PLOS Computational Biology

RESEARCH ARTICLE

Hierarchical recurrent temporal prediction as a model of the mammalian dorsal visual pathway

Sebastian Klavinskis-Whiting, Andrew J. King*, Nicol S. Harper*

Department of Physiology, Anatomy and Genetics, University of Oxford, Oxford, United Kingdom

* nicol.harper@dpag.ox.ac.uk (NSH); andrew.king@dpag.ox.ac.uk (AJK)

Abstract

A major goal of neuroscience is to identify general principles that can explain the diverse structures and functions of the brain. The principle of temporal prediction provides one approach, arguing that the sensory brain is optimized to represent stimulus features that efficiently predict the immediate future input. Previous work has demonstrated that feedforward hierarchical temporal prediction models can capture the tuning properties of neurons along the visual pathway, and that recurrent temporal prediction models can explain local functional connectivity within primary visual cortex. However, the visual system is also characterized by extensive inter-areal feedback recurrency, which existing models lack. We aimed to better account for the dynamic features of neurons in the visual cortex by incorporating both local recurrency and inter-areal feedback connectivity into a hierarchical temporal prediction model. The resulting model captured tuning properties along the dorsal visual pathway, including pattern motion selectivity and surround suppression, and the contribution of inter-areal connectivity to these properties. Moreover, compared with several alternative normative models, the hierarchical recurrent temporal prediction model provided the closest fit to these tuning properties and was best able to explain neuronal responses to natural stimuli. Accordingly, temporal prediction accounts well for information processing along the visual pathway.

Author summary

A goal of neuroscience is to identify general principles that explain how neural circuits represent information. Feedforward and locally recurrent models have shown that temporal prediction, the idea that sensory systems are optimized to predict immediate future input, can account for many aspects of the response properties and connectivity of sensory neurons. However, sensory pathways are also shaped by extensive feedback connections between brain regions, whose computational role remains poorly understood. Here, we address this gap by introducing a hierarchical recurrent temporal prediction model that explicitly incorporates inter-areal feedback, reflecting the bidirectional organization of

OPEN ACCESS

Citation: Klavinskis-Whiting S, King AJ, Harper NS (2026) Hierarchical recurrent temporal prediction as a model of the mammalian dorsal visual pathway. PLoS Comput Biol 22(5): e1013138. https://doi.org/10.1371/journal.pcbi.1013138

Editor: Tim Christian Kietzmann, University of Osnabrück: Universitat Osnabruck, GERMANY

Received: May 18, 2025

Accepted: May 7, 2026

Published: May 28, 2026

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcbi.1013138

Data availability statement: All code supporting this study is publicly available at https://github.com/sebbkw/hierarchical_temporal_prediction. The model was trained using a novel dataset drawn from around 2.5 hours of wildlife videos, which were obtained from public sources, and are available at https://osf.io/hf2pj/ (comprising https://doi.org/10.5281/zenodo.20041063; https://doi.org/10.5281/zenodo.20041567; https://doi.org/10.5281/zenodo.20041596).

Funding: This study was funded by Wellcome (WT108369/Z/2015/Z to AK). Salary support was provided by Wellcome to AJK and NSH, and SKW received a studentship stipend from the Nuffield Department of Clinical Neurosciences as part of the Neuroscience Doctoral Training Program at the University of Oxford. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

mammalian cortex. Trained on natural movies, the model reproduces stimulus feature selectivity and local connectivity patterns observed in the dorsal visual cortical pathway. Critically, it also explains neural phenomena that depend on feedback, including enhanced memory, surround suppression, and the emergence of pattern-motion selectivity across the cortical processing hierarchy. Removing the feedback selectively disrupts these properties, matching results from inactivation studies in animals. Compared with feedforward temporal prediction models and other unsupervised approaches, the feedback-enabled model better matches neuronal tuning and population-level stimulus representations across multiple visual cortical areas. These findings therefore suggest that inter-areal feedback may be organized to support temporal prediction in the brain.

Introduction

Classical theories of vision posit a series of hierarchically organized processing stages that gradually extract higher-level visual features [1,2]. In this way, the brain is thought to decompose the patterns of light that impinge on the retina into meaningful representations to be further processed in downstream areas, and ultimately, to guide behavior. A key question, then, is how to understand these computations and whether an underlying principle of organization can explain how the resulting representations change across the visual pathway [3,4].

One promising principle is that of temporal prediction, which argues that the sensory brain is optimized to represent stimulus features that are predictive of the immediate future [4–8]. This is likely to be useful for extracting underlying variables in the input, eliminating unnecessary information based on the behavioral relevance of stimuli, and creating a representation that is useful for guiding future action given sensory and motor delays [4–6,8]. Temporal prediction can be applied in a hierarchical manner to capture spatiotemporal receptive field properties across multiple stages of the visual pathway [4]. Importantly, as an unsupervised principle, temporal prediction does not rely on hand-labeled examples, as in many deep learning models of the visual system [9,10]. Together, these features make temporal prediction a promising candidate for modeling neural processing in sensory systems.

In addition to demonstrating that the tuning properties of neurons along the visual pathway can be reproduced by applying temporal prediction in a hierarchical manner [4,5], we have recently shown that the functional specificity of connections within the primary visual cortex (V1) can be captured by applying temporal prediction to a locally-recurrent network [6]. However, these models have so far neglected the substantial role of long-range feedback connectivity from higher-order visual areas, which has been implicated in a range of visual processing functions [11]. Feedback connectivity in the visual system appears to impart both modulatory and driver roles on downstream neural activity [11,12]. Among other roles, feedback from higher visual areas is believed to mediate extra-classical receptive field effects, such as surround suppression [13–16], maintain and gate working memory representations

[17–19], and, within the predictive coding framework, convey predictions that may drive learning via bottom-up prediction errors [20,21]. Accordingly, incorporating these elements into normative models, including temporal prediction, is likely to be an important step in explaining the structural and functional properties of the visual pathway.

Here, we demonstrate that a hierarchical recurrent temporal prediction model trained on movies of dynamic natural scenes better captures the tuning properties of visual neurons along the mammalian visual pathway than implementing hierarchy or local recurrency alone. The architectural addition of inter-areal recurrency and the network's optimization for temporal prediction jointly improved the model's similarity to the tuning properties of visual cortical neurons measured in different studies. Moreover, feedback connectivity significantly improved the network's memory capacity and frame prediction performance, and recapitulated the dependence on feedback of surround suppression and pattern motion selectivity in the responses of V1 neurons [13–16,22]. Finally, compared with alternative normative models, the hierarchical recurrent temporal prediction model was more closely aligned with the stimulus representations that have been described in different areas of the visual cortex. Together, these results provide evidence that inter-areal feedback across the visual hierarchy may also be optimized for temporal prediction.

Results

The hierarchical recurrent model improves temporal prediction performance

The hierarchical recurrent model consisted of a recurrent neural network instantiating a hierarchical model of temporal prediction, where each group of units was trained to predict its lower-order inputs. Thus, the first group (G_1) was trained for next-frame prediction of video clips of dynamic natural scenes, while the second and third groups (G_2 and G_3) were trained to predict the future activity of groups G_1 and G_2, respectively (Fig 1A). Within the recurrent layer of the network, internal connectivity ensured that each group received 'feedforward' input from the previous group, 'feedback' input from the subsequent group and local recurrent input from itself. We conceptualized each group as modeling increasingly deep regions of the dorsal visual pathway, with groups G_1, G_2, and G_3 in particular corresponding to primary visual cortex (V1), secondary visual cortex (V2) and middle temporal area (MT), respectively (Fig 1B). However, the model structure and training data are not *per se* tailored to any specific mammalian species. Hence, while we primarily compare our model to macaque data, where such data are not available, we compare the model to mouse data.

More explicitly, the network is essentially a single-hidden-layer recurrent neural network (RNN), with various constraints applied to provide hierarchical structure. All the units of the model are rate-based. At time t, the hidden unit activity vector $\mathbf{h}[t]$ of the network is given by:

$$\mathbf{h}[t] = \mathrm{f}\left(\mathbf{R}\mathbf{h}[t-1] + \mathbf{W}\mathbf{u}[t] + \mathbf{b}\right) \tag{1}$$

where $\mathbf{R}$ is the recurrent weight matrix, $\mathbf{W}$ is the input weight matrix, $\mathbf{b}$ is the input bias vector, $\mathbf{u}[t]$ is the input vector, and f(•) is a rectified linear unit (ReLU) nonlinearity. The input vector consists of a frame at time t from a movie of natural scenes of 40 timesteps in duration, with the 20x40 pixel frame flattened into a length 800 vector. This equation, with constraints that are outlined below, comprises our model of the dorsal pathway, which is compared with neural data.

The output vector $\mathbf{y}[t]$ of the network, which is a prediction of future input and activity, is given by:

$$\mathbf{y}[t] = \mathbf{M}\mathbf{h}[t] + \mathbf{c} \tag{2}$$

where $\mathbf{M}$ is the output weight matrix and $\mathbf{c}$ is the output bias vector.

Hierarchy is imposed on the network in three ways. First, the hidden unit vector is divided into three equally sized groups of 800 units each, $\mathbf{h}[t] = (\mathbf{h}^1[t], \mathbf{h}^2[t], \mathbf{h}^3[t])$, where the superscript denotes group number. Hierarchy is encouraged

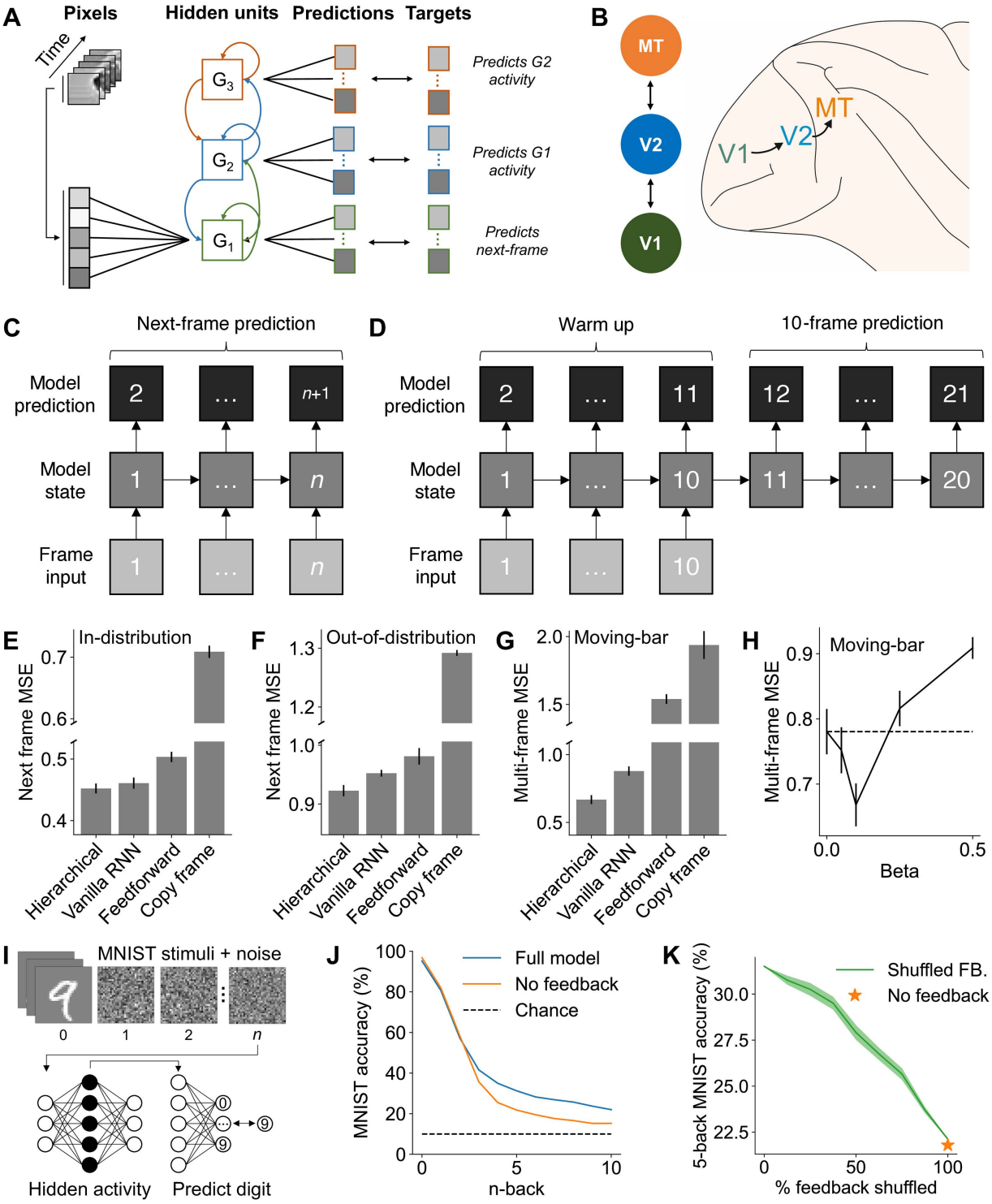

Fig 1. The hierarchical recurrent temporal prediction model shows improved frame prediction performance. (A) Schematic of the hierarchical recurrent temporal prediction model. **(B)** A subset of the dorsal visual pathway, with regions V1, V2 and MT highlighted in a posterolateral view of the macaque monkey brain, putatively corresponding to groups G_1, G_2 and G_3. **(C)** The next frame prediction procedure where the preceding frame and hidden state are inputted into the model at each timestep. **(D)** The multi-frame prediction procedure, where the model was run autoregressively such that the model predicted multiple frames, relying on its own predictions and hidden state alone. **(E-G)** Prediction performance across comparison models for

(E) in-distribution, (F) out-of-distribution and (G) multi-frame prediction of a moving bar. In all cases, the hierarchical recurrent model (labeled 'hierarchical') performed best, exceeding the performance of the comparison models. **(H)** Multi-frame prediction performance for a moving bar across differing beta values. The lowest mean squared error was found for a beta value of 0.1, implying that a hierarchically-predictive instantiation of the network can improve prediction performance. **(I)** Schematic of the *n*-back procedure. MNIST digits are used as input to the model followed by *n* frames of Gaussian noise. The hidden state is then linearly mapped to predict the digit identity to calculate an accuracy score, repeating the same procedure for all *n* 0-10. **(J)** MNIST accuracy declines more quickly when feedback is ablated, suggesting that model feedback increases the memory capacity of the network. **(K)** MNIST accuracy declines as the percentage of shuffled feedback (FB) weights increases, implying that the increase in memory capacity is not purely a feature of the model's architecture, but depends on the trained weights of the feedback connectivity. Note that every time the feedback weights were shuffled, we retrained the mapping weights.

