Hierarchical temporal prediction captures motion processing from retina to higher visual cortex

Yosef Singer*, Ben D. B. Willmore, Andrew J. King, Nicol S. Harper*

Department of Physiology, Anatomy and Genetics, University of Oxford, Sherrington Building, Parks Road, Oxford OX1 3PT, United Kingdom.

*Corresponding authors: yosef.singer@stcatz.ox.ac.uk; nicol.harper@dpag.ox.ac.uk.

Visual neurons respond selectively to features that become increasingly complex in their form and dynamics from the eyes to the cortex. These features take specific forms: retinal neurons prefer localized flashing dots[1], primary visual cortical (V1) neurons moving bars[2–4], and those in higher cortical areas, such as middle temporal (MT) cortex, favor complex features like moving textures[5–7]. Whether there are general principles behind this diverse complexity of response properties in the visual system has been an area of intense investigation. To date, no single normative model has been able to account for the hierarchy of tuning to dynamic inputs along the visual pathway. Here we show that hierarchical temporal prediction - representing features that efficiently predict future sensory input from past sensory input[8–11] - can explain how neuronal tuning properties, particularly those relating to motion, change from retina to higher visual cortex. In contrast to some other approaches[12–16], the temporal prediction framework learns to represent features of unlabeled and dynamic stimuli, an essential requirement of the real brain. This suggests that the brain may not have evolved to efficiently represent all incoming stimuli, as implied by some leading theories. Instead, the selective representation of sensory features that help in predicting the future may be a general coding principle for extracting temporally-structured features that depend on increasingly high-level statistics of the visual input.

Introduction

Temporal prediction[10] relates to a class of principles, such as information bottleneck[8,9,11] and slow feature analysis[17], that involve selectively encoding only those features that are efficiently predictive of the future. Such principles have value in finding features that can guide future action, uncovering underlying variables, and discarding irrelevant information[8,10]. This class of principles differs from others that are more typically used to explain sensory coding – efficient coding[15,18], sparse coding[13] and predictive coding[16,19] – that aim instead to efficiently represent all current and perhaps past input. Although these principles have been successful in accounting for visual receptive field properties in areas such as V1[10,11,13,15–17], it remains to be demonstrated whether a single principle can explain the diverse spatiotemporal tuning that emerges along the dorsal visual stream, which is responsible for the processing of complex object motion.

Any general principle of visual encoding needs to explain temporal aspects of neural tuning – the encoding of movies rather than static images. It is also important that any general principle is largely unsupervised. Some features of the visual system have been reproduced by deep supervised network models optimized for image classification on large labelled datasets[12]. While these models

can help explain the likely hard-wired retina[20], they are less informative if neuronal tuning is influenced by experience, as in cortex, since most sensory input is unlabeled except for sporadic reinforcement signals. The temporal prediction approach is unsupervised (requires no labelled data), inherently applies over the temporal domain, and can account for temporal aspects of V1 simple cell receptive fields[10]. We therefore investigated whether this principle can furthermore account for the emergence of motion processing along the visual pathway, from retina to higher cortical areas.

We instantiated temporal prediction as a hierarchical model consisting of stacked feedforward single-hidden-layer convolutional neural networks (Fig. 1). The first stack was trained to predict the immediate future frame (40 ms) of unfiltered natural video inputs from the previous 5 frames (200 ms). Each subsequent stack was then trained to predict the future hidden-unit activity of the stack below it from the past activity in response to the natural video inputs. The four stacks contained 50, 100, 200 and 400 hidden units, respectively. L1 regularization was applied to the weights of each stack, akin to a constraint on neural wiring.

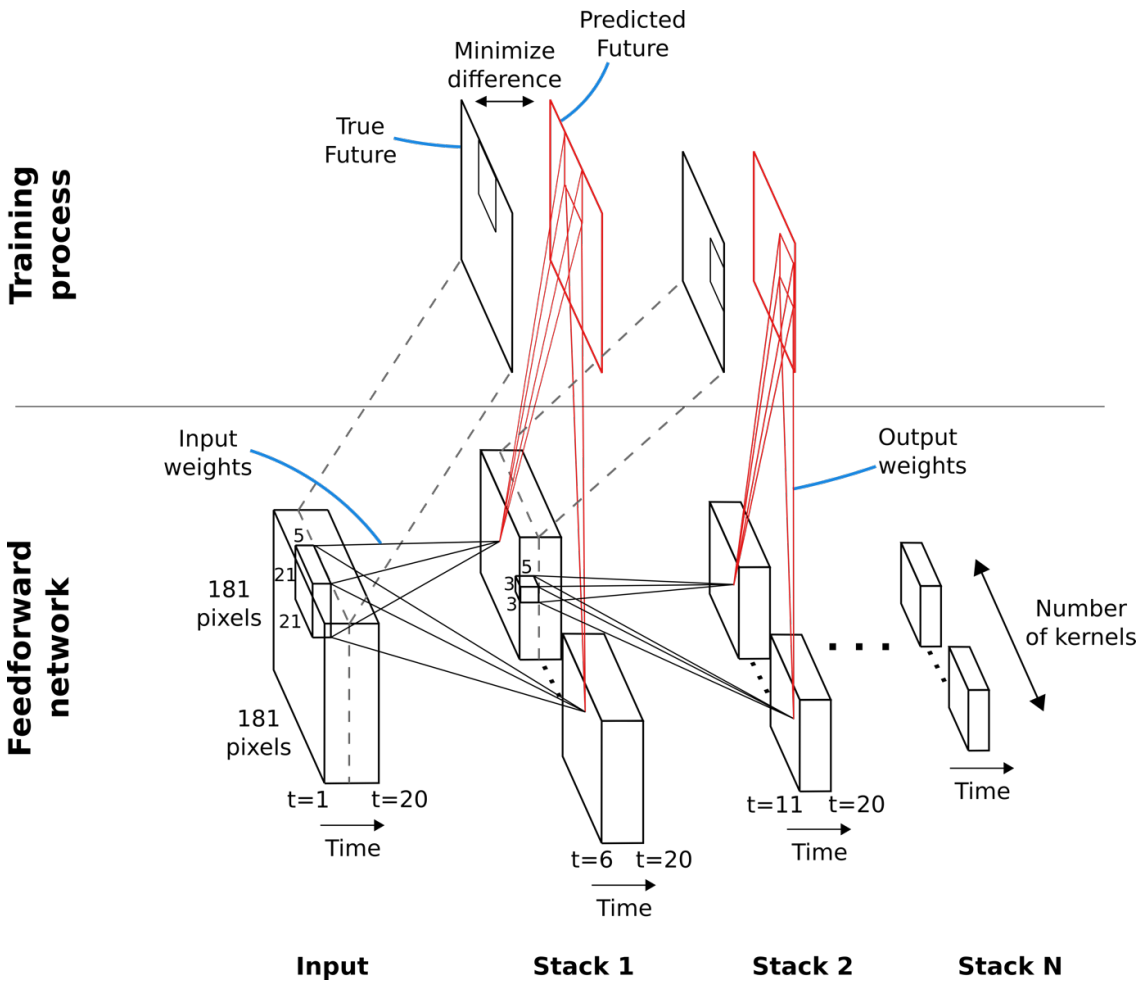


Figure 1 | Hierarchical temporal prediction model. Schematic of model architecture. Each stack is a single hidden-layer feedforward convolutional network, which is trained to predict the future time-step of its input from the previous 5 time-steps. The first stack is trained to predict future pixels of natural video inputs from their past. Subsequent stacks are trained to predict future time-steps of the hidden-layer activity in the stack below, given their past responses to the same natural video inputs.

Results and Discussion

After training, we examined the input weights of the units in the first stack. Each hidden unit can be viewed as a linear non-linear (LN) model[21,22], as commonly used to describe neuronal receptive fields. With L1 regularization slightly above the optimum for prediction, the receptive fields of the units showed spatially-localized center-surround tuning with a decaying temporal envelope, characteristic of retinal and lateral geniculate nucleus (LGN) neurons[1,23,24]. The model units have either an excitatory center with a suppressive surround (ON units) or vice versa (OFF units). Both ON and OFF units can have either small RFs that do not change polarity (Fig. 2a, units 3-6; Fig. 2b, bottom left) over time, or large RFs that switch polarity over time (Fig. 2a, units 1-2; Fig. 2b, top right). This is reminiscent of the four main cell types in the primate retina and LGN: the parvocellular-pathway ON/OFF neurons and the more change-sensitive magnocellular-pathway ON/OFF neurons, respectively[24].

