Many sensory-driven behaviors rely on predictions about future states of the environment. Visual input typically evolves along complex temporal trajectories that are difficult to extrapolate. We test the hypothesis that spatial processing mechanisms in the early visual system facilitate prediction by constructing neural representations that follow straighter temporal trajectories. We recorded V1 population activity in anesthetized macaques while presenting static frames taken from brief video clips, and developed a procedure to measure the curvature of the associated neural population trajectory. We found that V1 populations straighten naturally occurring image sequences, but entangle artificial sequences that contain unnatural temporal transformations. We show that these effects arise in part from computational mechanisms that underlie the stimulus selectivity of V1 cells. Together, our findings reveal that the early visual system uses a set of specialized computations to build representations that can support prediction in the natural environment.
The computations implemented by sensory systems are shaped by the properties of the physical environment and the behavioral demands of organisms. Many essential tasks such as catching prey and escaping predators rely on predictions about the future state of the world. Naturally occurring sensory input typically has a complex relationship with future environmental states, and it has been hypothesized that organisms transform these inputs into internal representations that facilitate temporal prediction. Consider vision. Under natural circumstances, the patterns of light projected on the retina evolve according to nonlinear dynamics that are difficult to extrapolate. We hypothesize that the visual system transforms its input into a representation that follows a “straighter” trajectory through time (the temporal straightening hypothesis). In this representation, prediction can be achieved through simple linear extrapolation.

Testing the temporal straightening hypothesis requires comparing the straightness (conversely, curvature) of the trajectory of a sequence of images in the pixel domain with that of the visual system’s internal representation. We previously developed a method for estimating the curvature of perceptual representations of video sequences and used it to demonstrate that the human visual system selectively straightens natural videos. Is this effect directly observable in the activity of neural populations, and if so, what are the response properties that underlie it? Here, we examine these questions in macaque primary visual cortex (V1).

We hypothesize that the visual system transforms its input into a representation that follows a “straighter” trajectory through time (the temporal straightening hypothesis). In this representation, prediction can be achieved through simple linear extrapolation.

Our analysis revealed that the early visual system transforms the frames of natural videos so as to straighten their V1 population trajectories. This temporal straightening process is specific to natural sequences: the same analysis of neural responses to frames of synthetic videos that contain unnatural transformations revealed that V1 populations substantially entangle such sequences. To gain insight into the computations underlying these effects, we examined the behavior of a stimulus response model built from operations of linear filtering, nonlinear pooling, and divisive normalization. This model captured the selectivity of individual V1 neurons for elementary stimulus attributes (orientation, scale, and phase), but only partially explained responses of single units to natural images. Despite this, the model largely replicated differences in population-level straightening effects across natural videos. We show that this behavior critically depends on the model’s nonlinear pooling. Together, these findings establish that the early visual system processes its inputs using a set of specialized computations that can support perceptual prediction in the natural environment.

**Results**

We compared the curvatures of brief video clips in the pixel and neural domains. In the pixel domain, each video frame corresponds to a point in a high-dimensional space whose coordinates are the brightness of individual pixels. A video clip corresponds to a sequence of such points. Consider three consecutive frames (Fig. 1a). A natural measure of local curvature is the unsigned angle between the segments that connect the middle frame to adjacent frames (Fig. 1b). We define the global curvature of a video clip as the average of these local curvatures. This measure, known as discrete curvature, is positive-valued and achieves a minimal value of zero only for straight sequences. The larger its value, the more curved a sequence is, and the more error-prone a linear extrapolation of the trajectory would be.

**Estimating neural curvature.** We recorded the activity of populations of up to a few dozen V1 neurons while presenting frames taken from brief video clips (see “Methods”). In our previous perceptual experiments, we used an experimental...

---

**Fig. 1 Measuring curvature in the pixel and neural domains.** a An example sequence of three movie frames. b Visualization of a high-dimensional representation of this sequence. Each point corresponds to a single movie frame, with each coordinate specifying the brightness of a pixel within that frame. Discrete curvature $c_{\text{pixel}}$ is defined as the unsigned angle between two segments connecting adjacent frames. c The curvature of a sequence is fully determined by the collection of pairwise distances. d Simulated responses of two neurons to a sequence of three movie frames. Error bars illustrate ± 1 standard deviation. e Joint response probabilities for these two neurons, for each of the three frames. f Representation of the same sequence of frames in a two-dimensional neural distance space. The distance metric is frame discriminability (in units of $d'=$Euclidian distance divided by standard deviation). g Comparison of neural curvature estimates obtained under a response model that includes correlations (abscissa) to one that assumes independence (ordinate).
protocol that relies on the presentation of static images, thereby isolating the contribution of spatial processing mechanisms\(^7,8\). Here, we seek to study the neural basis of these perceptual effects. We therefore used the same stimulus presentation method. Movie frames were presented in randomized order, interleaved with a blank screen. These manipulations removed any systematic contribution from fast motion-selective mechanisms, as well as from history-dependent mechanisms such as response adaptation. We then obtained neural response trajectories by arranging the data in the movies’ natural temporal order.

The curvature of a trajectory is fully determined by the distances between all pairs of frames. Moreover, the pairwise distances between a set of \(T\) points are uniquely constrained by their relative positions within a \(T\)-dimensional space (Fig. 1c). Mapping population representations into a neural distance space of this dimensionality (i.e., one less than the number of movie frames) is therefore sufficient for inferring neural curvature. Consider the responses of two neurons to a sequence of three movie frames (Fig. 1d). Their joint activity constitutes a population trajectory (Fig. 1e). Neural responses are noisy—repeated presentations of the same stimulus elicit different patterns of population activity\(^7,8\)—and this noise is generally believed to limit perceptual discriminability. We therefore used frame discriminability as neural distance metric. Specifically, the overlap of the population response distributions elicited by each frame provides a principled measure of this statistic (Fig. 1e).

The number of measurements obtained in typical experiments is insufficient to describe the joint population response distribution, even for populations of moderate size. For this reason, we fit a descriptive response model to the observed population activity and computed frame discriminability under this fitted model. We then mapped the population responses for each frame into a neural distance space such that pairwise Euclidean distances were equal to the discriminability of the model-predicted response distributions (Fig. 1f). Finally, we obtained a curvature estimate by inferring a distribution of neural trajectories that were consistent with the data, and computed the value of curvature that was most likely under this distribution (see “Methods”). This procedure allowed us to obtain largely unbiased neural curvature estimates (Supplementary Fig. 1).

Neural trajectories are not only shaped by the mean pattern of neural activity, but also by the variability of this pattern. Variations in response strength across repeated presentations of the same stimulus are often correlated across neurons\(^9\). Such “noise correlations” can impact stimulus discriminability by altering the overlap of population response distributions\(^10,11\). For the two-neuron example, the impact of noise correlations is considerable. Inspection of the average responses suggests that the neural trajectory is modestly curved (Fig. 1e). Yet, in neural distance space, the trajectory is highly curved (Fig. 1f). Moreover, shared variability affects the distribution of plausible neural trajectories and can therefore affect our curvature estimates. We wondered whether a simpler estimation procedure that ignores these correlations would produce the same answer. We found this not to be the case: for many recorded trajectories, treating populations as pseudo-populations by ignoring noise correlations gave a dramatically different estimate of curvature. For some trajectories, this increased the estimate, but for others, it decreased the estimate (Fig. 1g)\(^11\).

