center of the feature map of each channel, mirroring the process of us presenting the image at the receptive field (RF) of recorded neurons.

For each unit, we first measured its receptive field by backpropagating its activation back to the image, where we defined a square box around the pixels that contributed to its activation. This "receptive field" was in turn used to resize the images. Next, we used the activity of this unit to drive the *Evolution* experiment, using the same CMA-ES algorithm and generator. Finally, we used the same

method to analyze the trajectory and to sample images on 2d hemispheres around the prototype as we did in the *Manifold* experiments. We measured the activations of the recorded units responding to the manifold image set, thus obtaining the tuning maps of these CNN units.

For the reduced-dimension *Evolution*, we replicated these experiments using units from the eight layers of CaffeNet and AlexNet, with 100 units selected from the feature map center of the first 100 channels of each layer (except for the first layer, conv1, which has fewer channels). For each unit, we performed a reduced-dimension *Evolution* in a 50, 100, 200, and 400-dimensional random subspace and the 4096d full space. For each dimensionality for each unit, the *Evolution* was repeated 10 times with independently sampled random subspace. We used the same optimization algorithm (SphereCMA) *in vivo* and *in silico*. The score trajectories were analyzed in the same fashion *in vivo* and *in silico*.

QUANTIFICATION AND STATISTICAL ANALYSIS

Fitting tuning maps with kent function

By visual inspection, neuronal tuning functions were usually unimodal and likely to be well-fitted with a Gaussian function, but since the domain of the sampled tuning function was not a flat Euclidean space, it was not strictly correct to fit these responses using a Gaussian function. Thus, we used the Kent function, which is a natural analogue to the Gaussian function on the 2-D sphere S^2:

$$f(\boldsymbol{x}) = A\exp\left\{\kappa\,\boldsymbol{\gamma}_1^T\cdot\boldsymbol{x} + \beta\left[\left(\boldsymbol{\gamma}_2^T\cdot\boldsymbol{x}\right)^2 - \left(\boldsymbol{\gamma}_3^T\cdot\boldsymbol{x}\right)^2\right]\right\} + b$$

where $\boldsymbol{x}$ is a 3D vector, which, in our case, is the 3D coordinate of the latent vector on the PC basis of the *Manifold* experiments. $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \boldsymbol{\gamma}_3$ form an orthonormal basis set in 3D space, in which $\boldsymbol{\gamma}_1$ is the direction of the center of the distribution; and $\boldsymbol{\gamma}_2, \boldsymbol{\gamma}_3$ are analogues to the maximal and minimal covariance axes, if the peak is anisotropic. We parametrized the 3D orthonormal basis set using the following angle convention: θ, φ represent the azimuth and elevation angles corresponding to $\boldsymbol{\gamma}_1$ vector, while ψ is the angle of $\boldsymbol{\gamma}_2$ to the equator of the sphere. Thus, when the function is unimodal (single peak), θ, φ parametrize the location of the peak on the sphere, and ψ encodes the direction of elongation around the peak. Beyond these, A, b control the scaling and baseline, κ controls the concentration ("peaked-ness") of the function, β controls the degree of anisotropy around the peak. Overall, this function comprises 7 parameters $A, b, \kappa, \beta, \theta, \varphi, \psi$. We fit the Kent function to the tuning maps from *in silico* experiments or *in vivo* experiments using 'curve_fit' from 'SciPy' in 'Python' or 'fit' function in MATLAB.

Noise ceiling of explained variance

To control trial-to-trial variability in tuning map fitting, we computed the noise ceiling of explained variance R^2 as follows. For each *Manifold* session, we resampled the single trial neuronal responses to each image to get a bootstrapped mean tuning map $r'[i]$. Then we computed the explained variance of the original mean tuning map $r[i]$ by the bootstrapped one $r'[i]$.

$$R_{bstrp} = 1 - \frac{\sum_i (r[i] - r'[i])^2}{\sum_i (r[i] - \bar{r})^2}$$

This procedure was replicated 500 times to estimate the average $\overline{R}_{bstrp}$, which we regarded as the ceiling of explainable variance when fitting the noisy tuning map.