https://doi.org/10.1371/journal.pcbi.1013138.g001

by having no connections between group 1 and group 3 - that is by fixing the corresponding regions of the **R** matrix to zero. The output vector $\mathbf{y}[t]$ is also divided into three groups, $\mathbf{y}[t] = (\mathbf{y}^1[t], \mathbf{y}^2[t], \mathbf{y}^3[t])$, of corresponding lengths $K^1 = K^2 = K^3 = 800$ and the **M** matrix is fixed to zero except for the weights from $\mathbf{h}^1[t]$ to $\mathbf{y}^1[t]$, $\mathbf{h}^2[t]$ to $\mathbf{y}^2[t]$ and $\mathbf{h}^3[t]$ to $\mathbf{y}^3[t]$. Second, the input matrix **W** only connects to group 1, that is all weights are fixed at zero except those going into group 1. Finally, the most important source of hierarchy comes from the loss function used to train the network. The loss L was given by:

$$L = \lambda Z + \left\langle \frac{1-\beta}{K^1} \left\| \mathbf{y}^1[t] - \mathbf{u}[t+1] \right\|_2^2 + \frac{\beta}{K^2} \left\| \mathbf{y}^2[t] - \mathbf{h}^1[t+1] \right\|_2^2 + \frac{\beta}{K^2} \left\| \mathbf{y}^3[t] - \mathbf{h}^2[t+1] \right\|_2^2 \right\rangle_{n,t} \quad (3)$$

Here, group 1 output predicts the input, group 2 output predicts group 1 hidden unit activity, and group 3 output predicts group 2 hidden unit activity. The L_2 norm is given by $\|\cdot\|_2$ and the expectation over all clips and time given by $\langle\cdot\rangle_{n,t}$. The clip subscript n is left off the vectors for simplicity. The regularizer Z is the sum of the absolute values of all the weights in the network (**W**, **R** and **M**), and can be seen as a constraint on connectivity, and is scaled by λ. Crucially, the hyperparameter β governs the relative importance in the loss function of predicting the input versus higher groups predicting lower groups. Unless otherwise noted, λ was set to 10^{-6} and β to 0.1 (see Methods).

An additional constraint is that 20% of the hidden units in each group were constrained to be inhibitory; that is, the weights from them in the recurrent connectivity matrix (**R**) were constrained to be negative, and they only make intra-group connections. All other hidden units were constrained to be excitatory; that is, the weights from them in the **R** matrix were constrained to be positive. For a visual depiction of the constraints on the network see S1 Fig.

We trained the network by minimizing the loss function with respect to the weights (**W**, **R**, **M**) and biases (**b**, **c**) using the ADAM backpropagation algorithm, for a large dataset of movies of natural scenes. We used backpropagation as our focus was on whether the functional organization of the dorsal pathway is consistent with being optimized for hierarchical recurrent temporal prediction, and not the exact learning mechanism by which this is achieved. The number of active hidden units in the trained network, as measured by the response to a visual noise stimulus, was 800 units in G_1, 97 units in G_2, and 42 units in G_3. Note that this is consistent with primary visual cortex being typically larger and having greater neuronal density than higher visual areas [23]. We take the active hidden units of the trained network to correspond to neurons in the dorsal pathway.

We first analyzed the temporal prediction performance of the hierarchical recurrent model and compared this to several baseline models. Specifically, we compared the hierarchical recurrent temporal prediction model to a single-layer recurrent temporal prediction model without the addition of hierarchical groups (a vanilla RNN), a purely feedforward model that omitted both local and feedback recurrency, and a baseline comparison where we compared the predictions to simply copying the input frame. The copy frame comparison ensured that each model was learning to predict the next frame, rather than merely reproducing its inputs as a trivial solution. For next frame prediction (Fig 1C), we calculated the mean squared error between the true and predicted next frame of the stimulus. Conversely, for 10-frame prediction (Fig 1D), we

measured the mean squared error across 10 frames generated in an autoregressive manner. This involved repeatedly predicting the next frame using the model's own preceding predictions as input, rather than the true frame inputs. For next frame prediction, we assessed the model's performance both in-distribution (the held-out test set from the stimuli used for training; Fig 1E) and out-of-distribution (a novel dataset to that used during training; Fig 1F) to assess the generality of these findings. In the multi-frame case, the input consisted of oriented moving bar stimuli where there was an unambiguous 'true future' given the preceding input (Fig 1G, 1H).

Across all conditions, we found that the hierarchical recurrent model performed best, with the overall lowest mean squared error for temporal prediction (paired t-test, all $p < 0.0001$) (Fig 1E-1G). Thus, the addition of hierarchy in the form of additional groups improved prediction performance compared with both the vanilla RNN and the feedforward model trained for temporal prediction. Finally, all three models exceeded the baseline performance for copying the preceding frame (paired t-test, all $p < 0.0001$).

We also varied the beta parameter, which controls the influence of higher groups on the hierarchical recurrent model's loss function during training, to assess its impact on temporal prediction performance (Fig 1H). A higher beta value indicates a greater weighting applied to higher groups, with $\beta = 0.5$ indicating an equal weighting. Although higher beta values impaired next-frame prediction performance across the in- and out-of-distribution cases (S2 Fig), we found that a small beta improved multi-frame prediction on the moving bar stimuli set ($\beta = 0$ vs $\beta = 0.1$: paired t-test, $t(71)=3.20$, $p = 0.0021$). Thus, the addition of hierarchy to the loss function could improve the model's capacity to generalize for multi-frame prediction across novel stimuli.

Given the importance of hierarchy in the model for temporal prediction, we next asked whether this might in part be supported by increasing the memory capacity of the model, mediated by the network's feedback connectivity. In the visual cortex, feedback from higher visual areas to V1 is selectively recruited during working memory tasks [18] while disruption of corticocortical feedback to higher visual areas has been shown to impair performance in working memory-dependent behavior [19]. Thus, we reasoned that feedback in the hierarchical recurrent temporal prediction model might similarly support the network's memory capacity.

To test this hypothesis, we trained a linear decoder to predict the identity of hand-written MNIST digits from the network's hidden state. To probe how well the model's hidden state could maintain the digit representations, the decoder was trained based on the hidden state after n frames of Gaussian noise input (Fig 1I). For each n-back decoder, we then calculated the decoder's accuracy on the MNIST test set for the model with and without feedback. We retrained the decoder for each n-back condition since a fixed decoder is insufficient (S3 Fig). Ablating feedback significantly reduced the decoder accuracy from 3-back onwards (z-test, $z = 6.54$, $p < 0.0001$), implying that feedback improved the model's capacity to maintain representations over longer timescales (Fig 1J). To test whether this improvement was related to the model's training or whether any form of feedback could account for the difference, we assessed the 5-back accuracy when shuffling an increasing proportion of the model's feedback weights (Fig 1K). As the proportion of shuffled weights increased, the accuracy decreased until it was non-significantly different from the model where feedback was ablated entirely (z-test, $z = 0.494$, $p = 0.621$). Thus, randomized feedback was ineffective, and feedback weights required the structure imposed by training to maintain network representations.

Hierarchical recurrent temporal prediction captures tuning properties across the visual hierarchy

We next investigated how model hidden unit response properties varied across the hierarchical recurrent model's groups relative to different stages of the visual system. To make these comparisons, we estimated the model units' receptive fields using the response-weighted average of each unit to Gaussian random noise. Units in the first group generally had a clear Gabor-like structure with alternating excitatory and inhibitory regions akin to the receptive fields of V1 simple cells (Fig 2Ai). In contrast, units in higher-order groups generally displayed little spatial structure (Fig 2Aii, iii), consistent with a more complex-cell-like response that is not well estimated by the response-weighted averaging procedure [28].

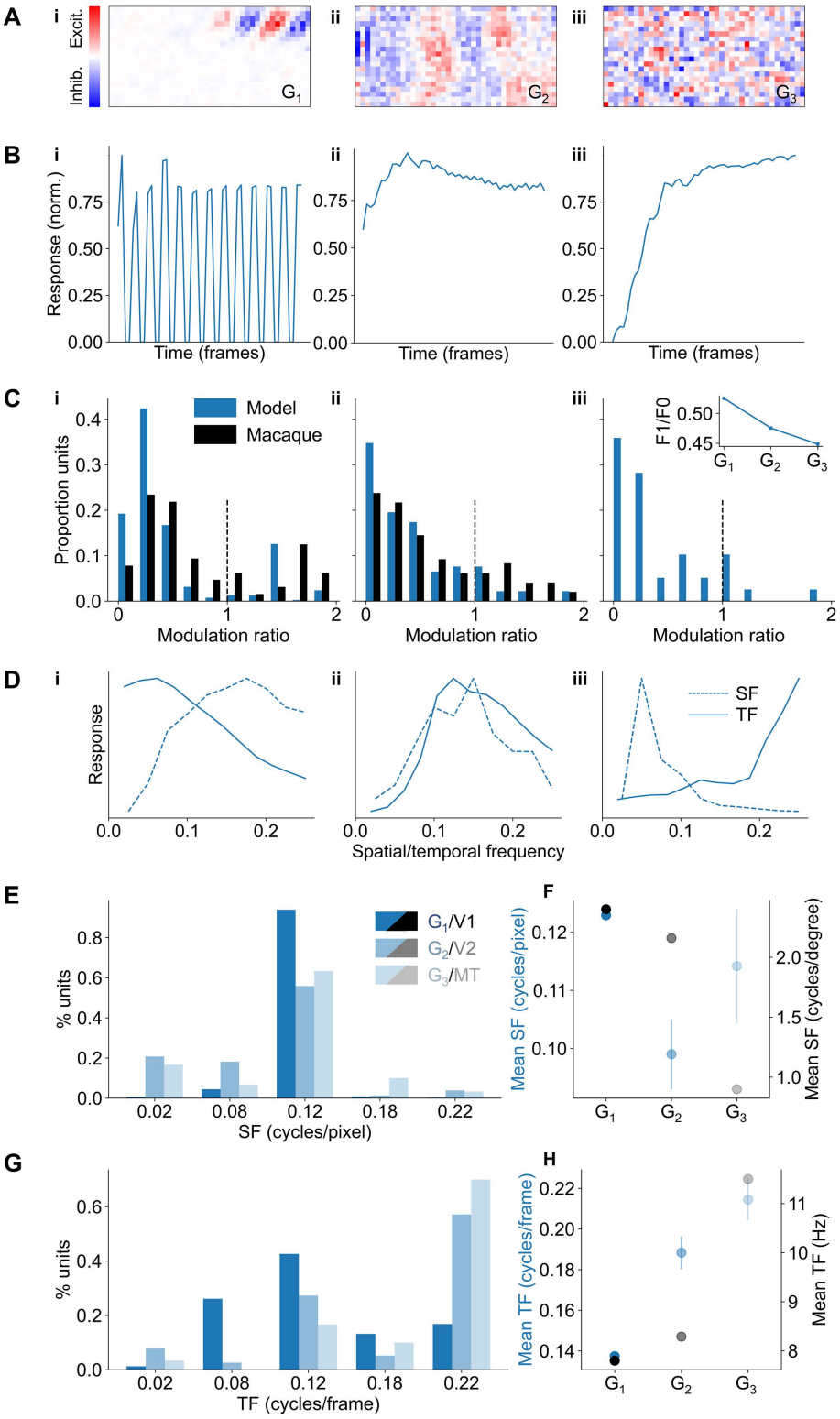

Fig 2. Hierarchical recurrent temporal prediction captures tuning properties across the visual hierarchy. (A) Exemplar model unit receptive fields from group G_1 **(i)**, G_2 (ii) and G_3 **(iii)**. **(B)** Normalized response to the preferred grating stimulus for the same model units as in **(A)**. **(C)** Distribution of

modulation ratio values across model groups G_1 **(i)**, G_2 (ii) and G_3 **(iii)** (blue bars). Comparison data for G_1 from macaque V1 and for G_2 from macaque V2 [24] (black bars) are also shown. Inset shows the average modulation value for each model group. **(D)** Exemplar spatial and temporal frequency tuning curves for three units from G_1 **(i)**, G_2 (ii) and G_3 **(iii)**. **(E)** Distribution of preferred spatial frequencies (SF) for model units across each group. **(F)** Mean preferred spatial frequency for each group compared with macaque V1 [25], V2 [26], and MT [27]. **(G)** Distribution of preferred temporal frequencies (TF) for model units across each group. **(H)** Mean preferred temporal frequency for each group compared with macaque V1 [27], V2 [26], and MT [27].

https://doi.org/10.1371/journal.pcbi.1013138.g002

To further characterize the receptive field properties of model units, we recorded their responses to full-field sinusoidal gratings of varying temporal frequency, spatial frequency, and orientation. The preferred stimulus was defined as the drifting grating that maximally stimulated each unit, which we used to classify model units as simple-cell- or complex-cell-like in their responses [28,29]. In visual cortex, the activity of simple cells is heavily modulated by a drifting grating, whereas complex-cell responses are phase invariant and only minimally modulated by the moving stimulus. We found model units that exhibited both simple-cell- (Fig 2Bi) and complex-cell-like (Fig 2Bii, 2Biii) responses.

We quantified these responses using the modulation ratio – the ratio of the response modulation amplitude to the average response – with a larger modulation ratio indicating a more simple-cell-like response (Fig 2C). G_1 units (mean modulation ratio = 0.53) had a bimodal-like distribution similar to that of macaque V1 [30] (mean = 0.79), indicating two subpopulations of units (Fig 2Ci). Higher groups (G_2 mean = 0.48, G_3 mean = 0.45) had a more unimodal-like distribution, which tended to monotonically decline with increasing modulation ratio, similarly to the true distribution found in V2 [24] (mean = 0.63) (Fig 2Cii, 2Ciii). This was reflected in the average modulation ratio across groups (Fig 2Ciii inset), which declined for higher model groups, indicating a larger proportion of complex-cell-like responses. Finally, we calculated the joint distributions of the modulation ratio with the model unit's orientation selectivity (S4A Fig), as well as how well each unit was modelled as a Gabor (S4B Fig). As expected, and in tandem with the biology, simple-cell-like model units were more orientation selective (mean orientation selectivity index, OSI = 0.84) and had a higher r^2 for their Gabor fits (mean = 0.60) than complex-cell-like units (mean OSI = 0.77; t-test, $t(181)=-3.84$, $p = 0.0002$; mean Gabor fit = 0.54 t-test, $t(251)=-4.23$, $p < 0.0001$). Furthermore, consistent with our qualitative assessment in Fig 2A, the Gabor fit r^2 tended to decrease the higher the group and, for the units that could be well fit, the size of the fitted receptive field tended to increase (S5 Fig).