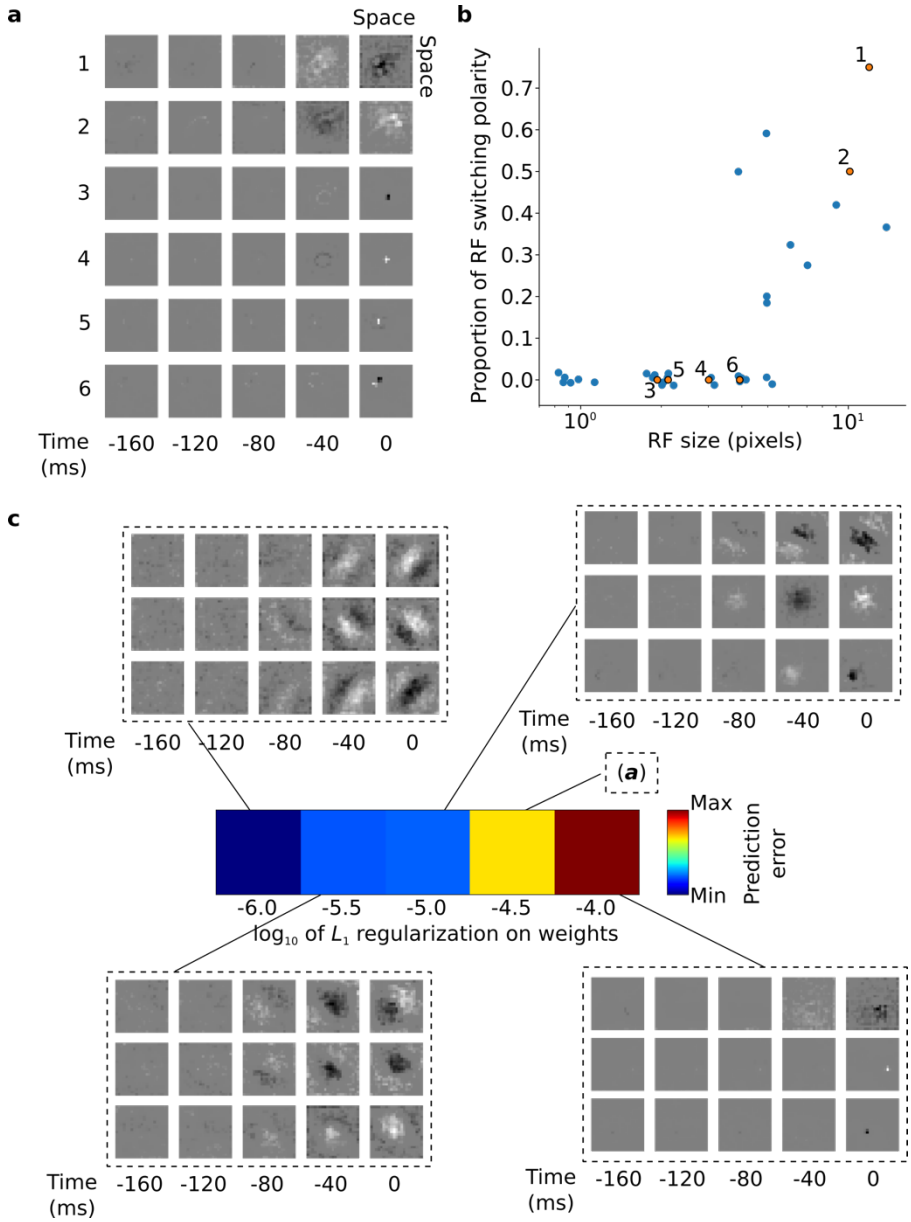


Figure 2 | RFs of trained first stack of the model show retina-like tuning. a, Example RFs with center-surround tuning characteristic of neurons in retina and LGN. RFs are large and switch polarity over time (units

1-2) or small and do not switch polarity (units 3-6), resembling cells along the magnocellular and parvocellular pathways, respectively. **b,** RF size plotted against proportion of the pixels in the RF that switch polarity over the course of the most recent two timesteps. Units in (a) labelled and shown in orange. **c,** Effect of changing regularization strength on the qualitative properties of RFs.

Interestingly, simply decreasing L1-regularization strength causes the model RFs to change from center-surround tuning to Gabor-like tuning, resembling localized oriented bars that shift over time (Fig. 2c). It is possible that this balance, between a code that is optimal for prediction and one that prioritizes efficient wiring, might underlie differences in the retina and LGN of different species. The retina of mice and rabbits contains many neurons with oriented and direction-tuned RFs, whereas cats and macaques mostly have center-surround RFs[25]. Efficient retinal wiring may be more important in some species, due, for example, to different constraints on the width of the optic nerve or different impacts of light scattering by superficial retinal cell layers.

Using the trained center-surround-tuned network as the first stack, a second stack was added to the model and trained. The output of each second stack unit results from a linear-nonlinear-linear-nonlinear transformation of the visual input, and hence we estimated their RFs by reverse correlation with Gaussian noise input. The resulting RFs were Gabor-like over space, resembling those of V1 simple cells[2–4]. The RF envelopes decayed into the past, and often showed spatial shifts or polarity changes over time, indicating direction or flicker sensitivity, as is also seen in V1[26] (Fig. 3a,b, I; Supplementary Fig. 1). Using full-field drifting sinusoidal gratings (Fig. 3a,b II), we found that most units were selective for stimulus orientation, spatial and temporal frequency (Fig. 3a,b, IV-VI), and some were also direction selective (Fig. 3b). Responses to the optimum grating typically oscillate over time between a maximum when the grating is in phase with the RF and 0 when the grating is out of phase (Fig. 3a,b, III). These response characteristics are typical of V1 simple cells[27].

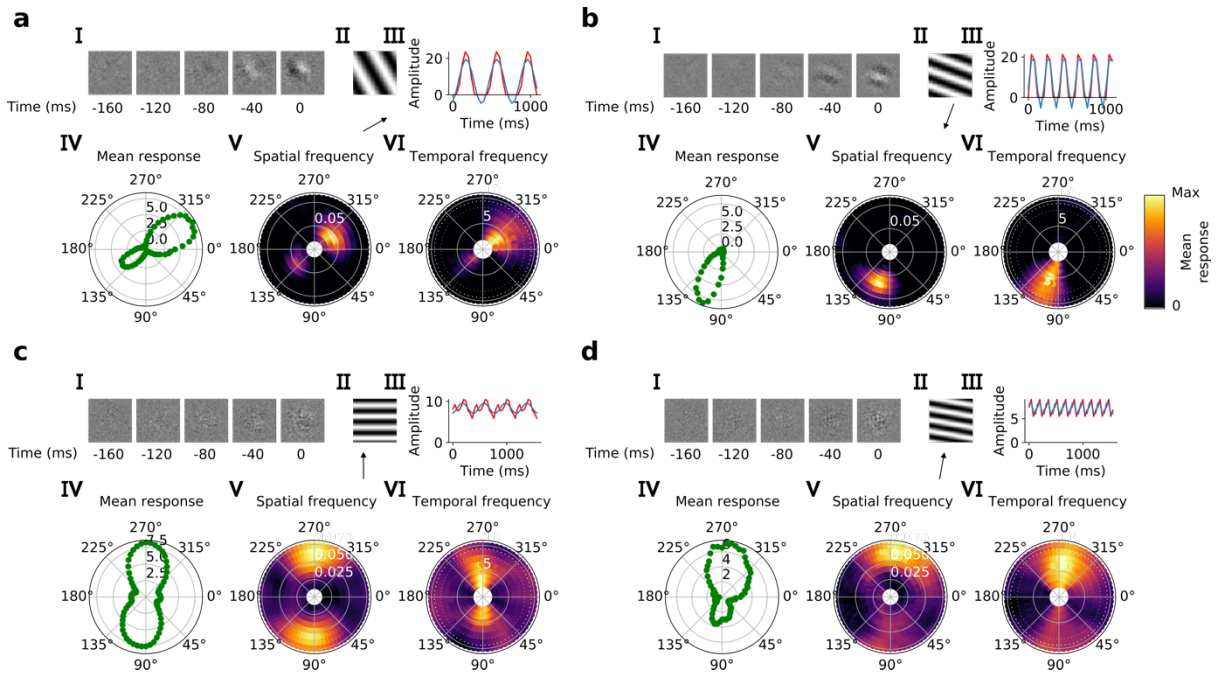


Figure 3 | Qualitative tuning properties of model units in stacks 2-3. **a,b,** Tuning properties of example units from the 2nd stack of the model, including (I) the linear RF, (II) the drifting grating that best stimulates this unit and (III) the amplitude of the unit's response to this grating over time. Red: unit response; blue: best-fitting sinusoid. (IV) The unit's mean response over time plotted against orientation for gratings presented at the optimal spatial and temporal frequency of this unit. (V,VI) Tuning curves showing the joint distribution of

responses to (V) orientation and spatial frequency at the preferred temporal frequency and to (VI) orientation and temporal frequency at the preferred spatial frequency. In V and VI the color represents the mean response over time to the grating presented. **c,d**, As in (a,b) for example units in stack 3.

In the third and fourth stack, we followed the same procedures as in the second stack. Most of these units are also tuned for orientation, spatial frequency, temporal frequency and in some cases for direction (Fig. 3c,d, IV-VI; Supplementary Figs. 2-3). However, while some units resembled simple cells, most resembled the complex cells of V1 and secondary visual areas (V2/V3)[3]. Complex cells are tuned for orientation and spatial and temporal frequency, but are relatively invariant to the phase of the optimum grating[28]; each cell's response to its optimum grating has a high mean value and changes little with the grating's phase (Fig. 3c,d, II,III). Whether a neuron is assigned as simple or complex is typically based on the modulation ratio in such plots (<1 indicates complex)[29]. Model units with low modulation ratios had little discernible structure in their RFs (Fig. 3c,d, I), another characteristic feature of V1 complex cells[30,31].