**Fig. 2 Testing the temporal straightening hypothesis.** a We presented individual frames from different video clips in randomized order. Each frame was shown for 200 ms, followed by 100 ms of constant luminance. We used multi-electrode arrays to record V1 population activity. The raster illustrates spiking activity recorded over a 1.3 sec period from an example V1 population (“Population 7”, consisting of 39 units). b Mean response for three units in the example population, to frames of “Movie 7”. We fit a descriptive response model to each unit. Points indicate mean spike counts, lines illustrate predicted responses of the fitted model. c We evaluated the model’s goodness of fit by computing the correlation between predicted and measured response mean, variance, and covariance across the units and frames in each dataset. Each point corresponds to a dataset (red points indicate the example dataset). Blue lines indicate the inclusion criteria independently applied to these three statistics (included datasets are black points, excluded datasets are transparent green points). d Two-dimensional projections of the trajectory for the example dataset, in the pixel domain (left), and in the neural domain (right). e Estimated neural curvature of the example dataset (white triangle) and its expected distribution under the null model (gray histogram, see “Methods”). The black and gray triangle indicate the pixel-domain value and the null model’s mean, respectively. Relative curvature is \(-69^\circ\), \(P < 0.05\), non-parametric test.
domain counterpart ($c_{\text{pixel}} = 34^\circ; c_{\text{neural}} = 19^\circ$). This difference is not due to dimensionality reduction: in their native 5-dimensional spaces the difference is even more substantial ($c_{\text{pixel}} = 83^\circ; c_{\text{neural}} = 19^\circ$). To evaluate the statistical significance of this effect, we computed the distribution of neural curvature estimates under a null model. We simulated a hypothetical population that was identical to the observed population in all regards, but whose neural trajectory preserved the pixel-domain curvature (see "Methods"). For the example dataset of Fig. 2, the empirically observed curvature fell well beyond the central 95% interval of the null distribution (Fig. 2e). Neural curvature estimates can be biased, causing the average estimate under the null model to deviate from the pixel-domain value (Fig. 2e, black vs gray triangle). To take this potential bias into account, our analysis hereafter focuses on the difference between the empirical curvature estimate and the average estimate of the null model. We refer to this as "relative curvature". For the example dataset, its value is $-69^\circ$.

V1 trajectories elicited by natural movies were typically straighter than their pixel-domain inputs (median relative curvature $= -9.9^\circ, P < 0.001, n = 127$ datasets, Wilcoxon signed-rank test), but the effect varied substantially across datasets (Fig. 3a; 23 of 127 datasets were significantly straighter than the null model). This variability existed across neural populations (Fig. 3b; $P < 0.001, F_{1,123} = 11.87, n = 7$, analysis of variance), as well as across movies (Fig. 3c; $P < 0.01, F_{1,123} = 7.77, n = 20$). These effects did not interact statistically ($P = 0.28, F_{1,123} = 1.17$). Visual inspection of the movies revealed that straightening was generally stronger for movies containing dense textures rather than isolated objects (Fig. 3c inset, Supplementary Fig. 3). Finally, we verified that our conclusions were robust to the choice of null model, by comparing neural curvatures directly to pixel-domain curvature (Supplementary Fig. 4).

Neural entangling of unnatural videos. We have shown that V1 populations represent images in a manner that tends to straighten natural videos. This does not imply, however, that this representation is specifically tailored for this purpose. Following the approach of our perceptual experiments$^9$, our stimulus set therefore included twenty unnatural image sequences that fade from the first to the final frame of each of the natural videos (see Fig. 4a and Supplementary Fig. 3 for examples). Such sequences are unlikely to occur in the real world (e.g., the middle frame contains the average of the first and last frames—a "double exposure"). Perceptually, we found that their curvature typically increased$^9$, presumably because they are entangled by the non-linear transformations of the early visual system. We fit the response model to the population activity elicited by the unnatural movies (Fig. 4b, c). The projection of the pixel-domain and neural-domain trajectories of an unnatural movie onto

---

**Fig. 3** Curvature reduction for natural movie sequences. **a** Relative curvature for seven V1 populations probed with twenty natural movies. Blue colors indicate neural straightening, red colors indicate neural entangling, and gray indicates missing data (see "Methods"). **b** Relative curvature for each V1 population, averaged across all movies. Error bars indicate s.e.m. across movies, $n$ values indicate population size. **c** Relative curvature for twenty movies, averaged across all populations. Error bars indicate s.e.m. across populations. Insets illustrate a single frame from three movies that elicited no, medium, and strong straightening (left, middle, and right, respectively; see "smile", "walking" and "prairie" in "Methods").

**Fig. 4** Testing the specificity of temporal straightening for natural sequences. **a** An example sequence illustrating the first, middle, and last frame of an "unnatural" movie, constrained to be straight in the pixel domain. **b** Mean responses of three example units (Population 1) to frames of Movie 1. Points indicate mean spike counts, lines illustrate predicted responses of the descriptive response model. Prediction line for Neuron 2 is shown in red for clarity. **c** The correlation between predicted and measured response mean, variance, and covariance across all units and frames in each dataset (red points correspond to the example dataset). Blue lines indicate the inclusion criteria independently applied to these three statistics (included datasets are black points, excluded datasets are transparent green points). **d** Two-dimensional projections of trajectory of the example dataset, in the pixel domain (left), and in the neural domain (right). **e** Neural curvature of the example dataset (white triangle) and its expected distribution under the null model (gray histogram). The black and gray triangle indicate the pixel-domain value and the null model’s mean, respectively. Relative curvature is $82^\circ$. $^*P < 0.05$, non-parametric test.
the first two principal components reveal a very different outcome (Fig. 4d). As can be seen, the neural trajectory is much more curved than its pixel-domain counterpart ($c_{\text{pixel}} = 1$; $c_{\text{neural}} = 80\)°). This visual impression was confirmed by calculating the relative curvature in the full-dimensional neural response domain (Fig. 4e; mean $c_{\text{null}} = 6$; $c_{\text{neural}} = 88\)°, $P < 0.05$). V1 trajectories elicited by unnatural movies were typically more curved than their pixel-domain inputs (Fig. 5a; estimated curvature fell outside the central 95% interval of the null distribution for 25% of datasets; Supplementary Fig. 5). For natural movies, there was also a clear association of the neural representations across spatial scales: movies that elicited stronger straightening at the coarser scale tended to do so as well at the finer one (Fig. 6a; $r = 0.80$, $P < 0.01$, $n = 10$). Similarly, there was a strong association of relative curvature across temporal scales (Fig. 6b; $r = 0.71$, $P < 0.001$, $n = 20$ movies).

The second principal components reveal a different relationship between the movie and neural response. For the unnatural videos, we found no significant correlation in relative curvature across spatial (Fig. 6a; $r = 0.23$, $P = 0.53$, $n = 10$) and temporal scales (Fig. 6b; $r = 0.14$, $P = 0.56$, $n = 20$).