The distribution of spatial and temporal frequency tuning properties also varied systematically across the model. For each unit, we measured the tuning curve for that unit's response as a function of temporal and spatial frequency, with the preferred frequency taken as the tuning curve's peak. Units were generally well tuned, with tuning curves spanning a wide range of temporal and spatial frequencies (Fig 2D, 2E). The second group G_2 had a lower average preferred spatial frequency than the first group G_1 (t-test, $t(79.4)=3.91$, $p = .0002$; Fig 2F). The same trend has been found in macaque visual cortex where the average preferred spatial frequency for units recorded in V1 [25] was greater than in higher-order areas V2 [26] and MT [27] (Fig 2F). However, whereas the preferred spatial frequency monotonically declined for the macaque visual system, there was a U-shaped profile for the model, with the average preferred spatial frequency increasing from G_2 to G_3. In contrast, the mean preferred temporal frequency increased across the model's groups similarly to the macaque visual system, with higher model groups and higher-order visual areas both showing greater selectivity to higher temporal frequency stimuli than the first model group (G_1 vs. G_2, t-test, $t(87.7)=-6.07$, $p < 0.0001$; G_1 vs. G_3, t-test, $t(31.7)=-7.32$, $p < 0.0001$) and macaque V1 [27], respectively (Fig 2G, 2H).

Finally, we analyzed the local connectivity of G_1 units and found that the network replicated the results previously described for the single-layer recurrent temporal prediction model (S6 Fig) [6]. Specifically, we found that short-range connections were more common for pairs of orientation-selective G_1 units that shared a similar orientation preference (S6A Fig) and for pairs of direction-selective G_1 units when they shared similar or opposite preferred directions of motion (S6B Fig), as is found in mouse V1 [31]. For connections between units whose receptive fields were displaced over a larger region of visual space ('long-range' connections), we found that the connection probability was greatest between

units with a similar orientation preference whose receptive fields were aligned co-axially in visual space, again as is found in mouse V1 [32] (S6C-S6E Fig).

Thus, overall, the model qualitatively matched the receptive field structure, distribution of modulation ratios and the change in preferred spatial and temporal frequencies across the macaque visual hierarchy as well as the local functional connectivity rules that have been found in V1.

Model units mirror the visual system's hierarchy of 2D motion sensitivity

Although direction selectivity is present in V1, neurons in this area are generally unable to represent the complex motion of two-dimensional moving surfaces or patterns [33]. This sensitivity to two-dimensional motion can be probed using plaid stimuli, which consist of two overlaid full-field gratings moving in different directions (Fig 3Ai). Perceptually, the direction

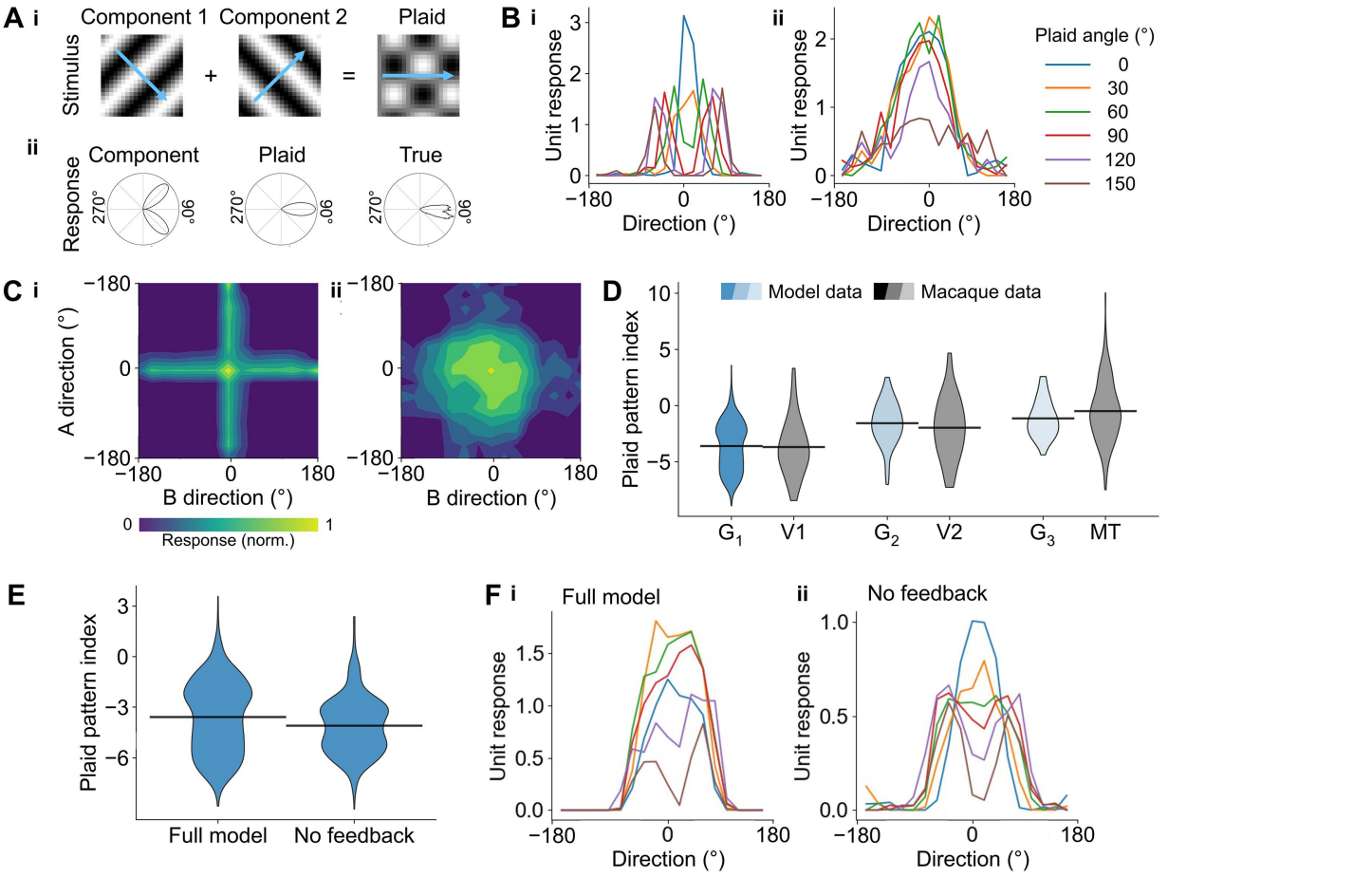

Fig 3. Model units mirror the visual system's hierarchy of 2D motion sensitivity. (A) Construction of plaid stimuli through the additive combination of two grating stimuli **(i)**. The plaid pattern index is computed by measuring the Fisher *z*-transformed difference in the correlation between the true response of the model unit or visual neuron and an idealized component or plaid response **(ii)**. **(B)** Example direction tuning curves to plaid stimuli with different plaid angles (the angular separation between the two gratings) for component- (i) and pattern-selective (ii) units. **(C)** Example contour plots for component (i) and pattern-selective (ii) units where the angle of each component is independently varied. **(D)** Plaid pattern index across the 3 model groups, G_1, G_2 and G_3, and in macaque V1 [34], V2 [26], and MT [35]. A larger plaid pattern index indicates greater response correlation to the theoretical plaid response. **(E)** The mean plaid pattern index for G_1 units is reduced when feedback in the model is abolished. **(F)** Example direction tuning curves for plaid stimuli across different plaid angles for the same model unit with (i) and without (ii) feedback, demonstrating a change from a pattern-like to a component-like response when feedback is abolished.

https://doi.org/10.1371/journal.pcbi.1013138.g003

of motion corresponds to the average of the direction of the two components. Accordingly, representing this plaid pattern motion requires integrating the motion signals from the individual grating components such that neurons across the visual system are differentially sensitive to the two-dimensional motion of the plaid versus its individual component gratings (Fig 3Aii). In macaque V1, neurons generally respond to the motion of individual components making up the plaid pattern, whereas a proportion of MT neurons signal the overall motion of the plaid pattern [34].

Individual model units included examples of both kinds of response to plaid stimuli. For component-selective units, the tuning curve displays two peaks corresponding to the direction of motion of the components of a plaid (Fig 3Bi), resulting in a cross-like contour plot when the direction of motion of each component is varied independently (Fig 3Ci). In contrast, pattern-selective units respond with a single peak corresponding to the overall direction of motion of the plaid stimulus (Fig 3Bii), resulting in a single response region in the contour plot (Fig 3Cii).

The plaid pattern index quantifies this preference [34,35], based on the difference in correlation between the unit's true response and an ideal component response versus an ideal plaid pattern response (Fig 3Aii). Thus, a higher plaid pattern index indicates greater plaid selectivity. The distribution of plaid pattern index values across model groups recapitulated the hierarchy found in the macaque visual system (V1 [34], V2 [26], and MT [34]), with the plaid selectivity monotonically increasing from lower to higher groups (G_1 mean = -3.60, G_2 mean = -1.56, G_3 mean = -1.12; one-way ANOVA, $F(2, 931)=61.6$, $p < 0.0001$) (Fig 3D). Indeed, no significant difference was found between the means of each group and the corresponding region of the macaque visual cortex (t-test, all $p > 0.153$), indicating that progressively greater selectivity for the plaid stimulus compared with the individual grating components was present across both the hierarchical recurrent temporal prediction model and the macaque visual system.

To test whether model units were specifically selective to the direction of plaid motion – as opposed to the plaid's orientation, for example – we measured the plaid pattern direction selectivity of each pattern-selective model unit (plaid pattern index > 0). As expected, the majority (73.2%) of pattern-selective units were also pattern-direction-selective, defined as having a direction selectivity index (DSI) greater than 0.3 [6] (mean DSI = 0.65). Thus, these units were specifically tuned to the direction of pattern motion.

Finally, we investigated the role of feedback in supporting pattern-like responses in model units. Although pattern direction selectivity is more weakly represented in V1, pattern-like motion responses are not entirely absent, which may result in part from feedback to V1 from higher cortical areas. Indeed, suppressing feedback from the middle suprasylvian gyrus has been shown to reduce pattern-like responses in early visual cortex of the cat [22]. In line with these experimental data, abolishing feedback in the model resulted in a significantly lower mean plaid pattern index in G_1 (t-test, $t(1519)=5.33$, $p < 0.0001$; Fig 3E). At the level of individual model units, this often resulted in a change from a plaid-like response (Fig 3Fi) to one that more closely resembled a component-like response (Fig 3Fii, S7i-ii Fig) though other units had more varied responses and appeared to show more generally degraded motion tuning following the abolition of feedback (S7iii Fig). Furthermore, we find that G_1 units show a mean magnitude change in direction tuning to simple moving gratings of only 6.6° after we ablated feedback ($t(579)= -4.4$, $p < 0.001$), indicating that tuning to the underlying grating stimuli remained relatively consistent. This is consistent with the impact of the feedback ablation on the plaid pattern index not simply being a consequence of a general disruption of direction tuning.

Model units recapitulate feedback-dependent surround suppression

We next investigated the functional role of the model's feedback connectivity with respect to surround suppression. Surround suppression is a hallmark nonlinearity in the responses of V1 neurons and occurs when the neural response is inhibited by stimuli extending beyond the neuron's classical receptive field [14,15]. In the context of predictive coding, surround suppression has been interpreted as a consequence of the statistics of the animal's natural environment [21]. As contours are generally continuous in the environment, they can be consistent with the visual system's higher-order predictions. By contrast, short discontinuous contours violate this internal model and produce a larger response, leading to the

surround suppression effect. In support of this interpretation, reducing or ablating feedback from higher visual areas to V1 reduces the magnitude of surround suppression [14,15]. Although the temporal prediction model is not trained with explicit error propagation between groups, as in predictive coding models, we asked whether ablating feedback would similarly reduce the surround suppression effect.

We examined surround suppression by playing to each unit its preferred drifting grating limited by a window of increasing diameter centered on the unit's receptive field. The mean response of the G_1 units increased with the window diameter and then decreased, indicating surround suppression (Fig 4Ai). Also, in line with the experimental data, surround

