Neuron

Interpreting the retinal neural code for natural scenes: From computations to neurons

Highlights

- A three-layer model captures retinal natural scene responses and phenomena
- Models have internal units highly correlated with interneuron recordings
- A general approach reveals how model interneuron pathways encode any stimulus
- Model analysis yields new automatic circuit hypotheses for neural computations

Authors

Niru Maheswaranathan, Lane T. McIntosh, Hidenori Tanaka, Satchel Grant, ..., Surya Ganguli, Stephen A. Baccus

Correspondence

baccus@stanford.edu

In brief

Maheswaranathan et al. create a three-layer network model that captures retinal encoding of natural scenes and many ethological phenomena. The model's structure is interpretable in terms of the actions of real interneurons. A new computational approach automatically generates hypotheses for how interneurons generate ethological computations under natural visual scenes.

Maheswaranathan et al., 2023, Neuron *111*, 1–14
September 6, 2023
https://doi.org/10.1016/j.neuron.2023.06.007

CellPress

Neuron

CellPress
OPEN ACCESS

Article

Interpreting the retinal neural code for natural scenes: From computations to neurons

Niru Maheswaranathan,[1,7,8] Lane T. McIntosh,[1,7,9] Hidenori Tanaka,[3,5,6,7] Satchel Grant,[4,7] David B. Kastner,[1,10] Joshua B. Melander,[1] Aran Nayebi,[1,11] Luke E. Brezovec,[1] Julia H. Wang,[2] Surya Ganguli,[3] and Stephen A. Baccus[4,12,*]

[1]Neuroscience Program, Stanford University School of Medicine, Stanford, CA, USA
[2]Stanford University, Stanford, CA, USA
[3]Department of Applied Physics, Stanford University, Stanford, CA, USA
[4]Department of Neurobiology, Stanford University, Stanford, CA, USA
[5]Physics & Informatics Laboratories, NTT Research, Inc., Sunnyvale, CA, USA
[6]Center for Brain Science, Harvard University, Cambridge, MA, USA
[7]These authors contributed equally
[8]Present address: Meta Reality Labs, Burlingame, CA, USA
[9]Present address: Tesla, Inc., Palo Alto, CA, USA
[10]Present address: Department of Psychiatry and Behavioral Sciences, University of California, San Francisco, San Francisco, CA, USA
[11]Present address: McGovern Institute for Brain Research, MIT Cambridge, MA 02139
[12]Lead contact
*Correspondence: baccus@stanford.edu
https://doi.org/10.1016/j.neuron.2023.06.007

SUMMARY

Understanding the circuit mechanisms of the visual code for natural scenes is a central goal of sensory neuroscience. We show that a three-layer network model predicts retinal natural scene responses with an accuracy nearing experimental limits. The model's internal structure is interpretable, as interneurons recorded separately and not modeled directly are highly correlated with model interneurons. Models fitted only to natural scenes reproduce a diverse set of phenomena related to motion encoding, adaptation, and predictive coding, establishing their ethological relevance to natural visual computation. A new approach decomposes the computations of model ganglion cells into the contributions of model interneurons, allowing automatic generation of new hypotheses for how interneurons with different spatiotemporal responses are combined to generate retinal computations, including predictive phenomena currently lacking an explanation. Our results demonstrate a unified and general approach to study the circuit mechanisms of ethological retinal computations under natural visual scenes.

INTRODUCTION

How neural circuit functions emerge in natural, ethologically relevant settings from the activity of multiple cell types is a fundamental open question in neuroscience. The retina has evolved to convey information about natural visual scenes to the brain[1] using a large, diverse set of interneurons.[2] Yet, because of the inability to model natural scene responses,[3,4] we neither understand the neural code for natural scenes nor how interneurons generate that code. Nearly all of our understanding of retinal computations and circuit mechanisms comes from artificial stimuli, such as flashing spots, drifting gratings, and white noise,[5,6] which have unknown relevance to natural visual processing. Although numerous computations have been identified by such methods, including various types of motion selectivity, adaptation, and prediction of visual features,[5] the number of interneurons (>50) is even greater, suggesting an undiscovered complexity in retinal processing.

A major barrier toward understanding how neural circuits function is the lack of models that can achieve three objectives: (1) capture the input-output relationship of neural circuits under natural inputs, (2) have an interpretable computational structure to enable analyses of how those computations are performed, and (3) relate the internal structure of the model to real interneurons to yield a mechanistic explanation of circuit function.

Deep neural network models have excelled at capturing complex phenomena, including how the ventral visual stream performs object recognition[7–9] and how the auditory cortex performs sound discrimination.[10] However, it is unclear how to interpret these models in terms of their contribution to neural computation or relationship to individual biological neurons. Simple linear-nonlinear (LN) models,[11] generalized linear models (GLMs),[6] or two-layer LN-LN models with nonlinear subunits,[12–16] show higher computational interpretability—the ability to understand the mathematical components of the

Please cite this article in press as: Maheswaranathan et al., Interpreting the retinal neural code for natural scenes: From computations to neurons, Neuron (2023), https://doi.org/10.1016/j.neuron.2023.06.007

model—yet an LN model's single spatiotemporal filter, or the two sequential stages of an LN-LN model, are inadequate to capture the complex visual processing of natural stimuli.

Here, we analyze neural network models of the retina that fulfill the goals of capturing the neural code for natural scenes, computational interpretability, and mechanistic interpretability, and use these models to generate new testable hypotheses for how interneurons generate specific ethological computations. In the salamander retina, we use three-layer convolutional neural network (CNN) models that accurately capture natural scene responses—nearly to within the limits set by the variability of retinal ganglion cells. These models, fitted only to natural scenes, capture a broad range of previously described phenomena related to motion encoding, adaptation, and predictive coding, defined using only artificially structured stimuli, thereby establishing the ethological relevance of these phenomena to natural visual processing. Models fitted to white noise do not capture all of these phenomena, pointing to the critical need to study natural scenes. These models have mechanistic interpretability in that models fitted only to ganglion cell responses nevertheless have internal units that are highly correlated with interneuron recordings from separate preparations.

Finally, we achieve the goal of computational interpretability using a novel general approach to attribute any retinal computation to the actions of model interneurons, analyzing several ethological computations. This analysis confirms existing mechanistic hypotheses and automatically generates new hypotheses, including those for predictive visual computations currently without explanation. We further find that, compared with white noise, natural scene responses are generated by a broader range of model interneuron pathways so as to engage a greater set of ethological phenomena. These findings present a unifying approach to study the retinal neural code and to generate hypotheses for the specific actions of interneuron pathways for any and all stimuli, including natural scenes.

RESULTS

The first goal was to create a minimal CNN model that could capture the retinal neural code for natural scenes. Between photoreceptors and ganglion cells, there are three levels of strong rectification, the bipolar cell terminal, the amacrine cell terminal, and ganglion cell spiking. Although there is greater complexity in the retina, one might expect that a minimum of three network layers consisting of filtering followed by a threshold might be needed to approximate the set of retinal computations. We therefore tested whether CNN models of up to three synaptic layers (Figure 1) could predict the responses of populations of salamander retinal ganglion cells responding to a 50-min sequence of either natural images or white noise. Natural scene images changed every second and were jittered every 30-Hz frame in a random walk with the statistics of fixational eye movements,[17,18] creating a spatiotemporal stimulus. The final chosen structure of our model had eight different cell types in each of two convolutional layers that tiled the visual field, with each cell type receiving input from the previous layer that was limited in spatial extent, up to 750 μm for the first spatiotemporal layer and 550 μm for the second spatial layer. In the final synaptic layer, model ganglion cells received input across the entire spatial extent of the model's preceding layer (up to 1.3 mm). Each synaptic layer also had a threshold following its spatiotemporal or spatial filter. Models were initialized randomly and then optimized via a gradient descent algorithm (see STAR Methods) so that the model's output most closely matched the responses of ganglion cells responding to a set of training stimuli. The model was then evaluated on the average response to repeated trials of a separate test dataset not used in training.

We found that CNN models could predict ganglion cell responses to either natural scenes or white noise substantially more accurately than LN models[11] or GLMs that incorporate an additional spike history feedback term[6] (Figures 1B, 1C, and S1). Although LN models can accurately predict ganglion cell responses to uniform-field stimuli at a fixed contrast[19] and GLMs can accurately predict responses to small patches of white noise with larger squares,[6] the strong spatial nonlinearities of salamander ganglion cells require models with more than a single spatiotemporal stage.[13] The accuracy of the prediction approached the limit of precision set by intrinsic neural variability, given the number of test trials (5–10), although this limit would increase with more trials. We also tested a range of different models by varying the number of model cell types in the first two layers and chose eight cell types as the minimum number that achieved near-maximal model performance (Figures 1D and 1E).

CNNs' internal units are highly correlated with interneuron responses

A model that captures the neural code for natural scenes is by itself useful for analyzing how visual features are sensed and to serve as a compact description of retinal output for studies of the higher visual system. However, another important goal was to examine whether the model's internal computational structure showed similarity to that of the retina.

To examine whether the internal computations of CNN models approximated those of the retina, we computed receptive fields (RFs) for layer 1 and 2 model interneurons of CNNs trained using natural scenes. We compared model responses to interneuron soma voltage recordings by taking the layer 1 and 2 signals after the filter but before the threshold. For example, the bipolar cell soma is the last retinal stage before strong rectification at the bipolar cell synaptic terminal, corresponding to the model's layer 1 signal before the threshold.

We found that the RFs of CNN model interneurons had the well-known structure of retinal interneurons,[20,21] with a spatially localized center-surround structure as first appears in bipolar cells (Figures 2A and 2B), ON and OFF responses, monophasic and biphasic temporal filters, and rectified nonlinear responses in the second layer (Figure S2), all properties that have been explained by principles of efficient coding in the retina.[22–25]

In the inferotemporal cortex, CNN units have been shown to be correlated with activity of a linear combination of neurons,[7] making it difficult to draw conclusions about individual neurons. We compared our CNN interneuron activity directly without modification to interneuron recordings performed on separate retinae that the model was never fitted to (Figure 3A). The stimulus presented to the retina—and separately to the

Please cite this article in press as: Maheswaranathan et al., Interpreting the retinal neural code for natural scenes: From computations to neurons, Neuron (2023), https://doi.org/10.1016/j.neuron.2023.06.007