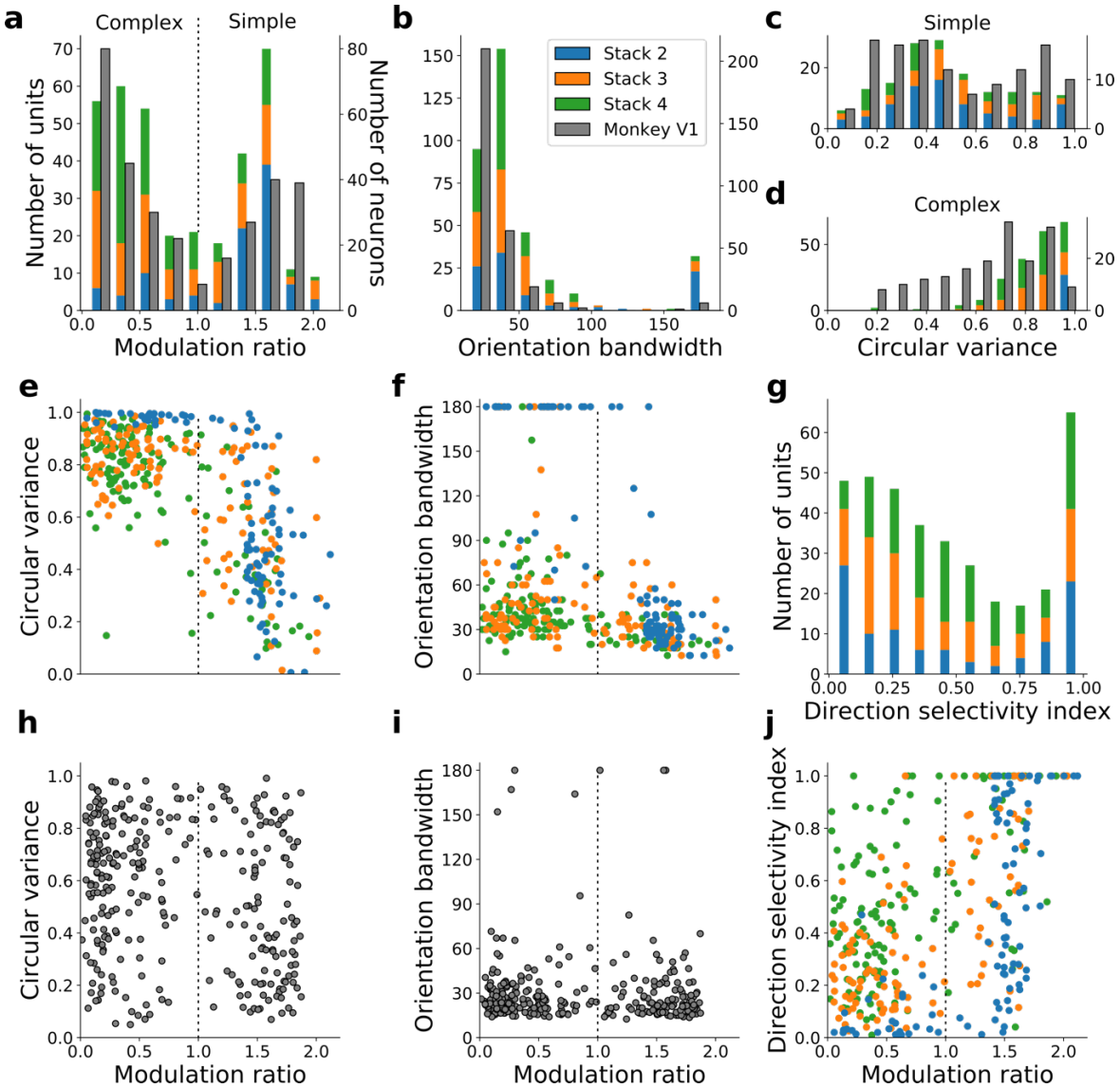


Figure 4 | Quantitative tuning properties of model units in stacks 2-4 in response to drifting sinusoidal gratings and corresponding measures of macaque V1 neurons. a-d, g, Histograms showing tuning properties of model and macaque V1 neurons as measured using drifting gratings. **e, f, h-j,** joint distributions for tuning measures.

We quantified the tuning characteristics of units in stacks 2-4 and compared them to published V1 data[32] (Fig. 4a-j). The distribution of modulation ratios is bimodal in both V1[29,32] and our model (Fig. 4a). Both model and real neurons were typically orientation selective with the model units more biased towards weaker tuning, as measured by orientation bandwidth (median data[32]: 23.5°, model: 37.5°; Fig 4b) and circular variance (median data[32]: simple cells 0.44, complex cells 0.69; median model: simple cells 0.46, complex cells 0.87; Fig 4c,d). Orientation-tuned units (circular variance < 0.9) units in the second stack were exclusively simple (modulation ratios > 1), whereas those in subsequent stacks become increasingly complex (Fig. 4a,c-f). In both model and data, circular variance is inversely correlated with the modulation ratio (Fig. 4c-e,h). As in V1, model units showed a range of direction selectivity preferences, with simple cells biased towards being direction tuned (direction selectivity index, DSI ≥ 0.5; layer 4 cat V1[33] 64%, n=220, model 68%, n=156), while the DSI values of complex cells tend to be more evenly distributed in V1[33] and lower in the model (22% with DSI ≥ 0.5, n=205; Fig. 4g,j).

Simple and complex cells extract many dynamic features from natural scenes. However, their small RFs prevent individual neurons from tracking the motion of complex objects because of the aperture problem; the direction of motion of an edge is ambiguous, with only the component of motion perpendicular to the cell's preferred orientation being represented. Two classes of neurons exist that can recover 2-dimensional motion information and overcome the aperture problem. End-stopped neurons, found in primary and secondary visual areas, respond unambiguously to the direction of motion of endpoints of restricted moving contours[34]. Pattern-selective neurons, in MT of primates, solve the problem for the more general case, likely by integrating over input from many direction-selective V1 complex cells[5,7,35–37], and hence respond selectively to the motion of patterns as a whole.

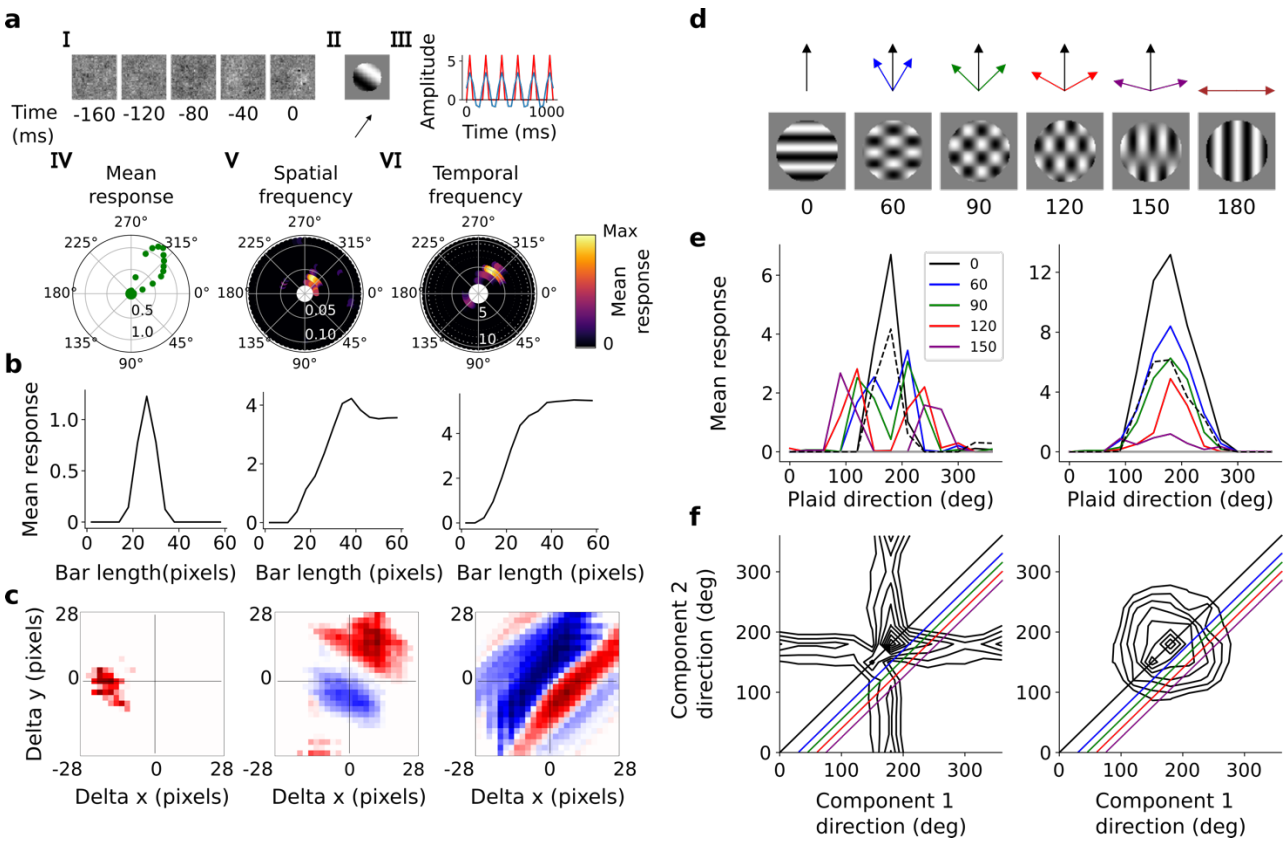


Figure 5 | Tuning to complex motion. a, Example end-stopped model unit. I-VI as in Figure 3. **b**, Response

as a function of bar length for the unit in *a* (left) and two other example units (middle, right). **c**, 2-bar maps of units with corresponding bar-length tuning plots shown in **b**. **d,** Example plaid stimuli used to measure pattern selectivity. Black arrow, direction of pattern motion. Colored arrows, directions of component motion. **e,** Direction tuning curves showing the response of an example component-selective (left) and pattern-selective (right) unit to grating and plaid stimuli. Colored lines, response to plaid stimuli composed of gratings at the given angle. Black solid line, unit's response to double intensity grating moving in the same direction as plaids. Dotted line, response to single intensity grating moving in the same direction. **f,** Surface contour plots showing response of units in **e** to plaids as a function of the direction of the grating components. Colored lines denote loci of plaids whose responses are shown in the same colors in **e**. Contour lines range from 20% of the maximum response to the maximum in steps of 10%. For clarity, all direction tuning curves are rotated so that the preferred direction of the response to the optimal grating is at 180˚. Responses (except **a**, III) are mean amplitudes over time.