Quantifying tuning width using volume under the surface

We generalized the idea of *area under the curve* (AUC) to *the volume under the surface* (VUS) for the 2-D tuning map in our scenario. The tuning map was defined on a hemisphere $A[\theta, \varphi]$, with $\theta \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$, $\varphi \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$; we integrated the evoked firing rate over the hemisphere and normalized its peak firing rate to 1. This statistic has a theoretical maximum of $2\pi \approx 6.28$ which is obtained when the neuron responds equally and above baseline over the whole hemisphere; it also has a theoretical minimum of 0, where no place on the hemisphere (i.e., no image) evoked activity above baseline. The value of VUS can be interpreted as the equivalent area on the hemisphere that evoked the maximal firing rate in this space:

$$VUS = \frac{1}{\overline{A} - b}\int d\theta d\varphi \max(A[\theta, \varphi] - b, 0),\ \overline{A} = \max A[\theta, \varphi]$$

The baseline b is the mean firing rate in the [0,50]-ms period after stimulus onset, across all trials in the experiment. $A[\theta, \varphi]$ is computed as the mean firing rate in [51,200] milliseconds. Numerical integration on the sphere was performed using the method in Rafaely (2019) (Rafaely, 2019).

Quantifying tuning map smoothness by dirichlet energy

To quantify the smoothness of a tuning map, we adapted a functional from mathematics, the Dirichlet energy of a function. This functional integrates the squared norm of gradient in the domain of the function, which, in our case, is a spherical manifold. Here we used finite difference estimation of gradient and discrete quadrature over sphere for integration. The distortion effect of the non-uniform sampling on sphere is accounted for in this calculation.

$$DE = \int_M \|\nabla f(x)\|^2 dx$$

To estimate how trial-to-trial variability affected *DE*, we performed 1000 trials resampling for each image, thus we obtained a distribution of average tuning maps and a distribution of Dirichlet energy. To estimate the null distribution, we shuffled the correspondence of the response and the position on the image manifold and computed the Dirichlet energy again (Figure S1B). We used the *t* statistic and Cohen's *d′* of these two distributions to quantify the smoothness of a tuning map. Recall that Cohen's *d′* is the difference between the mean of two distributions measured in the unit of combined standard deviation. So, a wider null distribution or a wider trial resampled distribution will make *d′* smaller, while a larger separation between the Dirichlet energy of the shuffled map and the resampled map will create a larger smoothness measure (Figure S1E).

Similarly, total variation energy is defined as follows, i.e., integrating the norm of the gradient (without squaring) on the manifold.

$$TVE = \int_M \|\nabla f(x)\| dx$$

We replicated our analysis of smoothness using this TVE measure, with equivalent results: 94/110 experiment with smaller total variation energy than shuffled control ($P < 0.001$), range of $T \in [-277.6, -5.8]$, Cohen's $d' \in [-12.41, -0.26]$.

Quantifying activation increase in *evolution* experiments

We wanted to design a statistic to quantify the activation increase during *Evolution* experiments. Inspired by delta fluorescence over fluorescence (DFOF) in two-photon imaging, we used the difference of activation over initial activation (DAOA) to quantify the relative increase.

$$DAOA = \frac{\bar{r}_{end,50:200} - \bar{r}_{init,50:200}}{\bar{r}_{init,50:200}}$$