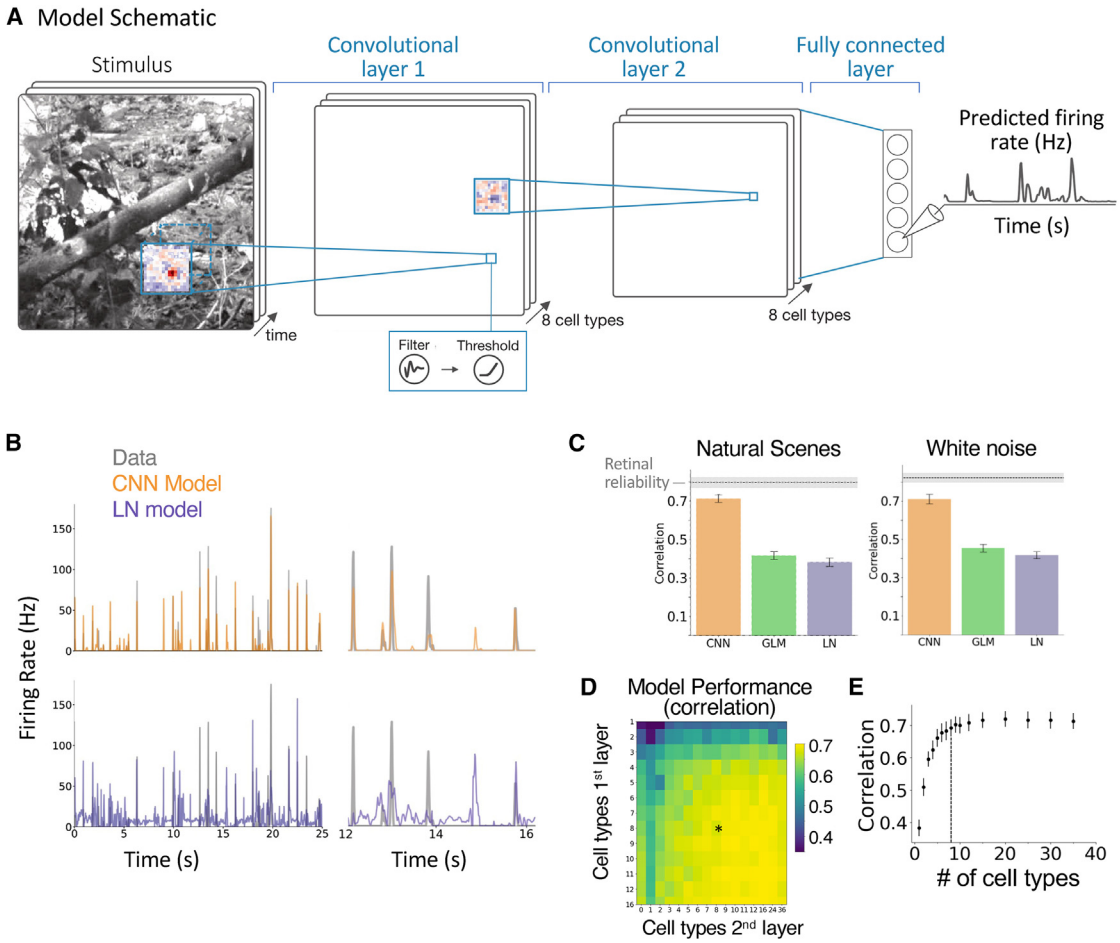

Figure 1. Convolutional neural networks provide accurate models of the retinal response to natural scenes

(A) Convolutional neural network model trained to predict the firing rate of simultaneously recorded ganglion cells. The first layer is a spatiotemporal convolution, the second is a spatial convolution, and the third is a fully connected (dense) layer, with rectifying nonlinearities in between each layer.

(B) Peristimulus time histograms (PSTHs) comparing recorded data and LN or CNN models for the test dataset.

(C) Correlation coefficients for LN, GLM, and CNN model predictions for 60 s of natural scenes test data. Dotted line is mean reliability of ganglion cell PSTHs correlated between different sets of trials, gray bar indicates one SEM.

(D) Correlation coefficient between model and test data for different numbers of cell types in layers 1 and 2. Star indicates 8 cell types, the value chosen for further analysis. Left column showing zero second layer cell types indicates a two-layer (LN-LN) model.

(E) Same as (D) for equal numbers of cell types in both first and second layers. Dashed line indicates 8 cell types. Error bars indicate mean ± S.E.M.

model—was a spatiotemporal white noise checkerboard having no spatiotemporal correlations except for the 50-μm size and 33-ms duration of square stimulus regions. We correlated each interneuron recording with 8 model interneurons of both layer 1 and 2 at each location to find the most-correlated model interneuron at the location of the cell. We always searched only over a single model, 16 cell types for the standard CNN model, and report all results separately for each model. We found that each recorded interneuron showed a high and spatially localized correlation with a particular model interneuron type (Figures 3B–3E). Spatiotemporal RFs were similar in appearance between recorded interneurons and their most-correlated model interneuron type (Figure 3B), including the time course of center and surround, although model interneurons appeared to have a systematically stronger surround.

To assess the similarity between model and real interneurons, we compared the correlations between cells in our set of recorded interneurons and between interneurons and model interneurons (Figures 3D and 3E). The comparison between different recorded interneurons gives a similarity measure for interneurons of the same cell type, which is the maximum correlation that we would expect between our model and the recorded interneurons. We found that the best match within our interneuron set (interneuron-interneuron) had a similar correlation to the best match between model cells and interneurons (model-interneuron). Thus, model interneurons revealed the responses of interneurons nearly to within the variability of those interneurons, despite having never been fit directly to neural responses. The range of correlations were similar in the model to those in our interneuron set, although there were fewer examples of negative interneuron-interneuron correlations. This result was likely

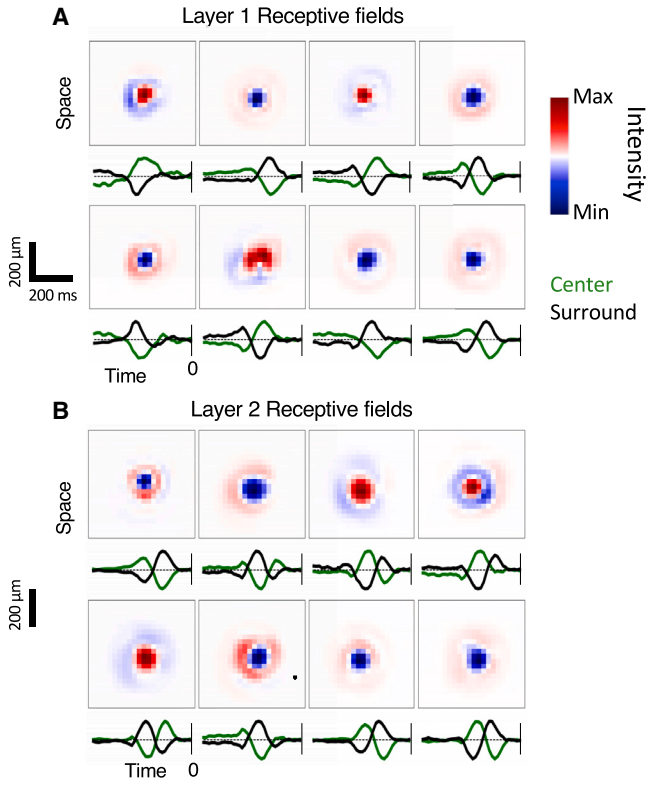

Figure 2. Structure of receptive fields of model cell types

(A) RFs of layer 1 cells computed by reverse correlation of a white noise stimulus presented to the model, shown as the spatial average (top), and the time course of the RF center and surround (bottom).

(B) Same for layer 2.

because our interneuron dataset consisted primarily of OFF type cells recorded for other studies and did not fully represent the true interneuron distribution.

To assess the similarity of different models, we correlated model interneurons with their best match in different models (model-model), finding that on average the model-model correlation was similar to the model-interneuron correlation (Figure 3F). We further tested the reproducibility of these correlations by initializing the model with different random seeds. We found high reproducibility when random seed was varied in the responses of model interneurons (Figure 3G), in the correlation between model interneurons and real interneurons (Figure S3A) and in the model's prediction of ganglion cells (Figure S3A).

To assess whether the number of cell types might influence whether those cells were correlated with real interneurons, we varied the number of model cell types and found that as the number increased, the model first produced cell types that were correlated with recorded interneurons (Figure S3B). This indicates that to capture the ganglion cell response, as the model was allowed to have more cell types, it rapidly and consistently found solutions that relied on model interneurons that were correlated with real interneurons. We further found that the model interneurons most highly correlated with real interneurons were also highly correlated across models fit to different retinae (Figure 3H), further indicating the importance that the model contain interneurons that correlate with real interneurons. Furthermore, because individual models are highly reproducible with different random seeds, the model interneurons less highly correlated with interneurons in our dataset likely indicate real preparation-to-preparation differences. Therefore, fitting a CNN model to ganglion cell natural scene responses alone models an entire population of interneurons, many of which have high correlation with measured interneuron responses to a different stimulus in a different retina. Although we do not think that there is necessarily a direct correspondence between CNN layer and cell type (bipolar vs. amacrine cell), model interneurons have a high enough similarity to real interneurons to give interpretable mechanistic insight as to the internal structure of retinal computations.

A wide range of retinal phenomena are engaged by natural stimuli

Numerous nonlinear retinal computations have been identified using artificial stimuli, including flashing spots, moving bars, and white noise. However, we neither understand to what degree natural vision engages these computations nor how retinal circuitry implements them under natural scenes. We tested models fit to either natural scenes or white noise by exposing them to a battery of structured artificial stimuli. We focused on effects shorter than 400 ms—the longest timescale our model could reproduce as limited by the first layer spatiotemporal filter. Remarkably, the CNN model exhibited all of these computations, described below.

In response to a uniform-field flickering stimulus with a constant mean, the retina adapts to temporal contrast, defined as the ratio of the standard deviation to the mean intensity. High contrast stimuli cause ganglion cells to respond with a time course that is faster, more biphasic, and less sensitive than during low contrast.[19,26,27] These properties can be assessed with an LN model consisting of a linear temporal filter and static nonlinearity fit at different contrasts. We found that both natural scene and white noise models reproduced temporal processing changes that occur during contrast adaptation (CA) (Figure 4A). However, only models fit to natural scenes correctly reproduced the decrease in sensitivity—the average slope of the nonlinearity—at high contrast, likely because white noise stimuli did not explore a sufficient range of contrasts.

A second ganglion cell phenomenon is that for flashed stimuli, as stronger stimuli decrease the response latency, allowing decoding of stimulus intensity using latency[31] (Figure 4B). Natural scenes models reproduced latency coding, but white noise models did so much less, indicating that white noise did not generate this property in a manner that could be captured by the model. Both models reproduced frequency doubling (FD), observed when ganglion cells respond at 2 Hz to a 1-Hz periodically reversing grating[12] (Figure 4C). Additionally, ON-OFF cells had RFs that reversed in polarity during a dynamic stimulus[32] (Figure 4D).

The model also reproduced several predictive phenomena. When a moving bar reverses in direction, the ganglion cell population responds synchronously, signaling the violation of the prediction that motion will continue smoothly[28] (Figure 4E). The model reproduced this reversal response with high accuracy, having a time course that closely matched published data.