To investigate end-stopping in our model units, we applied circular masks of different sizes to the grating stimuli. Some units displayed end-stopping, responding most strongly to gratings with an intermediate mask radius, with the response decreasing as the radius increased beyond this (Fig. 5a,b). To determine whether these end-stopped units unambiguously represent the direction of motion of end-points, we measured two-bar response maps[34], which determine response dependence on the horizontal and vertical components of motion (see Methods). Recordings from V1 indicate that more strongly end-stopped neurons have a weak tendency for less ambiguous motion tuning in these maps[34]. Consistent with this, our model produces examples of end-stopped units with less ambiguous motion tuning (Fig. 5b,c, first two panels) and non-end-stopped units with more ambiguous motion tuning (Fig. 5b,c, last panel).

To investigate pattern selectivity in our model units, we measured their responses to drifting plaids, comprising two superimposed drifting sinusoidal gratings with different orientations. The net direction of the plaid movement lies midway between these two orientations (Fig. 5d). In V1 and MT, component-selective cells respond maximally when the plaid is oriented such that either one of its component gratings moves in the preferred direction of the cell (as measured by a drifting grating). This results in two peaks in plaid-direction tuning curves[5–7]. Conversely, pattern-selective cells have a single peak in their direction tuning curves, when the plaid's direction of movement aligns with the preferred direction of the cell[5–7]. We see examples of both component-selective units (in stacks 2-4) and pattern-selective units (only in stack 4) in our model, as indicated by plaid-direction tuning curves (Fig. 5e) and plots of response as a function of the directions of each component (Fig. 5f).

Previous non-hierarchical models of retina or primary cortex alone have demonstrated retina-like[38], simple-cell like[10], or primary auditory cortical features[10] using principles related to temporal prediction. However, any general principle of neural representation should be extendable to a hierarchical form and not tailored to the region it is attempting to explain. Here we show that temporal prediction can indeed be made hierarchical and so reproduce the major motion-tuning properties that emerge along the dorsal visual pathway from retina to MT. This includes center-surround retinal features, direction-selective simple and complex cells, and complex-motion sensitive units resembling end-stopped and pattern-sensitive neurons. This suggests that, by learning behaviorally-useful features from dynamic unlabeled data, temporal prediction may represent a fundamental coding principle in the brain.

References

1. Kuffler, S. W. Discharge patterns and functional organization of mammalian retina. *J. Neurophysiol.* **16,** 37–68 (1953).

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

3. Hubel, D. H. & Wiesel, T. N. Receptive fields and the functional architecture in two striate visual areas (18 and 19) of the cat. *J. Neurophysiol.* **28,** 229–89 (1965).

4. Hubel, D. H. & Wiesel, T. N. Receptive fields of single neurones in the cat's striate cortex. *J. Physiol.* **148,** 574–591 (1959).

5. Movshon, J. A., Adelson, E. H., Gizzi, M. S. & Newsome, W. T. The analysis of moving visual patterns. In *Pattern Recognition Mechanisms* (eds. Chagas, C., Gattass, R. & Gross, C.) 117–151 (Vatican Press, 1985).

6. 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).

7. Rust, N. C., Mante, V., Simoncelli, E. P. & Movshon, J. A. How MT cells analyze the motion of visual patterns. *Nat. Neurosci.* **9,** 1421–1431 (2006).

8. Bialek, W., Nemenman, I. & Tishby, N. Predictability, complexity, and learning. *Neural Comput.* **13,** 2409–63 (2001).

9. Salisbury, J. M. & Palmer, S. E. Optimal prediction in the retina and natural motion statistics. *J. Stat. Phys.* **162,** 1309–1323 (2016).

10. Singer, Y. *et al.* Sensory cortex is optimized for prediction of future input. *Elife* **7,** e31557 (2018).

11. Chalk, M., Marre, O. & Tkačik, G. Toward a unified theory of efficient, predictive, and sparse coding. *Proc. Natl. Acad. Sci. U. S. A.* **115,** 186–191 (2018).

12. Yamins, D. L. K. & DiCarlo, J. J. Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* **19,** 356–365 (2016).

13. Olshausen, B. A. & Field, D. J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **381,** 607–609 (1996).

14. Olshausen, B. A. & Field, D. J. Sparse coding with an overcomplete basis set: a strategy employed by V1? *Vision Research* **37,** 3311–3325 (1997).

15. Bell, A. J. & Sejnowski, T. J. The 'independent components' of natural scenes are edge filters. *Vision Res.* **37,** 3327–3338 (1997).

16. Rao, R. P. & Ballard, D. H. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* **2,** 79–87 (1999).

17. Berkes, P. & Wiskott, L. Slow feature analysis yields a rich repertoire of complex cell properties. *J. Vis.* **5,** 9–9 (2005).

18. Barlow, H. B. in *The Mechanisation of Thought Processes* (eds. Blake, D. V. & Uttley, A. M.) 535–539 (H. M. Stationary Office, 1959).

19. Huang, Y. & Rao, R. P. N. Predictive coding. *WIREs.* **2,** 580–593 (2011).

20. Lindsey, J., Ocko, S. A., Ganguli, S. & Deny, S. A unified theory of early visual representations from retina to cortex through anatomically constrained deep CNNs. *arXiv:1901.00945* (2019).

21. Chichilnisky, E. J. A simple white noise analysis of neuronal light responses. *Network* **12,** 199–213 (2001).

22. Simoncelli, E., Pillow, J. W., Paninski, L. & Schwartz, O. Characterization of neural responses with stochastic stimuli. In *The cognitive neurosciences, III* (ed. Gazzaniga, M.) 327–338 (MIT Press, 2004).

23. Barlow, B. Y. H. B. Summation and Inhibition in the Frog's Retina. *J. Physiol* 69–88 (1953).

24. Shapley, R. & Hugh Perry, V. Cat and monkey retinal ganglion cells and their visual functional roles. *Trends Neurosci.* **9,** 229–235 (1986).

25. Scholl, B., Tan, A. Y. Y., Corey, J. & Priebe, N. J. Emergence of orientation selectivity in the mammalian visual pathway. *J. Neurosci.* **33,** 10616–10624 (2013).

26. DeAngelis, G. C., Ohzawa, I. & Freeman, R. D. Spatiotemporal organization of simple-cell receptive fields in the cat's striate cortex. I. General characteristics and postnatal development. *J. Neurophysiol.* **69,** 1091–1117 (1993).

27. Movshon, J. A., Thompson, I. D. & Tolhurst, D. J. Spatial summation in the receptive fields of simple cells in the cat's striate cortex. *J. Physiol.* **283,** 53–77 (1978).

28. 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).

29. Skottun, B. C. *et al.* Classifying simple and complex cells on the basis of response modulation. *Vision Res.* **31,** 1079–1086 (1991).

30. Rust, N. C., Schwartz, O., Movshon, J. A. & Simoncelli, E. P. Spatiotemporal elements of macaque V1 receptive fields. *Neuron* **46,** 945–956 (2005).

31. Schwartz, O., Pillow, J. W., Rust, N. C. & Simoncelli, E. P. Spike-triggered neural characterization. *J. Vis.* **6,** 13 (2006).

32. Ringach, D. L., Shapley, R. M. & Hawken, M. J. Orientation selectivity in macaque V1: diversity and laminar dependence. *J. Neurosci.* **22,** 5639–51 (2002).

33. Kim, T. & Freeman, R. D. Direction selectivity of neurons in the visual cortex is non-linear and lamina-dependent. *Eur. J. Neurosci.* **43,** 1389–99 (2016).

34. Pack, C. C., Livingstone, M. S., Duffy, K. R. & Born, R. T. End-stopping and the aperture problem: two-dimensional motion signals in macaque V1. *Neuron* **39,** 671–680 (2003).

35. Movshon, J. A. & Newsome, W. T. Visual response properties of striate cortical neurons projecting to area MT in macaque monkeys. *J. Neurosci.* **16,** 7733–7741 (1996).

36. Zeki, S. M. Functional organization of a visual area in the posterior bank of the superior temporal sulcus of the rhesus monkey. *J. Physiol.* **236,** 549–573 (1974).

37. Simoncelli, E. P. & Heeger, D. J. A model of neuronal responses in visual area MT. *Vision Res.* **38,** 743–761 (1998).

38. Ocko, S. A., Lindsey, J., Ganguli, S. & Deny, S. The emergence of multiple retinal cell types through efficient coding of natural movies. *bioRxiv* 458737 (2018).

39. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980* (2014).

Methods

Data used for model training and testing

Visual inputs

Videos (grayscale, without sound, sampled at 25 fps) of wildlife in natural settings were used to create visual stimuli for training the artificial neural network. The videos were obtained from http://www.arkive.org/species, contributed by: BBC Natural History Unit, http://www.gettyimages.co.uk/footage/bbcmotiongallery; BBC Natural History Unit & Discovery Communications Inc., http://www.bbcmotiongallery.com; Granada Wild, http://www.itnsource.com; Mark Deeble & Victoria Stone Flat Dog Productions Ltd., http://www.deeblestone.com; Getty Images, http://www.gettyimages.com; National Geographic Digital Motion, http://www.ngdigitalmotion.com. The longest dimension of each video frame was clipped to form a square image. Each frame was then down-sampled (using bilinear interpolation) over space, to provide 180 x 180 pixel frames. The video patches were cut into non-overlapping clips, each of 20 frames duration (800 ms). We used a training set of N = ~1305 clips from around 17 min of video, and a validation set of N = ~145 clips. Finally, each clip was normalized by subtracting the mean and dividing by the standard deviation of that clip.

Hierarchical temporal prediction model

The model and cost function

The hierarchical temporal prediction model consisted of stacked feedforward single-hidden-layer 3D convolutional neural networks. Each stack consisted of an input layer, a convolutional hidden layer and a 'transposed convolutional' output layer. Each unit (convolutional kernel) in the hidden layer performed 3D convolution over its inputs (over time and 2D space; Figure 1) and its output was determined by passing the result of this operation through a rectified linear function. Following the hidden layer there was a 'transposed convolutional' output layer, which again performed convolution (and dilation for stride >1). Each stack was trained to minimize the difference between its output and its target. The target was the input at the immediate future time-step.