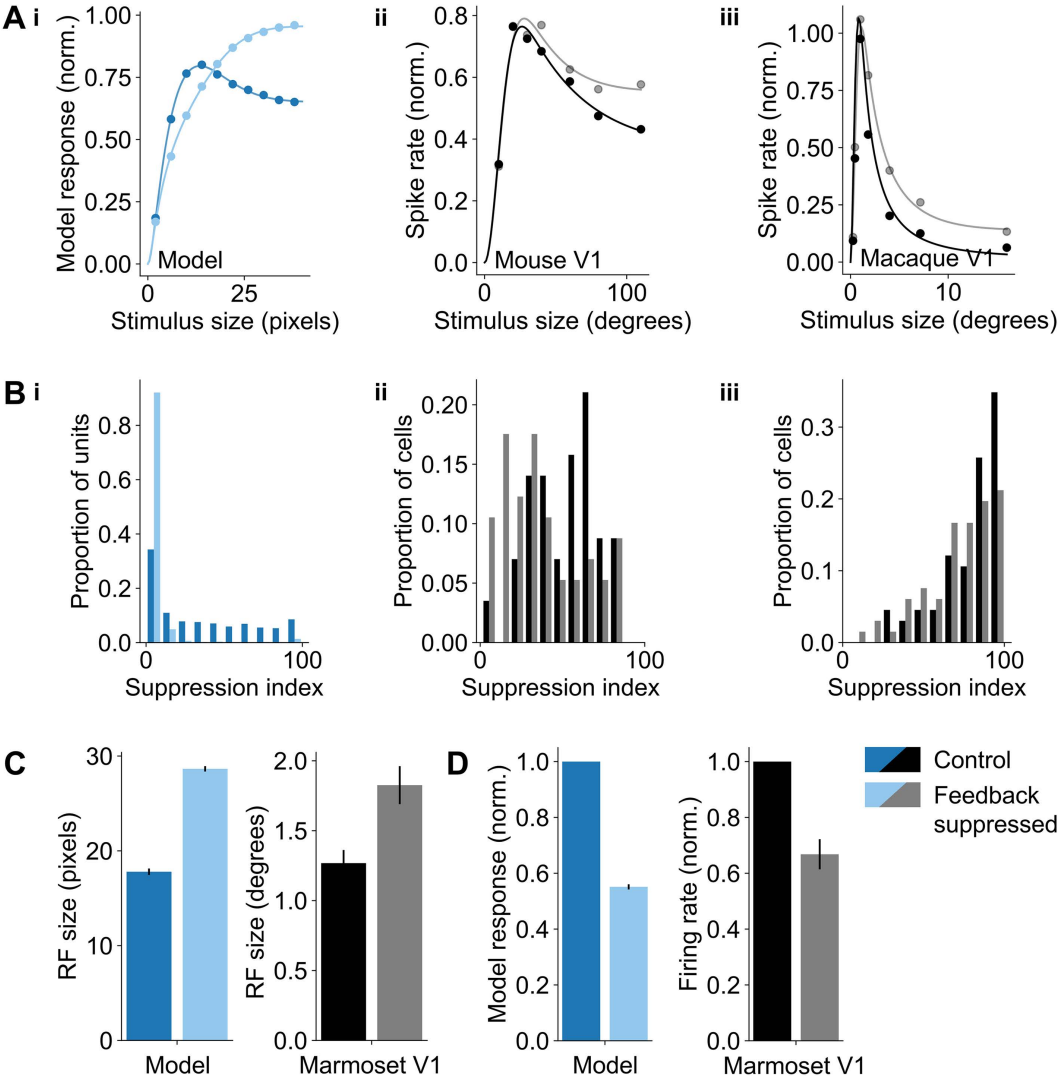

Fig 4. The role of feedback connectivity in surround suppression. (A) Mean response across model G_1 units **(i)**, mouse V1 neurons [15] (ii) and macaque V1 neurons [14] (iii) as a function of stimulus size with and without feedback for model units or with and without inactivation of higher visual areas (mouse) or V2 (macaque). **(B)** Distribution of suppression index values for model G_1 units **(i)**, mouse V1 neurons [15] (ii) and macaque V1 neurons [14] (iii) as a function of stimulus size with and without feedback for model units or with and without inactivation of higher visual areas (mouse) or V2 (macaque). A higher index indicates greater surround suppression. **(C)** Receptive field size is larger when higher-order feedback is suppressed for both model G_1 units (left) and marmoset V1 [13] (right). **(D)** Model responses in the classical receptive field are suppressed when higher-order feedback is suppressed for both model G_1 units (left) and marmoset V1 [13] (right).

https://doi.org/10.1371/journal.pcbi.1013138.g004

suppression was significantly reduced when model feedback was abolished (Fig 4Ai-iii). The average G_1 unit activity for the largest stimulus size was 24.8% smaller for the full model compared with the model where feedback was abolished (Fig 4Ai; t-test, $t(1432)=-4.48$, $p<0.0001$), which is comparable to mouse V1 data [15] (Fig 4Aii, 33.6% smaller response) though less than that for macaque V1 data [14] (Fig 4Aiii, 52% smaller response). This effect was confirmed by considering the suppression index, which describes the extent to which each unit or neuron is suppressed with increasing stimulus size. The average suppression index was significantly lower for the full model (median = 51.6) compared with the model without feedback (median = 19.8; Mann-Whitney U test, $U=519075$, $p<0.0001$), indicating greater surround suppression with the full model (Fig 4Bi). This effect was comparable for both mouse V1 (Fig 4Bii) and macaque V1 (Fig 4Biii), though the reduction in surround suppression was more modest in the biology.

We further analyzed surround suppression in the model with respect to the receptive field properties with and without feedback. Specifically, we analyzed the model's G_1 unit responses to the above stimuli within each unit's classical receptive field (defined as the maximally exciting stimulus size), as well as to such stimuli extending into the proximal receptive field (defined as the ring of visual space area beyond the unit's classical receptive field). In line with data from marmoset V1 [13], the average G_1 model unit's classical receptive field size was significantly larger when feedback was abolished (paired t-test, $t(790)=-29.9$, $p<0.0001$; Fig 4C). We next analyzed the average response of G_1 model units for stimuli located in each unit's classical receptive field, as well as for stimuli that spanned both the classical and proximal receptive fields. As for marmoset V1, the normalized response was significantly smaller for stimuli within each unit's classical receptive field when feedback was abolished (one-way t-test, $t(790)=-48.3$, $p<0.0001$; Fig 4D). However, unlike in marmoset V1, we found that model G_1 units' average response was suppressed for stimuli spanning the classical and proximal receptive field when feedback was abolished, whereas it increased slightly in marmoset V1 (S8 Fig).

The hierarchical recurrent temporal prediction model matches cortical stimulus representations better than other models

We next compared how well different normative models could recapitulate the response properties across the visual pathway in terms of the modulation ratio, plaid motion selectivity and the presence of surround suppression. To probe the impact of recurrency while maintaining a hierarchical representation, we compared the model to a purely feedforward hierarchical temporal prediction network. Similarly, to assess the role of the training objective, we compared the hierarchical recurrent temporal prediction model with a hierarchical recurrent autoencoder. For each model, we quantified the difference between the experimental and model distributions of each measure using the Kolmogorov-Smirnov distance, where a lower value indicates that the two distributions are more similar.

The distribution of modulation ratio values across the model groups was closer to those measured for macaque V1 and V2 neurons in the hierarchical recurrent temporal prediction than in the comparison models (Fig 5Ai, ii). The autoencoder exhibited weak bimodality among the first group of units but, unlike macaque V2 responses, was skewed towards a modulation ratio greater than one in the second group. In contrast, the feedforward temporal prediction model produced only simple-cell-like units in the first group, with complex-cell-like responses emerging only in the second group. Thus, the recurrency in the full model was required for the emergence of simple- and complex-cell-like units within the first group.

A monotonic increase in the mean plaid pattern index was found across the groups in both the feedforward and hierarchical recurrent temporal prediction models, which is consistent with a hierarchical organization of global motion sensitivity (Fig 5Bi, 5Bii). However, the population mean across groups was significantly lower for the feedforward model (t-test, $t(1790)=28.1$, $p<0.0001$), which produced very few plaid-like responses. In contrast, there was relatively little variation in pattern selectivity over component selectivity as a function of model group in the autoencoder model, and the mean pattern index was overall greater than that found in macaque V1 or V2 (t-test, $p<0.022$; Fig 5Biii). Thus, the emergence of pattern-selectivity across the visual pathway was best captured by the hierarchical recurrent temporal prediction model.

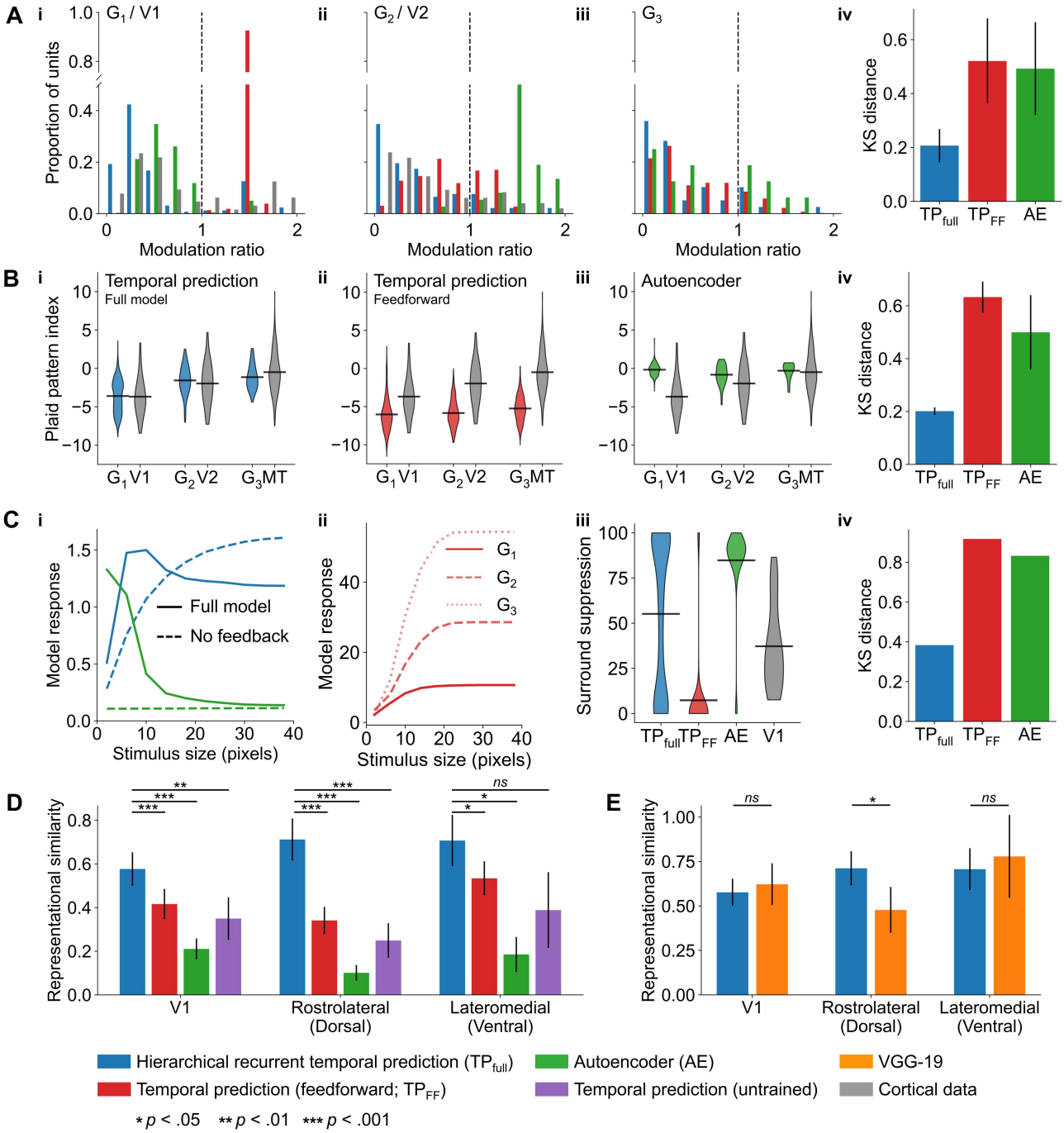

Fig 5. Comparison of tuning properties across the models. (A) Distribution of modulation ratio values across different models for group G_1 and macaque V1 **(i)**, group G_2 and macaque V2 **(ii)**, and group G_3 **(iii)**, plus the Kolmogorov-Smirnov (KS) distance between each of these models and the experimental data **(iv)**. **(B)** Distribution of plaid pattern index values in macaque V1, V2 and MT and across groups for the hierarchical recurrent temporal prediction model (full model) **(i)**, feedforward temporal prediction model **(ii)**, and the autoencoder model **(iii)**, plus the KS distance between these models and macaque data **(iv)**. **(C)** Surround suppression among G_1 units for the hierarchical recurrent temporal prediction and autoencoder models with and without feedback ablated **(i)**, for each group of the feedforward temporal prediction model **(ii)**, the distribution of surround suppression index values for each model and mouse V1 **(iii)**, and the KS distance between each model and mouse V1 **(iv)**. Experimental data in **A-C** as in Fig 2–4. **(D-E)** Representational similarity of the recurrent temporal prediction model relative to comparison normative models with mouse V1, rostrolateral visual cortex and lateromedial visual cortex.

https://doi.org/10.1371/journal.pcbi.1013138.g005

For surround suppression, we first compared the mean responses in G_1 as a function of stimulus size with and without feedback for the hierarchical recurrent temporal prediction and autoencoder models (Fig 5Ci). For the full models with feedback, surround suppression was greater for the autoencoder model than for the temporal prediction model, although the response for the autoencoder was very weak once feedback was abolished. Indeed, when feedback is abolished in the autoencoder model, the units' responses became much weaker but tended to remain above zero (mean unit response of 0.11). Thus, in this context, feedback in the autoencoder model appears to have a disinhibiting effect on the network compared with the more suppressive effect in the hierarchical temporal prediction model. In contrast, for the feedforward temporal prediction model, the mean tuning curve showed no evidence of surround suppression for any group, though the mean response increased with model depth (Fig 5Cii). These results were confirmed by comparing the distribution over units of surround suppression across all models (Fig 5Ciii). Although the bimodal distribution of the hierarchical recurrent temporal prediction model was qualitatively distinct from the distribution of these values in mouse V1, the mean suppression index (mean = 55.1) was closer to the V1 data (mean = 37.2) and the Kolmogorov-Smirnov distance was much smaller than for the other models (Fig 5Civ). The mean suppression index was larger for the autoencoder model (mean = 84.8), where fully suppressed units were overrepresented, whereas only a few units showed any degree of surround suppression in the feedforward model (mean = 7.3).

We also performed a comparison of spatial and temporal frequency tuning across model types. The relative spatial frequency tuning across model groups was relatively consistent across the different models (S9Ai-Ci Fig), showing a U-shaped profile with spatial frequency tuning lowest in G_2. In contrast, temporal frequency tuning was more similar to the experimental data for the hierarchical temporal prediction model (S9Aii Fig) than for either the feedforward model (Fig S9Bii) or the autoencoder model (S9Cii Fig).

To further relate the networks' learned representations to the visual system, we quantified the similarity of the responses of each model to the responses recorded in different regions of the mouse visual cortex to dynamic natural scenes. For each model, we computed the representational similarity by first calculating the response similarity matrices – the correlation of model or neural activity across pairs of stimuli – before measuring the distance between these similarity matrices (S10 Fig). In this way, representational similarity gives a population-level measure of model-brain alignment. For all representational similarity analyses, the reported values are for the highest performing group or layer across each network (these results did not qualitatively change when all units were combined across groups either for the hierarchical recurrent temporal prediction model or its variants) (Fig 5D, 5E).