This statistic was key to connect the *Evolution* experiment to Manifold tuning maps (Figure S2E). We used the κ in Kent fitting to quantify sharpness of tuning maps. In all experiments where the tuning maps were well-fit by a Kent function ($R^2 > 0.5$), the Spearman correlation between DAOA and κ was 0.609 ($P = 1.0 \times 10^{-8}$, $N = 76$). One possibility was that this correlation was mediated by the areal difference of tuning width: we found that neuronal sites in higher visual cortex had sharper tuning maps and larger activation increase in *Evolution* experiments. We tested this by computing the correlation for neuronal sites in three cortical areas separately, and we found these correlation values were also statistically significant: V1 ($\rho = 0.564, P = 2.6 \times 10^{-3}, N = 27$), V4 ($\rho = 0.492, P = 0.029, N = 20$), IT ($\rho = 0.586, P = 1.0 \times 10^{-3}, N = 29$). Pictorially, the neuron-guided *Evolution* that "climbed" to a taller peak ($\bar{r}_{end,50:200}$) on the tuning landscape or started at a lower level ($\bar{r}_{init,50:200}$), tended to reach a sharper peak measured by the *Manifold* experiment. This relationship was consistent with the picture of the *Evolution* experiment allowing neuronal searches for a peak on the tuning landscape, and the *Manifold* experiment characterizing the tuning around the peak.

Comparing tuning across image space via radial tuning curve analysis

In this experiment, we measured site responses to images to different image sets: Manifold, gratings, curved objects, and photographs, resulting in image-response pairs $\{I_i, \bar{r}_i\}$ for each image set. We first computed the image distance matrices between all pairs of images using the LPIPS distance *D* (Zhang et al., 2018)

$$d[i,j] = D(I_i, I_j)$$

Then, we found the image evoking the highest response in the neuron, i.e., the tuning peak in that image space.

$$k = \arg\max_i \bar{r}_i,\ r_{MAX} = \max_i \bar{r}_i$$

Finally, we fit the neuronal response as a function $\hat{f}$ of the image distance to the highest activating image I_k, this is the *radial tuning function*. Specifically, we used a non-parametric fitting method, Gaussian process regression (MATLAB function *fitrgp.m*).

$$\hat{f} = \text{fitrgp}(d[k,:], r[:])$$

To compute the area under the curve, we integrated the area under the estimated function; the definite integral is performed from 0 to the maximal distance from the peak image D_{MAX} (Figure 3A right). Normalized AUC was computed by dividing the area under the curve by the peak activation.

$$AUC = \int_0^{D_{MAX}} \hat{f}(x) dx,\ D_{MAX} = \max_i d[k,i]$$

$$normAUC = AUC/r_{MAX}$$

As the distance matrix was computed between all image pairs, the *normAUC* statistic quantified how fast the response decreased from a peak on the tuning landscape, averaging the different directions of deviating from the peak.

Convergence speeds of *evolution* experiments

For each *Evolution* experiment, we computed the response firing rate in the evoked time window [50, 200] ms for every image. Then we used Gaussian process regression (Rasmussen and Williams, 2006) (*fitrgp.m* in MATLAB) to obtain the smoothed and averaged optimization trajectory $\hat{f}(t)$. Next, we computed the generation number C_{63} when 63.2% of maximal activation r_{max}, (i.e., $0.632(r_{max} - r_{init}) + r_{init}$) was reached, quantifying the timescale of convergence per *Evolution* experiment.

Effect of dimensionality restriction on *evolutions*

For the reduced-dimension *Evolution* experiments, we measured the effect of constraining search dimensionality on the activation increase for units in all three cortical areas (V1/V2, V4, and IT). We defined a ratio R between the activation increases in 4096D and 50D as follows:

$$R = \frac{\overline{r_{50d,\,end}} - \overline{r_{50d,init}}}{\overline{r_{4096d,\,end}} - \overline{r_{4096d,init}}}$$

$\overline{r_{50d,init}}$ and $\overline{r_{50d,\,end}}$ are the average firing rate for all images in the first and the last generation in the 50D reduced-dimension *Evolution*; $\overline{r_{4096d,init}}$ and $\overline{r_{4096d,\,end}}$ are those average firing rates for the full space *Evolution*.