Note that on average, the zoom x1 condition elicited stronger straightening for natural movies than zoom x2 (Fig. 6a; $P < 0.05$, $n = 10$ movies, Wilcoxon signed-rank test of difference in relative curvature). Why might this be? V1 neurons are selective for the spatial scale of visual input. The zoom x1 condition elicited stronger responses than zoom x2 at the level of populations (Fig. 6c, mean response histograms; $P = 0.01$, $n = 61$ paired datasets, two-sided Wilcoxon signed-rank test), and individual units ($P < 0.001$, $n = 2614$ units, two-sided Wilcoxon signed-rank test), indicating that those videos were better matched to the preferred spatial scale of the recorded populations. We therefore asked whether firing rate was associated with straightening, and found this to indeed be the case (Fig. 6c; $P < 0.001$, $F_{1,123} = 12.59$, ANCOVA). After controlling for this relationship, there was no effect of spatial scale on relative curvature ($P = 0.60$, $F_{1,123} = 0.28$). The relationship between straightening and mean firing could potentially be explained by movies containing cells’ preferred image features also engaging additional non-linearities of V1 circuits (Supplementary Fig. 6). This interpretation comes with an important implication. These stimuli activate millions of cortical neurons, and it is unlikely that our recordings include the most strongly driven cells. As such, our estimates of relative curvature might understate the temporal straightening occurring in the full cortical population.

In summary, the straightening of natural sequences appears to be an intrinsic property of each sequence, whereas the entangling of unnatural sequences is more erratic.

**Effects of spatial and temporal scale.** So far, we have presented evidence that V1 population representations straighten naturally occurring image sequences and distort unnatural ones. We hypothesize that the straightening or entangling of a sequence depends on its probability of occurring in the real world. If so, sequences which have equal real world probability should display similar amounts of straightening. Natural scenes are approximately scale-invariant, both spatially and temporally. Modest changes in either the spatial or temporal scale of a video should preserve its real world probability and it should therefore undergo a similar amount of straightening. Our experimental stimuli were constructed to test this prediction. Specifically, our stimulus set of twenty natural and unnatural movies included ten unique movies that were each displayed at two spatial scales: zoom x1 (the original scale) and zoom x2 (created by upscaling the central portion of each frame—see Fig. 6a for an example). Likewise, for each 6-frame movie, we also included 5 intermediate frames, enabling us to compare two temporal scales (frame rate x1 and frame rate x2—see Fig. 6b). Consistent with natural videos’ statistical scale-invariance, we found that the pixel-domain curvature of these sequences was largely preserved across spatial and temporal scales ($r = 0.96$ across spatial scales, $r = 0.91$ across temporal scales; Supplementary Fig. 5). For natural movies, there was also a clear association of the neural representations across spatial scales: movies that elicited stronger straightening at the coarser scale tended to do so as well at the finer one (Fig. 6a; $r = 0.80$, $P < 0.01$, $n = 10$). Similarly, there was a strong association of relative curvature across temporal scales (Fig. 6b; $r = 0.71$, $P < 0.001$, $n = 20$ movies).

Unnatural sequences do not occur in the real world, and their entangling cannot therefore not be a design goal, but must instead be a byproduct of the non-linearity of the visual system. As such, we would not expect entangling to be a robust phenomenon. For example, two sequences that are displayed at different scales need not be equally entangled by neural transformations. Indeed, for the unnatural videos, we found no significant correlations in relative curvature across spatial (Fig. 6a; $r = 0.23$, $P = 0.53$, $n = 10$) and temporal scales (Fig. 6b; $r = 0.14$, $P = 0.56$, $n = 20$).

**Computational basis of neural straightening.** Which neural computations give rise to the perceptual straightening of natural videos? To explore this, we constructed a computational model for V1 neurons that is simple enough to be fit to data, yet powerful enough to describe basic feature selectivity and nonlinear response properties. This model has its roots in the work of Hubel and Wiesel, and incorporates elements introduced by many others. Neural responses are described using a cascade of two linear-nonlinear (LN) operations. Visual stimuli are transformed by calculating the relative curvature for each V1 population, averaged across all movies. Error bars indicate s.e.m. across movies. Insets illustrate a single frame from three movies that elicited strong, mild, and no entangling (left, middle, and right, respectively; see “bees”, “walking”, and “carnegie dam” in “Methods”).

For the unnatural videos, we found no significant correlation in relative curvature across spatial (Fig. 6a; $r = 0.23$, $P = 0.53$, $n = 10$) and temporal scales (Fig. 6b; $r = 0.14$, $P = 0.56$, $n = 20$).
selectivity for spatial phase. The filter responses are nonlinearly transformed (halfwave rectified and squared), and then locally pooled. Finally, this pooled signal is subject to untuned divisive normalization to generate a response rate (Fig. 7a), which is used to generate spikes from a modulated Poisson process.

We probed all neurons within each V1 population with both white noise stimuli and drifting gratings (see "Methods"). Responses to the white noise stimuli enabled us to determine each neuron’s receptive field location; responses to the drifting gratings revealed their selectivity for orientation, scale, and phase. For each neuron, we fit the parameters of the stimulus-response model to these data (see "Methods"). Consider an example neuron, for which the model provided a good account of basic stimulus selectivity (Fig. 7b). We quantified goodness of fit by...
computing the correlation between the model’s predicted responses and the measured tuning for orientation, spatial frequency, and temporal modulation while the other stimulus parameters were at the cells’ preferred value. We found that the model typically performed well in describing these classic forms of stimulus selectivity (Fig. 7c; median $r = 0.89, 0.95$, and $0.60$ for orientation, spatial frequency, and temporal modulation, respectively).

We also examined how well this model predicts individual cells’ responses to natural and unnatural movie frames. In general, performance was much worse, at times close to chance (Fig. 7b, c; median $r = 0.07, P < 0.001, n = 440$ frames, Wilcoxon signed-rank test for natural frames; median $r = 0.08, P < 0.001$ for unnatural frames). These failures may indicate that the model lacks computational ingredients that are essential for explaining V1 activity in natural settings. If so, we might expect it to make poor predictions of the curvature of population trajectories. Alternatively, the model may fail to capture idiosyncratic response properties of individual neurons, but still provide a good summary of population trajectories.

For each dataset, we computed the population trajectory predicted by the stimulus-response model and compared its curvature to the empirically obtained estimate (see Fig. 7d for an example). Across all movies, there was a robust relationship between predicted and measured relative curvature (Fig. 7c; $r = 0.79, P < 0.001, n = 40$). Separated by movie type, this relationship remained substantial for natural movies ($r = 0.63, P = 0.003, n = 20$), but not for unnatural ones ($r = 0.02, P = 0.928, n = 20$). This is consistent with our results from previous analyses: the straightening of natural sequences displays a strong variation across videos (Fig. 3) and a strong association across spatial and temporal scales (Fig. 6). The entangling of unnatural sequences lacks both of these properties, indicating that it contains less explainable variability (the fraction of cross-movie variance is $26.6\%$ for natural movies and $10.7\%$ for unnatural movies).

Which elements of the stimulus-response model enable it to capture population-level straightening effects? To answer this question, we created restricted model versions by eliminating specific components. Removing untuned divisive normalization largely preserved the model’s explanatory power (Fig. 8). This was also the case when we eliminated the simple cells (see “Methods”, for reference: $60\%$ of cells were “complex” when categorized by the relative temporal modulation of their firing rate to the preferred grating ($F1/F0 < 1$)\(^{30}\)). On the other hand, when we eliminated complex cells, the model’s explanatory power was substantially diminished (Fig. 8). We conclude that, as with perceptual effects\(^6\), the nonlinear pooling mechanism that underlies the phase-invariance of complex cells is largely responsible for the model’s ability to explain patterns of straightening at the population level. This is broadly consistent with the view that complex cell pooling reduces temporal image fluctuations due to translation or other deformations\(^{120,22,23,31–34}\).