In terms of brain regions, we considered mouse V1 as well as two early regions of the dorsal (rostrolateral visual cortex) and ventral (lateromedial visual cortex) visual streams. In the primate visual system, the dorsal- and ventral-streams are characterized by their relative selectivity to motion and form, respectively [36]. The mouse visual system similarly may feature dorsal- and ventral-like streams, which have been suggested to respectively contain the rostolateral and lateromedial cortical regions [37–39]. However, the distinction between these streams is less well characterized than in macaque monkeys and perhaps less distinct. For example, one study found that while the rostrolateral region showed strong sensitivity to motion in natural scenes, it was only moderately sensitive to motion in random dot kinematograms [40]. Conversely, a study in anesthetized rats (also in the family *Muridae*) showed more pattern-motion-selective cells in lateromedial than rostrolateral cortex [41], the opposite of findings in mice where rostrolateral cortex was found to have more [42]. Nonetheless, the majority of studies in mice suggest rostrolateral visual cortex is more sensitive to motion than lateromedial visual cortex, broadly consistent with the functional distinction between cortical areas in the primate visual system [40]. Accordingly, mice provide an opportunity to investigate how the motion tuning of the temporal prediction model relates to the model-brain alignment in the stimulus representations across these three visual areas.

For each of these cortical regions, the representational similarity of the hierarchical recurrent temporal prediction model exceeded that of both the feedforward model (paired t-test, all $p < 0.023$) and the autoencoder model (paired t-test, all $p < 0.011$) (Fig 5D). We also compared the full model against an untrained variant and again found that the hierarchical

recurrent temporal prediction model had higher similarity scores in all cortical areas, with significant differences between the models in V1 (paired t-test, $t(13)=4.01$, $p=0.001$) and rostrolateral visual cortex (paired t-test, $t(12)=7.31$, $p<0.0001$), but not in lateromedial visual cortex (paired t-test, $t(9)=2.26$, $p=0.050$) (Fig 5D). Given the greater motion sensitivity of the dorsal stream, we hypothesized that the hierarchical recurrent temporal prediction model would perform better as a model of a dorsal region (rostrolateral visual cortex) versus a ventral region (lateromedial visual cortex) relative to a model optimized on static images (Fig 5E). Accordingly, we compared the hierarchical recurrent temporal prediction model to a model trained for object recognition – namely, VGG-19 [43]. In support of this hypothesis, we found no difference in the representational similarity score between the temporal prediction model and VGG-19 across V1 (paired t-test, $t(13)=2-0.70$, $p=0.499$) and lateromedial visual cortex (paired t-test, $t(12)=2.60$, $p=0.023$), whereas the temporal prediction model was significantly better for the more dorsal rostrolateral visual cortex (paired t-test, $t(9)=-0.41$, $p=0.694$).

Overall, these results show that the distribution of the tuning parameters examined and model-brain alignment were best captured by the hierarchical recurrent temporal prediction model.

Variations on the model

We explored the effects of varying various features of the hierarchical recurrent temporal prediction model – notably, the number of hidden units, the temporal statistics of the training set, the temporal offset of prediction, and the presence or absence of Dale's law. The number of hidden units per group, within a range of 200–1600 hidden units, affected each network's similarity to the visual cortex. For each network, we computed the mean KS distance between the distributions for the network and the brain across modulation ratio, plaid pattern index and surround suppression index (see Fig 5). The minimum KS distance was found for the model with 800 units per group, indicating that this network had the closest correspondence with the empirical distributions found in the biology (S11A Fig). Similarly, we found that the next frame prediction error reached a minimum at around 800 hidden units per group, while the same-frame prediction error reached a maximum at this same sized network (S11B Fig), indicating that greater similarity to the biology tended to be accompanied by improved temporal prediction performance.

We would expect that the statistics of the training dataset will have an impact on the representations learned by the network. Indeed, at an extreme, we found that training the network on the same dataset but removing temporal contiguities by shuffling the order of frames within a clip, prevents the network from learning useful representations. For example, compare the receptive fields learned by the network where the frames are shuffled (S12Ai Fig) with those learned by the standard temporally contiguous dataset, which show a clear Gabor-like structure (S12Aii Fig). Nevertheless, the network was trained on a large diverse dataset of naturalistic stimuli (S12B Fig; example frames) and we believe that any sufficiently rich, naturalistic dataset like the one used in this study would result in broadly similar network properties.

To better understand the specific role of temporal prediction in the network model, we interpolated between an autoencoder and temporal prediction model by shifting the temporal offset of the frame target from 0 to 42 ms into the future (see Methods). For each temporal offset, we computed an aggregate score as the mean Kolmogorov-Smirnov (KS) distance for each of the measures described in Fig 5 – namely, the modulation ratio, plaid pattern index and surround suppression index (S13A Fig). Overall, we found that the 0 ms temporal offset (autoencoder) model had the highest score – and was therefore most different from the comparison neural data – whereas the model trained to predict a target 42 ms into the future was most similar to the neural data.

We next fitted each G_1 unit with either a Gabor or Difference of Gaussians (DoG) function (S13B Fig). As the temporal offset increased, G_1 units became more Gabor-like, and less DoG-like. Finally, we investigated the representational similarity of each network to the mouse visual cortex as we increased the temporal offset (S13C Fig). Overall, we found that models with a temporal offset greater than 8 ms best captured the representational geometry of the mouse visual cortex across all three areas surveyed. Interestingly, we found that for both V1 and lateromedial visual cortex (part of the mouse ventral visual cortex), further increases in representational similarity saturated at 17 ms into the future. Conversely, we

found that for rostrolateral visual cortex (part of the mouse dorsal visual stream), there was a distinct benefit to predicting at 25 ms into the future, implying that this region of the dorsal stream may be optimized for a precise temporal horizon of prediction.

To probe the role of Dale's law in the network's learned properties, we compared two networks trained for hierarchical temporal prediction, differing only in the inclusion of Dale's law. Overall, including or excluding Dale's Law had almost no impact on next-frame prediction performance (S14A Fig), and, visually, the two networks' next frame predictions were very similar (S14B Fig). Nevertheless, in terms of the networks' learned properties, we found that the model with Dale's law was more like the neural data, implying that this added biological detail produced a more brain-like network. In particular, we found that the distributions of modulation ratios (S14C Fig), plaid pattern indices (S14D Fig) and surround suppression indices (S14E Fig) were more like those reported in the visual cortex for the model that included Dale's law.

Discussion

The principle of temporal prediction argues that the sensory brain is optimized to predict immediate future input based on the recent past. In line with this principle, we hypothesized that a hierarchical recurrent model optimized for temporal prediction should recapitulate the functional organization of visual cortex. With little fine tuning to the model, the network exhibited tuning properties akin to those found across several visual cortical areas, including not only the distribution of simple- and complex-like responses found in macaque V1 and V2, but also feedback-dependent surround suppression and the emergence of global motion sensitivity across those visual regions leading to area MT. Compared with alternative normative models, the response properties of visual cortical neurons were best captured by the hierarchical recurrent temporal prediction model, which similarly exhibited the closest alignment with the stimulus representations found in mouse visual cortex. Together, these results provide evidence for temporal prediction as an organizational principle across the visual cortex.

Relation to biology

As a normative model, the hierarchical recurrent temporal prediction model represents a fairly abstract representation of the visual brain [3,44], rather than focusing on low-level mechanistic details. It is important to find a careful balance between incorporating sufficient details to make meaningful comparisons with the biology possible, while avoiding potentially redundant features that could obscure the role of the more general normative principle of interest. In the case of the current temporal prediction model, the network's units obey Dale's law – projecting exclusively excitatory or inhibitory connections – and incorporate both local recurrent and feedback connectivity. These fundamental properties of neural circuits are generally omitted from normative models of sensory processing, particularly those implemented as deep object recognition networks [10,45].

However, as point-like neurons, the model's units are fairly abstract representations of true biological neurons. That is, they lack different neuronal compartments and their activity is rate based and omits spiking. Nevertheless, these neuronal properties are fully compatible with temporal prediction as a principle and implementing them in other studies has enabled the action potential firing patterns and membrane time constants of neurons to be reproduced [46,47]. Furthermore, the model is trained via backpropagation, which is generally considered to be biologically implausible [48]. However, the model itself is agnostic about the learning rule – certain elements of temporal prediction could be genetically hard-wired or arise via competitive activity-dependent mechanisms during development – and novel more biologically-plausible algorithms are being developed, which could, in principle, be applied to the current network [49,50].

As well as adding more biological elements to the temporal prediction model, we could compare the trained model to more diverse features of the biology, for example, frequency domain aspects. When a nonlinearly-integrated Gabor-filter bank model (inspired by earlier theoretical work, [51]) is fitted to MT neurons' responses to naturalistic stimuli, the excitatory spectral receptive fields of the fit tend to fall on a plane within the three-dimensional spatiotemporal frequency

domain, but the suppressive spectral receptive fields are found off this plane [52]. This ensures that MT neurons are velocity tuned but are insensitive to ambiguous direction-aligned static textures. There is also evidence of somewhat similar properties in a small fraction of V1 neurons [53]. An interesting future extension of our work would be to fit this model to the temporal prediction model so that these frequency domain properties could be examined, particularly for the higher groups in our model.

Comparison with alternative models

In the current study, we compared the hierarchical recurrent temporal prediction model to several alternative models to better understand how each model's features related to their capacity to capture different elements of the visual system.

The feedforward temporal prediction model incorporated both a hierarchical organization and the temporal prediction objective – that is, each higher group predicted the future activity of its lower order inputs, but in contrast to the full temporal prediction model, did not include local recurrency or feedback connectivity. As shown in previous work, the distribution of model response properties in the feedforward model mirrored some aspects of the dorsal visual pathway [4]. However, plaid motion selectivity and surround suppression were considerably weaker than in the recurrent model, and the feedforward model performed worse in terms of its representational similarity to the visual cortex. Accordingly, the architectural addition of local and long-range recurrency had a large impact in improving the model's fit to the visual system. Similarly, in terms of the model objective, the hierarchical recurrent autoencoder model performed worse on the response similarity measure and did not match the visual system's response properties as well as the temporal prediction model. Thus, both the temporal prediction objective, as well as the model's architecture, were important in producing a brain-like model.

The autoencoder used exactly the same architecture as the temporal prediction networks but was substantially worse at explaining the neural responses. This suggests that it is the temporal prediction objective that is particularly critical for reproducing the biological data. However, it is still the case that appropriate architecture, constraints and training data can help produce brain-like properties. This includes hierarchy and recurrency (Fig 5), Dale's law S14 Fig, appropriate prediction offset (S13 Fig), and temporally-structured training stimuli (S12 Fig). Similarly, we have shown in earlier work that such factors are important for simple feedforward [5,54], hierarchical feedforward [4], and non-hierarchical recurrent temporal prediction networks [6]. Our previous studies also explored the importance of other constraints, such as connectivity strength (L_1 regularization strength) and noise levels. Thus, these further constraints will also likely be relevant to the hierarchical recurrent temporal prediction model.

For temporal prediction models, these constraints are often tuned to optimize prediction, rather than simply to make the model look like the brain. Hence, temporal prediction has an independent criterion for tuning hyperparameters, something the autoencoder lacks. In contrast, an autoencoder with exactly these same constraints can only poorly replicate brain-like properties, even when these constraints, such as the L_1 regularization, are tuned to make the network more like the brain rather than an independent criterion. Furthermore, our earlier work [4–6] has also shown that temporal prediction can account for various neuronal properties better than sparse coding, predictive coding, slowness, inpainting or denoising models. Together, this suggests that temporal prediction represents a good normative principle for sensory processing in the brain.

Several other studies have also modeled the visual system by applying unsupervised learning principles to dynamic spatiotemporal inputs. In particular, a number of studies have investigated how well dual-stream networks can model both the dorsal and ventral streams when trained on dynamic moving visual inputs. Bakhtiari et al. [55] demonstrated that a dual-stream network trained for contrastive temporal prediction can develop dorsal stream-like representations that recapitulate some elements of motion processing, such as random dot kinematogram selectivity. Although more abstract in its form than the hierarchical recurrent temporal prediction model, their contrastive model underscores the value of temporal prediction as an unsupervised principle for modeling sensory areas. However, unlike in this study, their contrastive objective was trained to predict the *latent* state of the future frame rather than the pixel-wise future frame itself. Considering the successes of contrastive methods elsewhere [56], it will be instructive to explore this form of temporal prediction in future work.

Similarly, other studies employing a dual-stream network, such as Cadieu and Olshausen [57], have described how optimizing a dual-pathway network – in their case to decompose inputs into amplitude and phase information – can produce ventral- and dorsal-like representations. After training, the optimized network produces simple- and complex-cell-like responses, as well as units with various kinds of form and motion invariance. Although their model was assessed against fewer physiological properties than in the current study, it illustrates the utility of such dual-stream models, which could be integrated into the temporal prediction framework in future.

Two further studies examined whether properties of the dorsal pathway could be explained by networks optimized to estimate certain gross velocity measures from visual input. These measures were either the overall direction and speed of the whole moving scene [58] or head and body motion of agents in simulated environments [59]. The first study showed that their network could explain neural and perceptual biases for particular motion directions, the interrelation between speed and direction preferences in MT neurons, and certain motion illusions. The second study showed that their network to a degree captured in a hierarchical manner the responses of neurons along the visual pathway to naturalistic stimuli. An interesting extension of this would be to develop a multisensory hierarchical recurrent temporal prediction model that combines temporal prediction of these gross velocity measures, as might be provided by vestibular inputs, with the prediction of spatially detailed future visual input that we use here. We suspect that both objectives are important, with different properties of the dorsal pathway being more relevant to one or the other objective.

Another related body of work is the temporal straightening hypothesis. Human psychophysics [60], V1 neural recordings from macaques [61] and some feedforward mechanistic models of the visual system [60] suggest that the visual system transforms natural visual inputs to follow straighter temporal trajectories. Such a transformation has been argued to support temporal prediction in natural environments [60,61]. A feedforward visual processing model optimized for temporal straightening was found to increase recognition robustness more than a similar invariance-optimized model [62]. Conversely, visual processing models adversarially optimized for robustness were found to increase temporal straightening, something not necessarily found in other visual processing models [63]. It would be interesting to examine in future work the effect of the hierarchical temporal prediction objective on both recognition robustness and temporal straightening in our recurrent networks.

Finally, predictive coding represents another hierarchical normative framework, which has been very influential when applied to sensory processing. Predictive coding argues that the brain is optimized to reduce statistical redundancies by passing forward only the residual prediction errors that cannot be “explained away” by the brain’s internal model [20]. Predictive coding is inherently hierarchical in that it is organized around a series of stacked prediction modules, which – like temporal prediction – is consistent with a hierarchical view of cortical organization. The original model by Rao and Ballard [21] was able to account for non-linear effects in V1, such as surround suppression, though it was only trained on static images. In contrast, a more recent variant on predictive coding – PredNet [64] – has also been trained on dynamic natural movies. While PredNet could account for some properties such as surround suppression [65], it does not generally recapitulate low-level features of the visual cortex, such as the Gabor-like receptive fields of V1 [6], nor has it been shown to capture the hierarchical organization of the visual system.