Please cite this article in press as: Maheswaranathan et al., Interpreting the retinal neural code for natural scenes: From computations to neurons, Neuron (2023), https://doi.org/10.1016/j.neuron.2023.06.007

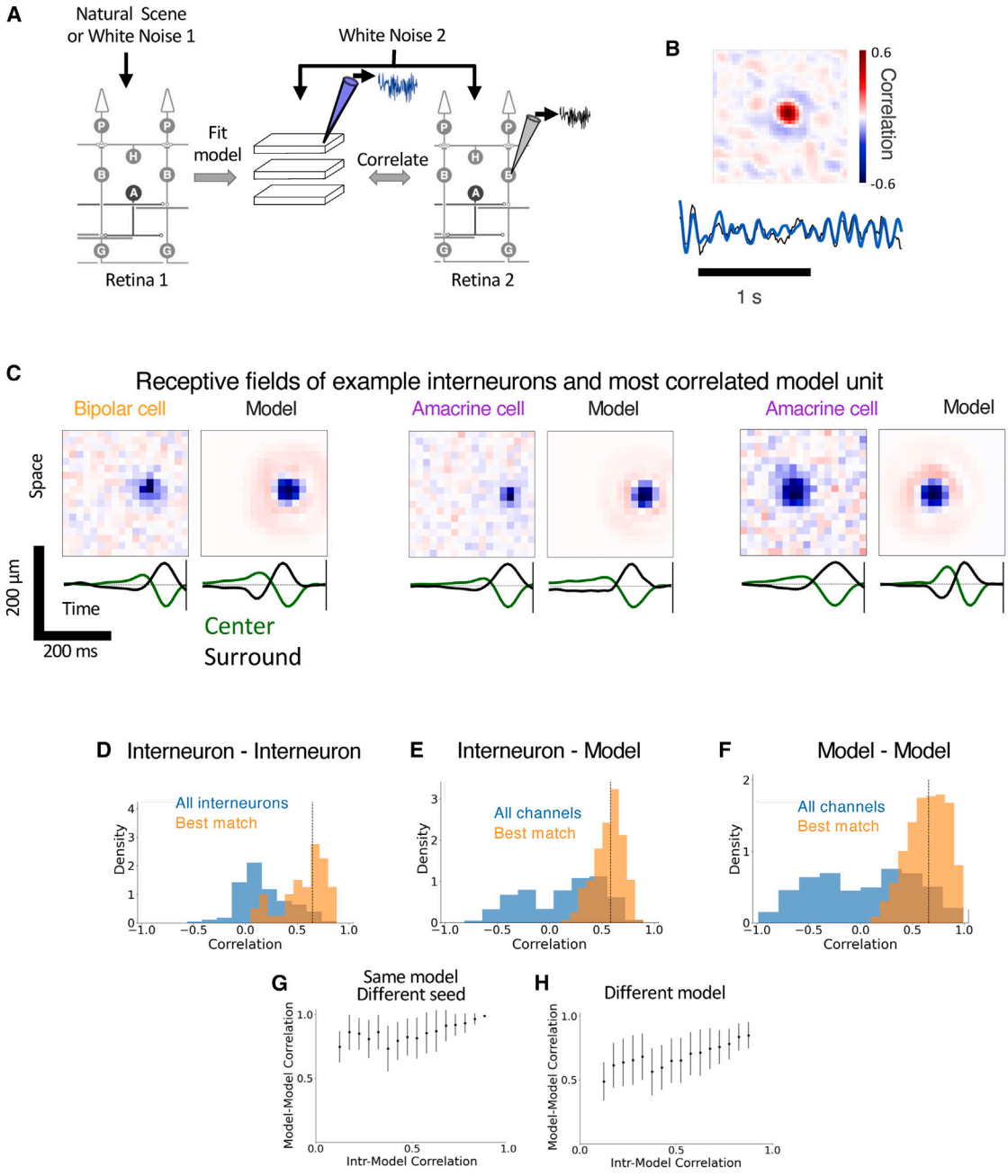

Figure 3. Model internal units are correlated with interneuron responses

(A) Schematic of experiment. White noise responses of bipolar or amacrine cells from a different retina were recorded intracellularly.

(B) Top: correlation map, where each pixel is the correlation between a different spatial location within a single model cell type and the interneuron recording from a different retina using the same stimulus. Bottom: responses of the interneuron and the most-correlated model interneuron.

(C) Spatiotemporal RFs of example interneurons recorded from a separate retina, and the model interneuron from Figure 2 that was most correlated with that recorded interneuron. The model was never fit to the interneuron's response. The time courses of the spatially integrated center and surround are shown scaled to the minimum and maximum.

(D) Histogram of correlation coefficients between interneurons for all interneuron-interneuron pairs, and the best match (7 bipolar cells, 26 amacrine cells, and 10 cells that were either bipolar or amacrine cells).

(E) Histogram of correlation coefficients between model interneurons of CNN models fit to ganglion cells and recorded interneurons. Shown are all model interneuron-interneuron pairs and the best match. Each of three natural scenes and three white noise CNN models were searched separately to find the best match, and the results of all models are shown (see STAR Methods).

(legend continued on next page)

Please cite this article in press as: Maheswaranathan et al., Interpreting the retinal neural code for natural scenes: From computations to neurons, Neuron (2023), https://doi.org/10.1016/j.neuron.2023.06.007

As a separate predictive phenomenon, ganglion cells anticipate the motion of a steadily moving bar, shifting the traveling wave of population activity in the direction of motion.[29] Motion anticipation (MA) is thought to compensate for the lagging representation of a moving object due to processing delays. The model ganglion cell population anticipated motion similarly to previously published results (Figure 4F).

A final predictive phenomenon is the omitted stimulus response[30] (OSR) (Figure 4G) in which a periodic flash sequence entrains the ganglion cell response, but when a single flash is omitted, the cell produces an even larger response at the expected time of the response to the omitted flash. Moreover, the OSR occurs at the expected time over a range of frequencies, suggesting that the retina somehow retains a memory trace of the flash period. Models fit to natural scenes, but not to white noise, reproduced the OSR over a range of frequencies.

These phenomena arose in a CNN model as a byproduct of optimization to natural scenes. Compared with natural scene models, white noise models showed a great reduction in CA, latency coding, and the OSR (Figure S4), as well as reduced polarity reversal (PR) and motion reversal (MR), slightly reduced MA, and a similar level of FD. These results indicate that natural scene statistics trigger nonlinear computations that white noise does not. Even though natural scenes consisted only of a sequence of jittered images with no explicit object motion or periodic patterns, models still exhibited MA and reversal responses, and the OSR.

The only phenomenon tested that was not captured by the model initially was the object motion sensitive (OMS) response,[18] which discriminates differential motion as caused by an object moving against a background from global motion, as occurs from eye movements. The OMS response was tested by presenting a jittering central object grating along with a jittering background grating, with the two gratings having either the same trajectory (global motion) or different trajectories (differential motion). We hypothesized that the model's lack of an OMS response was due to the absence of differential motion in the training stimulus. To test this idea, we trained additional models on the retinal response to movies of swimming fish that included differential motion. We found that these models did indeed exhibit an OMS response of a similar range as did previous experimental data, verifying that if a stimulus does not engage a computation, the model cannot capture it (Figure 4H). Thus, the model reveals that the nonlinear properties and circuit mechanisms of a broad range of ethological computations are engaged during natural scenes.

INCs to a dynamic visual code

The ability of the model to accurately capture retinal responses and ethological computations, and the correlation of model interneurons with real interneurons, creates the opportunity to analyze the model to generate hypotheses as to the circuit bases of these computations. To do this, we calculated for each stimulus the contribution of each model interneuron to the model's response, which we term the interneuron contribution (INC) (Figure 5A).[33] The INC for each model interneuron at each time point is the product of the interneuron's activity caused by the stimulus and the sensitivity of the model output to the model interneuron, which takes into account both an interneuron's input (RF) and its output (projective field),[34] where both input and output stages can be polysynaptic. The INC is the net excitation or inhibition that each interneuron contributes across the circuit to the model's firing rate, and the sum of all INCs is the model's firing rate output (Figure 5A). Therefore, INCs reveal which model interneurons create different computations.

If one considers a linear ganglion cell whose RF is stimulus-independent, this linear RF arises from a weighted sum of the RFs of the contributing interneurons. Each INC is then the product of the interneuron's response (the stimulus projected onto the interneuron's RF) and its output synaptic weight to the ganglion cell. In an arbitrarily nonlinear circuit, the ganglion cell's linear RF can change at each point in stimulus space (Figure S5). In this case, we can compute the instantaneous RF (IRF) as the direction of greatest sensitivity (steepest slope) or the gradient of the response with respect to the stimulus. Because of the local response surface, the IRF can change direction and magnitude, and is a weighted sum of the IRFs of the interneurons that create it. To compute the INCs to the response for a particular stimulus, we used the method of integrated gradients[33,35] (see STAR Methods), which computes the average INC as the stimulus is changed from zero stimulus (a gray screen) in a straight path to the chosen stimulus (Figure S5). A straight path is the one choice that satisfies the requirements that the method should be sensitive to any interneuron in the circuit that causes a change in the response, and it should be invariant to the particular implementation of the network.[35] We performed an exact decomposition of a ganglion cell's response into the unique INCs of each of the 8 model cell types in the first layer at each time point, using the method of integrated gradients.

The computation of latency coding

To assess how individual interneurons contributed to latency encoding, we computed INCs during sudden intensity decrements for different contrasts. Previous pharmacological studies demonstrated the effects of ON and OFF pathways, supporting a proposed model with two excitatory neural pathways, an OFF pathway and an ON pathway with delayed kinetics,[31] creating threshold crossings whose timing was stimulus-dependent.

To test whether our CNN model produced latency encoding in this simple manner, we first separated INCs that generated excitation and inhibition according to the sign of the final effect on the

(F) Histogram of correlation coefficients between model interneurons of CNN models from different preparations. Shown are all model-interneuron pairs and the best match for each model.

(G) Reproducibility of model interneuron-interneuron correlations, as assessed by comparing the best match between the same model initialized with different random seeds, which shows fitting reproducibility vs. the best match between model interneurons and interneurons.

(H) Same as (G) for a comparison between models from different retinae, showing cell type reproducibility across preparations. Error bars indicate mean ± S.E.M.

Please cite this article in press as: Maheswaranathan et al., Interpreting the retinal neural code for natural scenes: From computations to neurons, Neuron (2023), https://doi.org/10.1016/j.neuron.2023.06.007

CellPress
OPEN ACCESS

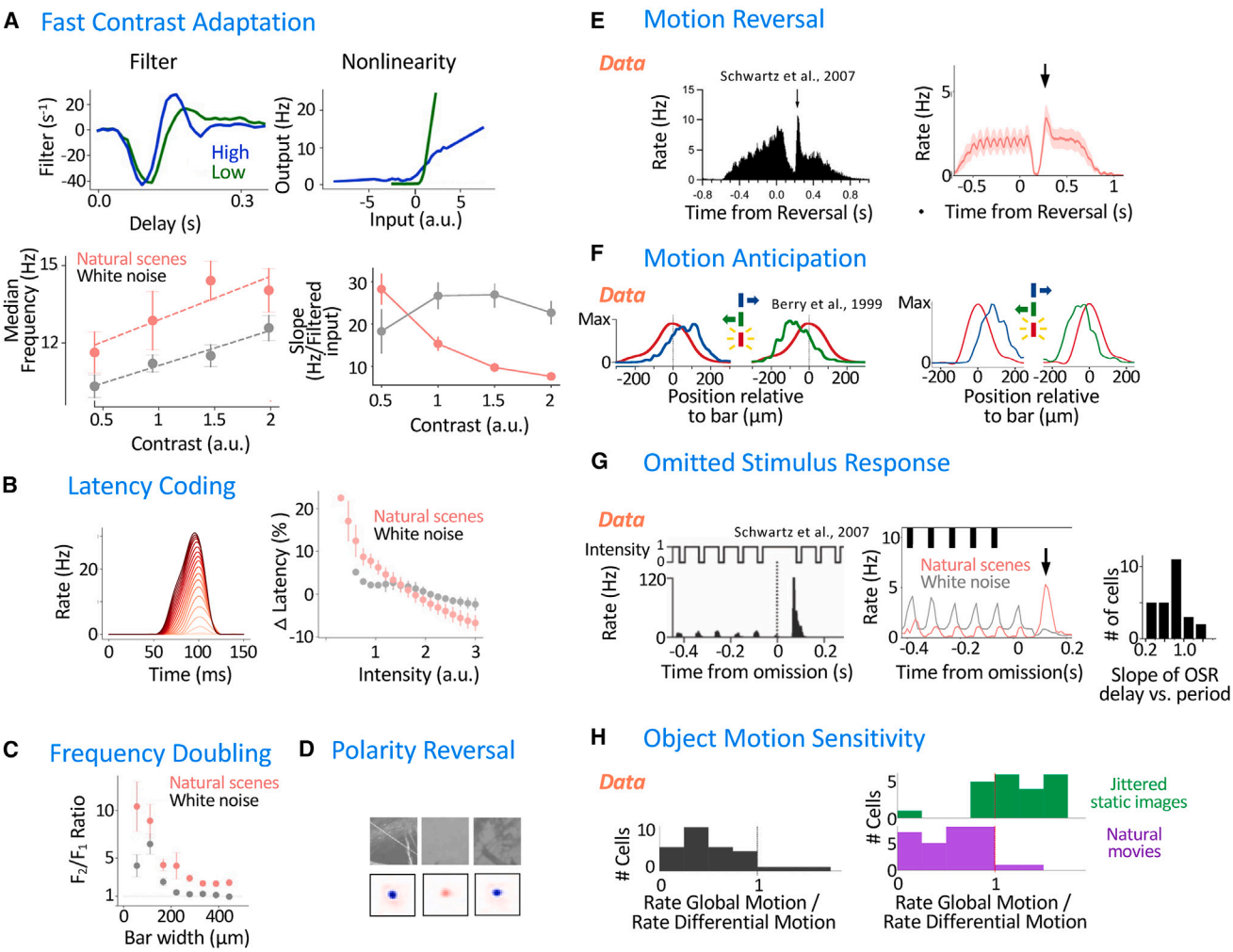

Figure 4. Models reveal that many nonlinear retinal computations are engaged in natural scenes