While this LN–LN model provides insight into the computational basis of neural straightening, note that it systematically underestimated the values obtained from V1 measurements (median predicted relative curvature across all natural movies $= -0.6\%$; median neural relative curvature $= -9.9\%$) and failed to capture cross-population differences in curvature (natural movies: $r = -0.71, P = 0.077, n = 7$; unnatural movies: $r = 0.02, P = 0.962, n = 7$). We conclude that simple stimuli only partially engage the mechanisms that govern population responses to natural signals, and that additional model elements are necessary to fully explain the data.

### Discussion

What do sensory neurons seek to accomplish when they transform their inputs? We have tested the proposal that neurons in the visual system transform their inputs so as to straighten their temporal trajectories\(^6\). We developed a method to measure the curvature of a neural population trajectory and used it to analyze physiologically recorded V1 responses elicited by natural and unnatural image sequences. Our analysis revealed that the spatial processing mechanisms of the early visual system selectively straighten naturally occurring sequences.

Estimating the curvature of a neural population trajectory is a challenging statistical problem. The difficulty arises from several factors. First, neural responses are highly variable, leading to substantial variance in estimates. Second, realistic numbers of experimental trials and neurons are insufficient to yield precise estimates of joint response distributions or correlations. And third, curvature measurements in high-dimensional neural spaces are plagued by severe estimation bias (Supplementary Fig. 1a). We found that the combination of a principled statistical model of neural population activity, a probabilistic inference procedure for curvature, and an explicit null model of expected curvature estimates was sufficient to overcome these challenges (Supplementary Fig. 1b).

Our experimental test of the temporal straightening hypothesis only partially engages the mechanisms that shape neural trajectories under natural viewing conditions. In particular, by presenting static movie frames in a randomized order, interleaved by blanks, we excluded systematic contributions from history-dependent mechanisms such as temporal filtering, motion selectivity, response adaptation and recurrent computation, all of which are known to influence V1 responses. Despite this restriction, we found V1 representations to behave as predicted, consistent with a simple static model. How would these results generalize to representations of continuous streams of images? If temporal straightening is a prominent goal of visual processing, we might expect stronger effects under more natural stimulus conditions. Indeed, the inertia of physical systems makes their three-dimensional physical trajectories locally linear, and thus predictable. Although linear motion trajectories display a stable distribution of spatiotemporal energy, they can yield highly curved image-domain trajectories, since the brightness of individual pixels can change abruptly due to the passage of sharp edges, sudden occlusion, or changes in illumination. Engaging motion-selective neural mechanisms may therefore further...
increase temporal straightening. This is an important question for future work.

In recent years, the neuroscience community has witnessed a growing interest in characterizing the geometric properties of population representations. For example, recent studies of the motor cortex and the frontal cortex sought to measure the tangling35, divergence36, and curvature37 of neural trajectories that unfold over hundreds of milliseconds. Each of these relied on analyses of trial-averaged trajectories and hence did not take response variability or correlation into account. While the simplicity of such analyses is appealing, we found that ignoring shared variability can dramatically alter neural curvature estimates in V1 (Fig. 1g). Our study is thus an example of how precise quantification of the geometry of neural trajectories may require the joint characterization of population activity11. In identifying the potential biases exhibited by population-level statistics, our results also highlight the need for an explicit null model when interpreting them12. The methodology we introduced for the estimation of neural curvature can be applied to other population datasets and generalized to other geometrical properties.

We also evaluated the predictions of a stimulus-response model that describes the primary selectivity properties of individual V1 neurons (in particular, for orientation, spatial frequency, and spatial phase) using linear filtering, the phase or position-invariance properties using nonlinear pooling, and the gain control properties using divisive normalization. Although this model did not accurately describe the responses of individual cells to natural videos38, it was sufficient to account for the population-level pattern of straightening across natural videos. Comparison of different model parameterizations revealed that the nonlinear pooling mechanism played the most important role. How could the model be further improved? Under natural stimulation, intricate center-surround interactions shape the suppression at play in the primary visual cortex39–41. It seems likely that enriching our model to include these normalization mechanisms or those of earlier processing stages6,42,43 could lead to a better account for our experimental findings.

We found that V1 population representations straighten natural videos, consistent with our previous perceptual findings6. But despite the use of largely identical stimuli, there was a substantial quantitative difference between these results (average straightening of ~10° in macaque V1, and ~30° in human perception), although the most responsive V1 populations did show effects nearly as large as the perceptual effect (Fig. 6c). One possible explanation for this discrepancy is that the best-driven subset of V1 neurons can fully explain the perceptual results, but that we typically failed to observe this subset because we did not tailor our stimuli to the recorded populations. Alternatively, it could be the case that downstream visual areas further straighten their inputs, and provide the substrate for perception. To examine this question, we have begun to assess the evolution of neural curvature along the visual hierarchy44.

Finally, our results suggest that straightening might serve as an objective for learning more biologically plausible models of the visual system. Deep neural networks trained for object recognition are currently the best predictors of high-level visual function45–47. But these are typically trained on large sets of labeled images, and more stringent tests of their ability to account for human perception have revealed systematic deficiencies43,48–50, including a failure to account for straightening. These studies imply that a robust metric for the similarity of natural images is an important property of biological systems, and that this is currently lacking in artificial systems. An unsupervised objective that can learn to represent image structures and the natural continuous perturbation of those structures as they evolve over time provides a promising alternative1,46,33,31. As such, temporal straightening could provide an effective principle underlying the emergence of visual representations that support the efficiency, flexibility, and robustness of biological intelligence.

Methods

Surgical preparation. We recorded from four anesthetized, paralyzed, adult macaque monkeys (Macaca cynomolgus). Before surgery, each animal was initially anesthetized with ketamine (10 mg/kg, IM injection) and pretreated with 1.5 mg/kg diazepam and atropine sulfate (0.04 mg/kg, IM). Surgery consisted of the placement of cannulate in the saphenous veins of both legs as well as installation of a tracheal cannula. The animal’s head was placed in a stereotaxic frame (David Kopf Instruments) and craniotomies and durotomies (2 × 2 cm) were performed to expose the brain. Experiments typically lasted 5–6 days. Throughout the experiment, anesthesia was maintained with sufentanil citrate (4–20 µg/kg/h, IV), sup- erine (0.5–2%) during surgery. Animals were recovered using pancuronium bromide (0.1–0.2 mg/kg/h, IV). We monitored vital signs (heart rate, blood pressure, lung pressure, end-tidal CO2, EEG, body temperature, urine flow and osmolality), and maintained them within appropriate physiological ranges. Pupils were dilated with topical atropine. The eyes were protected with gas-permeable contact lenses and refracted with supplementary lenses chosen through retinoscopy. All procedures were approved by the University of Texas Institutional Animal Care and Use Committee and conformed to National Institutes of Health standards.

Electrophysiology. We recorded extracellular activity in area V1 using multi-shank electrode arrays (32 or 64 channels; Neuronexus A32, 2 or 4 shanks; A64, 8 shanks), advanced into the brain through a craniotomy and small durotomy. Once the electrode array was in place, we covered the exposed brain with a mixture of agarose and artificial cerebrospinal fluid. We allowed 30 min for the cortical tissue to settle before recording began.

