Convolutional neural network models of the primate retina reveal adaptation to natural stimulus statistics

Michaela Vystrčilová[1] ✉, **Shashwat Sridhar**[2,3], **Max F. Burg**[1,6,7], **Tim Gollisch**[2,3,5], **Alexander S. Ecker**[1,4] ✉

[1]University of Göttingen, Institute of Computer Science and Campus Institute Data Science, Göttingen, Germany; [2]University Medical Center Göttingen, Department of Ophthalmology, Göttingen, Germany; [3]Bernstein Center for Computational Neuroscience Göttingen, Göttingen, Germany; [4]Max Planck Institute for Dynamics and Self-Organization, Göttingen, Germany; [5]Cluster of Excellence "Multiscale Bioimaging: from Molecular Machines to Networks of Excitable Cells" (MBExC), University of Göttingen, Göttingen, Germany; [6]International Max Planck Research School for Intelligent Systems, Tübingen, Germany; [7]Tübingen AI Center, University of Tübingen, Germany

✉ **For correspondence:** michaela.vystrcilova@uni-goettingen.de, ecker@cs.uni-goettingen.de

Data availability: All data and code will be made available upon publication.

Funding: This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project ID 432680300 (SFB 1456, project B05) – and by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement number 101041669). Computing time was made available on the high-performance computers HLRN-IV at GWDG at the NHR Center NHR@Göttingen.

Competing interests: The author declare no competing interests.

Abstract

The diverse nature of visual environments demands that the retina, the first stage of the visual system, encodes a vast range of stimuli with various statistics. The retina adapts its computations to some specific features of the input, such as brightness, contrast or motion. However, it is less clear whether it also adapts to the statistics of natural scenes compared to white noise, the latter of which is often used to infer models of retinal computation. To address this question, we analyzed neural activity of retinal ganglion cells (RGCs) in response to both white noise and naturalistic movie stimuli. We performed a systematic comparative analysis of traditional linear-nonlinear (LN) and recent convolutional neural network (CNN) models and tested their generalization across stimulus domains. We found that no model type trained on one stimulus ensemble was able to accurately predict neural activity on the other, suggesting that retinal processing depends on the stimulus statistics. Under white noise stimulation, the receptive fields of the neurons were mostly lowpass, while under natural image statistics they exhibited a more pronounced surround resembling the whitening filters predicted by efficient coding. Together, these results suggest that retinal processing dynamically adapts to the stimulus statistics.

Introduction

The retina is the direct interface between the visual environment and an animal's brain. It evolved for encoding stimuli characterized by different statistical properties. This encoding process is not static; the retina's computational strategies dynamically adjust to specific attributes of visual inputs. Adaptation to the mean light intensity (Dowling, 1960; Dowling, 1963) and contrast levels (Smirnakis et al., 1997; Kim and Rieke, 2001; Baccus and Meister, 2002) is well documented. Adap-

tation of RGCs to higher order statistics has been investigated, showing that RGCs are only mildly influenced by skew and kurtosis (Bonin, Mante, and Carandini, 2006; Tkačik et al., 2014). The RGCs of salamanders and rabbits adapt to correlated full-field flicker, negatively correlated checkerboard stimuli and oriented bars (Hosoya, Baccus, and Meister, 2005). Adaptation to motion statistics has been shown using moving gratings (Ölveczky, Baccus, and Meister, 2007) and parameterized motion clouds (Ravello et al., 2019). However, it is not known whether and how the retina adapts its processing between stimulation with noise stimuli and natural scenes. Noise stimuli are often used to characterize retinal computation experimentally. How much we can learn about retinal computation under natural conditions from white noise stimulation depends on whether the retina adapts between stimulus statistics.

One approach to investigating adaptation in RGCs is through training predictive models on different stimulus statistics and comparing the resulting model parameters. Linear-nonlinear (LN) models, which have been used in most of the aforementioned adaptation studies are a well established model for linearly characterizing RGC responses (Marmarelis and Naka, 1972; Korenberg and Hunter, 1986; Sakai, Ken-Ichi, and Korenberg, 1988; Chichilnisky, 2001; Pillow et al., 2008). However, they cannot capture non-linear computations such as contrast adaptation, motion anticipation and omitted stimulus response (Schwartz, Harris, et al., 2007; Schwartz, Okawa, et al., 2012; Kastner and Baccus, 2014; Karamanlis and Gollisch, 2021; Maheswaranathan et al., 2023). These limitations have been addressed through augmenting such models with non-linear, yet interpretable features that capture responses more accurately (Baccus, Ölveczky, et al., 2008; McFarland, Cui, and Butts, 2013; Liu et al., 2017; Shi et al., 2019; Sridhar et al., 2024).