The first stack of the model was trained to predict the immediate future frame (40 ms) of unfiltered natural video inputs from the previous 5 frames (200 ms). Each subsequent stack was then trained to predict the immediate future hidden-unit activity of the stack below it from the past hidden-unit activity in response to the natural video inputs. This process was repeated until 4 stacks had been trained. The first stack used 50 hidden units and this number was doubled with each added stack, until we had 400 units in the 4th stack.

More formally, each stack of model can be described by a network of the same form. The input to the network has i = 1 to to I input channels. For channel i, for clip n, the input $\underline{\mathbf{U}}_{in}$ is a rank-3 tensor spanning time and 2D-space with x = 1 to X and y = 1 to Y spatial positions, and t = 1 to T time steps. Throughout the Methods, capital, bold and underlined variables are rank-3 tensors over time and 2D-space, otherwise variables are scalars. The first stack has only a single input channel (the grayscale input frames, I = 1). Each subsequent stack had as many input channels (I) as the number of hidden units (feature maps) in the previous stack.

The network has a single hidden layer of $j = 1$ to J convolutional kernels. For clip n and kernel j, the output of each kernel is a rank-3 tensor over time and 2D-space, $\underline{\mathbf{H}}_{jn}$:

$$\underline{\mathbf{H}}_{jn} = \mathrm{f}\left(b_j + \sum_{i=1}^{I} \underline{\mathbf{U}}_{in} * \underline{\mathbf{W}}_{ji} \right) \tag{1}$$

The parameters in Equation 1 are the connective input weights of kernels $\underline{\mathbf{W}}_{ji}$ (between each input channel i and hidden unit j) and the bias b_j (of hidden unit j). f( ) is the rectified linear function and * is the 3D convolutional operator over the 2 spatial and 1 temporal dimensions of the input, with stride (s_1, s_2, s_3). Each hidden layer kernel $\underline{\mathbf{W}}_{ji}$ is 3D with size (X', Y', T'). No zero padding is applied to the input.

The output of the network predicts the future activity of the input. Hence, the number of input channels (I) always equals the number of output channels (K) for each stack. To ensure that the predicted output has the same size as the input when a stride >1 is used, the hidden layer representation is dilated by adding s-1 zeros between adjacent input elements, where $s = (s_1, s_2, s_3)$ is the stride of the convolutional operator in the hidden layer. The dilated hidden-unit outputs is $\underline{\mathbf{H}}_{jn}^{\mathrm{dil}}$. When stride=1, $\underline{\mathbf{H}}_{jn}^{\mathrm{dil}} = \underline{\mathbf{H}}_{jn}$.

The activity $\underline{\widehat{\mathbf{V}}}_{kn}$ of each output channel k is the estimate of the true future $\underline{\mathbf{V}}_{kn}$ given the past $\underline{\mathbf{U}}_{in}$. $\underline{\mathbf{V}}_{kn}$ is simply $\underline{\mathbf{U}}_{in}$ shifted 1 time step into the future, and $k = i$. This prediction $\underline{\widehat{\mathbf{V}}}_{kn}$ is estimated from the hidden unit output $\underline{\mathbf{H}}_{jn}^{\mathrm{dil}}$ by:

$$\underline{\widehat{\mathbf{V}}}_{kn} = b_k + \sum_{j=1}^{J} \underline{\mathbf{H}}_{jn}^{\mathrm{dil}} * \underline{\mathbf{W}}_{kj} \tag{2}$$

The parameters in Equation 2 are the connective output kernels $\underline{\mathbf{W}}_{kj}$ (the weights between each hidden unit j and output channel k) and the bias b_k. Each output kernel $\underline{\mathbf{W}}_{kj}$ is 3D with size ($X', Y', 1$), predicting a single time-step into the future based on hidden layer activity in that portion of space.

The parameters $\underline{\mathbf{W}}_{ji}$, $\underline{\mathbf{W}}_{kj}$, b_j, and b_k were optimized for the training set by minimizing the cost function given by:

$$E = \frac{1}{NKXYT} \sum_{n=1}^{N} \sum_{k=1}^{K} \left\| \underline{\widehat{\mathbf{V}}}_{kn} - \underline{\mathbf{V}}_{kn} \right\|_2^2 + \lambda \left(\sum_{i=1}^{I} \sum_{j=1}^{J} \left\| \underline{\mathbf{W}}_{\mathrm{ji}} \right\|_1 + \sum_{j=1}^{J} \sum_{k=1}^{K} \left\| \underline{\mathbf{W}}_{kj} \right\|_1 \right) \tag{3}$$

Where $\| \ \|_p$ is the entrywise p-norm of the tensor over time and 2D-space, p = 2 is the sqrt of the sum of squares of all values in the tensor, and p = 1 is the sum of absolute values. Thus, E is the sum of the squared error between the prediction $\underline{\widehat{\mathbf{V}}}_{kn}$ and the target $\underline{\mathbf{V}}_{kn}$, plus an L_1-norm regularization term, which is proportional to the sum of absolute values of all weights in the network and its strength is determined by the hyper-parameter λ. This regularization tends to drive redundant weights to near zero and provides a parsimonious network.

Implementation details

The networks were implemented in Python (https://lasagne.readthedocs.io/en/latest/; http://deeplearning.net/software/theano/). The objective function was minimized using backpropagation as performed by the Adam optimization method[39] with hyperparameters β1 and β2 kept at their default settings of 0.9 and 0.999, respectively, and the learning rate (α) varied as detailed below. Training examples were split into minibatches of 32 training examples each.

During model network training, several hyperparameters were varied, including the regularization strength (λ) and the learning rate (α). For each hyperparameter setting, the training algorithm was run for 1000 iterations. The effect of varying λ on the prediction error (the first term of Equation 3) and receptive field structure of the first stack is shown in Fig. 2. For all subsequent stacks, the results presented are the networks that predicted best on the validation set after 1000 iterations through the training data. The settings for each stack are presented in Table 1:

Table 1: Model parameter settings for each stack

Stack	Input size (X,Y,T,I)	Hidden layer size	Kernel size (X',Y',T')	Spatial and temporal extent	Stride (s_1,s_2,s_3)	Number of kernels	Learning rate (α)	L1 Regularization strength (λ)
1	181x181x20x1	17x17x16x50	21x21x5	21x21x5	10x10x1	50	10^{-2}	$10^{-4.5}$
2	17x17x16x50	15x15x12x100	3x3x5	41x41x9	1x1x1	100	10^{-4}	10^{-6}
3	15x15x12x100	13x13x8x200	3x3x5	61x61x13	1x1x1	200	10^{-4}	10^{-6}
4	13x13x8x200	11x11x4x400	3x3x5	81x81x17	1x1x1	400	10^{-4}	10^{-6}

Model unit spatiotemporal extent and receptive fields

Due to the convolutional form of the hidden layer, each hidden unit can potentially receive from a certain span over space and time. We call this the unit's spatial and temporal extent. For stack 1, this this extent is given by the kernel size (21 x 21 x 5, space x space x time). For stack 2, the extent of each hidden unit is a function of its kernel size and the kernel size and stride of the hidden units in the previous stack, resulting in an extent of 41 x 41 x 9. Similarly, the extent of each hidden unit in stack 3 is 61 x 61 x 13 and in 4 is 81 x 81 x 17. The receptive field size of a unit can be considerably smaller than the unit's extent.

In the first stack of the model, the combination of linear weights and nonlinear activation function are similar to the basic linear non-linear (LN) model[21,22] commonly used to describe neuronal RFs. Hence, the input weights between the input layer and a hidden unit of the model network are taken directly to represent the unit's RF, indicating the features of the input that are important to that unit. The output activities of hidden units in stacks 2-4 are transformations with multiple linear and nonlinear stages, and hence we estimated their RFs by applying reverse correlation to 100,000 responses to Gaussian noise input with mean 0 and standard deviation 1.5 to stack 1.

In vivo V1 RF data

Responses to drifting gratings measured using recordings from V1 simple and complex cells were compared against the model (Fig. 4). The *in vivo* data were taken from http://www.ringachlab.net/lab/Data.html[32].

RF size vs proportion of RF switching polarity

We measured the size of the receptive fields of the units in the first stack and examined the relationship between the RF size and the proportion of the RF switching polarity. For each unit, all pixels in the most recent time-step of the RF with intensities ≥50% of the maximum pixel intensity in that time-step are included in the RF. The RF size was determined by counting the number of pixels fitting this criterion. We then counted the proportion of pixels included in the RF that changed sign (either positive to negative or vice versa) between the two most recent timesteps. The relationship between these two properties for the units in the first stack is shown in Fig. 2b.

Drifting sinusoidal gratings

In order to characterize the tuning properties of the model's visual RFs, we measured the responses of each unit to full-field drifting sinusoidal gratings. For each unit, we measured the response to gratings with a wide range of orientations, spatial frequencies and temporal frequencies until we found the parameters that maximally stimulated that unit (giving rise to the highest mean response over time). We define this as the optimal grating for that unit. In cases where orientation or tuning curves were measured, the gratings with optimal spatial and temporal frequency for that unit were used and were varied over orientation. Each grating alternated between an amplitude of ±3 on a gray (0) background. Some units, especially in higher stacks, had weak or no responses to drifting sinusoidal gratings. To account for this, we excluded any units with a mean response (over time) <1% of the maximum mean response of all the units in that stack. As a result of this, 0/100, 88/200 and 261/400 units were excluded from the 2nd, 3rd and 4th stacks, respectively.