In conclusion, we have shown that a hierarchical recurrent network optimized for temporal prediction replicates many aspects of the neuronal response properties and functional organization of different areas within the mammalian visual cortex. These findings add to the growing evidence that temporal prediction can account for information processing across the sensory hierarchy.

Materials and methods

Model dataset

The model was trained using a novel dataset drawn from around 2.5 hours of wildlife videos (obtained from public sources, which are available via the Open Science Framework wrapper https://osf.io/hf2pj/) down sampled to 25–30 frames per

second. Videos were pre-processed by converting them to grayscale, resizing each clip to a width and height of 600 pixels using bilinear interpolation and applying a bandpass filter. Each video was then cropped into 1200 spatially overlapping clips of 20x40 pixels at 40 contiguous frames each. Finally, each video clip was normalized by subtracting its mean and dividing by its standard deviation. This produced a training, test and validation dataset of 1000k, 200k and 200k clips, respectively.

The hierarchical recurrent temporal prediction model

The model is implemented as a recurrent network consisting of a series of hierarchically organized groups of units (Fig 1A). During training, each group in the model is optimized to predict the future value of its inputs – specifically, the future video frame for group one (G_1) units or the future internal state of the lower-order group for group two (G_2) and above. Internally, the model is implemented as a single recurrent layer, with the hierarchy defined by 1) the hierarchy of prediction whereby higher-order areas are trained to predict the future value of their lower-order inputs, 2) the restricted internal connectivity, such that each group receives inputs from and projects to its immediate higher- and lower-order groups only, and 3) only G_1 receiving direct video input. This internal hidden state is then mapped to the predicted future state by a linear output layer, with the prediction error between the true and expected future state minimized during training via backpropagation. Finally, an L_1 regularization penalty in the cost function ensures that only weights that contribute to network performance are retained. Below we will describe the hierarchical recurrent temporal model in detail and more formally, largely using scalar notation rather than the vector notation used in the Results.

The model receives as input a tensor U of shape $T \times I$ where $t = 1$ to 40 timesteps and $i = 1$ to 800 pixels as a flattened 20×40 pixel image. The network itself consists of 3 groups, $g = 1$ to 3, where each group g comprised $j = 1$ to 800 units. At time t, the activity of each hidden unit $h_j^g[t]$ is defined as:

$$a_j^g[t] = b_j^g + \sum_{j'=1}^{J^g} r_{j,j'}^{g,g} h_{j'}^g[t-1] + \begin{cases} \sum_{i=1}^{I} w_{j,i} u_i[t] + \sum_{j'=1}^{J^2} r_{j,j'}^{1,2} h_{j'}^2[t-1] & g = 1 \\ \sum_{j'=1}^{J^{g-1}} r_{j,j'}^{g,g-1} h_{j'}^{g-1}[t-1] + \sum_{j'=1}^{J^{g+1}} r_{j,j'}^{g,g+1} h_{j'}^{g+1}[t-1] & g > 1 \\ \sum_{j'=1}^{J^{g-1}} r_{j,j'}^{g,g-1} h_{j'}^{g-1}[t-1] & g = G \end{cases}$$

$$h_j^g[t] = \mathrm{f}\left(a_j^g[t]\right)$$

where b_j^g is the bias of group g unit j, $w_{j,i}$ is the input weight between pixel i and group one unit j, $u_i[t]$ is the activity of pixel i at timestep t, $r_{j,j'}^{g,g'}$ is the recurrent weight from presynaptic unit j' in group g' to postsynaptic unit j in group g, and $h_{j'}^{g'}[t-1]$ is the activity at the previous timestep of presynaptic unit j' in group g'. The presynaptic group g' is either the postsynaptic group g for local recurency, or the lower group $g-1$ or the upper group $g+1$. The function f(•) is a ReLU.

In addition, to enforce Dale's Law whereby units have exclusively excitatory or inhibitory connections, each recurrent weight $r_{j,j'}^{g,g'}$ was clamped after each forward pass as:

$$r_{j,j'}^{g,g'} \rightarrow \begin{cases} + \left| r_{j,j'}^{g,g'} \right| & \text{if excitatory} \\ - \left| r_{j,j'}^{g,g'} \right| & \text{if inhibitory} \end{cases}$$

where 20% of units in each group were set as inhibitory (based on the approximate percentage of cortical inhibitory interneurons found in the cortex [66,67]), with the additional constraint that inhibitory units could only project locally.

Finally, the network's internal hidden state was mapped to the output predictions as $y_k^g[t]$ for group g, unit k at time t, which was defined as:

$$y_k^g[t] = c_k^g + \sum_{j=1}^{J^g} m_{k,j}^{g,g} h_j^g[t]$$

where c_k^g is the bias for unit k of group g and $m_{k,j}^{g,g}$ is the weight between hidden unit $h_j^g[t]$ and output unit $y_k^g[t]$.

The trainable parameters in the network, b_j^g, $w_{j,i}$, $r_{j,j'}^{g,g'}$, c_k^g and $m_{k,j}^{g,g}$, were optimized by minimizing the loss function:

$$\mathrm{L} = \lambda Z + (1-\beta)E^1 + \sum_{g=2}^{G} \beta E^g$$

where Z is the sum of the absolute values of all weights in the network (the sum of elementwise L_1-norms of each weight matrix). The parameter λ is a weighting parameter that describes the degree of regularization, and β is a weighting parameter that determines the relative contribution of the prediction error between the first and higher-order groups. Finally, E^g is the mean squared error between the true and predicted future value of the lower-order group $g-1$:

$$E^g = \frac{1}{N(T-4)K^g} \sum_{n=1}^{N} \sum_{t=4}^{T-1} \sum_{k=1}^{K^g} \begin{cases} \left(y_{k,n}^1[t] - u_{k,n}[t+1]\right)^2 & g = 1 \\ \left(y_{k,n}^g[t] - h_{k,n}^{g-1}[t+1]\right)^2 & g > 1 \end{cases}$$

where, for group one, $y_{k,n}^1[t] - u_{k,n}[t+1]$ is the difference between the predicted and true future pixel value, and, for higher order groups, $y_{k,n}^g[t] - h_{k,n}^{g-1}[t+1]$ is the difference between the predicted and true future value of the lower order group. Finally, E^g is averaged across all clips N, time steps T and pixels I. Subscript n indicating clip number was left off from all previous equations for brevity.

Comparison models

Three comparison models were developed in addition to the hierarchical temporal prediction model as described above.

1. Feedforward temporal prediction model – this model was trained to maximize the same loss function as the hierarchical temporal prediction model but using a strictly feedforward architecture. The training and architectural details were adapted from previous work [4], though we only trained three stacks (groups), and adapted the kernel sizes as detailed in Table 1 to equate the receptive field sizes in the hierarchical recurrent models:

2. Hierarchical recurrent autoencoder – the same architecture as the hierarchical recurrent temporal prediction network but with the loss function altered to predict the activity of the current rather than future lower-order group, with an

Table 1. Details of the feedforward temporal prediction model.

Stack	Input size	Hidden layer size	Kernel size
0	800x40x1	800x36x1	5x800
1	800x36x1	800x32x1	5x1
2	800x32x1	800x28x1	5x1

https://doi.org/10.1371/journal.pcbi.1013138.t001

additional regularization term based on the summed activity in the network to ensure the sparsity of the network's internal representations (λ_α was set to 1×10^{-10}).

$$\text{Loss} = \lambda Z + \lambda_\alpha \frac{1}{N(T-4)} \sum_{g=1}^{G} \sum_{n=1}^{N} \sum_{t=4}^{T-1} \sum_{j=1}^{J^g} h_{j,n}^{g}[t] + (1-\beta)E^1 + \sum_{g=2}^{G} \beta E^g$$

$$E^g = \frac{1}{N(T-4)K^g} \sum_{n=1}^{N} \sum_{t=4}^{T-1} \sum_{k=1}^{K^g} \begin{cases} \left(y_{k,n}^{1}[t] - u_{k,n}[t]\right)^2 & g = 1 \\ \left(y_{k,n}^{g}[t] - h_{k,n}^{g-1}[t]\right)^2 & g > 1 \end{cases}$$

3. Untrained hierarchical recurrent temporal prediction model – the same architecture as described above, but with no training and hence randomly initialized weights.

One other pre-trained model developed by another group was also included to compare its capacity to capture the neural representation of naturalistic visual stimuli:

1. VGG-19 [43] – a deep convolutional network trained on the ImageNet database to classify images into one of 1000 categories. This network is purely feedforward and does not incorporate any form of recurrency.

Finally, for S9 Fig, we trained a series of hierarchical recurrent temporal prediction networks by varying the temporal horizon of the network – that is, how far into the future the network was trained to predict the next frame. For these networks, although the network architecture was identical to the hierarchical recurrent temporal prediction model, we used a high-frame-rate (120 Hz) naturalistic dataset which allowed us to vary the target from 0 to 42 ms into the future. Note that due to the smaller size of this dataset (40k clips total), we used a smaller 20x20 pixel clip size for model training. For more details about the dataset implementation, see [6].

Implementation

The temporal prediction model and its variants were all implemented in PyTorch, with gradient descent performed using the ADAM optimizer set at a learning rate of $\alpha = 10^{-4}$. The hyperparameter β was set at 0.1, as this value gave cortex-like properties for the different model types and good out-of-distribution prediction for moving bars. More systematic exploration of hyperparameter λ was performed, which was then set as 10^{-6}, close to the global minimum that minimized the mean squared error for next frame prediction on the validation dataset while producing the most biologically realistic receptive fields. For the temporal prediction model, we found that the optimal λ value produced receptive fields that were similar to but not as spatially distinct as those found in the cortex, so we chose a nearby higher value that gave more spatially defined receptive fields while remaining close to the normative-driven optimum (i.e., the prediction error remained small). For the autoencoder model, where there is a less well-defined objective value for the hyperparameters (the reconstruction error almost always improves for a less 'constrained' network), we chose a value for λ that gave the most spatially well-defined receptive fields and which best represented what is found in the visual cortex. Note that for Fig 1E-1K, where a large number of model variants were investigated, for computational efficiency, these models were trained on a smaller 20x20 pixel dataset [6], rather than the 20x40 pixel dataset used throughout the rest of the study. All networks were trained for a maximum of 4000 epochs or until the loss function converged on a steady state.

The recurrent autoencoder and recurrent temporal prediction networks had exactly the same parameter counts (13.4 million trainable parameters) versus a slightly lower number for the feedforward temporal prediction network (11.5 million trainable parameters). This represents a small (14%) reduction in parameters, largely due to the absence of feedback.

While we could have produced a feedforward model that equalized the number of parameters, this would have come at the expense of increasing the number of hidden units in the feedforward model. Thus, there is an inevitable trade-off between equalizing the number of total parameters versus the total number of hidden units across networks.

In terms of the experiments where we ablated feedback from a model that initially included feedback connections, this necessarily reduced the total parameter count. However, feedback was removed via ablation, analogous to the experimental data, where reducing feedback in the actual brain involves reducing the number of connections. Thus, we believe this ablation approach is the most appropriate for the questions we were trying to answer about the biology.

All code supporting this study is publicly available at https://github.com/sebbkw/hierarchical_temporal_prediction.

Data analysis

Receptive field estimation. Receptive fields were estimated by the response-weighted average [6], using the response of each unit in the network to 200,000 frames of random Gaussian noise.

Unit tuning characteristics. To assess each unit's tuning properties, we measured its response to sinusoidal gratings varying in temporal frequency, spatial frequency and orientation for 50 frames. Each unit's preferred temporal frequency, spatial frequency and orientation was taken as the parameter or parameter combination that maximized the unit's mean response across frames.

Because the input visual stimuli used to train and analyze the network have no fixed reference point, we do not think there is a reliable way of converting between pixel space and degrees in visual space. Thus, we take the comparison between experimental measurements in cycles/degree and model results in cycles/pixel as best understood on a relative, rather than absolute, basis.

Receptive field parameterization. For S9B Fig, we fitted a Difference-of-Gaussians (DoG) or Gabor function (equated to the same number of free parameters) to each receptive field estimate. We then defined each unit as DoG- or Gabor-like if the receptive field estimate had a higher pixel-wise Pearson correlation coefficient with the DoG or Gabor function, respectively, and if the correlation coefficient was at least 0.8.

Modulation ratio. The modulation ratio was computed as $F = F_1/F_0$ where F_0 is the mean response of the neuron to its preferred stimulus and F_1 is the amplitude of the fitted sinusoid to the neuron's response to its preferred stimulus.

Orientation and direction selectivity. Orientation selectivity was quantified via the orientation selectivity index (OSI):

$$OSI = \frac{R^{\mathrm{Or}}_{\mathrm{pref}} - R^{\mathrm{Or}}_{\mathrm{orth}}}{R^{\mathrm{Or}}_{\mathrm{pref}} + R^{\mathrm{Or}}_{\mathrm{orth}}}$$

where $R^{\mathrm{Or}}_{\mathrm{pref}}$ and $R^{\mathrm{Or}}_{\mathrm{orth}}$ are respectively the unit responses at the preferred and orthogonal orientations for that unit. Similarly, direction selectivity was quantified via the direction selectivity index (DSI):

$$DSI = \frac{R^{\mathrm{Dir}}_{\mathrm{pref}} - R^{\mathrm{Dir}}_{\mathrm{opp}}}{R^{\mathrm{Dir}}_{\mathrm{pref}} + R^{\mathrm{Dir}}_{\mathrm{opp}}}$$

where $R^{\mathrm{Dir}}_{\mathrm{pref}}$ and $R^{\mathrm{Dir}}_{\mathrm{opp}}$ are respectively the unit responses at the preferred and opposite directions for that unit.