Artificially structured stimuli were presented to models fit to natural scenes or models fit to white noise, where indicated. All panels show model results, except for the left panels of (E)–(H), which show new or previously published experimental data.

(A) Contrast adaptation. Left: LN model of a model ganglion cell responding to a uniform-field stimulus with low or high contrast. Middle: median temporal frequency taken from the Fourier transform of the temporal filter, averaged over a population of model ganglion cells. Right: averaged sensitivity measured as the slope of the nonlinearity.

(B) Latency encoding. Left: flash response at different intensities. Right: latency of the peak response vs. stimulus intensity for models trained on natural scenes or white noise.

(C) Frequency doubling in response to reversing gratings of different width, computed as the ratio of the response at twice the stimulus frequency (F2) and at the stimulus frequency (F1).

(D) Polarity reversal. Example reversal of polarity during a natural image sequence. Each panel shows the stimulus (top) and corresponding instantaneous RF (bottom) for an example model ganglion cell at a fixed delay (~100 ms) relative to the stimulus at different times, showing fast kernel reversal from an OFF feature (blue) to an ON feature (red), and back.

(E) Motion reversal. Stimulus consists of a moving bar that abruptly reverses direction at different positions. Left: published results[28] of a ganglion cell population showing a synchronous reversal response (arrow). Right: population response of CNN model ganglion cells.

(F) Motion anticipation. Population ganglion cell responses to a flashed bar and left or right motion, from published results[29] (left) or the CNN model (right). Note the x axis represents the ganglion cell position relative to the instantaneous position of the moving bar, and the shift in the population firing rate in the direction of motion indicates motion anticipation.

(G) Omitted stimulus response (OSR). Left: published results[30] showing the response to a missing stimulus following a train of flashes. Middle: CNN model response to a flash sequence, showing the OSR (arrow) for models trained on natural scenes but not white noise. Right: histogram for a population of model ganglion cells of the slope of the OSR delay as a function of the stimulus period, which is centered near one.

(H) Object motion sensitivity. CNN models were fit to either jittered static images or jittered natural movies consisting of swimming fish and abrupt image transitions representing saccades. Models were then shown a jittering central grating surrounded by a jittering background grating that moved either

(legend continued on next page)

Please cite this article in press as: Maheswaranathan et al., Interpreting the retinal neural code for natural scenes: From computations to neurons, Neuron (2023), https://doi.org/10.1016/j.neuron.2023.06.007

ganglion cell (Figure 5B). Under a moderate contrast decrement, excitatory contributions dominated, having a peak that matched the timing of the ganglion cell response (Figure 5B). But as the magnitude of the contrast decrement increased, delayed inhibition increased, truncating excitation and causing a shift in the location of the peak response to earlier times. Also, both excitation and inhibition arose at an earlier time at high contrast. The dual pathway mechanism is consistent with the previous model.

However, in our model we could examine not only excitation and inhibition onto the ganglion cell but also whether that excitation or inhibition originated from ON vs. OFF pathways, as determined by the filter of layer 1 cells (Figure 5C). We found that both ON and OFF pathways showed latency encoding due to increased inhibition at high contrast. Although this result at first seemed in conflict with previous pharmacological experiments, in fact a closer examination of previous results shows that when the ON pathway is blocked, the OFF pathway still exhibits latency encoding.[31] Thus, latency coding in our CNN model is consistent with the details of previous experimental results and, in fact, reproduces those results better than the previous model. It is well established that salamander ON-OFF ganglion cells receive excitation and inhibition that originate in both ON and OFF pathways.[36,37] Therefore, the CNN model's implementation of latency encoding is also more consistent with the physiological literature than the previous two pathway model. We propose that rather than relying solely on timing differences between ON and OFF pathways, latency coding is also generated by timing differences between excitatory and inhibitory pathways within both ON and OFF pathways. This example illustrates the power of training a CNN on natural scenes, followed by model analysis on other stimuli to automatically generate hypotheses.

Motion anticipation

For MA, it has been proposed that for a moving bar, delayed inhibition causes a delayed reduction in sensitivity, suppressing the lagging edge of the traveling wave of activity so its peak shifts toward the leading edge.[29,38] Our INC analysis revealed a computational mechanism consistent with this hypothesis (Figure 6A). However, excitation and inhibition largely overlapped in time, and different asymmetries in excitation and inhibition caused inhibition to shift population activity in the direction of motion. Although this model is qualitatively consistent with previous proposals,[29,38] it adds additional information as to the time course of the responsible neural pathways.

Motion reversal

When a moving bar suddenly reverses direction, ganglion cells near the reversal location synchronously fire. This burst occurs at a fixed latency after the reversal, rather than coinciding with the spatial re-entry of the bar into the RF center. What causes the burst and why does it occur at a fixed latency? The total INCs show that excitation drives the burst at a fixed latency. However, when the reversal is further from the RF, excitation is stronger and more prolonged. Therefore, excitation alone fails to account for the synchronous activation of the burst at different spatial locations. Inhibition, however, is delayed relative to excitation, and is also stronger when the reversal occurs further from the RF center (Figure 6B). The net effect is a response that occurs at approximately a fixed latency and duration relative to the reversal.

At different spatial locations, the linear activity of layer 1 interneurons fails to account for the nonlinear reversal response (Figure 6C).[28] However, the net contributions (sum of excitation and inhibition) at each spatial location show that interneurons generating the synchronous burst are located near the RF center. Furthermore, the net contribution of those interneurons to the total response switches between excitation and inhibition at different times. INC analysis reveals that the fixed temporal latency of the MR response can be explained by truncation of the lagging edge of excitation by inhibitory pathways. This mechanism is qualitatively consistent with a recent experimentally motivated model[39] that points out the crucial role of dual pathways of ON and OFF bipolar cells.

Omitted stimulus response

It is currently unclear how the OSR response might be created, although pharmacological experiments suggest that On bipolar cells are required.[40,41] For periodic flashes of 6–12 Hz, the latency between the omitted flash and the OSR is proportional to the period of the flash sequence.[30,40] Two fundamental questions are: what computational mechanism causes the large amplitude burst, and how is the timing of the peak sensitive to the period of the flashes? One model proposes that bipolar cell activity responds to each flash with an oscillatory response whose period adapts to the flash train period.[42] However, recent bipolar cell recordings suggest that such adaptation to the stimulus period is not present.[43] Another model proposes that ON and OFF pathways together can reproduce most experimental aspects of the phenomena.[41] However, the model only shifts at the onset of the burst with stimulus frequency, but not at the peak of the burst, which has the critical predictive latency.[42]

An analysis of the INCs of the CNN model (Figure 7A) revealed a more sophisticated mechanism than either prior model, consisting of three important cell types, including two OFF cell types and a biphasic ON cell type. This ON cell type was missing from the model fit to white noise, which failed to reproduce the OSR (Figure 7B). The periodic response preceding the OSR was caused by INCs from layer 1 OFF cell types, and the OSR itself was generated by different cell types (Figure 7A). In addition, the ratio of contributions between two main layer 1 model cell types is responsible for shifting OSR latency as a function of frequency (Figure 7C). We explicitly tested this idea by creating a reduced model with only three parallel pathways having only two layers, which is an LN-LN model (Figures 7D–7F). The three cell types of this reduced model were sufficient to reproduce the key features of the OSR.

synchronously (global motion), representing eye movements, or asynchronously (differential motion), representing object motion. Shown is the ratio of firing rates of global to differential motion (OMS index). A ratio much less than one indicates OMS. Left: OMS index computed directly from the spiking responses of recorded ganglion cells responding to jittering grating stimuli. Right: OMS index from CNN model ganglion cells fit to jittered natural scenes or natural movies. Results for (A)–(F) are from a population of 26 ganglion cells. Figures reproduced with permission from authors. Error bars indicate mean ± S.E.M.

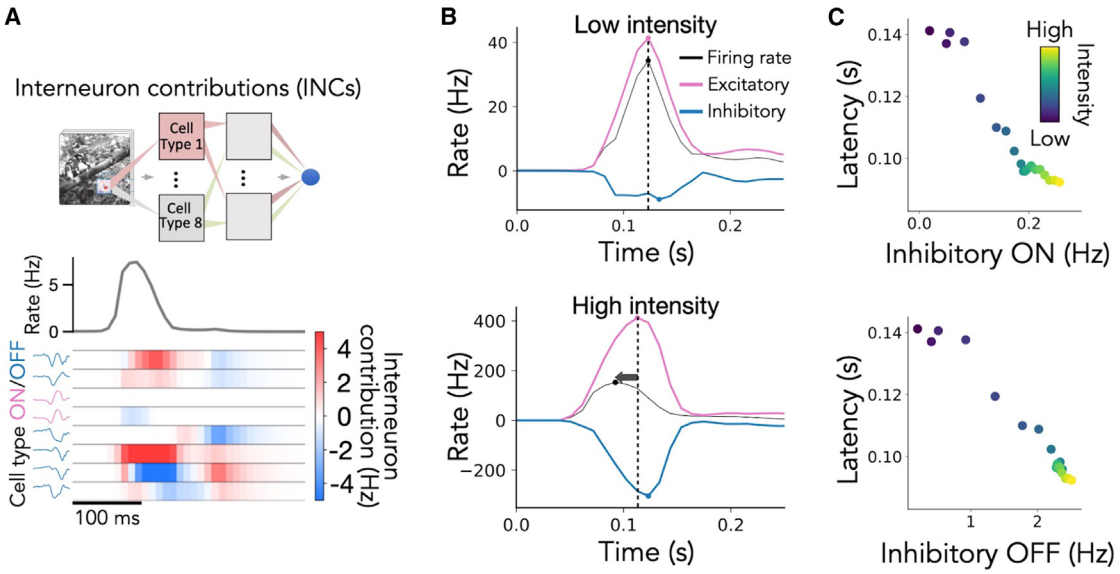

Figure 5. Interneuron contributions to the computation of latency encoding

(A) Top: conceptual diagram of interneuron contributions (INCs), which represent how much each model interneuron contributes to the model's output for each particular stimulus (see STAR Methods). Bottom: INCs for layer 1 model units, averaged over all units of a given type for the 8 cell types for the model's first layer for a short natural stimulus sequence. Each colored row shows the contribution of a layer 1 cell type.