We measured several aspects of the V1 neuron and model unit responses to the drifting gratings. For each unit we measured the circular variance, orientation bandwidth, modulation ratio and direction selectivity.

As a control, we examined the receptive fields and responses to drifting gratings of each unit with its immediate input weights shuffled. In this case, the receptive fields lacked discernable structure, with only patchy spatial frequency and orientation tuning in response to gratings. There were very few orientation tuned (circular variance < 0.9) units with modulation ratios <1.

Circular variance

Circular variance (CV) is a global measure of orientation selectivity. For a unit with mean response over time r_q to a grating with angle θ_q, with angles θ spanning the range of 0 to 360˚ in equally spaced intervals of 5˚ and measured in radians, the circular variance is defined as[32]:

$$CV = 1 - \frac{\sqrt{(\sum_q r_q \sin(2\theta_q))^2 + (\sum_q r_q \cos(2\theta_q))^2}}{\sum_q r_q} \quad (4)$$

Orientation bandwidth

We measured the orientation bandwidth[32], which is a more local measure of orientation selectivity. First, we smoothed the direction tuning curve with a Hanning window filter with a half width at half height of 13.5°. We then determined the peak of the orientation tuning curve. The orientation angles closest to the peak for which the response was $1/\sqrt{2}$ (or 70.7%) of the peak response were measured. The orientation bandwidth was defined as half of the difference between these two angles. We limited the maximum orientation bandwidth to 180°.

Modulation ratio

We measured the modulation ratio of each unit's response to its optimal sinusoidal grating. The modulation ratio is defined as:

$$M = F1/F0 \quad (5)$$

where F1 is the amplitude of the best-fitting sinusoid to the unit's response over time to the drifting grating. F0 is the mean response to the grating over time.

Direction selectivity index

To measure the direction selectivity index, we obtained each unit's direction tuning curve at its optimal spatial and temporal frequency. We measured the peak of the direction tuning curve, indicating the unit's response to gratings presented in the preferred direction (r_p) as well as the response to the grating presented in the opposite (non-preferred) direction (r_{np}). The direction selectivity index is then defined as:

$$DSI = 1 - (r_p - r_{np})/(r_p + r_{np}) \quad (6)$$

Measuring end-stopping

In order to measure the effects of end-stopping, we measured the responses of the hidden units to the same set of drifting gratings but with a circular mask applied to the inputs (e.g. Fig 5a, ii). Masks with a range of spatial extents were tested and the response of the units as a function of this spatial extent was measured (Fig. 5b).

Sparse noise stimuli and two-bar maps

We measured the responses of the hidden units to 'sparse noise' (moving two-bar) stimuli[34]. Each stimulus contained a single oriented bar over the two most recent time-steps and a blank stimulus in

the preceding time-steps. For each unit, the bar was oriented in the preferred orientation (as measured using drifting sinusoidal gratings) of the unit being probed. The length and width of the bar were limited to 50% and 10% of the unit's spatial extent, respectively. This was typically enough for the bar to be longer than the unit's spatial receptive field. In the first time-step with a bar, the center position of the bar (its x and y coordinate) was selected from a dense grid of spatial positions starting from the center position and covering 1/3 of the unit's spatial extent in each direction. In the second time-step, another center position was selected from the same grid. This way, displacement of the bar from each grid position to each other grid position was used to stimulate the unit. To generate two-bar response maps, we measured the response of the unit as a function of the vertical and horizontal displacement (the starting position minus the end position) and then averaged over starting position. This gives a map of the unit's response as a function of the displacement of the stimulus regardless of the starting position. We performed this procedure using all combinations of pairs of white (amplitude +3) and black (amplitude -3) bars on a gray (amplitude 0) background. This yielded four maps (white-to-white, black-to-black, white-to-black and black-to-white). We then summed the same contrast maps (white-to-white and black-to-black) and subtracted the opposite contrast maps (white-to-black and black-to-white) to yield the final two-bar map for each unit. This preserves directional responses while eliminating the responses that depend only on the spatial position of the bars in each frame[34].

Examining the two-bar maps, the position (0,0) indicates that the bar was in the same position in two successive frames, while the vertical and horizontal axes indicate movement in these directions. Positive activity indicates that the unit was excited by movement in that direction, while negative activity indicates inhibition of the unit to movement in the given direction. A non-end-stopped unit will respond to any movement with a component in the preferred direction of the cell. This results in an elongated response profile on the two-bar map (Fig. 5c, right). An end-stopped unit will only respond to movement in the cell's preferred direction, resulting in a two-bar map whose excitatory activity is limited to a more circumscribed region[34] (Fig. 5c, left).

Drifting plaid stimuli

In order to test whether units were pattern selective, we measured their responses to drifting plaid stimuli. Each plaid stimulus was composed of two superimposed half-intensity (amplitude 1.5) sinusoidal gratings with different orientations. The net direction of the plaid movement lies midway between these two orientations (Fig. 5d). As with the sinusoidal inputs, plaids with a variety of orientations, spatial frequencies, temporal frequencies and spatial extents (as defined by the extent of a circular mask) were tested. For each unit, the direction tuning curves of the optimal plaid stimulus (that giving rise to the largest mean response over time) was measured (Fig. 5d-f).

Code and data availability

All custom code used in this study was implemented in Python. We will upload the code to a public Github repository upon acceptance. The movies used for training the models are all publicly available at the websites detailed in the Methods. The V1 data used for comparison is available at http://www.ringachlab.net/lab/Data.html[32].

Supplementary Discussion

Here we have presented a simple model that hierarchically instantiates temporal prediction. This model has several advantages: it is unsupervised; it operates over spatiotemporal inputs and its hierarchical implementation is general, allowing the same network model to learn features that resemble the tuning seen at multiple stages of the visual hierarchy without fine-tuning the model structure to allow the model to reproduce properties of a particular set of neurons. As a result of this, our simple model accounts for many spatial and temporal properties of cells in the dorsal visual pathway, from the center-surround tuning of cells in the retina and LGN, to the spatiotemporal tuning and direction selectivity of V1 simple and complex cells, to the complex motion processing of end-stopped and pattern selective cells.

There are many other normative models of visual processing, based on a range of principles, which can account for some of these properties. Prominent theories include predictive coding, sparse coding, independent component analysis (ICA) and temporal coherence. Temporal prediction is related to each of these frameworks, but there are some important differences. The predictive coding framework postulates that sensory systems learn the statistical regularities present in natural inputs, feeding forward the errors caused by deviations from these regularities to higher areas[16,19,40]. In this process, the predictable components of the input signal are removed and only unexpected inputs are fed forward through the hierarchy. This should be distinguished from temporal prediction, which performs selective coding, where predictive elements of the input are explicitly represented in neuronal responses and non-predictable elements are discarded[8,9,41]. Sparse coding[13,14], which shares similarities with predictive coding[19], is built on the idea that an overcomplete set of neurons is optimized to represent inputs as accurately as possible using only few active neurons for a given input. ICA[15,42] is a related framework that finds maximally independent features of the inputs. Sparse coding and ICA are practically identical in cases where a critically complete code is used. In these frameworks, as with predictive coding, the aim is to encode all current or past input, whereas in temporal prediction, only features that are predictive of the future are encoded and other features are discarded. Slow feature analysis[43] (SFA), a prominent extension of the temporal coherence framework[44], suggests that a key goal of sensory processing is to identify slowly varying features of natural inputs. SFA is closely related to temporal prediction because features that vary slowly are likely to be predictive of the future. However, SFA and temporal prediction may give different weighting to the features that they find[45], and SFA could also fail to capture features that do not vary slowly, but are predictive of the future.

Here we will focus on unsupervised normative models (i.e. those trained on natural inputs) because they are the most relevant. Broadly, these models can be divided into several categories: local models, trained to represent features of a specific subset of neurons such as simple cells in V1, and hierarchical models, which attempt to explain features of more than one cell type (such as simple and complex cells) in a single model. These two categories can be further divided into models that are trained on natural spatial inputs (images) and those that are trained on natural spatiotemporal inputs (movies). Here we will consider a few of the most relevant models from each of these categories.

Local models trained on spatial or spatiotemporal inputs

Sparse coding and ICA are the standard normative models of V1 simple cell RFs[13–15,42,46,47]. Typically, these models are trained using still natural images and have shown remarkable success in accounting for the spatial features of V1 simple cell receptive fields[13–15,42]. However, models trained on static images are unable to account for temporal aspects of neuronal receptive fields, such as direction selectivity or complex motion processing.

The ICA and sparse coding frameworks have been extended to model features of spatiotemporal inputs[11,46,48]. While these models capture many of the spatial tuning properties of simple cells, they tend to produce symmetric temporal envelopes that do not match the asymmetric envelopes of real neurons[10]. Capturing temporal features is especially important when building a normative model of the dorsal visual stream, which is responsible for processing cues related to visual motion.

When trained to find slowly varying features in natural video inputs, SFA models[17] find features with tuning properties that closely resemble those of V1 complex cells, including phase invariance to drifting sinusoidal gratings and end- and side-inhibition. A sparse prior must be applied to the activities of the model units in order to produce spatial localization – a key feature of V1 complex cells[49]. Although SFA does provide a natural explanation for complex cell tuning, on its own this framework does not provide a normative explanation for simple cells.

Hierarchical models trained on spatial inputs

Predictive coding[16] provides a powerful framework for learning hierarchical structure from visual inputs in an unsupervised learning paradigm. When applied to natural images, predictive coding has been successfully used to explain the Gabor-like tuning of simple cells in V1 and some nonlinear tuning properties of neurons in this area, such as end-stopping[16]. However, it is not clear whether this framework can reproduce the phase-invariant tuning of complex cells. Nor has it been shown to account for direction selectivity, end-stopped tuning for motion[34], or complex motion sensitivity[5].