Plaid responses: Plaid stimuli were produced by summing two drifting gratings, each at half amplitude (0.5). The gratings differed by a 'plaid angle' – the angular separation of each plaid component ($|\alpha - \beta|$) around the overall direction of the plaid stimulus ($\frac{\alpha+\beta}{2}$). To identify the degree to which units responded to the direction of the individual plaid components versus the overall direction of the plaid stimulus, we computed the partial correlation of the unit's response to the plaid stimulus with an idealized response assuming the unit to be entirely component or plaid selective [35]. Thus, the plaid correlation r_p was defined as the correlation between the true plaid response and the predicted plaid response, controlling for

the predicted component response. Similarly, the component correlation r_c was taken as the correlation between the true component response and the predicted component response, controlling for the predicted plaid response. To compare across units, we computed the Fisher's *z*-transformations Z_p and Z_c of r_p and r_c, respectively. This process was repeated across component separation angles of 60, 90, 120 and 150 degrees. The plaid pattern index was then taken as the difference between the average Z_p and Z_c values across all component separation angles. A value greater than 0 indicates greater selectivity for the plaid stimulus over its individual components, whereas a value less than 0 indicates greater selectivity to the individual components over the composite plaid stimulus.

Surround suppression. For each G_1 model unit, we first fitted a Gabor function to each unit's response-weighted average to extract the receptive field center. We then presented a drifting grating stimulus determined by the unit's preferred orientation, spatial and temporal frequency, with a mask applied to localize the grating stimulus to the unit's receptive field center. We recorded the response of that unit as the diameter of the mask was increased from 2 to 40 pixels. We calculate the surround suppression index based on this surround suppression tuning curve as:

$$\text{surround suppression} = 100\frac{r_{\max} - r_{\text{large}}}{r_{\max}}$$

where r_{max} is the unit's maximum response across all grating diameters and r_{large} is the unit's response to the large diameter stimulus.

Model-brain alignment

Neural data were taken from the Allen Brain Institute's Neuropixels Visual Coding dataset of *in vivo* electrophysiological recordings of the mouse brain [68]. All recordings were pre-processed and spike-sorted by the Allen Institute. For these analyses, only data from the natural movie presentations were included ("Natural Movie One" and "Natural Movies Three", 150 seconds total). All recordings were from wildtype mice and data from V1 ($n = 15$ mice), rostrolateral visual cortex ($n = 14$ mice) and lateromedial visual cortex ($n = 11$ mice) were considered. In addition, only those units whose noise power to signal power ratio did not exceed 60 were included [69].

To calculate the representational similarity between each model and the neural data, we first computed the representational similarity matrix (RSM) [55,70]. For this, we binned the responses of the model or real neurons into 5-frame chunks and took the mean of each bin such that each model's output was represented as an $M \times N$ matrix with M rows of stimuli and N columns of units. The RSM was calculated as the correlation between each pair of columns in the matrix (the response to a given stimulus m), resulting in an $M \times M$ RSM. The representational similarity was then calculated by taking the distance between the flattened lower triangle of the RSMs of each model and the neural data using Spearman's correlation coefficient [71].

To compare across mice, this similarity measure was normalized by dividing by the noise ceiling for that recording, where a value ≥ 1 indicates that model similarity is as high as could be achieved once accounting for random noise [55]. The noise ceiling describes the theoretical maximum similarity that could be achieved by even a perfect model given the inherent noise present in the neural system. It was computed by randomly dividing the neural data into two halves of units and by taking the similarity of the RSMs for each half of units. This shuffling process was then repeated for 100 iterations, with the noise ceiling taken as the mean similarity value across these 100 shuffle iterations.

Supporting information

S1 Fig. Visual illustrations of the groupings and constraints applied to the weight matrices and activity vectors of a single hidden-layer recurrent neural network to produce the model. The squares and broad vertical rectangles illustrate the weight matrices. The shade (dark/mid/light) indicates what group (G_1/G_2/G_3) the weights in

that region of the matrix project from. The color (white/red/blue/gray) indicates whether those weights are constrained to be zero (white), positive (red), negative (blue) or unconstrained (gray). The narrow vertical rectangles illustrate the vectors, with shade indicating group and color indicating whether they are excitatory units (red), inhibitory units (blue) or unconstrained output units (gray). Vectors have been widened slightly for visualization purposes only. Bias vectors have been left off for simplicity. **(A)** Illustration of the equation for the hidden unit activity (Equation 1). This shows the input vector, $\mathbf{u}[t]$, the input weight matrix, $\mathbf{W}$, the hidden unit vector, $\mathbf{h}[t]$, and the recurrent weight matrix, $\mathbf{R}$. Note the inhibitory (blue) and excitatory (red) units of the hidden unit vector and their corresponding constrained weights in the recurrent matrix. **(B)** Illustration of the equation for the network output (Equation 2). This shows the output vector, $\mathbf{y}[t]$, and output weight matrix, $\mathbf{M}$, along with the hidden unit vector to which the output weights are applied. **(C)** Illustration of the loss function (Equation 3), using the input vector and groups ($\mathbf{h}^1[t]$, $\mathbf{h}^2[t]$, $\mathbf{y}^1[t]$, $\mathbf{y}^2[t]$, $\mathbf{y}^3[t]$) from the hidden and output vectors.
(TIFF)

S2 Fig. Next-frame prediction performance as a function of the beta hyperparameter. Next-frame prediction error increased as a function of beta for both the in-distribution **(A)** and out-of-distribution datasets **(B)**.
(TIFF)

S3 Fig. MNIST decoder accuracy across n-back frames when training on a 0-back decoder for all n-back conditions. Here we trained a 0-back decoder and tested this fixed decoder for each n-back condition. We found that the decoder – in both the full and no feedback conditions, as well the shuffled feedback condition – was unable to generalize from the 0-back condition and performed at chance. Given the performance of the model in Fig 1J-1K when trained and tested on the same n-back condition (with separate train and test splits), it is unlikely that these fixed 0-back decoder results reflect an absence of information in the network over time. Rather, this suggests that using a fixed decoder is not sufficient and that it needs to be re-trained for each n-back condition to extract the information, as we do in Fig 1J-1K. **(A)** MNIST accuracy declines to chance at decoding from the n-back condition if a fixed 0-back decoder is used, for both the full and no feedback models. **(B)** MNIST accuracy at 5-back performs at chance as the percentage of feedback is shuffled for a fixed 0-back decoder. In both cases, chance performance indicates that the decoder is unable to generalize from the 0-back condition.
(TIFF)

S4 Fig. Joint distribution of modulation ratios with the orientation selectivity index (OSI) and Gabor fit r^2 for the first model group (G_1). The simple-cell-like units (modulation ratio > 1) had higher mean orientation selectivity **(A)** and were better fit by the Gabor function **(B)** than the complex-cell-like units.
(TIFF)

S5 Fig. Gabor-like receptive field properties across the model's hierarchy. (A) Only units in the model's first group could be modeled as Gabors, with higher groups showing a poorer fit (Gabor r for G_1 vs G_2: $t(16.6)=7.43$, $p<0.0001$). **(B)** For the subset of units that could be modeled as a Gabor ($r>0.3$), the receptive field size – indexed by the product of the x and y Gabor standard deviation parameters – increased from group 1 to group 2 (Gabor size for G_1 vs G_2: $t(7.04)=-4.22$, $p=0.004$).
(TIFF)

S6 Fig. Functional connectivity of group one (G_1) model units. (A-B) Short-range connections between model units (black bars) are more prevalent for excitatory units that have similar orientation tuning **(A)** and direction-tuned units that have similar or opposite preferred directions of motion **(B)**, as is also the case in V1 neurons (gray bars) [31]. **(C-E)** In both the model and V1 [32], long-range connection probability is higher for units with similar orientation preferences when their receptive fields are located in co-axial **(C)** than in co-orthogonal **(D)** locations. Heatmap **(E)** shows the normalized

connection probability over visual space (relative to postsynaptic neuron position) across differences in orientation tuning for model units. For detailed methods, see Klavinskis-Whiting et al. [6].
(TIFF)

S7 Fig. Example model units (i-iii) showing changes in plaid tuning after the removal of feedback connections. Direction tuning curves for plaid stimuli across different plaid angles are shown for three example model units with (left) and without (right) feedback.
(TIF)

S8 Fig. Model and V1 responses to stimuli in the classical and proximal receptive field. (A) Diagram of the classical and proximal receptive field for an exemplar neuron. **(B)** The full model's average response rate is higher for stimuli spanning the classical and proximal receptive field than when feedback is suppressed. This contrasts with the data from marmoset V1, where the opposite trend is observed, such that the average neural response for stimuli spanning the classical and proximal receptive field is greater when feedback is suppressed [13].
(TIFF)

S9 Fig. Spatial and temporal frequency tuning across model types. Spatial frequency tuning (i) and temporal frequency tuning (ii) across the **(A)** hierarchical recurrent temporal prediction model (blue), **(B)** feedforward temporal prediction model (green) and **(C)** autoencoder (red) models compared with data from the macaque visual cortex (gray). Experimental data as in Fig 2F,H.
(TIFF)

S10 Fig. Schematic of the representational similarity analysis procedure.
(TIFF)

S11 Fig. Impact of varying the number of hidden units per group in the model. (A) Mean KS distance between model and neural distributions was lowest at 800 units per group. **(B)** As the number of hidden units increased, the model shifted from representing the current frame (same frame MSE, which increased as the number of hidden units increased) to predicting the upcoming frame (next-frame MSE, which decreased as the number of hidden units increased). Thus, a larger number of hidden units improved the network's capacity for temporal prediction.
(TIFF)

S12 Fig. Impact of network training dataset and exemplar frames. (A) A network trained on the same dataset but with temporal contiguities removed by shuffling the frame order fails to produce coherent receptive fields (i). Conversely, the standard temporally-contiguous dataset produces receptive fields with a clear Gabor-like structure (ii). **(B)** Example frames from the dataset used for model training.
(TIFF)

S13 Fig. Role of the temporal prediction offset in network properties. (A) The mean KS distance between the model and brain distributions of modulation ratios, plaid pattern indices and surround suppression indices. A lower score indicates a more brain-like distribution in the model, with the autoencoder (0 ms temporal offset/future prediction) having the least brain-like distribution. **(B)** The distribution of Gabor-like versus difference-of-Gaussian-like units in model group 1 (G_1) across temporal offsets, with the highest proportion of Gabor-like units found when predicting the next frame 42 ms into the future. **(C)** Representational similarity between the model and different regions of mouse visual cortex across temporal offsets.
(TIFF)

S14 Fig. Comparison of network properties when including or excluding Dale's Law. (A) Next-frame mean squared error for hierarchical temporal prediction networks that either included (TP_{Dale}) or excluded ($TP_{No\ Dale}$) Dale's Law. **(B)**

Example next-frame predictions for each network. **(C-E)** Distribution of model and visual cortical response properties for **(C)** modulation ratio, **(D)** plaid pattern indices and **(E)** surround suppression indices. These response properties more closely resembled those found in visual cortex (gray) when Dale's Law was included in the model, as shown by the lower mean Kolmogorov-Smirnov (KS) distance for each of these measures.
(TIFF)

Author contributions

Conceptualization: Sebastian Klavinskis-Whiting, Andrew J. King, Nicol S. Harper.

Data curation: Sebastian Klavinskis-Whiting, Nicol S. Harper.

Formal analysis: Sebastian Klavinskis-Whiting, Nicol S. Harper.

Funding acquisition: Andrew J. King.

Investigation: Sebastian Klavinskis-Whiting, Nicol S. Harper.

Methodology: Sebastian Klavinskis-Whiting, Nicol S. Harper.

Project administration: Andrew J. King, Nicol S. Harper.

Resources: Andrew J. King.

Software: Sebastian Klavinskis-Whiting, Nicol S. Harper.

Supervision: Andrew J. King, Nicol S. Harper.

Validation: Sebastian Klavinskis-Whiting.

Visualization: Sebastian Klavinskis-Whiting.

Writing – original draft: Sebastian Klavinskis-Whiting.

Writing – review & editing: Sebastian Klavinskis-Whiting, Andrew J. King, Nicol S. Harper.

References

1. Poggio T. Marr's computational approach to vision. Trends in Neurosciences. 1981;4:258–62. https://doi.org/10.1016/0166-2236(81)90081-3
2. Herzog MH, Clarke AM. Why vision is not both hierarchical and feedforward. Front Comput Neurosci. 2014;8:135. https://doi.org/10.3389/fncom.2014.00135 PMID: 25374535
3. Kanwisher N, Khosla M, Dobs K. Using artificial neural networks to ask "why" questions of minds and brains. Trends Neurosci. 2023;46(3):240–54. https://doi.org/10.1016/j.tins.2022.12.008 PMID: 36658072
4. Singer Y, Taylor L, Willmore BDB, King AJ, Harper NS. Hierarchical temporal prediction captures motion processing along the visual pathway. Elife. 2023;12:e52599. https://doi.org/10.7554/eLife.52599 PMID: 37844199
5. Singer Y, Teramoto Y, Willmore BD, Schnupp JW, King AJ, Harper NS. Sensory cortex is optimized for prediction of future input. Elife. 2018;7:e31557. https://doi.org/10.7554/eLife.31557 PMID: 29911971
6. Klavinskis-Whiting S, Fristed E, Singer Y, Iacaruso MF, King AJ, Harper NS. Prediction of future input explains lateral connectivity in primary visual cortex. Curr Biol. 2025;35(3):530-541.e5. https://doi.org/10.1016/j.cub.2024.11.073 PMID: 39798566
7. Nejad KK, Anastasiades P, Hertäg L, Costa RP. Self-supervised predictive learning accounts for cortical layer-specificity. Nat Commun. 2025;16(1):6178. https://doi.org/10.1038/s41467-025-61399-5 PMID: 40615428
8. Bialek W, Van Steveninck RRDR, Tishby N. Efficient representation as a design principle for neural coding and computation. IEEE, 2006. 659–63.
9. Kindel WF, Christensen ED, Zylberberg J. Using deep learning to probe the neural code for images in primary visual cortex. J Vis. 2019;19(4):29. https://doi.org/10.1167/19.4.29 PMID: 31026016
10. Yamins DLK, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ. Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc Natl Acad Sci U S A. 2014;111(23):8619–24. https://doi.org/10.1073/pnas.1403112111 PMID: 24812127
11. Muckli L, Petro LS. Network interactions: Non-geniculate input to V1. Curr Opin Neurobiol. 2013;23(2):195–201. https://doi.org/10.1016/j.conb.2013.01.020 PMID: 23402951
12. Briggs F. Role of feedback connections in central visual processing. Annu Rev Vis Sci. 2020;6:313–34. https://doi.org/10.1146/annurev-vision-121219-081716 PMID: 32552571