More recently, deep learning models, particularly convolutional neural networks (CNNs), have emerged as a tool for predicting RGC responses, outperforming LN models (Batty et al., 2016; McIntosh et al., 2017). Besides improving prediction accuracy, CNNs also facilitated new insights into retinal cell types (Hoefling et al., 2022), and have helped develop mechanistic theories of how non-linear computations in the retina are implemented (Tanaka et al., 2019; Maheswaranathan et al., 2023). While LN models have been challenged for their often poor response prediction to natural stimuli (Heitman et al., 2016; McIntosh et al., 2017), CNNs can predict responses to natural stimuli well (Batty et al., 2016; McIntosh et al., 2017; Hoefling et al., 2022).

Whether models trained on one stimulus capture aspects of processing on the other stimulus has been the subject of recent studies. In V1, it has been shown that models trained on naturalistic movie stimuli generalize well to white noise (Sinz et al., 2018) but not the other way around (David, Vinje, and Gallant, 2004; Sinz et al., 2018). In the context of the retina, poor generalization from white-noise to naturalistic movies has been reported (Heitman et al., 2016; McIntosh et al., 2017), which has been attributed to white noise lacking many important statistical features of natural scenes and therefore not revealing important nonlinear computations (Maheswaranathan et al., 2023). However, a systematic study of generalization abilities across stimuli and model types in the retina is currently missing.

In this paper, we study combinations of models and stimuli of various complexity to establish how suitable they are for modeling RGC responses and to quantify the extent to which the retina adapts to stimulus statistics by assessing how well models generalize. Specifically, we first train and evaluate LN and CNN models across white noise and naturalistic movie stimuli. We show that LN models can perform reasonably well on both noise and movie stimuli, even without the ability to capture non-linear characteristics and that non-linear multi-layer CNNs show better predictive capabilities across species and stimuli. Secondly, we investigate whether training on one of the stimulus types is sufficient for the model to generalize to the other and, if so, at which model complexity. We find that the retina adapts to the stimulus statistics, as in neither case do the models reach in-domain performance: the model trained and evaluated on the same stimulus type always outperforms the one trained on one stimulus type and evaluated on the other. Interestingly, we find that generalization works better from models trained on white noise to naturalistic movies than the other way around.

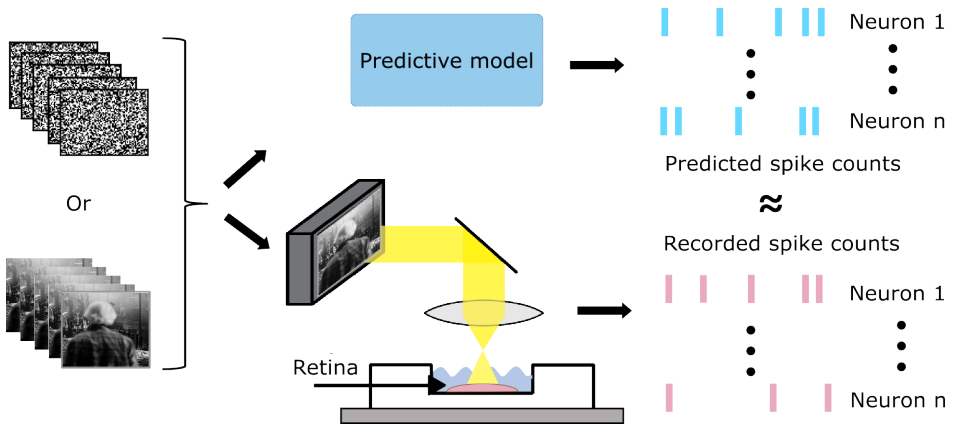

Figure 1. We present white noise (WN) or naturalistic movie (NM) stimuli to a retina while recording ganglion cell responses using a multielectrode array. We train a predictive model (LN and CNN models) to predict the average spike count of all neurons given the stimulus in a time window preceding the prediction.