(B) Top: INCs for net excitatory and inhibitory input to a ganglion cell for a low intensity flash. Excitation is defined functionally as a positive contribution to the ganglion cell firing rate, regardless of the sign of the model interneuron response or intervening circuitry. Bottom: same as for a high intensity flash, showing a shift in the peak of the firing rate consistent with latency encoding. The temporal shift is caused by both excitation and inhibition arising at an earlier time, and by asymmetric delayed inhibition that is proportionally larger for a strong flash.

(C) Response latency plotted against inhibitory contributions for layer 1 ON cells (top), and OFF cells (bottom), defined by the sign of the layer 1 cell flash response.

Thus, our INC analysis uncovers a new plausible mechanism of the OSR that cures important inadequacies of prior models. The model yields a new, experimentally testable scientific hypothesis that the OSR is an emergent property of at least two OFF bipolar pathways and a biphasic ON bipolar pathway with different timing.

Model generalization from natural scenes to ethological computations

The generalization of conclusions from one stimulus to another is an important topic, reflecting all of sensory neuroscience, as experiments must always choose one specific set of stimuli. Are phenomena identified with artificial stimuli relevant to natural scenes, and are their mechanisms also engaged during natural scenes? Why do models trained with white noise stimuli not generalize to certain artificial stimuli? To gain insight into these questions, we examined the internal states of the model by calculating the INC to model output. Points in the high dimensional space of INCs for different stimuli can be clustered to examine their similarity or difference. Across the entire natural scene response, we clustered the set of INCs to identify the different patterns of interneuron activity that were directly responsible for retinal output. To simplify the analysis, we chose the eight-dimensional space defined by the eight cell types in layer 1, averaged spatially. We found that different bursts were generated by different combinations of layer 1 cell types and that these patterns clustered into several different modes. Natural scenes activated a broader range of these modes than did white noise, which primarily activates a single mode (Figures 8A–8C), showing that natural scenes explore a wider range of internal states of the model than white noise.

We then examined the entire space of INCs for natural stimuli, white noise, and those that create ethological computations by a t-distributed stochastic neighbor embedding (t-SNE) analysis. The space of INCs for natural scenes and for white noise was in part distinct, which is expected from the two different stimulus distributions. We observed that the natural scenes' INC distribution wholly encompassed the smaller white noise distribution. However, we found, surprisingly, that the interneuron patterns generating responses to some artificial stimuli lie within the space of those elicited by natural stimuli but not within the space of white noise (Figure 8D). Latency coding, OSR, and CA, which occurred much more in natural scenes models than for white noise models (Figure S4), created INCs that were positioned outside the space occupied by white noise (Figure 8D). For MR, MA, and PR, which were more similar between white noise and natural scenes models, INCs occurred only partially outside the space covered by white noise. FD, which was produced by both white noise and natural scenes models, occurred within the space covered by white noise.

This result explains why models fit to natural scenes, but not white noise, recapitulated some phenomena triggered by structured stimuli. White noise does not explore the stimulus space that generates these phenomena, but natural scenes do. We conclude that natural scenes drive the pattern of INCs into

Please cite this article in press as: Maheswaranathan et al., Interpreting the retinal neural code for natural scenes: From computations to neurons, Neuron (2023), https://doi.org/10.1016/j.neuron.2023.06.007

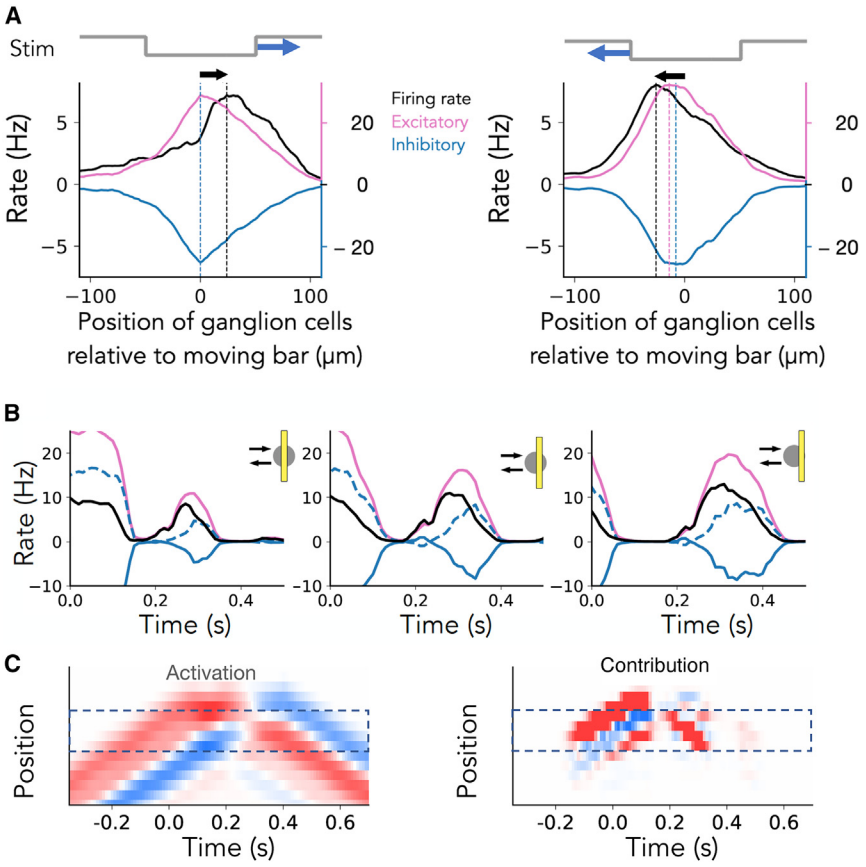

Figure 6. Interneuron contributions to predictive motion computations

(A) Interneuron contributions for motion anticipation, showing excitation and inhibition when a bar moves in different directions. Excitation and inhibition largely overlapped in time, but temporal asymmetry that differed for inhibition and excitation shifted the firing rate in the direction of motion. As in Figure 4F, the x axis represents ganglion cell position relative to the instantaneous moving bar position. A shift of the population firing rate in the direction of motion indicates motion anticipation.

(B) INCs for reversal of a moving bar at t = 0 for three different spatial locations. Delayed inhibition (blue solid line, dashed line is inverted inhibition) is greater at a more distal location (left plot), causing a greater truncation of the firing rate. Across the population, this differential response truncation at different locations synchronizes the response.

(C) Left: activation map of layer 1 cells for one cell type, showing a linear response. Right: spatial map of INCs. At the RF center (dashed box), cells switch the sign of their contribution over time, with the synchronous burst after the reversal comprised predominantly of net excitation. Note that this is only the net effect (excitation plus inhibition) arising from a single cell type. The total excitation and inhibition shown in (B) arises from multiple layer 1 cell types.

a set of states that encompasses previously explored artificial stimuli, showing the ethological relevance of these stimuli to natural scenes, including those of unknown functional importance, such as the OSR.[30]

DISCUSSION

These results capture for the first time the retinal neural code for natural scenes, enabling analyses of that code in terms of the circuit elements that generate any response or computation. By connecting the actions of model interneurons to the responses of real interneurons, and by assigning the responsibility for specific computations to those model interneurons, our approach provides a roadmap to understand how interneurons implement that neural code under a wide range of stimuli.

Understanding ethological phenomena

The process of assigning ethological functions to interneurons has typically taken an ad hoc approach, using fortuitous pharmacology[44,45] or prior knowledge of anatomy or cellular responses.[20,46] One successful approach directly tests computational models of hypothetical neurons proposed to perform a function against interneuron recordings.[18] Our approach generalizes this strategy across a wide range of stimuli and interneurons, yielding an automatic approach to hypothesizing specific roles for interneurons in the response to any stimulus. Guided by an attribution analysis that reveals the internal model states (Figure 8), a future goal will be defining a minimal sufficient stimulus set, including natural images, movies, and potentially artificial stimuli that engage all circuit properties, thus efficiently generating a model that captures all nonlinear properties and phenomena of the retinal circuit.

Correspondence of model and retinal architecture and future extensions

The correspondence of CNN model architecture to that of the retina is not exact, and future versions will more directly match neural and model synaptic connections. Because the first CNN synaptic layer is linear and has spatiotemporal center-surround RF properties found in bipolar cells (the last stage prior to strong rectification), it is reasonable to propose a correspondence between the first CNN layer and the photoreceptor to bipolar cell transformation or to the outer retinal synaptic layer. Although we found that some amacrine cells were highly correlated with layer 1 CNN cells, some amacrine cells are linear and some are strongly nonlinear.[47] Further studies are required to determine whether more nonlinear amacrine cells are more similar to layer 2 in the CNN, or even require three layers. A simplified hypothesis is that the second CNN layer corresponds to the bipolar-cell-to-amacrine cell transformation, and that the last layer adds additional processing found in the amacrine-cell-to-ganglion cell transformation. In our model, all transmission goes through the second layer, whereas a closer correspondence to bipolar-to-ganglion-cell transmission would have direct layer 1 to 3 transmission. This model architecture is known as a skip connection and implies that some layer 2 neurons in our current

Please cite this article in press as: Maheswaranathan et al., Interpreting the retinal neural code for natural scenes: From computations to neurons, Neuron (2023), https://doi.org/10.1016/j.neuron.2023.06.007

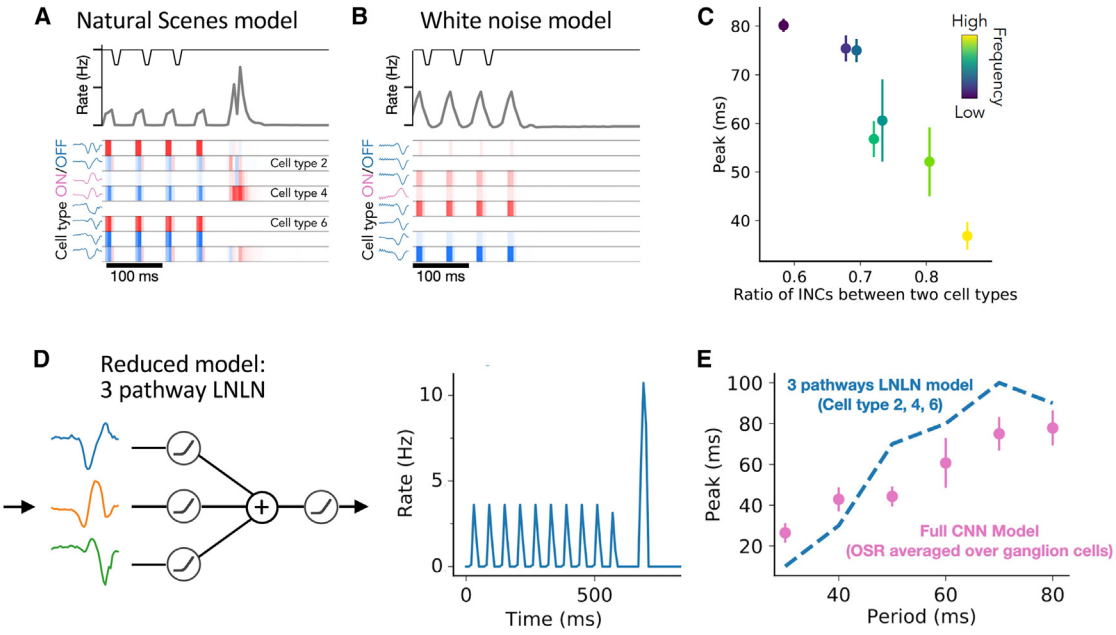

Figure 7. New hypothesis for the omitted stimulus response

(A) Top: strong omitted stimulus response for an example model ganglion cell, consisting of a small periodic response and a much larger burst at the missing flash. Bottom: interneuron contributions from layer 1, showing that excitation from one OFF and a biphasic ON cell drives the burst (cell types 2 and 4), whereas other cells drive the response to periodic flashes.