Visual stimulation. We presented visual stimuli on a 20 inch gamma-corrected CRT monitor at a resolution of 1024 × 768 pixels with a refresh rate of 60 Hz. Visual stimuli were generated and presented using the Psychophysics Toolbox extensions for Matlab25,26. For each population, we first determined eye dominance, occluded the non-preferred eye, and estimated the population analyzed using panceuronium bromide (0.1–0.2 mg/kg/h, IV). We monitored vital signs (heart rate, blood pressure, lung pressure, end-tidal CO2, EEG, body temperature, urine flow and osmolality), and maintained them within appropriate physiological ranges. Pupils were dilated with topical atropine. The eyes were protected with gas-permeable contact lenses and refracted with supplementary lenses chosen through retinoscopy. All procedures were approved by the University of Texas Institutional Animal Care and Use Committee and conformed to National Institutes of Health standards.

In the first experiment, we presented drifting sinusoidal gratings with a large circular aperture (diameter: 15 visual degrees). The stimulus set consisted of the combination of 16 equi-spaced directions of motion and 6 logarithmically spaced spatial frequencies (0.25–8 cycles/deg). Stimuli were presented for 1000 ms each, interleaved with a blank screen for 500 ms, and each was repeated 50 times.

In the second experiment, we presented modified sparse-noise stimuli, following the approach of ref. 34. This stimulus set consisted of a collection of static bars (2 × 0.01 degrees) with randomly chosen orientation, position, and polarity. Stimuli were presented for 153 ms each for approximately 20 min in total.

In the third experiment, we presented frames drawn randomly from 40 sequences. Ten of these were 11 frame clips extracted from movies: ‘chironomus’, ‘bees’, ‘egomotion’, ‘prairie’, ‘carnegie dam’, ‘walking’, ‘water’, and ‘leaves-stand’ (Chicago Motion Database, https://cmd.rcc.uchicago.edu), ‘dogville’ (a feature film, from Lions Gate Entertainment), and ‘smile’ (LIVE Video Quality Database17). All frames were achromatic and presented through a large aperture (diameter: 15 visual degrees). We generated rescaled versions of each original clip by extracting the central square of pixels, and upsampling and interpolating by a factor of two in both directions. We also designed artificial control sequences by linearly interpolating between the first and last frames of the corresponding natural sequence. In summary, 10 original movie clips produced 20 natural image sequences and 20 artificial image sequences. Single frames were presented for 200 ms each and interleaved with a blank screen for 100 ms. Each trial, all frames of all movies were presented exactly once, in randomized order. The experiment consisted of 50 trials in total.

Extraction of neuronal response. We used multi-electrode arrays to make extracellular recordings from V1. To extract responses of individual units, we first automatically spike-sorted these data with Kilosort257 (https://github.com/kwikteam/phy). For each identified unit, we calculated its response by counting spikes in a time window whose duration matched that of the stimulus presentation. We chose a response latency for each unit by maximizing the stimulus-associated response variance58.
Measuring pixel-domain curvature. Given an image sequence, we wish to compare its curvature in the domain of pixel-intensities with that of neural responses. Measuring the pixel domain curvature is straightforward. Let $x_{\text{pixel}}^\text{local}$ be the vector of normalized displacement vectors $v_{\text{pixel}}^\text{local}$:

$$v_{\text{pixel}}^\text{local} = x_{\text{pixel}}^\text{local} - x_{\text{pixel}}^{\text{local}-1}$$

and the curvature at time $t$ is simply the angle between two such vectors, which can be computed from their dot product:

$$c_t^\text{pixel} = \arccos \left( v_{\text{pixel}}^\text{local} \cdot v_{\text{pixel}}^{\text{local}-1} \right)$$

The global curvature of the sequence $c_{\text{pixel}}$ is the average of these local curvature values, in degrees.

Modulated Poisson model of response distributions. On a given trial $k$, the image $x_{\text{pixel}}^\text{local}$ induces a pattern of recorded spiking activity which is summarized as a vector of spike counts $n_t^\text{local}$, whose dimensionality $D$ is equal to the number of cells being recorded. We start by building a parametric model in which this spiking activity arises from a vector of spike rates $\lambda_t$. Images are presented in random order, interleaved with a blank screen, and we assume that the spike counts in response to different images are independent from one another:

$$P(n_t^\text{local}|\lambda_t^k) = \prod_{t=1}^T \prod_{i=1}^D \exp(-\lambda_t^i) n_t^i!$$

We enrich this description by modeling spike-count correlations across cells, assuming their rates are modulated by a common gain factor. Specifically, we assume they are independent and Poisson-distributed when conditioned on both the stimulus-driven rate, $\lambda_t$, and an unknown multiplicative gain $g$. We model the gain as a multivariate log-normal random variable $g = \exp(\epsilon)$, where $\epsilon \sim N(\mu_g, \Sigma_g)$. The covariance matrix $\Sigma_g$ thus captures the dependencies across cells, and the mean vector $\mu_g$ is chosen such that the gain variable has unit mean (i.e. $\mu_g = \frac{1}{2} \text{diag}(\Sigma_g)$, where diag extracts the diagonal of a matrix). The probability of a vector of spike counts is then:

$$P(n_t^\text{local}|\lambda_t, \Sigma_g) = \int \prod_{i=1}^D \exp(-\lambda_t^i) n_t^i!$$

where $\otimes$ denotes the element-wise product between two vectors. We constrain $\Sigma_g$ to be the sum of a diagonal matrix (representing private gain modulation) and a low-rank matrix (representing shared gain modulation). We chose the rank of this matrix by cross-validation, resulting in the use of a rank-2 matrix throughout the analysis. The covariance matrix $\Sigma_g$ represents a stable property of the neural population, and we therefore share its parameters across all frames and videos.

Summarizing responses with neural embeddings. Having described the neural responses to each frame with a parametric distribution, we now summarize these distributions with a sequence of neural embeddings. We choose these embeddings such that their pairwise Euclidean distances are equal to the discriminability of the corresponding multi-dimensional response distributions. Specifically, we use the Fisher information metric $I(\lambda)$ to approximate the discriminability of the rate vector $\lambda$ from nearby vectors:

$$[I(\lambda)]_{ij} = E_{\lambda} \left[ \frac{\partial}{\partial \lambda^i} \log P(n|\lambda) \right] \left[ \frac{\partial}{\partial \lambda^j} \log P(n|\lambda) \right]$$

We simplify our analysis by first considering the Fisher information metric resulting from a modulated Poisson model with independent gain modulation for each neuron (the general case is described in the section “Numerical curvature calculation”). Under this approximation, the metric is diagonal with entries:

$$[I(\lambda)]_{ii} = \frac{1}{\lambda^i} + \sigma_{\lambda,i}^2$$

where $\sigma_{\lambda,i}^2 = \exp(\Sigma_g)_{ii} - 1$. Thus, the discriminability of the rate vector $\lambda$ is heterogeneous (i.e. it is a function of $\lambda$).