To shed light on how the adaptation is realized, we compare non-linear processing across stimulus statistics by synthesizing optimal stimuli for the CNN models. We find systematic differences across different cell types, showing that the center size of the optimal stimulus increases and the surround becomes more pronounced when modeling responses to naturalistic movies compared to white noise. Both of these changes are a sign of retinal adaptation to stimulus statistics, with the more pronounced surround for natural scenes following the prediction of the efficient coding hypothesis, which postulates that retinal processing should remove the correlations in natural stimuli to reduce redundant information before transmitting it to the brain through the bottleneck of the optic nerve (Barlow, 1961; Linsker, 1988; Atick and Redlich, 1992).

Results

We analyzed the activity of marmoset (*Callithrix Jacchus*) and salamander (*Ambystoma Mexicanum*) retinal ganglion cells (RGCs) in response to white noise (WN) and naturalistic movie (NM) visual stimuli using microelectrode arrays (Fig. 1). We will refer to these stimuli simply as “noise” and “movies”. Our datasets contained the activity of 327 RGCs across three marmoset retinas which were shown both movie and noise stimuli, and the activity of 236 RGCs across five salamander retinas which were shown only noise stimuli. We selected these cells based on the reliability of their responses and included only those cells in the analysis for which at least 15% of the total response variance could be attributed to the stimulus. Further details on the methods can be found in the Methods section.

We split our stimuli into non-repeating, unique sequences used for training models (“training set”) and a smaller number of repeated segments for evaluating the final model performance (“test set”). A sequence of a non-repeating segment followed by a repeated segment comprised a “trial”. In our data we showed 10 to 248 stimulus trials to the retina. The non-repeating segment had 50–300 seconds, the repeating segment 10–60 seconds depending on the stimulus type and species.

We trained Linear-Nonlinear (LN) and Convolutional Neural Network (CNN) models to predict the neural responses in these datasets. To evaluate our models’ performance, we computed the correlation coefficient between the predicted spike counts and trial-averaged neuronal response on the test segments (Eq. 3).

LN models can predict responses to white noise as well as natural movie stimuli

LN models are a standard way of modeling RGC responses to white noise. We start by establishing their predictive performance not only on white noise but also on natural movie stimuli to find out, whether they are able to predict RGC responses to movie stimuli and to serve as a baseline for