(B) Same as (A) for a model fit to white noise, which lacks an OSR response and a biphasic ON cell.

(C) The OSR peak time as a function of the ratio of the contribution of two cell types corresponding to cell types 2 and 4 for the natural scenes model. Cell type 2 has a larger contribution at higher frequency. Average is shown over three models.

(D) Left: reduced LN-LN model consisting of three pathways corresponding to cell types 2, 4, and 6 in (A). Compared with the full CNN, this model has three layer 1 cell types, and then the result is summed and then rectified. Right: OSR response from the reduced model.

(E) Comparison of reduced model with full CNN. Error bars indicate mean ± S.E.M.

model acts as relay cells. However, the present analysis of computations is directed toward layer 1 output neurons, and we expect that this more correct architecture would only affect analyses of layer 2 neurons.

A second deviation of the model from the retina affecting primarily the second layer is that our model neurons can have both positive and negative synaptic weights. Retinal synaptic effects are generally restricted to a single sign for each cell, although counterexamples exist, such as photoreceptors having opposite effects on ON and OFF bipolar cells due to different glutamate receptors, and amacrine cells making both chemical inhibitory and electrical excitatory synapses.[48] The addition of this sign constraint could potentially increase the number of cell types required. There are not necessarily the same number of response types as anatomical cell types, as different cells may have the same response type but different targets. In a previous reanalysis of the kinetics of published amacrine responses,[49,50] we concluded that there were approximately eight kinetic classes of salamander amacrine cells, even though there are anatomically more cell types, matching well our conclusion here that eight cell types in the second layer are sufficient to capture amacrine-cell-layer dynamics. However, we expect that a sign constraint on synapses will result in more required model interneuron types than eight in each layer, as separate cells will be needed to create excitation and inhibition, perhaps matching more closely the larger number of anatomical bipolar and amacrine cell types.

Further variations of the model could include recurrent connections to account for horizontal cell connections to the photoreceptor synaptic terminal and recurrence among amacrine cells. Such architectures would necessarily induce additional complexity, which should be evaluated as to tradeoffs in mechanistic understanding vs. simplicity in analysis.

Because the current model has a 400-ms integration time, it cannot capture prolonged phenomena such as slow CA and sensitization.[27,51] However, we have begun to extend the model to longer timescales by incorporating biophysical components that capture the slow synaptic dynamics thought to underlie these phenomena.[52]

Modes in RFs and neural populations

The different INC modes (Figure 8) indicate different network states for different stimuli and are pooled spatially over all of each cell type. A greater number of modes may exist when spatial variation is considered. A possibly related phenomenon to these modes has been observed in ganglion cells, whereby population responses occupy distinct modes[53] under natural visual stimuli. A potential connection between the two observations may exist, in that the INC modes that we observe may underlie distinct clusters of ganglion cell population activity. Similarly, one might predict

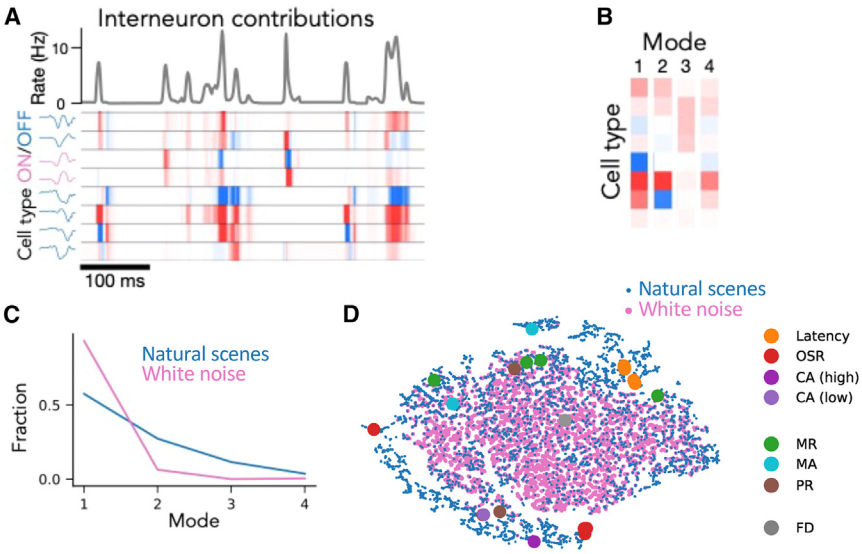

Figure 8. Structure of interneuron contributions under natural scenes, white noise, and ethological computations

(A) Layer 1 INCs, averaged over all units of a given type for a natural stimulus sequence. Each colored row shows the contribution of one cell type.

(B) Different modes of INC combinations identified by k-means clustering (k = 4).

(C) Fraction of INCs occupying each cluster for models fit to natural scenes or white noise.

(D) t-SNE plot of INCs, including natural scenes, white noise, and artificially structured stimuli. OSR, omitted stimulus response; CA, contrast adaptation; MR, motion reversal; MA, motion anticipation; PR, polarity reversal; FD, frequency doubling. Each point represents the vector of interneurons in layer 1 contributing to the response at a single time point. For artificial stimuli, points are taken during the peak response, and 1–4 different conditions are shown for each stimulus type.

that ganglion cell population modes lead to dynamic RF modes in the higher brain. Further work to connect these two observations may connect the proposed functions of error-correcting codes[53] and dynamic changes in INCs.

The ability to analyze the state of the neural code and of the circuit at each time point allows a new level of access to the neural code and its construction for any arbitrary stimulus, allowing the statement of detailed quantitative hypotheses for natural scenes and ethological computations that include specific interneuron types. These hypotheses will serve as the foundation for future directed experiments to define how interneuron patterns generate the dynamic neural code for natural scenes.

STAR★METHODS

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

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - Electrophysiology
  - Visual Stimuli
  - Model training
  - Response reliability
  - Interneuron correlations
  - Interneuron contributions

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.neuron.2023.06.007.

ACKNOWLEDGMENTS

The authors wish to acknowledge William Newsome, Jennifer Raymond, Tom Clandinin, and Leonidas Guibas for helpful discussions. This work was supported by grants from the NEI (R01EY022933, R01EY025087, and P30-EY026877), Pew Charitable Trusts, McKnight Endowment Fund for Neuroscience, and the Ziegler Foundation (S.A.B.); Burroughs Wellcome, McKnight, James S. McDonnnell, Simons Foundations, and the Office of Naval Research (S. Ganguli); an NSF fellowship (N.M.) and an NIH NRSA (L.T.M.), by the Stanford Medical Scientist Training Program (D.B.K.); and an NSF IGERT graduate fellowship (D.B.K.).

AUTHOR CONTRIBUTIONS

All authors participated in the overall design of the study. N.M., L.T.M., D.B.K., J.B.M., S. Ganguli, and S.A.B. participated in the design of experiments. N.M., L.T.M., D.B.K., and J.B.M. performed biological experiments. N.M., L.T.M., S. Grant, J.B.M., A.N., and J.H.W. participated in the development and fitting of computational models. N.M., L.T.M., H.T., S. Grant, J.B.M., L.E.B., and J.H.W. performed *in silico* experiments. N.M., L.T.M., H.T., S. Grant, J.B.M., L.E.B., J.H.W., and S.A.B. participated in computational analyses. H.T. designed and performed integrated gradients analyses. N.M., L.T.M., H.T., S. Grant, S. Ganguli, and S.A.B. contributed to the writing of the paper.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: May 20, 2022
Revised: January 30, 2023
Accepted: June 14, 2023
Published: July 13, 2023

REFERENCES

1. Simoncelli, E.P., and Olshausen, B.A. (2001). Natural image statistics and neural representation. Annu. Rev. Neurosci. *24*, 1193–1216.
2. Masland, R.H. (2012). The neuronal organization of the retina. Neuron *76*, 266–280. https://doi.org/10.1016/j.neuron.2012.10.002.
3. Felsen, G., and Dan, Y. (2005). A natural approach to studying vision. Nat. Neurosci. *8*, 1643–1646. https://doi.org/10.1038/nn1608.
4. Rust, N.C., and Movshon, J.A. (2005). In praise of artifice. Nat. Neurosci. *8*, 1647–1650. https://doi.org/10.1038/nn1606.

5. Gollisch, T., and Meister, M. (2010). Eye smarter than scientists believed: neural computations in circuits of the retina. Neuron *65*, 150–164. https://doi.org/10.1016/j.neuron.2009.12.009.

6. Pillow, J.W., Shlens, J., Paninski, L., Sher, A., Litke, A.M., Chichilnisky, E.J., and Simoncelli, E.P. (2008). Spatio-temporal correlations and visual signalling in a complete neuronal population. Nature *454*, 995–999. https://doi.org/10.1038/nature07140.

7. Yamins, D.L.K., Hong, H., Cadieu, C.F., Solomon, E.A., Seibert, D., and DiCarlo, J.J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc. Natl. Acad. Sci. USA *111*, 8619–8624. https://doi.org/10.1073/pnas.1403112111.

8. Khaligh-Razavi, S.-M., and Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS Comput. Biol. *10*, e1003915. https://doi.org/10.1371/journal.pcbi.1003915.

9. Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. Annu. Rev. Vis. Sci. *1*, 417–446. https://doi.org/10.1146/annurev-vision-082114-035447.

10. Kell, A.J.E., Yamins, D.L.K., Shook, E.N., Norman-Haignere, S.V., and McDermott, J.H. (2018). A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. Neuron *98*, 630–644.e16. https://doi.org/10.1016/j.neuron.2018.03.044.

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

12. Hochstein, S., and Shapley, R.M. (1976). Linear and nonlinear spatial subunits in Y cat retinal ganglion cells. J. Physiol. *262*, 265–284.

13. Maheswaranathan, N., Kastner, D.B., Baccus, S.A., and Ganguli, S. (2018). Inferring hidden structure in multilayered neural circuits. PLoS Comput. Biol. *14*, e1006291. https://doi.org/10.1371/journal.pcbi.1006291.

14. Li, P.H., Field, G.D., Greschner, M., Ahn, D., Gunning, D.E., Mathieson, K., Sher, A., Litke, A.M., and Chichilnisky, E.J. (2014). Retinal representation of the elementary visual signal. Neuron *81*, 130–139. https://doi.org/10.1016/j.neuron.2013.10.043.

15. Vintch, B., Movshon, J.A., and Simoncelli, E.P. (2015). A convolutional subunit model for neuronal responses in macaque V1. J. Neurosci. *35*, 14829–14841. https://doi.org/10.1523/JNEUROSCI.2815-13.2015.

16. Wu, A., Park, I.M., and Pillow, J.W. (2015). Convolutional spike-triggered covariance analysis for neural subunit models. In Advances in Neural Information Processing Systems 28 (NIPS 2015).