We devised a monotonic nonlinear transformation that maps the rates into a space that is homogeneous in discriminability. Specifically, consider the following transformation:

$$y_t = \frac{2}{\sigma_t} \sinh^{-1}(\sigma_t \sqrt{\lambda_t})$$

which can be inverted to yield:

$$\lambda_t = 1 - \sigma_t^{-2} \sinh^2 \left( \frac{\sigma_t y_t}{2} \right)$$

The Fisher information metric for the transformed rate variables $y_t$ may be computed using the chain rule:

$$[I(y)]_{ij} = E_{y} \left[ \frac{\partial}{\partial y^i} \log P(n|y) \right] \left[ \frac{\partial}{\partial y^j} \log P(n|y) \right]$$

$$= \frac{\partial}{\partial y^i} E_{y} \left[ \frac{\partial}{\partial y^j} \log P(n|y) \right] \frac{\partial}{\partial y^j} E_{y} \left[ \frac{\partial}{\partial y^i} \log P(n|y) \right]$$

$$= \frac{1}{\lambda^i} + \sigma_{\lambda,i}^2$$

Thus, the discriminability between response distributions for frames $t$ and $t'$ is simply the Euclidean distance between transformed rate vectors $y_t$ and $y_{t'}$. As such, the curvature of a neural trajectory may be estimated by computing the curvature of the transformed trajectory $[y_t]_{t=1:T}$ using the same procedure as in the pixel-domain.

Variational Bayesian inference of neural curvature. Next, we turn to the question of inferring the curvature of the sequence $[y_t]_{t=1:T}$ from data. It is tempting to solve this sequentially, by first estimating the parameters of each response distribution by maximizing their likelihood (Eq. (5)) given the dataset of all spike counts:

$$[\lambda_t], [\Sigma_g] = \arg \max_{\lambda_t, \Sigma_g} \prod_{t=1}^T \prod_{i=1}^D \exp(-\lambda_t^i) n_t^i!$$

and then using these to estimate the most likely neural embedding $[y_t]_{t=1:T}$, and taking the curvature of this. However this two-step procedure is plagued by substantial estimation bias when used with the amounts of data available in our experiments (Supplementary Fig. 1a). We therefore sought to compute the most likely curvature over a distribution of plausible neural embeddings.

Since a set of $T+1$ points lie within a $T$-dimensional subspace, we start by parameterizing the $D$-dimensional rate vectors $\lambda_t$ as a function of $T$-dimensional vectors $x_{\text{pixel}}^\text{local}$ and an orthogonal embedding matrix $E$:

$$y_t = E x_{\text{pixel}}^\text{local}$$

Since distances are preserved in this lower-dimensional subspace, so is the curvature of the sequence. We parameterize the trajectory in terms of the distances between successive points $d_t$, local curvatures $\gamma_t$ and directions of curvature $a_t$:

$$x_{\text{pixel}}^\text{local} = y_t + d_t a_t$$

$$\gamma_t = \cos(c_t) = y_t^2 + 1 - 2 y_t a_t$$

These are themselves parameterized in terms of real-valued variables, on which we place the following priors:

$$d_t = f_g(c_t^\text{local}), \quad x^2_c \sim N(0,\Sigma_g)$$

$$\gamma_t = f_g(c_t^\text{local}), \quad a_t = f_g(c_t^\text{local})$$

where $f_g$ is a smooth rectifying function, $f_g$ ensures that $a_t$ is of unit length and orthogonal to $y_t$. In reality, we assume that the columns of $E$ are orthonormal functions of the vector $D$-dimensional neural space. This guarantees that the metric properties of the neural trajectory, and $z = [c_t, z_c, z^2] \in Z^2$ be those specifying its local properties. We wish to estimate the most likely value of the global parameters (including the global curvature) given the dataset of all spike counts $N = [n_t^k]_{k=1:T}, t$:

$$\theta^* = \arg \max_{\theta} \log P(N|\theta) = \arg \max_{\theta} \int P(N|\theta) P(\theta) d\theta$$

Intuitively, these are the global parameters that are most consistent with the family of neural trajectories $z$ that are supported by the data. However, this quantity is
impossible to compute in practice, given the dimensionality of \( z \). We therefore derive a lower bound on the likelihood, using an approximate posterior distribution over local variables \( \mathcal{P}(z|\theta) \).

\[
\log \mathbb{P}(\mathbf{N}|\theta) = \log \int \mathcal{P}(\mathbf{N}|\theta) \mathcal{P}(\mathbf{N}) \mathcal{P}(\mathbf{z}|\theta) d\mathbf{z} \\
= \log \mathbb{E}_{\mathcal{P}(\mathbf{N})} \left[ \log \mathcal{P}(\mathbf{N}|\theta) \right] + \mathcal{D}(\mathcal{P}(\mathbf{z}|\theta) || \mathcal{P}(\mathbf{N}|\theta))
\]

where \( \mathcal{D}(\cdot) \) denotes the Kullback-Leibler divergence. Equation (16) can be expressed in closed form. We use a Monte Carlo approximation of the expectation by computing the model-predicted gain covariance. Given that we enforce the gain to have unit-mean, this covariance is diagonal.

\[
\mathcal{D}(\mathcal{P}(\mathbf{z}|\theta) || \mathcal{P}(\mathbf{N}|\theta)) = \frac{1}{2} \text{tr} \left[ \Sigma + \Sigma^T \right] - \log \left( \sqrt{2\pi e} \right)
\]

where \( \Sigma = \text{diag} \{ \lambda_1, \ldots, \lambda_L \} \) is the covariance matrix of the local variables. The diagonal covariance is

\[
\Sigma_i = \sum_j \sum_k \mathbf{M}_{ij} \mathbf{M}_{jk}^T
\]

where \( \mathbf{M}_{ij} \) is the jth column of the jth neuron's weight matrix, and \( \mathbf{M}_{jk} \) is the kth column of the kth neuron's weight matrix. The cross-covariance is

\[
\Sigma_{ij} = \sum_j \sum_k \mathbf{M}_{ij} \mathbf{M}_{jk}^T
\]

and the mean vector

\[
\mu = \sum_j \sum_k \mathbf{M}_{ij} \mathbf{M}_{jk}^T
\]

The discriminability of two nearby distributions can be expressed as

\[
\text{discr}(\mathbf{N}_1, \mathbf{N}_2) = \frac{1}{2} \text{tr} \left[ \Sigma + \Sigma^T \right] - \log \left( \sqrt{2\pi e} \right)
\]

where \( \Sigma = \text{diag} \{ \lambda_1, \ldots, \lambda_L \} \) is the covariance matrix of the local variables. The diagonal covariance is

\[
\Sigma_i = \sum_j \sum_k \mathbf{M}_{ij} \mathbf{M}_{jk}^T
\]

and the cross-covariance is

\[
\Sigma_{ij} = \sum_j \sum_k \mathbf{M}_{ij} \mathbf{M}_{jk}^T
\]

and the mean vector

\[
\mu = \sum_j \sum_k \mathbf{M}_{ij} \mathbf{M}_{jk}^T
\]

The discriminability of two nearby distributions can be expressed as

\[
\text{discr}(\mathbf{N}_1, \mathbf{N}_2) = \frac{1}{2} \text{tr} \left[ \Sigma + \Sigma^T \right] - \log \left( \sqrt{2\pi e} \right)
\]

where \( \Sigma = \text{diag} \{ \lambda_1, \ldots, \lambda_L \} \) is the covariance matrix of the local variables. The diagonal covariance is

\[
\Sigma_i = \sum_j \sum_k \mathbf{M}_{ij} \mathbf{M}_{jk}^T
\]

and the cross-covariance is

\[
\Sigma_{ij} = \sum_j \sum_k \mathbf{M}_{ij} \mathbf{M}_{jk}^T
\]

and the mean vector

\[
\mu = \sum_j \sum_k \mathbf{M}_{ij} \mathbf{M}_{jk}^T
\]
and a response scalar:

\[ R = \epsilon + \beta \frac{\sum_{i=1}^{n} w_i \max(0, L_i)}{\sigma^2 + \sum_{i=1}^{n} L_i^2} \]  

(24)

where \( \epsilon \) is spontaneous discharge rate, \( \beta \) controls the response range, \( w_i \) is a pooling weight, and \( L_i \) a linear filter’s response. The normalization signal consists of the sum of a stimulus-independent normalization constant \( \sigma^2 \) and the pooled activity of a diverse set of neurons. We set \( \sigma = 0.15 \), and approximated the latter term with the root mean square contrast of the stimulus.