To reduce the noise in the single-trial neural responses, we also used the integrated d' to quantify the difference: we calculated d' between the set of single trial firing rates in each generation of 4096D vs 50D *Evolution* (i in the following equation), and then averaged the d' over all generations.

$$\overline{d'} = \frac{1}{N}\sum_{i}^{N} d'(r_{50d,i}[:], r_{4096d,i}[:])$$

Correlated feature attribution

The goal of this model was to localize the visual attributes — shapes, colors, and textures — selected by neurons during the *Evolution* experiments, within the whole evolved image. While this can be approximated by superimposing the independently measured receptive fields of the neurons, this analysis presents an alternative that works using only the *Evolution* data itself. We encourage the readers to read our paper dedicated to this method (Wang and Ponce, 2022) and our code (https://github.com/Animadversio/Neuronal_Feature_Attribution_Model). Briefly, our tactic was to rely on CNN feature (hidden) units sharing similar response properties as the recorded neuron. First, we picked a convolutional layer in a pretrained CNN (e.g., AlexNet, VGG-16, ResNet-50, ResNet-50-robust) and computed the correlation and covariance of each unit with the observed (V1/V2, V4 or IT) neuronal responses across all *Evolution* images $\{I_i^e\}$. The correlation and covariance were both tensors with the same shape as the activation tensor of a layer to a single image $F[c,x,y]$. Using Python convention of indexing, the tensors were

$$C[c,x,y] = corr(r[:], F[:,c,x,y])$$

$$Q[c,x,y] = cov(r[:], F[:,c,x,y])$$

Note that thousands of images were displayed in the *Evolution* experiments. Given the hundreds of thousands of features in the convolutional layer, the subsequent memory cost required us to compute correlation values via online updates. We built a custom pipeline to compute the correlation and covariance value for each feature unit when propagating images through the neural network, in a batch-update fashion. Next, we selected the most correlated feature units by thresholding the t statistics associated with correlation value $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$. We set the threshold as $t \geq 3$ for the presented results, although we varied this threshold to test the robustness with no changes in overall conclusions. The units with correlation values less than this threshold were excluded by setting their weight to zero. Because CNNs usually exhibit boundary artifacts, we also excluded the feature units at the border of feature maps. The original covariance tensor $Q[c,x,y]$ became the sparse tensor $\tilde{Q}[c,x,y]$ after unit exclusion. Note, we chose to include only the positively correlated units in the model, because we found the feature visualization from it more intuitive.

Next, we factorized the sparse $\tilde{Q}$ tensor through non-negative matrix factorization (Lee and Seung, 1999, 2001) (decomposing a spatial-channel tensor into the product of a few non-negative spatial masks and channel vectors).

$$\underset{V_r,S_r}{\operatorname{argmin}} \left\| \tilde{Q}[c,x,y] - \sum_r V_r[c]S_r[x,y] \right\|_2^2,\ S_r[x,y] \geq 0, V_r[c] \geq 0$$

We used the coordinate-descent solver with Frobenius norm minimization as objective and *nndsvda* initialization as implemented in 'scikit-learn' package (Pedregosa et al., 2011). Here we kept the factor number to three. The order of the factors was arbitrary, but the magnitude of each factor signified how much it contributed to the whole tensor, potentially representing the feature's importance to the *Evolution* process. As an example, factorizing the $\tilde{Q}$ tensor in layer 3 of ResNet50-robust model to 3 components usually accounted for 0.132 ± 0.003 variance of the original tensor ($N = 90$).

Finally, we used penalized regression (Ridge) to determine the weights w_r of the three factors, such that these weights multiplying the CNN activations predict the neuronal activity the best.

$$\min_{w_r} \sum_i \left\| r[i] - \sum_{c,x,y,r} F[i,c,x,y] V_r[c] S_r[x,y] w_r \right\|_2^2$$

This model could predict the neuronal response to *Evolution* and *Manifold* experiment well, Pearson correlation between predicted and actual neuronal response to Manifold images was 0.70 ± 0.02 across $N = 90$ sessions; correlation was 0.67 ± 0.03 when including all reference images. We visualized the spatial masks combined by weights $\left| \sum_{r=1}^{3} w_r S_r[x,y] \right|$ as the feature attribution mask (Figure 4A (ii)), with the brighter region representing the image area correlated with and probably contributing to higher neuronal activation.