13. Nurminen L, Merlin S, Bijanzadeh M, Federer F, Angelucci A. Top-down feedback controls spatial summation and response amplitude in primate visual cortex. Nat Commun. 2018;9(1):2281. https://doi.org/10.1038/s41467-018-04500-5 PMID: 29892057
14. Nassi JJ, Lomber SG, Born RT. Corticocortical feedback contributes to surround suppression in V1 of the alert primate. J Neurosci. 2013;33(19):8504–17. https://doi.org/10.1523/JNEUROSCI.5124-12.2013 PMID: 23658187
15. Vangeneugden J, van Beest EH, Cohen MX, Lorteije JAM, Mukherjee S, Kirchberger L, et al. Activity in lateral visual areas contributes to surround suppression in awake mouse V1. Curr Biol. 2019;29(24):4268-4275.e7. https://doi.org/10.1016/j.cub.2019.10.037 PMID: 31786063
16. Angelucci A, Bijanzadeh M, Nurminen L, Federer F, Merlin S, Bressloff PC. Circuits and mechanisms for surround modulation in visual cortex. Annu Rev Neurosci. 2017;40:425–51. https://doi.org/10.1146/annurev-neuro-072116-031418 PMID: 28471714
17. Merrikhi Y, Clark K, Albarran E, Parsa M, Zirnsak M, Moore T, et al. Spatial working memory alters the efficacy of input to visual cortex. Nat Commun. 2017;8:15041. https://doi.org/10.1038/ncomms15041 PMID: 28447609
18. Lawrence SJD, van Mourik T, Kok P, Koopmans PJ, Norris DG, de Lange FP. Laminar organization of working memory signals in human visual cortex. Curr Biol. 2018;28(21):3435-3440.e4. https://doi.org/10.1016/j.cub.2018.08.043 PMID: 30344121
19. Voitov I, Mrsic-Flogel TD. Cortical feedback loops bind distributed representations of working memory. Nature. 2022;608(7922):381–9. https://doi.org/10.1038/s41586-022-05014-3 PMID: 35896749
20. Bastos AM, Usrey WM, Adams RA, Mangun GR, Fries P, Friston KJ. Canonical microcircuits for predictive coding. Neuron. 2012;76(4):695–711. https://doi.org/10.1016/j.neuron.2012.10.038 PMID: 23177956
21. Rao RP, Ballard DH. Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. Nat Neurosci. 1999;2(1):79–87. https://doi.org/10.1038/4580 PMID: 10195184
22. Schmidt KE, Lomber SG, Payne BR, Galuske RAW. Pattern motion representation in primary visual cortex is mediated by transcortical feedback. NeuroImage. 2010;54(1):474–84. https://doi.org/10.1016/j.neuroimage.2010.08.017
23. Collins CE, Airey DC, Young NA, Leitch DB, Kaas JH. Neuron densities vary across and within cortical areas in primates. Proc Natl Acad Sci U S A. 2010;107(36):15927–32. https://doi.org/10.1073/pnas.1010356107 PMID: 20798050
24. Levitt JB, Kiper DC, Movshon JA. Receptive fields and functional architecture of macaque V2. J Neurophysiol. 1994;71(6):2517–42. https://doi.org/10.1152/jn.1994.71.6.2517 PMID: 7931532
25. Jones HE, Grieve KL, Wang W, Sillito AM. Surround suppression in primate V1. J Neurophysiol. 2001;86(4):2011–28. https://doi.org/10.1152/jn.2001.86.4.2011 PMID: 11600658
26. Hu J, Ma H, Zhu S, Li P, Xu H, Fang Y, et al. Visual motion processing in macaque V2. Cell Rep. 2018;25(1):157-167.e5. https://doi.org/10.1016/j.celrep.2018.09.014 PMID: 30282025
27. Bair W, Movshon JA. Adaptive temporal integration of motion in direction-selective neurons in macaque visual cortex. J Neurosci. 2004;24(33):7305–23. https://doi.org/10.1523/JNEUROSCI.0554-04.2004 PMID: 15317857
28. Rust NC, Schwartz O, Movshon JA, Simoncelli EP. Spatiotemporal elements of macaque v1 receptive fields. Neuron. 2005;46(6):945–56. https://doi.org/10.1016/j.neuron.2005.05.021 PMID: 15953422
29. Adelson EH, Bergen JR. Spatiotemporal energy models for the perception of motion. J Opt Soc Am A. 1985;2(2):284–99. https://doi.org/10.1364/josaa.2.000284 PMID: 3973762
30. Ringach DL, Shapley RM, Hawken MJ. Orientation selectivity in macaque V1: Diversity and laminar dependence. J Neurosci. 2002;22(13):5639–51. https://doi.org/10.1523/JNEUROSCI.22-13-05639.2002 PMID: 12097515
31. Ko H, Hofer SB, Pichler B, Buchanan KA, Sjöström PJ, Mrsic-Flogel TD. Functional specificity of local synaptic connections in neocortical networks. Nature. 2011;473(7345):87–91. https://doi.org/10.1038/nature09880 PMID: 21478872
32. Iacaruso MF, Gasler IT, Hofer SB. Synaptic organization of visual space in primary visual cortex. Nature. 2017;547(7664):449–52. https://doi.org/10.1038/nature23019 PMID: 28700575
33. Movshon JA, Adelson EH, Gizzi MS, Newsome WT. The analysis of moving visual patterns. Experimental Brain Research Supplementum. Springer Berlin Heidelberg. 1985. 117–51. https://doi.org/10.1007/978-3-662-09224-8_7
34. Movshon JA, Newsome WT. Visual response properties of striate cortical neurons projecting to area MT in macaque monkeys. J Neurosci. 1996;16(23):7733–41. https://doi.org/10.1523/JNEUROSCI.16-23-07733.1996 PMID: 8922429
35. Rust NC, Mante V, Simoncelli EP, Movshon JA. How MT cells analyze the motion of visual patterns. Nat Neurosci. 2006;9(11):1421–31. https://doi.org/10.1038/nn1786 PMID: 17041595
36. Mishkin M, Ungerleider LG, Macko KA. Object vision and spatial vision: Two cortical pathways. Trends in Neurosciences. 1983;6:414–7. https://doi.org/10.1016/0166-2236(83)90190-x
37. Wang Q, Gao E, Burkhalter A. Gateways of ventral and dorsal streams in mouse visual cortex. J Neurosci. 2011;31(5):1905–18. https://doi.org/10.1523/JNEUROSCI.3488-10.2011 PMID: 21289200
38. Wang Q, Sporns O, Burkhalter A. Network analysis of corticocortical connections reveals ventral and dorsal processing streams in mouse visual cortex. J Neurosci. 2012;32(13):4386–99. https://doi.org/10.1523/JNEUROSCI.6063-11.2012 PMID: 22457489
39. Yu Y, Stirman JN, Dorsett CR, Smith SL. Selective representations of texture and motion in mouse higher visual areas. Curr Biol. 2022;32(13):2810-2820.e5. https://doi.org/10.1016/j.cub.2022.04.091 PMID: 35609609

40. Sit KK, Goard MJ. Distributed and retinotopically asymmetric processing of coherent motion in mouse visual cortex. Nat Commun. 2020;11(1):3565. https://doi.org/10.1038/s41467-020-17283-5 PMID: 32678087
41. Matteucci G, Bellacosa Marotti R, Zattera B, Zoccolan D. Truly pattern: Nonlinear integration of motion signals is required to account for the responses of pattern cells in rat visual cortex. Sci Adv. 2023;9(45):eadh4690. https://doi.org/10.1126/sciadv.adh4690 PMID: 37939191
42. Juavinett AL, Callaway EM. Pattern and component motion responses in mouse visual cortical areas. Curr Biol. 2015;25(13):1759–64. https://doi.org/10.1016/j.cub.2015.05.028 PMID: 26073133
43. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv 1409.1556v6 2015. https://doi.org/10.48550/arXiv.1409.1556
44. Levenstein D, Alvarez VA, Amarasingham A, Azab H, Chen ZS, Gerkin RC, et al. On the role of theory and modeling in neuroscience. J Neurosci. 2023;43(7):1074–88. https://doi.org/10.1523/JNEUROSCI.1179-22.2022 PMID: 36796842
45. Yamins DL, DiCarlo JJ. Using goal-driven deep learning models to understand sensory cortex. Nat Neurosci. 2016;19:356–65. https://doi.org/10.1038/nn.4244
46. Taylor L, Zenke F, King AJ, Harper NS. Temporal prediction captures key differences between spiking excitatory and inhibitory V1 neurons. 2024; 2024–05. https://doi.org/10.1101/2024.05.12.593763
47. Taylor L, Zenke F, King AJ, Harper NS. Temporal prediction captures retinal spiking responses across animal species. bioRxiv. 2024. https://doi.org/10.1101/2024.03.26.586771
48. Whittington JCR, Bogacz R. Theories of error back-propagation in the brain. Trends Cogn Sci. 2019;23(3):235–50. https://doi.org/10.1016/j.tics.2018.12.005 PMID: 30704969
49. Bengio Y, Lee D-H, Bornschein J, Mesnard T, Lin Z. Towards biologically plausible deep learning. 2016. http://arxiv.org/abs/1502.04156
50. Millidge B, Tang M, Osanlouy M, Harper NS, Bogacz R. Predictive coding networks for temporal prediction. PLoS Comput Biol. 2024;20(4):e1011183. https://doi.org/10.1371/journal.pcbi.1011183 PMID: 38557984
51. Simoncelli EP, Heeger DJ. A model of neuronal responses in visual area MT. Vision Res. 1998;38(5):743–61. https://doi.org/10.1016/s0042-6989(97)00183-1 PMID: 9604103
52. Nishimoto S, Gallant JL. A three-dimensional spatiotemporal receptive field model explains responses of area MT neurons to naturalistic movies. J Neurosci. 2011;31(41):14551–64. https://doi.org/10.1523/JNEUROSCI.6801-10.2011 PMID: 21994372
53. Zaharia AD, Goris RLT, Movshon JA, Simoncelli EP. Compound stimuli reveal the structure of visual motion selectivity in macaque MT neurons. eNeuro. 2019;6(6):ENEURO.0258-19.2019. https://doi.org/10.1523/ENEURO.0258-19.2019 PMID: 31604815
54. Trinh F, King AJ, Willmore BD, Harper N. Cochlear tuning characteristics arise from temporal prediction of natural sounds. bioRxiv. 2023. https://doi.org/10.1101/2023.10.02.560418
55. Bakhtiari S, Mineault P, Lillicrap T, Pack C, Richards B. The functional specialization of visual cortex emerges from training parallel pathways with self-supervised predictive learning. Advances in Neural Information Processing Systems. Curran Associates, Inc.; 2021. 25164–25178. https://proceedings.neurips.cc/paper/2021/hash/d384dec9f5f7a64a36b5c8f03b8a6d92-Abstract.html
56. Nayebi A, Kong NCL, Zhuang C, Gardner JL, Norcia AM, Yamins DLK. Mouse visual cortex as a limited resource system that self-learns an ecologically-general representation. PLoS Comput Biol. 2023;19(10):e1011506. https://doi.org/10.1371/journal.pcbi.1011506 PMID: 37782673
57. Cadieu CF, Olshausen BA. Learning intermediate-level representations of form and motion from natural movies. Neural Comput. 2012;24(4):827–66. https://doi.org/10.1162/NECO_a_00247 PMID: 22168556
58. Rideaux R, Welchman AE. But still it moves: Static image statistics underlie how we see motion. J Neurosci. 2020;40(12):2538–52. https://doi.org/10.1523/JNEUROSCI.2760-19.2020 PMID: 32054676
59. Mineault P, Bakhtiari S, Richards B, Pack C. Your head is there to move you around: Goal-driven models of the primate dorsal pathway. Adv Neural Inf Process Syst. 2021;34:28757–71.
60. Hénaff OJ, Goris RLT, Simoncelli EP. Perceptual straightening of natural videos. Nat Neurosci. 2019;22(6):984–91. https://doi.org/10.1038/s41593-019-0377-4 PMID: 31036946
61. Hénaff OJ, Boundy-Singer ZM, Meding K, Ziemba CM, Goris RLT. Representation of visual uncertainty through neural gain variability. Nat Commun. 2020;11(1):2513. https://doi.org/10.1038/s41467-020-15533-0 PMID: 32427825
62. Niu X, Savin C, Simoncelli E. Learning predictable and robust neural representations by straightening image sequences. Advances in Neural Information Processing Systems 37, 2024. 40316–35. https://doi.org/10.52202/079017-1275
63. Harrington A, DuTell V, Tewari A, Hamilton M, Stent S, Rosenholtz R, Freeman WT. Exploring perceptual straightness in learned visual representations. ICLR 2023.
64. Lotter W, Kreiman G, Cox D. Deep predictive coding networks for video prediction and unsupervised learning. 2017. http://arxiv.org/abs/1605.08104
65. Lotter W, Kreiman G, Cox D. A neural network trained for prediction mimics diverse features of biological neurons and perception. Nat Mach Intell. 2020;2(4):210–9. https://doi.org/10.1038/s42256-020-0170-9 PMID: 34291193
66. Meyer HS, Schwarz D, Wimmer VC, Schmitt AC, Kerr JND, Sakmann B, et al. Inhibitory interneurons in a cortical column form hot zones of inhibition in layers 2 and 5A. Proc Natl Acad Sci U S A. 2011;108(40):16807–12. https://doi.org/10.1073/pnas.1113648108 PMID: 21949377

67. van Versendaal D, Levelt CN. Inhibitory interneurons in visual cortical plasticity. Cell Mol Life Sci. 2016;73(19):3677–91. https://doi.org/10.1007/s00018-016-2264-4 PMID: 27193323

68. Siegle JH, Jia X, Durand S, Gale S, Bennett C, Graddis N, et al. Survey of spiking in the mouse visual system reveals functional hierarchy. Nature. 2021;592(7852):86–92. https://doi.org/10.1038/s41586-020-03171-x PMID: 33473216

69. Sahani M, Linden J. How linear are auditory cortical responses?. Adv Neural Inf Process Syst. 2002;15.

70. Deitch D, Rubin A, Ziv Y. Representational drift in the mouse visual cortex. Curr Biol. 2021;31(19):4327-4339.e6. https://doi.org/10.1016/j.cub.2021.07.062 PMID: 34433077

71. Nili H, Wingfield C, Walther A, Su L, Marslen-Wilson W, Kriegeskorte N. A toolbox for representational similarity analysis. PLoS Comput Biol. 2014;10(4):e1003553. https://doi.org/10.1371/journal.pcbi.1003553 PMID: 24743308