In total, the LN–LN model has 11 free parameters: five filter parameters (preferred orientation \( \theta_i \), preferred spatial frequency \( v_i \), and phase \( \phi_i \), aspect ratio \( \alpha_i \), derivative order \( d \)), and direction selectivity \( \beta_i \), four pooling weights (\( w_i \), constrained to sum to one, thus yielding three free parameters), and two parameters controlling response range and amplitude (spontaneous discharge rate \( \epsilon \) and response scale \( \beta \)). For each cell, we first optimized the filter and response range parameters by maximizing the likelihood over the observed grating responses calculated by counting spikes in a time window whose duration matched that of the stimulus presentation (1000 ms). We obtained a likelihood by assuming that spike counts arise from a modulated Poisson process. We used a simplex algorithm and multi-start fitting procedure with semi-randomized starting values to find the best fitting parameters. Next, we estimated the four pooling weights \( w_i \) by fitting the temporal response modulation exhibited during stimulation with the most effective grating (see Fig. 7b for an example). We estimated the temporal evolution of the firing rate by counting spikes in 12.5 ms time windows, and used least-squares regression to find the best fitting pooling weights.

We analyzed the responses to the modified sparse-noise stimuli shown in experiment 2 to obtain an estimate of the spatial location of each cell’s receptive field center. Following the approach of ref. 5, we fit a 2-D Gaussian to the spatial response profile measured for black and white bars whose orientation approximated the cell’s preferred value. We used the center of the Gaussian as an estimate of the cell’s receptive field center.

Finally, we combined all of these parameter estimates to generate model predictions for the natural and unnatural stimuli of experiment 3. For each cell, we evaluated the model’s performance by computing the correlation between the predicted and observed curvature for each natural and unnatural movie, averaged across all populations (Fig. 7c).

Dissection of LN–LN model. To interrogate the role of each of the model’s elements, we created multiple versions of the model by manipulating its parameter values (Fig. 8). Specifically, we created a version without divisive normalization by adopting all the parameter estimates of the fitted model, but setting the denominator in Eq. (24) to one. We also created a version in which all cells were turned into phase-invariant complex cells by setting all the pooling weights \( w_i \) to 0. Finally, we created a version in which all cells were turned into simple cells by setting each cell’s highest pooling weight \( w_i \) to one and the three other weights to zero.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The data that support the findings of this study are available in the following repository: https://doi.org/10.17605/OSF.IO/VF2XX. Source data are provided with this paper.

Code availability

The analysis code that supports the findings of this study is available from the corresponding authors upon reasonable request.

Received: 24 July 2020; Accepted: 8 September 2021; Published online: 13 October 2021