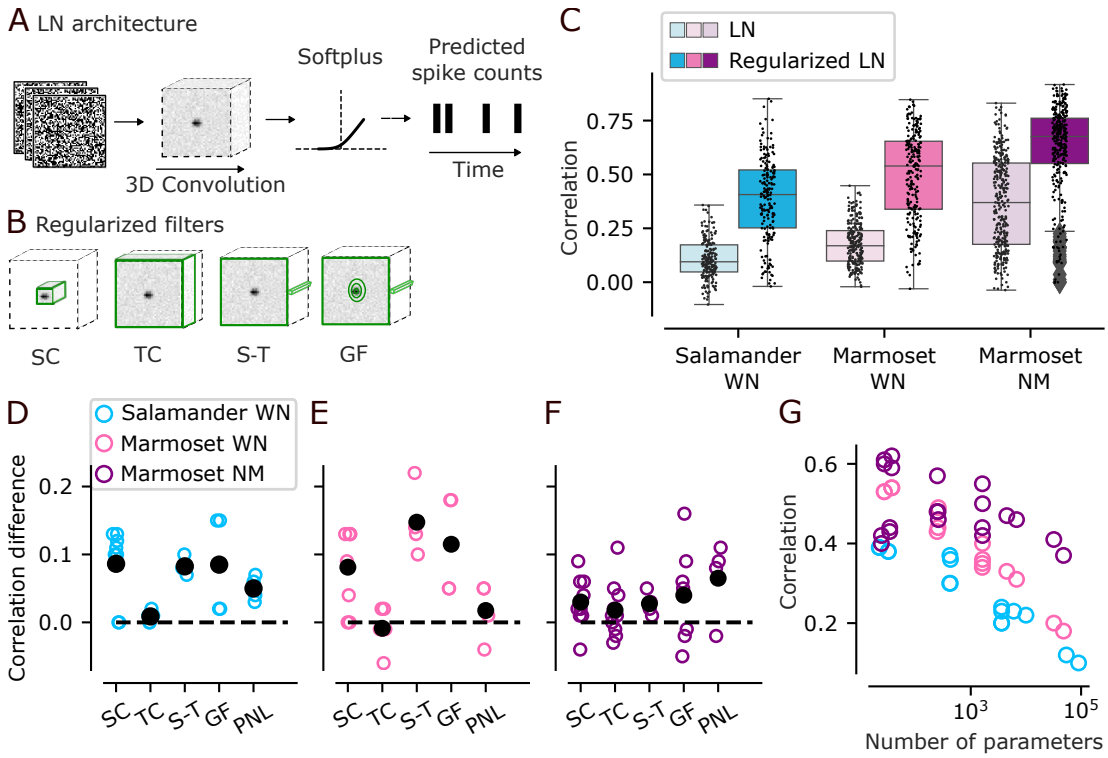

Figure 2. A. Basic LN model architecture without any of the regularization methods applied. **B.** Green outlines show how the LN model kernels change based on the regularization methods – *spatial-crop (SC) temporal-crop (TC)*, *space-time separation (S-T)*, *Gaussian fit (GF)*. **C.** Performance comparison between non-regularized LN and one with the best combination of regularizers (inductive biases). **D.–F.** Each coloured circle represents the difference that adding the inductive bias on the x-axis makes in performance when keeping all other constant. The black dots are the mean improvement for a given inductive bias over all the scenarios. **G.** Each circle is a single model's performance vs. the number of parameters it has.

more complex non-linear models in our analyses. The LN models learned a linear spatio-temporal filter capturing the neuron's linear receptive field. To obtain the neurons' spiking responses, we took the inner product between the stimulus and this linear filter and applied a standard softplus nonlinearity (Fig. 2A).

A simple way of obtaining the linear filters for white noise stimuli is spike-triggered averaging (STA), which we used for the noise stimuli models. For natural movie stimuli, STA does not recover the correct filter because of the correlations within the stimuli. We therefore fit the LN model directly using maximum likelihood estimation by stochastic gradient descent on naturalistic movie stimuli.

This basic LN model architecture resulted in 0.09 median (IQR 0.05–0.17) correlation (Eq. 3) over cells on salamander noise data, 0.17 (IQR 0.1–0.24) of median correlation over cells on marmoset noise data and 0.37 of median (IQR 0.17–0.55) correlation over cells on marmoset movie data. The marmoset movie models performed notably better, presumably because a larger fraction of the response variance to natural movies is caused by local brightness fluctuations, which are captured well by a linear model.

We suspected the poor performance of these basic LN architectures to be caused by overfitting, as they have too many parameters for the given amount of recorded data. To prevent overfitting, we turned to regularization – reducing the number of parameters or adding smoothing priors restricting the values of the parameters. We used various regularization techniques on the LN models to improve performance. All of them can be viewed as inductive biases that we impose on our model based on known properties of RFs. We examined how strongly the different regu-

larization methods affect the performance, namely: (1) *spatial-crop (SC)*: cropping of the model's spatial weights to the neurons STA-based receptive fields, (2) *temporal-crop (TC)*: cropping temporal weights to the STA-based RF, (3) *space-time separation (S-T)*: decomposing the spatio-temporal filter into separate spatial and temporal filters, and (4) *Gaussian fit (GF)*: fitting a two-dimensional Gaussian to the spatial component of the space-time separated filter of the models (Fig. 2B). For noise stimuli, the Gaussian fit was done on the STA, for movies we fit the parameters of a Gaussian using gradient descent from scratch. We added an additional inductive bias on top of the regularization techniques. We (5) modified the final default softplus non-linearity by adding learnable parameters to it: *parameterized non-linearity (PNL)* (Eq. 1), allowing the model to learn a more nuanced relationship between linear filter output and spiking response.

On all datasets, we drastically improved performance by adding combinations of the inductive biases (1)–(5) to the models (Fig. 2C). The best resulting model on salamander noise data incorporated all inductive biases and reached a median correlation of 0.41 (IQR 0.25–0.52) which is 0.32 points better than the unregularized model. For the marmoset data, models incorporating all inductive biases except temporal cropping performed best. While temporal cropping did not improve performance when all other biases were employed, it also did not reduce it significantly. The best model for the noise stimulus improved from the baseline median correlation of 0.17 to 0.54 (IQR 0.34–0.65). For the movie stimulus, the best model strongly improved from a baseline median correlation of 0.37 to 0.68 median (IQR 0.55–0.76). Thus, the LN models are decent predictive models for both noise and movie stimuli when properly regularized. The complete tables of results for the given species and stimulus combinations can be found in the Appendix in Tab. 4–Tab. 6.