Hierarchical ICA and related models consist of two-layer networks that are trained on natural images with an independence prior placed on the unit activities[50–53]. They have been shown to produce simple cell subunits in the first layer of the network and phase-invariant tuning reminiscent of complex cells in the second layer. However, these models typically incorporate features specifically included to encourage complex cell-like characteristics[54] in the form of a quadratic nonlinearity resembling the complex cell energy model[55]. In some models[56], phase invariance is enforced by requiring the outputs of individual subunits to be uncorrelated[57]. This is in contrast to our model where the phase invariance is learned as a consequence of the general principle of finding features that can efficiently predict future input. An advantage to learning the invariance rather than hand-crafting it is that the identical model architecture can then be applied hierarchically to explain features in higher visual areas without changing the form of the model.

Hierarchical models trained on spatiotemporal inputs

Hierarchical models based on temporal coherence have been shown to capture properties of both simple and complex cells when trained on natural video inputs[57–59]. Typically, these share a similar structure and assumptions with the hierarchical ICA models outlined above, consisting of a two-layer model where the outputs of one or more simple cell subunits are squared and passed forward to a complex cell layer.

Other models have combined sparsity/independence priors with a temporal slowness constraint in a hierarchical model[60–62]. Since sparsity constraints tend to produce simple cell tuning and slowness constraints result in complex cell tuning, these models produce units with both types of selectivity. This contrasts with our model which produces both types of selectivity with a single objective and also accounts for tuning in higher visual areas.

Supplementary References

1. Srinivasan, M. V, Laughlin, S. B. & Dubs, A. Predictive coding: a fresh view of inhibition in the retina. *Proc. R. Soc. London. Ser. B, Biol. Sci.* **216,** 427–59 (1982).

2. Rao, R. P. & Ballard, D. H. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* **2,** 79–87 (1999).

3. Huang, Y. & Rao, R. P. N. Predictive coding. *Wiley Interdiscip. Rev. Cogn. Sci.* **2,** 580–593 (2011).

4. Salisbury, J. M. & Palmer, S. E. Optimal prediction in the retina and natural motion statistics. *J. Stat. Phys.* **162,** 1309–1323 (2016).

5. Bialek, W., Nemenman, I. & Tishby, N. Predictability, complexity, and learning. *Neural Comput.* **13,** 2409–63 (2001).

6. Chalk, M., Marre, O. & Tkačik, G. Toward a unified theory of efficient, predictive, and sparse coding. *Proc. Natl. Acad. Sci. U. S. A.* **115,** 186–191 (2018).

7. Olshausen, B. A. & Field, D. J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **381,** 607–609 (1996).

8. Olshausen, B. A. & Field, D. J. Sparse coding with an overcomplete basis set: a strategy employed by V1? *Vision Research* **37,** 3311–3325 (1997).

9. Bell, A. J. & Sejnowski, T. J. The 'independent components' of natural scenes are edge filters. *Vision Res.* **37,** 3327–3338 (1997).

10. van Hateren, J. H. & van der Schaaf, A. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proc. R. Soc. London B* **265,** 359–366 (1998).

11. Wiskott, L. & Sejnowski, T. J. Slow feature analysis: unsupervised learning of invariances. *Neural Comput.* **14,** 715–770 (2002).

12. Földiák, P. Learning Invariance from Transformation Sequences. *Neural Comput.* **3,** 194–200 (1991).

13. Creutzig, F. & Sprekeler, H. Predictive coding and the slowness principle: an information-

theoretic approach. *Neural Comput.* **20,** 1026–1041 (2008).

14. van Hateren, J. H. & Ruderman, D. L. Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex. *Proc. R. Soc. Lond. B* **265,** 2315–2320 (1998).

15. Olshausen, B. A. Learning sparse, overcomplete representations of time-varying natural images. *IEEE Int. Conf. Image Process.* (2003).

16. Olshausen, B. A. Sparse coding of time-varying natural images. *J. Vis.* **2,** 130 (2002).

17. Chalk, M., Marre, O. & Tkačik, G. Toward a unified theory of efficient, predictive, and sparse coding. *Proc. Natl. Acad. Sci. U. S. A.* **115,** 186–191 (2018).

18. Singer, Y. *et al.* Sensory cortex is optimized for prediction of future input. *Elife* **7,** e31557 (2018).

19. Berkes, P. & Wiskott, L. Slow feature analysis yields a rich repertoire of complex cell properties. *J. Vis.* **5,** 9–9 (2005).

20. Lies, J. P., Häfner, R. M. & Bethge, M. Slowness and Sparseness Have Diverging Effects on Complex Cell Learning. *PLoS Comput. Biol.* **10,** 9–12 (2014).

21. Pack, C. C., Livingstone, M. S., Duffy, K. R. & Born, R. T. End-stopping and the aperture problem: two-dimensional motion signals in macaque V1. *Neuron* **39,** 671–680 (2003).

22. Movshon, J. A., Adelson, E. H., Gizzi, M. S. & Newsome, W. T. The analysis of moving visual patterns. In *Pattern Recognition Mechanisms* (eds. Chagas, C., Gattass, R. & Gross, C.) 117–151 (Vatican Press, 1985).

23. Hyvärinen, A. & Hoyer, P. Emergence of phase- and shift-invariant features by decomposition of natural images into independent feature subspaces. *Neural Comput.* **12,** 1705–1720 (2000).

24. Hyvärinen, A. & Hoyer, P. A two-layer sparse coding model learns simple and complex cell receptive fields and topography from natural images. *Vision Res.* **41,** 2413–2423 (2001).

25. Karklin, Y. & Lewicki, M. S. Emergence of complex cell properties by learning to generalize in natural scenes. *Nature* **457,** 83–86 (2009).

26. Osindero, S., Welling, M. & Hinton, G. E. Topographic product models applied to natural scene statistics. *Neural Comput.* **18,** 381–414 (2006).

27. Kayser, C., Körding, K. P. & König, P. Learning the nonlinearity of neurons from natural visual stimuli. *Neural Comput.* **15,** 1751–1759 (2003).

28. Adelson, E. H. & Bergen, J. R. Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Am. A* **2,** 284 (1985).

29. Hyvärinen, a & Hoyer, P. Emergence of phase- and shift-invariant features by decomposition of natural images into independent feature subspaces. *Neural Comput.* **12,** 1705–1720 (2000).

30. Kording, K. P. How are complex cell properties adapted to the statistics of natural stimuli? *J. Neurophysiol.* **91,** 206–212 (2003).

31. Kayser, C., Einhäuser, W., Dümmer, O., König, P. & Körding, K. Extracting slow subspaces from natural videos leads to complex cells. In *ICANN 2001. LNCS* (eds. Dorffner, G., Bischof, H. & Hornik, K.) 1075–1080 (Springer, Berlin, Heidelberg, 2001). doi:10.1007/3-540-44668-0_149

32. Hurri, J. & Hyvärinen, A. Simple-cell-like receptive fields maximize temporal coherence in

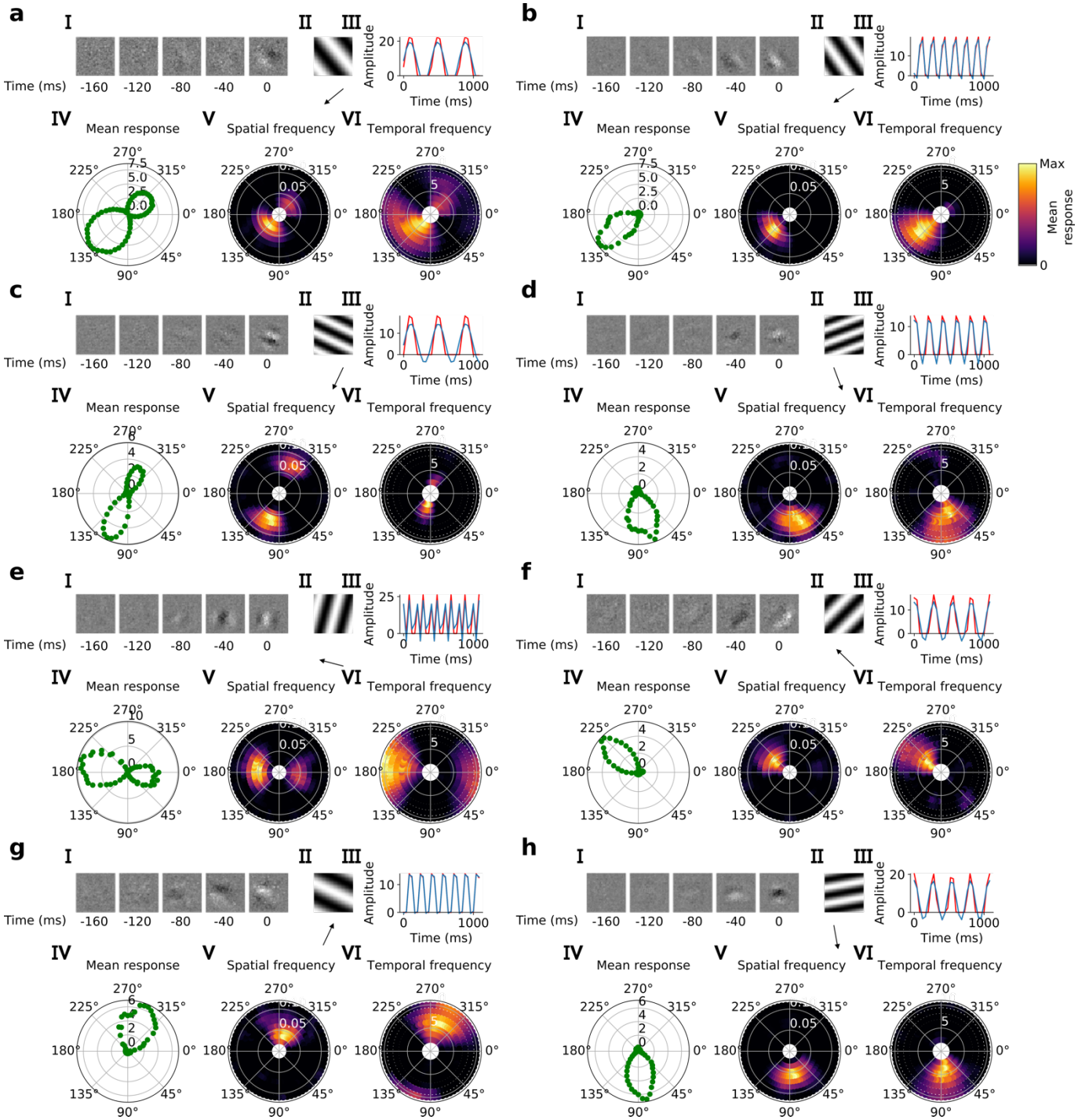
natural video. *Neural Comput.* **15,** 663–691 (2003).