Because the goal of this analysis was to find an image-level attribution instead of just predicting neural responses, we used correlation rather than regression to identify image regions related to changes in neural activity. Directly using penalized regression on the full feature tensor or dimension reduced feature tensor (per sparse random projection or principal component analysis) often resulted in an overly sparse weight tensor with no coherent spatial structure, inadequate for interpretation (Wang and Ponce, 2022). Moreover, we chose to use matrix factorization instead of just mean or max compression of the weight tensor across the channel dimension, since this disambiguated unique features and highlighted the spatially coherent features more clearly. Visually, it broke down a complex feature into spatial composition of several simpler ones.

Inclusion criteria for non-driving units

In addition to the driving units, we also recorded the activities of other neuronal sites in our three electrode arrays. We used one-way ANOVA tests to select neuronal sites that were modulated in the Manifold image space, using image identity as the main factor with the criterion of $p < 0.001$. These well-modulated sites were included in other analyses, comprising $N = 3427$ non-driving- and 104 driving sites; these constituted 43.4% and 1.3% of all recorded units.

Measuring tuning map similarity

We measured the similarity of tuning maps based on *functional correlation* (Wikipedia Editors, n.d.) on the manifold, which generalized the correlation of functions in Euclidean space. This could be interpreted as the angular similarity between two mean-subtracted functions defined on the manifold. Our rationale for using this method is as follows. As the tuning maps we measured were defined over a hemisphere, some points were sampled closer to each other than to the others (e.g., around the hemisphere "poles"), thus the similarity of the response at those points in any tuning maps was not surprising. Functional correlation tackles this problem elegantly by considering the underlying metric structure over which we are sampling the map.

The equations for this correlation follow below; the integrals were evaluated using discrete quadrature on the hemisphere (Rafaely, 2019). J is the Jacobian of the map from $[\theta, \varphi]$ parameter to the Euclidean coordinate of the hemisphere.

$$\bar{f} = \int_M dx |\det J| f(x), \; Var[f] = \int_M dx |\det J| (f(x) - \bar{f})^2$$

$$< f, g > = \int_M dx |\det J| \, f(x) g(x)$$

$$corr(f, g) = \frac{< f - \bar{f}, g - \bar{g} >}{\sqrt{Var[f] \cdot Var[g]}}$$

Note that correlation in Euclidean space yields an even higher correlation value, so our results were not artifacts created by the correlation calculated on a spherical domain.

Naive bayes decoding for population neural activity

We used Naive Bayes decoding (Rish and Rish, 2001) method to assess how variability affected decoding. We assumed the response of each neuronal site was an independent normal variable with a mean $\mu_i(I)$ and variance $\sigma_i^2(I)$ depending on the image identity I. Thus, we could write the likelihood of an image identity given a vector of neural responses as

$$\log\mathcal{L}(I|\boldsymbol{r}) = \sum_i -\frac{1}{2}\left(\frac{r_i - \mu_i(I)}{\sigma_i(I)}\right)^2 - \log\sigma_i$$

Then assuming a uniform prior on the image identity, we could get the distribution of image identity based on a response vector using Bayes' rule.

$$p(I_j|\boldsymbol{r}) = softmax(\log\mathcal{L}(I|\boldsymbol{r}))_j$$

With this conditional distribution, we estimated the expected decoding error in terms of the L2 or angular distance between the latent vectors corresponding to the image

$$\mathbb{E}_{L2} = \sum_j \|z_j - z_{real}\| p(I_j|\boldsymbol{r})$$

$$\mathbb{E}_{ang} = \sum_j \arccos <z_j, z_{real}> p(I_j|\boldsymbol{r})$$