We also examined how each single bias (1)–(5) improved the performance of LN models. For this, we calculated the difference between the performance of two models, one with and one without applying a given bias while all other biases were the same between the two models. We found that all biases improved performance across species and stimulus-types at least in certain combinations (Fig. 2D–F). For salamander noise, spatial cropping, space-time separation and Gaussian fitting are all effective regularization methods. Marmoset noise models mostly benefit from space-time separation and also Gaussian fits and spacial cropping. Marmoset movie models mostly benefit from a parameterized non-linearity.

The effect of specific methods varies depending on what other inductive biases methods are already present. For all species and stimulus types, reducing the overall number of parameters is the most important factor, whereas the specific method of parameter reduction not as critical (Fig. 2G).

Non-linear CNNs improve predictive performance

Based on the results in the previous sections, we saw that reducing the number of parameters boosts performance. While we explored a range of regularization techniques on the level of single cell models, another possibility is to use a different model design which allows us to model a population of neurons simultaneously. In such a case a portion of the parameters does not have to be learned for each neuron individually, but can be shared by all neurons. To achieve this, we used CNN models with a *core-readout* architecture (Fig. 3A). Such models have been used successfully in V1 as well as for RGCs cell prediction (McIntosh et al., 2017; Klindt et al., 2017; Lurz et al., 2021; Hoefling et al., 2022; Maheswaranathan et al., 2023). They consist of one or more layers of convolutions – the *core* – which creates a feature space shared by all neurons. The *readout* then allows each neuron to learn its receptive field position in this feature space and weight the shared features at this position (Fig. 3A).

We first built an LN model in this fashion which we refer to as a population LN model. It has only one space-time separated convolutional layer in the core directly followed by the readout. The only non-linearity is at the very end after the readout. Since there are no non-linearities between the spatial and temporal convolutions and between core and readout, this model is effectively an LN model as well, with the restriction that the linear filter is constructed as a linear combination of a

basis set shared by all the modeled cells. We trained this model end-to-end for all three of the data types and found that in all three cases, a population LN outperformed the best regularized single cell LN model, improving from median correlation of 0.39 (IQR 0.23–0.54) to 0.45 (IQR 0.31–0.58) on salamander noise, from 0.52 (IQR 0.32–0.64) to 0.56 (IQR 37–68) on marmoset noise and from 0.67 (IQR 0.55–0.76) to 0.72 (IQR 0.62–0.8) on marmoset movies (Fig. 3B–D).

While the predictive performance of the LN and population LN models is good, there remains a substantial gap to predicting the RGCs responses perfectly. So far we tried to improve performance by regularizing linear models. However, numerous studies show that the retina performs non-linear computations (Baccus, Ölveczky, et al., 2008; Gollisch and Meister, 2010; Kastner and Baccus, 2014; Kuo, Schwartz, and Rieke, 2016; Turner and Rieke, 2016; Karamanlis and Gollisch, 2021). To push performance further, we therefore moved beyond LN models and implemented models with multiple non-linearities allowing them to learn these non-linear computations. The core-readout CNN model structure allowed us to easily incorporate the non-linearities, placing them after each space-time convolutional layer within the core and between the core and the readout. We implemented networks of various depths to establish the complexity that yields the highest performance and establish benchmarks at multiple complexity levels. Our CNNs ranged from one to four layers.

The complexity of the CNN architecture, particularly the number of layers and subsequent non-linearities, directly impacted performance (Fig. 3B–D). Models with a single non-linearity after the core already demonstrated a significant improvement, indicating the value of non-linear processing in these predictive tasks and thus in the retinal processing. As the number of layers – and therefore also, the number of non-linear transformations – increased, we observed a corresponding increase in performance. For salamander data, the performance kept increasing up to four layers to 0.71 median (IQR 0.55–0.8) correlation across cells. For marmoset data, performance saturated at three layers at 0.83 median (IQR 0.7–0.86) correlation across cells for noise and 0.83 median (IQR 0.77–0.86) correlation across cells for movies. This trend suggests that the depth of the network, which relates to its ability to form more complex representations and feature hierarchies, is a critical factor in achieving accurate predictions of ganglion cell responses. Notably, the performance gain between linear and non-linear models on naturalistic movies is smaller than on white noise.