33. Hyvärinen, A., Hurri, J. & Väyrynen, J. Bubbles: a unifying framework for low-level statistical properties of natural image sequences. *J. Opt. Soc. Am. A* **20,** 1237–1252 (2003).

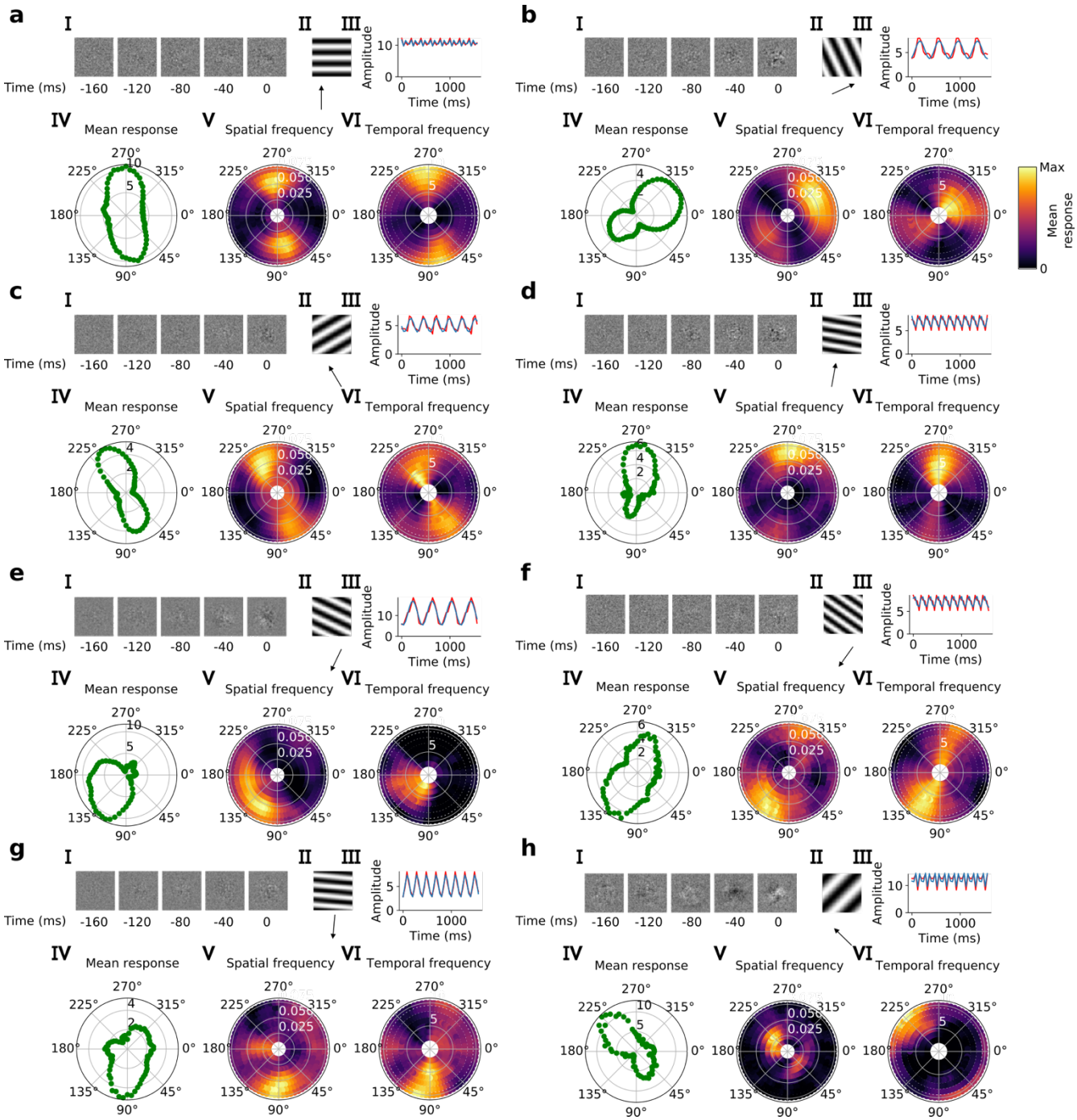
34. Berkes, P., Turner, R. E. & Sahani, M. A structured model of video reproduces primary visual cortical organisation. *PLoS Comput. Biol.* **5,** e1000495 (2009).

35. Cadieu, C. F. & Olshausen, B. A. Learning intermediate-level representations of form and motion from natural movies. *Neural Comput.* **24,** 827–866 (2012).

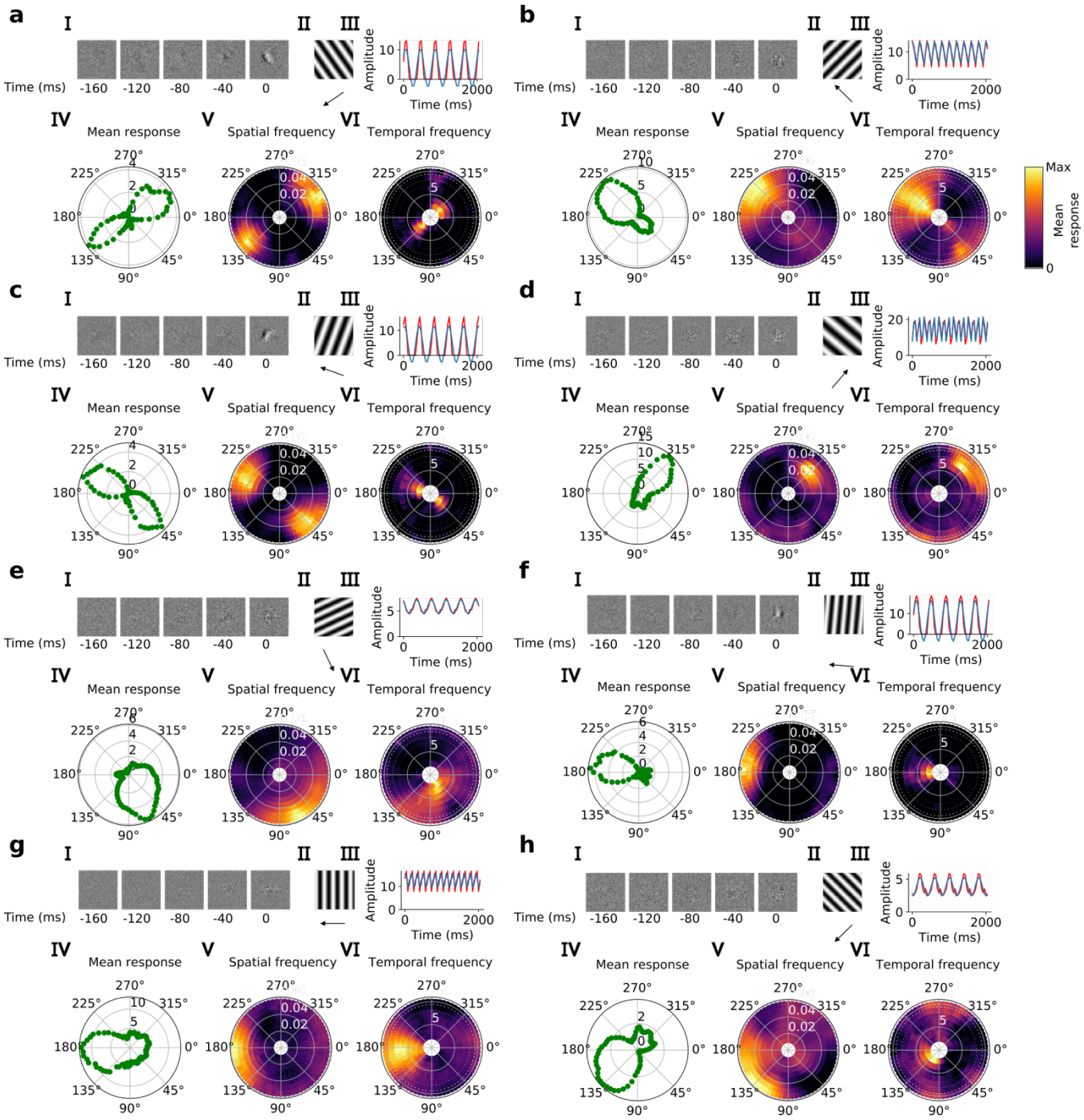
Supplementary Figures



Supplementary Figure 1 | Tuning properties of example units in stack 2. ***a-h, I-IV*** as in Fig. 3*a-d*.



Supplementary Figure 2 | Tuning properties of example units in stack 3. ***a-h, I-IV*** as in Fig. 3*a-d*.



Supplementary Figure 3 | Tuning properties of example units in stack 4. ***a-h, I-IV*** as in Fig. 3*a-d*.