17. Martinez-Conde, S., and Macknik, S.L. (2008). Fixational eye movements across vertebrates: comparative dynamics, physiology, and perception. J. Vis. *8*, 28.1–16. https://doi.org/10.1167/8.14.28.

18. Olveczky, B.P., Baccus, S.A., and Meister, M. (2003). Segregation of object and background motion in the retina. Nature *423*, 401–408. https://doi.org/10.1038/nature01652.

19. Baccus, S.A., and Meister, M. (2002). Fast and slow contrast adaptation in retinal circuitry. Neuron *36*, 909–919.

20. Diamond, J.S. (2017). Inhibitory interneurons in the retina: types, circuitry, and function. Annu. Rev. Vis. Sci. *3*, 1–24. https://doi.org/10.1146/annurev-vision-102016-061345.

21. Kaneko, A. (1973). Receptive field organization of bipolar and amacrine cells in the goldfish retina. J. Physiol. *235*, 133–153. https://doi.org/10.1113/jphysiol.1973.sp010381.

22. Atick, J. (1992). Could information theory provide an ecological theory of sensory processing? Netw.: Comput. Neural Syst. *3*, 213–251.

23. Doi, E., Gauthier, J.L., Field, G.D., Shlens, J., Sher, A., Greschner, M., Machado, T.A., Jepson, L.H., Mathieson, K., Gunning, D.E., et al. (2012). Efficient coding of spatial information in the primate retina. J. Neurosci. *32*, 16256–16264.

24. Gjorgjieva, J., Sompolinsky, H., and Meister, M. (2014). Benefits of pathway splitting in sensory coding. J. Neurosci. *34*, 12127–12144. https://doi.org/10.1523/JNEUROSCI.1032-14.2014.

25. Pitkow, X., and Meister, M. (2012). Decorrelation and efficient coding by retinal ganglion cells. Nat. Neurosci. *15*, 628–635. https://doi.org/10.1038/nn.3064.

26. Kim, K.J., and Rieke, F. (2001). Temporal contrast adaptation in the input and output signals of salamander retinal ganglion cells. J. Neurosci. *21*, 287–299.

27. Smirnakis, S.M., Berry, M.J., Warland, D.K., Bialek, W., and Meister, M. (1997). Adaptation of retinal processing to image contrast and spatial scale. Nature *386*, 69–73. https://doi.org/10.1038/386069a0.

28. Schwartz, G., Taylor, S., Fisher, C., Harris, R., and Berry, M.J. (2007). Synchronized firing among retinal ganglion cells signals motion reversal. Neuron *55*, 958–969. https://doi.org/10.1016/j.neuron.2007.07.042.

29. Berry, M.J., Brivanlou, I.H., Jordan, T.A., and Meister, M. (1999). Anticipation of moving stimuli by the retina. Nature *398*, 334–338. https://doi.org/10.1038/18678.

30. Schwartz, G., Harris, R., Shrom, D., and Berry, M.J. (2007). Detection and prediction of periodic patterns by the retina. Nat. Neurosci. *10*, 552–554. https://doi.org/10.1038/nn1887.

31. Gollisch, T., and Meister, M. (2008). Rapid neural coding in the retina with relative spike latencies. Science *319*, 1108–1111. https://doi.org/10.1126/science.1149639.

32. Geffen, M.N., de Vries, S.E.J., and Meister, M. (2007). Retinal ganglion cells can rapidly change polarity from Off to On. PLoS Biol. *5*, e65. https://doi.org/10.1371/journal.pbio.0050065.

33. Tanaka, H., Nayebi, A., Maheswaranathan, N., McIntosh, L., Baccus, S.A., and Ganguli, S. (2019). From deep learning to mechanistic understanding in neuroscience: the structure of retinal prediction. Adv. Neural Inf. Process. Syst. *32*, 8537–8547.

34. Lehky, S.R., and Sejnowski, T.J. (1988). Network model of shape-from-shading: neural function arises from both receptive and projective fields. Nature *333*, 452–454.

35. Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. In Proceedings of the International Conference on Machine Learning, pp. 3319–3328.

36. Pang, J.-J., Gao, F., and Wu, S.M. (2002). Relative contributions of bipolar cell and amacrine cell inputs to light responses of ON, OFF and ON-OFF retinal ganglion cells. Vision Res. *42*, 19–27.

37. Pang, J.-J., Gao, F., and Wu, S.M. (2007). Cross-talk between ON and OFF channels in the salamander retina: indirect bipolar cell inputs to ON-OFF ganglion cells. Vision Res. *47*, 384–392. https://doi.org/10.1016/j.visres.2006.09.021.

38. Johnston, J., and Lagnado, L. (2015). General features of the retinal connectome determine the computation of motion anticipation. eLife *4*, 1581. https://doi.org/10.7554/eLife.06250.

39. Chen, E.Y., Chou, J., Park, J., Schwartz, G., and Berry, M.J. (2014). The neural circuit mechanisms underlying the retinal response to motion reversal. J. Neurosci. *34*, 15557–15575.

40. Schwartz, G., and Berry, M.J. (2008). Sophisticated temporal pattern recognition in retinal ganglion cells. J. Neurophysiol. *99*, 1787–1798.

41. Werner, B., Cook, P.B., and Passaglia, C.L. (2008). Complex temporal response patterns with a simple retinal circuit. J. Neurophysiol. *100*, 1087–1097.

42. Gao, J., Schwartz, G., Berry, M.J., and Holmes, P. (2009). An oscillatory circuit underlying the detection of disruptions in temporally-periodic patterns. Network *20*, 106–135.

43. Deshmukh, N.R. (2015). Complex computation in the retina. Doctoral dissertation (Princeton University).

44. Kastner, D.B., Ozuysal, Y., Panagiotakos, G., and Baccus, S.A. (2019). Adaptation of inhibition mediates retinal sensitization. Curr. Biol. *29*, 2640–2651.e4. https://doi.org/10.1016/j.cub.2019.06.081.

45. Vaney, D.I., Sivyer, B., and Taylor, W.R. (2012). Direction selectivity in the retina: symmetry and asymmetry in structure and function. Nat. Rev. Neurosci. *13*, 194–208. https://doi.org/10.1038/nrn3165.

46. Vaney, D.I., Young, H.M., and Gynther, I.C. (1991). The rod circuit in the rabbit retina. Vis. Neurosci. *7*, 141–154.

47. Manu, M., and Baccus, S.A. (2011). Disinhibitory gating of retinal output by transmission from an amacrine cell. Proc. Natl. Acad. Sci. USA *108*, 18447–18452. https://doi.org/10.1073/pnas.1107994108.

48. Masland, R.H. (2012). The tasks of amacrine cells. Vis. Neurosci. *29*, 3–9.

49. Baccus, S.A. (2007). Timing and computation in inner retinal circuitry. Annu. Rev. Physiol. *69*, 271–290.

50. Pang, J.-J., Gao, F., and Wu, S.M. (2002). Segregation and integration of visual channels: layer-by-layer computation of ON-OFF signals by amacrine cell dendrites. J. Neurosci. *22*, 4693–4701.

51. Kastner, D.B., and Baccus, S.A. (2011). Coordinated dynamic encoding in the retina using opposing forms of plasticity. Nat. Neurosci. *14*, 1317–1322. https://doi.org/10.1038/nn.2906.

52. Ding, X., Lee, D., Grant, S., Stein, H., McIntosh, L.T., Maheswaranathan, N., and Baccus, S.A. (2021). A mechanistically interpretable model of the retinal neural code for natural scenes with multiscale adaptive dynamics. In 55th Asilomar Conference on Signals, Systems, and Computers (IEEE Publications), pp. 287–291.

53. Prentice, J.S., Marre, O., Ioffe, M.L., Loback, A.R., Tkačik, G., and Berry, M.J. (2016). Error-robust modes of the retinal population code. PLoS Comput. Biol. *12*, e1005148. https://doi.org/10.1371/journal.pcbi.1005148.

54. Brainard, D.H. (1997). The psychophysics toolbox. Spat. Vis. *10*, 433–436.

55. Tkačik, G., Garrigan, P., Ratliff, C., Milčinski, G., Klein, J.M., Seyfarth, L.H., Sterling, P., Brainard, D.H., and Balasubramanian, V. (2011). Natural images from the birthplace of the human eye. PLoS One *6*, e20409. https://doi.org/10.1371/journal.pone.0020409.

56. McIntosh, L.T., Maheswaranathan, N., Nayebi, A., Ganguli, S., and Baccus, S.A. (2016). Deep learning models of the retinal response to natural scenes. Adv. Neural Inf. Process. Syst. *29*, 1369–1377.

57. Kingma, D.P., and Ba, J. (2014). Adam: A method for stochastic optimization. Preprint at arXiv. https://doi.org/10.48550/arXiv.1412.6980.

58. Abadi, M. (2016). TensorFlow: A system for large-scale machine learning. Preprint at arXiv. https://doi.org/10.48550/arXiv.1605.08695.

59. Ketkar, N., and Moolayil, J. (1997). Introduction to Pytorch. In Deep Learning with Python (Apress), pp. 195–208.

60. Ioffe, S., and Szegedy, C. (2015). Batch normalization: accelerating deep network training by reducing internal covariate shift. In International Conference on Machine Learning (PMLR), pp. 448–456.

Please cite this article in press as: Maheswaranathan et al., Interpreting the retinal neural code for natural scenes: From computations to neurons, Neuron (2023), https://doi.org/10.1016/j.neuron.2023.06.007

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chemicals, peptides, and recombinant proteins		
Sodium Chloride	Fisher	S271-500
Potassium Chloride	Fisher	P217-500
Calcium Chloride	Fisher	C69-500
Magnesium Chloride	Fisher	M33-500
Sodium Bicarbonate	Fisher	S233-500
D-Glucose	Fisher	D16-500
Experimental models: Organisms/strains		
Larval tiger salamanders, *Ambystoma tigrinum*	Wadelco, TX	N/A
Software and algorithms		
Igor	Wavemetrics, Inc.	Igor 6, 7
Matlab	Mathworks	N/A
Custom Software for model fitting and data analysis	Baccus and Ganguli laboratories	Zenodo: https://doi.org/10.5281/zenodo.8001612
Deposited data		
Experimental data	Baccus laboratory	Stanford Digital Repository: https://doi.org/10.25740/rk663dm5577

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Stephen Baccus (baccus@stanford.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

The data used in this study has been deposited at Stanford Digital Repository: rk663dm5577. Original code used in this study has been deposited at Zenodo:8001612.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Larval tiger salamanders of either sex were used.

METHOD DETAILS

Electrophysiology

Retinal ganglion cells were recorded using an array of 60 electrodes (Multichannel Systems) as previously described.[51] Intracellular recordings were performed using sharp microelectrodes as previously described.[47]

Visual Stimuli

A video monitor projected the visual stimuli at 30 Hz controlled by Matlab (Mathworks), using Psychophysics Toolbox.[54] Stimuli had a constant mean intensity of 10 mW/m^2. Images were presented in a 50 x 50 grid with a square size of 50 μm. Static natural jittered scenes consisted of images drawn from a natural image database.[55] To create images for presentation to the retina, original color images were converted to grayscale, and were scaled to have minimum and maximum pixel intensities that matched that of the monitor. Pixel regions of 50 x 50 size were then selected from each image at a random location without spatial averaging for

presentation. Images drifted in two dimensions in a random walk,[18] moving with a standard deviation of 0.5 pixels per video frame horizontally and vertically. The image also abruptly changed in a single frame to a different location every one second, representing saccades, although such transitions did not contain a sweeping shift in the image. Transitions occurred both between different locations in the same image and between different images. White noise stimuli were binary (black and white) stimuli with the same frame rate, spatial size and duration as natural scene stimuli. Natural movies consisted of fish swimming in an aquarium and contained both drift and abrupt image transitions that matched static jittered natural scenes. For analysis of model responses to artificial stimuli (Figure 4), unless otherwise stated stimuli were chosen to match published values for each phenomenon.