Retinal ganglion cells adapt their RFs to stimulus statistics

We have established that LN models and more so CNNs are able to predict RGCs responses to both noise and movies with high accuracy mimicking retinal computations under the given stimulus statistics. We next asked, whether these computations are adapted specifically to the stimulus statistics. To do so, we tested how well do models trained on one stimulus predict the activity of RGCs in response to the other. Should the models maintain their predictive performance across both stimuli, it would imply that the retinal computations are universal, without specific adaptation to the type of stimulus. If prediction accuracy is maintained in only one direction, it would suggest that the given training stimulus encompasses a more universal set of features, capable of capturing aspects of both stimulus types. If model generalization is unsuccessful in both directions, it would indicate a tailored adaptation of the RFs to the unique statistics of each stimulus.

For this analysis, we used the marmoset data where have recorded the same neurons under both stimulus conditions and trained all models included in this analysis on the same amount of noise or movie data. We took models trained on one of the stimuli and evaluated them on the other in two scenarios. The first scenario we evaluated was unadapted prediction on the other stimulus, meaning we evaluated the model directly on the stimulus it has not seen during training (Fig. 4). Under these conditions, neither stimulus ensemble allowed models to reach in-domain performance on the other stimulus (Fig. 4A,B; white vs. colored bars). This was true across all models, but the gap was larger for the CNN models than the LN models. These results suggest that the processing of the neurons adapts to the stimulus statistics, and the adaptation is strongest for

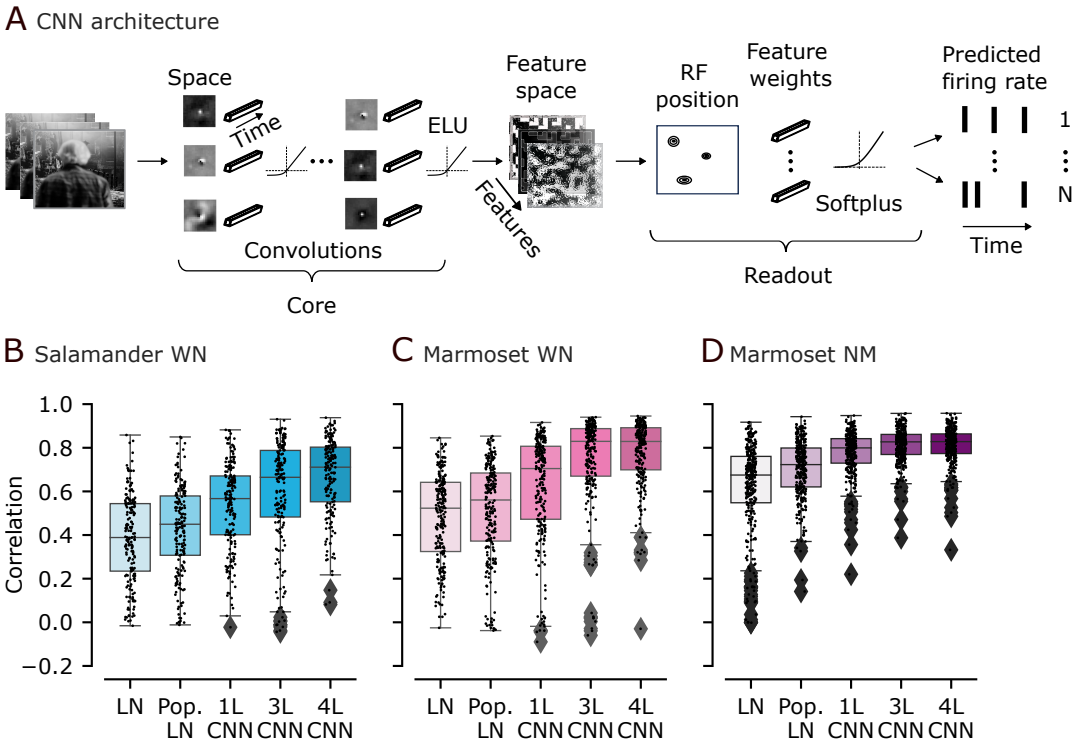

Figure 3. A. The core-readout architecture of the population LN and CNN models. N equals the number of neurons. **B.-D.** The IQ1-IQ3 range (boxplots) with median (black line) of model types of increasing complexity on **B.** salamander cells being shown noise stimuli **C.** marmoset cells being shown noise stimuli **D.** marmoset cells being shown movie stimuli. Small black dots represent cells.