References

1. Foldiák, P. Learning invariance from transformation sequences. Neural Comput. 3, 194–200 (1991).
2. Tishby, N., Pereira, F. C. & Bialek, W. The information bottleneck method. In Proc. 37th Annual Allerton Conference on Communication, Control and Computing (University of Illinois, Urbana, IL) Vol 37, 368–377, pages 1–16 (1999).
3. Li, N. & DiCarlo, J. J. Unsupervised natural visual experience rapidly reshapes size-invariant object representation in inferior temporal cortex. Neuron 67, 1062–1075 (2010).
4. Goroshin, R., Mathieu, M. & LeCun, Y. Learning to Linearize Under Uncertainty (NIPS, 2015).
5. Palmer, S. E., Marre, O., Berry, M. J. & Bialek, W. Predictive information in a sensory population. Proc. Natl Acad. Sci. USA 112, 6908–6913 (2015).
6. Hénafl, O. J., Goris, R. L. T. & Simoncelli, E. P. Perceptual straightening of natural videos. Nat. Neurosci. 22, 984–991 (2019).
7. Tolhurst, D. J., Movshon, J. A. & Dean, A. F. The statistical reliability of signals in single neurons in cat and monkey visual cortex. Vision Res. 23, 775–785 (1983).
8. Goris, R. L. T., Movshon, J. A. & Simoncelli, E. P. Partitioning neuronal variability. Nat. Neurosci. 17, 858–865 (2014).
9. Cohen, M. R. & Kohn, A. Measuring and interpreting neuronal correlations. Nat. Neurosci. 14, 811 (2011).
10. Moreno-Bote, R. et al. Information-limiting correlations. Nat. Neurosci. 17, 1410 (2014).
11. Elsayed, G. F. & Cunningham, J. P. Structure in neural population recordings: an expected byproduct of simpler phenomena? Nat. Neurosci. 20, 1310 (2017).
12. Movshon, N. C., Goris, R. L., Cohen, M. & Simoncelli, E. P. Slow gain fluctuations limit benefits of temporal integration in visual cortex. J. Vision 18, 8–8 (2018).
13. Poole, B., Lahiri, S., Raghu, M., Solo-Dickstein, J. & Ganguli, S. In Advances in Neural Information Processing Systems Vol 29 (Curran Associates, Inc., 2016).
14. Field, D. J. Relations between the statistics of natural images and the response properties of cortical cells. J. Opt. Soc. Am. A 4, 2379–2394 (1987).
15. Ruderman, D. L. & Bialek, W. In Advances in Neural Information Processing Systems 551–558 (1994).
16. Dong, D. W. & Atick, J. J. Statistics of natural time-varying images. Network: Comput. Neural Syst. 6, 345–358 (1995).
17. De Volois, R. L., Albrecht, D. G. & Thorell, L. G. Spatial frequency selectivity of cells in macaque visual cortex. Vision Res. 22, 545–559 (1982).
18. Toruyian, J., Felsen, G. & Dan, Y. Spatial structure of complex cell receptive fields measured with natural images. Neuron 45, 781–791 (2005).
19. Hubel, D. H. & Wiesel, T. N. Receptive fields, binocular interaction and functional architecture in the cat’s visual cortex. J. Physiol. 160, 106–154.2 (1962).
20. Movshon, J. A., Thompson, I. D. & Tolhurst, D. J. Receptive field organization of complex cells in the cat’s striate cortex. J. Physiol. 283, 79–99 (1978).
21. Adelson, E. H. & Bergen, J. R. Spatiotemporal energy models for the perception of motion. J. Opt. Soc. Am. A 2, 284 (1985).
22. Heeger, D. J. Normalization of cell responses in cat striate cortex. Visual Neurosci. 9, 181–197 (1992).
23. Carandini, M., Heeger, D. J. & Movshon, J. A. Linearity and normalization in simple cells of the macaque primary visual cortex. J. Neurosci. 17, 8621–8644 (1997).
24. Toruyian, J., Lau, B. & Dan, Y. Isolation of relevant visual features from random stimuli for cortical complex cells. J. Neurosci. 22, 10811–10818 (2002).
25. Sharp, J., Rust, N. C. & Bialek, W. Analyzing neural responses to natural signals: maximally informative dimensions. Neural Comput. 16, 223–250 (2004).
26. Rust, N. C., Schwartz, O., Movshon, J. A. & Simoncelli, E. P. Spatiotemporal elements of macaque V1 receptive fields. Neuron 46, 945–956 (2005).
27. Vintch, B., Movshon, J. A. & Simoncelli, E. P. A convolutional subunit model for neuronal responses in macaque v1. J. Neurosci. 35, 14829–14841 (2015).
28. Goris, R. L. T., Simoncelli, E. P. & Movshon, J. A. Origin and function of tuning diversity in macaque visual cortex. Neuron 88, 819–831 (2015).
29. Skottun, B. C. et al. Classifying simple and complex cells on the basis of response modulation. Vision Res. 31, 1078–1086 (1991).
30. Fukushima, K. Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol. Cybernetics 36, 193–202 (1980).
31. Riesenhuber, M. & Poggio, T. Hierarchical models of object recognition in cortex. Nat. Neurosci. 2, 1019–1025 (1999).
32. Wiskott, L. & Sejnowski, T. J. Slow feature analysis: unsupervised learning of invariances. Neural Comput. 14, 715–770 (2002).
33. Bruna, J. & Mallat, S. Invariant scattering convolution networks. IEEE Trans. Pattern Analysis Mach. Intell. 35, 1872–1886 (2013).
34. Russo, A. A. et al. Motor cortex embeds muscle-like commands in an untangled population response. Neuron 97, 953–966 (2018).
36. Russo, A. A. et al. Neural trajectories in the supplementary motor area and motor cortex exhibit distinct geometries, compatible with different classes of computation. *Neuron* **107**, 745–758 (2020).
37. Sohn, H., Narain, D., Meirhaeghe, N. & Jazayeri, M. Bayesian computation through cortical latent dynamics. *Neuron* **103**, 934–947 (2019).
38. Carandini, M. et al. Do we know what the early visual system does? *J. Neurosci.* **25**, 10577–10597 (2005).
39. Cavanaugh, J. R., Bair, W. & Movshon, J. A. Nature and interaction of signals from the receptive field center and surround in macaque v1 neurons. *J. Neurophysiol.* **88**, 2530–2546 (2002).
40. Schwartz, O. & Simoncelli, E. P. Natural signal statistics and sensory gain control. *Nat. Neurosci.* **4**, 819–825 (2001).
41. Coen-Cagli, R., Kohn, A. & Schwartz, O. Flexible gating of contextual influences in natural vision. *Nat. Neurosci.* **18**, 1648 (2015).
42. Mante, V., Bonin, V. & Carandini, M. Functional mechanisms shaping lateral geniculate responses to artificial and natural stimuli. *Neuron* **58**, 625–638 (2008).
43. Berardino, A., Ballé, J., Laparra, V. & Simoncelli, E. P. Eigen-distortions of hierarchical representations. *Advances in Neural Information Processing Systems* 30 (eds. Guyon, I. et al.) (Curran Associates, Inc., 2017).
44. Goris, R. L. T., Bai, Y. H., Henaff, O. J. & Ziemba, C. M. Perceptual straightening of natural videos arises from a cascaded computation. *CoSyNe Abstracts*, (III-27) (2020).
45. Yamins, D. L. K. et al. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl Acad. Sci. USA* **111**, 8619–8624 (2014).
46. Khaligh-Razavi, S. M. & Kriegeskorte, N. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput. Biol.* **10**, e1003915 (2014).
47. Zhuang, C. et al. Unsupervised neural network models of the ventral visual stream. *Proc. Natl Acad. Sci. USA* **118**, e2014196118 (2021).
48. Szegedy, C. et al. Intriguing properties of neural networks. In *Proc. 2nd International Conference on Learning Representations (ICLR)* (2013).
49. Henaff, O. J. & Simoncelli, E. P. Geodesics of learned representations. In *Proc. 4th International Conference on Learning Representations (ICLR)* (2015).
50. Feather, J., Durango, A., Gonzalez, R. & McDermott, J. In *Advances in Neural Information Processing Systems* (eds. Wallach, H. et al.) 10078–10089 (Curran Associates, Inc., 2019).
51. Singer, Y. et al. Sensory cortex is optimized for prediction of future input. *Elife* **7**, e31557 (2018).
52. Brainard, D. H. The psychophysics toolbox. *Spatial Vision* **10**, 433–466 (1997).
53. Pelli, D. G. The video toolbox software for visual psychophysics: transforming numbers into movies. *Spatial Vision* **10**, 437–442 (1997).
54. Nauhaus, I., Nielsen, K. J. & Callaway, E. M. Efficient receptive field tiling in primate v1. *Neuron* **91**, 893–904 (2016).
55. Seshadrinathan, K., Soundararajan, R., Bovik, A. C. & Cormack, L. K. A Subjective study to evaluate video quality assessment algorithms. In *SPIE Proceedings Human Vision and Electronic Imaging* (2010).
56. Seshadrinathan, K., Soundararajan, R., Bovik, A. C. & Cormack, L. K. Study of subjective and objective quality assessment of video. *IEEE Trans. Image Process.* **19**, 1427–1441 (2010).
57. Pachitariu, M., Steinmetz, N. A., Kadir, S. N., Carandini, M. & Harris, K. D. In *Advances in Neural Information Processing Systems* (eds. Lee, D. D. et al.) 4448–4456 (Curran Associates, Inc., 2016).
58. Smith, M. A., Majaj, N. J. & Movshon, J. A. Dynamics of motion signaling by neurons in macaque area mt. *Nat. Neurosci.* **8**, 220–228 (2005).
59. Jordan, M. I., Ghahramani, Z., Jaakkola, T. S. & Saul, L. K. Introduction to variational methods for graphical models. *Mach. Learning* **37**, 183–233 (1999).
60. Kingma, D. P. & Welling, M. Auto-encoding variational Bayes. In *Proc. 2nd International Conference on Learning Representations (ICLR)* (2013).
61. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. In *Proc. 3rd International Conference on Learning Representations (ICLR)* (2014).

**Acknowledgements**

The authors thank Corey Ziemba and Zoe Boundy-Singer for help with data-collection and valuable discussions. This work was supported by UT Austin (startup funds provided to RLTG and IN), and the Howard Hughes Medical Institute (Investigatorship to EPS).

**Author contributions**

O.J.H., Y.H.B., E.P.S., and R.L.T.G. conceived and designed the study. J.N. developed the physiology setup and performed experimental preparations. Y.H.B., J.A.C., and R.L.T.G. collected data. O.J.H. and E.P.S. developed the curvature estimation procedure. O.J.H. and Y.H.B. analyzed data. O.J.H. and R.L.T.G. wrote the manuscript with input from Y.H.B. and E.P.S.

**Competing interests**

The authors declare no competing interests.

**Additional information**

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41467-021-25939-z.

**Correspondence** and requests for materials should be addressed to Olivier J. Henaff, Yoon Bai or Robbe L. T. Goris.

**Peer review information** *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

**Reprints and permission information** is available at http://www.nature.com/reprints

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

---

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2021