Model training

We trained Convolutional Neural Network (CNN) models to predict retinal ganglion cell responses to either a white noise or natural scenes stimulus, simultaneously for all cells in the recorded population of a given retina.[56] All results, unless otherwise specified, are reported as an average over 26 retinal ganglion cells split between three different datasets. The datasets consisted of 5, 4, and 17 cell recordings, each with white noise and natural scenes stimulus segments. A single model was trained for each dataset and each stimulus type resulting in 6 models total.

Model parameters were set to a random initial condition and then optimized to minimize a loss function corresponding to the negative log-likelihood under Poisson spike generation,

$$L(y_t, \hat{y}_t) = \frac{1}{T}\sum_{t=0}^{T} \hat{y}_t - y_t \log \hat{y}_t \qquad \text{(Equation 1)}$$

where y_t and $\hat{y}_t$ are the experimental and predicted firing rates of the retinal ganglion cells at time t, respectively with a batch size of T, chosen to be 5000 samples. Models were optimized to fit all cells in a preparation simultaneously by minimizing the sum of negative log-likelihood over all cells. To help with model fitting, we smoothed retinal ganglion responses during training with a 10 ms standard deviation Gaussian, the size of a single time bin in our model.

The architecture of the CNN consisted of two convolutional layers, with 8 cell types (or channels) per layer, followed by a fully connected layer. Each convolutional layer consisted of a linear spatiotemporal filter and a rectification using a rectified linear unit (ReLU). The final layer consisted of a spatial filter, a scaling and shifting parameter for each ganglion cell, and a nonlinearity using a softplus function.

CNN model optimization was performed using Adam,[57] a variant of stochastic gradient descent. Models were trained using TensorFlow[58] or PyTorch[59] on NVIDIA Titan X GPUs. Training an individual model to convergence required ~8 hours on a single GPU. The networks were regularized with an L2 weight penalty at each layer and an L1 activity penalty at the final layer, which helped maintain a baseline firing rate near 0 Hz.

The response and stimulus were binned in 10 ms time bins. To present images to the model, the mean intensity across all images was subtracted from the ensemble of images, and then the ensemble was divided by its standard deviation. We split our dataset into training (323,786 ten ms samples, ~ 54 min.), validation (35,976 samples, ~6 min.), and test sets (5,957 samples, ~ 5 min.) The number of 30 Hz stimulus frames was one third of these values. We chose the number of layers, number of filters per layer, the type of layer (convolutional or fully connected), filter kernel sizes, regularization hyperparameters (L1=0.0001, L2=0.001), and learning rate (0.005) based on performance on the validation set. Training and validation sets were single trials of stimuli. The number of training, validation and test samples were the same for natural scenes and white noise.

The test set included 5 – 10 sixty second repeats of an identical stimulus different from those used for training, from which we computed an average firing rate for comparing to the model. We found that increasing the number of layers beyond three did not improve performance, and we settled on eight filter types in both the first and second layers, with filters that were much larger (Layer 1,15 x 15 and Layer 2, 11 x 11) compared to traditional deep learning networks used for image classification (usually 5 x 5 or smaller).

The spatial components of the convolutional filters were implemented as a series of stacked linear convolutions, each consisting of a series of 3 x 3 filters with 8 channels. Thus seven 3 x 3 filters were applied in sequence to generate a 15 x 15 filter. After optimization, we collapsed this series of 3 x 3 convolutional filters into a single larger convolutional filter (15 x 15 or 11 x 11). Therefore, this procedure did not change the final architecture of the model, but improved the model's performance, presumably by reducing the number of parameters and centering the features of the filter. Model interneurons were present at each location. Because the size of model RFs was unrelated to their spacing, the tiling of cells did not influence filter parameters. The number of model interneurons was reduced at each layer as a consequence of the filter sizes and edge effects, so that the stimulus input was 50 x 50 regions, the first convolutional layer output had 36 x 36 units, the second convolutional layer output had 26 x 26 units, and the third fully connected layer had an output equal to the number of recorded ganglion cells. The first layer consisted of eight 15 x 15 x 40 spatiotemporal filters, which were fit as 8 channels each having a 3 x 3 x 40 spatiotemporal (x,y,t) filter followed by 6 sublayers each having 8 input channels, 8 output channels and a 3 x 3 spatial filter. The number of filter parameters in the first layer was 8 x 3 x 3 x 40 + 6 x 8 x 8 x 3 x 3 = 6336 parameters. The second layer consisted of 8 channels of 11 x 11 spatial filters, and was fit using five sublayers each having eight input channels, eight output channels and a 3 x 3 spatial filter for 8 x 8 x 3 x 3 = 576 filter parameters. The third fully connected layer had 8 x 26 x 26 = 5408 filter parameters for each ganglion cell.

Please cite this article in press as: Maheswaranathan et al., Interpreting the retinal neural code for natural scenes: From computations to neurons, Neuron (2023), https://doi.org/10.1016/j.neuron.2023.06.007

Several strategies were added to improve optimization. During optimization, independent gaussian noise with zero mean a standard deviation of 0.05 was added to each activation following the convolutional filter. Batch normalization[60] was applied at each layer, which normalizes the model activations using the mean and standard deviation of inputs collected during training. For the first two convolutional layers, Batchnorm2D was applied which uses the full spatial dimension of each channel for the normalization. In the last layer, a Batchnorm1D was applied which is specific to each ganglion cell. Values quoted are mean ± s.e.m. unless otherwise stated.

Linear-Nonlinear Models

Linear-nonlinear models were fit by minimizing the same objective as used for the CNN, the Poisson log-likelihood of data under the model. Each model, however, was only trained on an individual ganglion cell firing response. We found that these were highly susceptible to overfitting the training dataset, and imposed an additional regularization procedure of zeroing out the stimulus outside of a 500 μm window centered on the cell's receptive field. The nonlinearity of the LN model was a soft-plus function.

Generalized Linear Models

Generalized linear models (GLMs) were fit by minimizing the same objective as used for the CNN, the Poisson log-likelihood of data under the model. We performed the same cutout regularization procedure of only keeping the stimulus within a 500 μm region around the receptive field, which was critical for performance. The GLMs differed from the linear-nonlinear models in that they have an additional spike history feedback term used to predict the cell's response.[6] Instead of the standard exponential nonlinearity, we found that using soft rectified functions log(1+exp(x)) gave better performance.

Response reliability

The reliability of recorded ganglion cells over the course of each experiment was measured by computing the correlation coefficient between a cell's average response to the same stimulus on different blocks of trials. We analyzed only those cells with a correlation exceeding 0.3. This measure of a cell's reliability is not the cell's trial to trial variability, but is an estimate of how closely the average data matches itself given the number of trials that we have for the purpose of estimating the maximum that a model of the average response could be expected to match the average of the data. This maximum is limited by the number of trials (5 – 10). Of 26 cells that met our criteria of reliability, 17 were fast OFF-type cells (which are On-Off cells), 5 were medium OFF, 3 were slow OFF, and one was an On cell. Our dataset also included a small percentage of additional ON ganglion cells, but these did not meet the reliability criterion.

Interneuron correlations

We compared the similarity of responses of recorded interneurons with other interneurons in order to reveal both the diversity of the interneuron responses and assess the similarity between cells of the same type. However, one cannot directly compare the recordings of interneurons because each interneuron is recorded at a different position with respect to the stimulus. However, by fitting a model to each interneuron recording, the stimulus could then be presented at the same location relative to each model, thus allowing a direct comparison of the response of a recorded interneuron to the model of another recorded interneuron. To choose a model for recorded interneurons, we first fit different types of CNN models including single layer, LNLN models, two and three layer CNN models, and then chose the best performing model individually to represent each interneuron. We then computed the correlation coefficient for all interneuron pairs, and the best match for each interneuron.

To compare the similarity of recorded interneurons with model interneurons, for each model interneuron type in layer 1 and layer 2 it was necessary to find the location that corresponded to the location of the recorded interneuron. To do this, we computed the correlation coefficient between the actual interneuron recording (not the model of the recorded interneuron) and each model interneuron at each spatial location and, and then chose the spatial location with the highest absolute value correlation. We used the absolute value to account for the fact that Off cells might be compared with On cells, although we report the actual correlation coefficient, not the absolute value.

To compare model interneurons with model interneurons of another model, we adopted a similar procedure as used for comparing recorded interneurons with model interneurons. We chose the center of the visual field for one model, and found the corresponding spatial location in another model by finding the model interneuron with the highest absolute value of the correlation coefficient.

When finding the best match between recorded interneurons and model interneurons, or between model interneurons in different CNN models, we always searched only over a single model, 16 unit types for the standard model and varying numbers of units for different architectures. Results are reported across all models, either separately or by averaging.

Interneuron contributions

In this section, we outline a method to determine how much each model interneuron contributes to the firing rate of the output neuron. To do so, we apply a method called Integrated Gradients,[33,35] which was originally developed for interpretability in machine learning. In order to determine which part of the image the model's decisions originate from, Integrated Gradients decomposes the model's output into attributes for each pixel in the input image. Analogously, we apply the chain rule to decompose the firing rate of ganglion cells into Interneuron Contributions (INCs) for each of the eight model cells by performing path integration.

Mathematically, the trained deep learning model represents a nonlinear function $r(t) = \mathcal{F}[s(t)]$, where $r(t)$ is the output firing rate and $s(t) \in \mathbb{R}^{50\times50\times40}$ is the movie input. Using the line integral $s(t,\alpha) = \alpha s(t)$ where the path takes a straight line $\alpha : 0 \rightarrow 1$, and assuming $\mathcal{F}[s(t;0)] = 0$, we obtain an equality

$$\mathcal{F}[s(t,1)] = \int_0^1 d\alpha \frac{\partial \mathcal{F}}{\partial s}\Big|_{s(t,\alpha)} \cdot \frac{\partial s(t,\alpha)}{\partial \alpha} = s(t) \cdot \int_0^1 d\alpha \frac{\partial \mathcal{F}}{\partial s}\Big|_{s(t,\alpha)} \quad \text{(Equation 2)}$$

Our goal is to quantify the contributions of the first layer model units, whose responses to the stimulus are $z_c^{[1]} = W_c^{[1]} * s(t) + b_c$, where [1] refers to an index of the layer, c refers to channel or model cell type, $W_c^{[1]}$ is the linear convolutional filter, * indicates the convolution operation and b_c is the bias or offset parameter.

Therefore we further apply the chain rule to define A_c, the INC of the cth channel as

$$r(t) = \sum_c \left(W_c^{[1]} * s\right) \cdot \int_0^1 d\alpha \frac{\partial \mathcal{F}}{\partial z_c^{[1]}}\Big|_{s(t,\alpha)} \equiv \sum_c A_c \quad \text{(Equation 3)}$$

with $\partial \mathcal{F}/\partial z$ for each model interneuron computed as the gradient of the model output with respect to the model interneuron's activation under the stimulus $s(t,\alpha) = \alpha s(t)$. Finally, the spatially averaged INCs $A_{c:1-8}$ form a vector with eight elements, which is taken as the contribution of that model cell type to the model output at that instant of time.