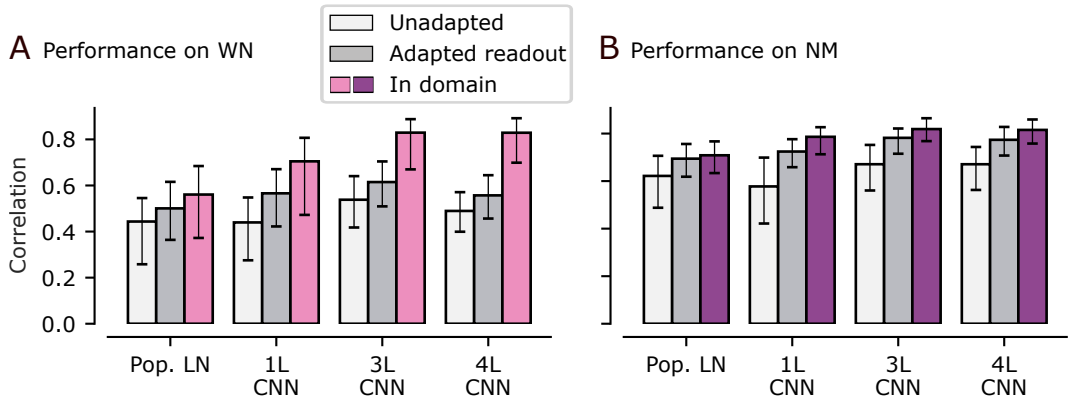

Figure 4. A. Models evaluated on white-noise. White: Undadapted performance – i.e. models trained on naturalistic movies applied directly to white-noise data. Gray: performance with adapted readout: model trained on naturalistic move, readout and parameterized final non-linearity adapted by training on noise data. Pink: In domain: Model trained on white-noise evaluated on white-noise. **B.** Models evaluated on naturalistic movies. White: Unadapted performance - i.e. model trained on white-noise applied directly to naturalistic movie. Gray: performance with adapted readout: model trained on white-noise, readout and parameterized final non-linearity adapted by training on naturalistic movie data. Purple: In domain: Model trained on naturalistic movie evaluated on naturalistic movies.

the non-linear response properties, as they do not generalize across stimulus statistics.

Next we asked whether the features learned by the shared core generalized across stimulus statistics, and only the way they are combined changes. To test this idea, we allowed the models to adapt their linear readout. To do so, we kept the core fixed and fine-tuned the weights of the readout and the output non-linearity on the other stimulus (Fig. 4). Adapting the readout improved

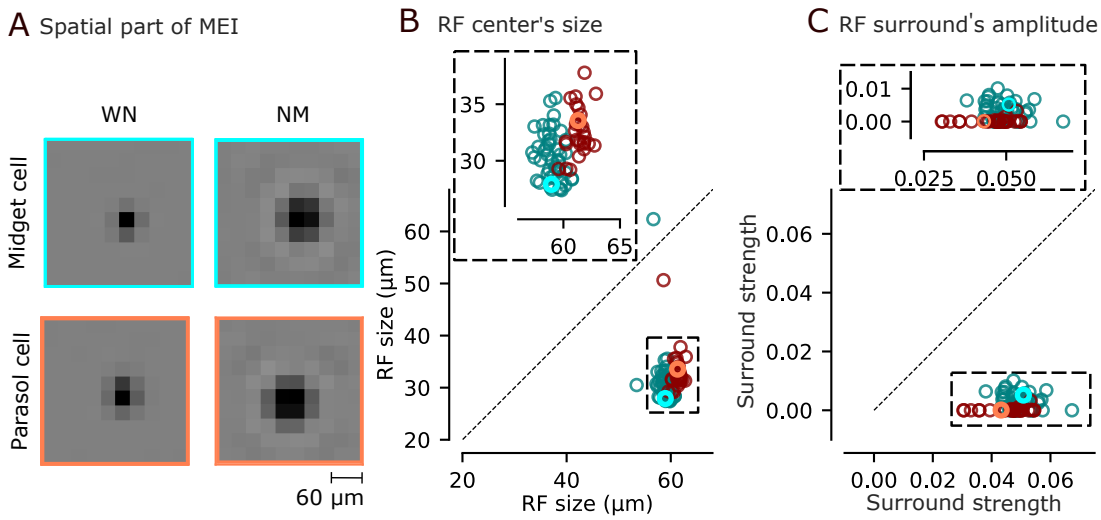

Figure 5. A. Comparison of the spatial parts of RFs of two example cells, one classified as a midget cell and the other as parasol cell on white noise and naturalistic stimulus. **B.** Quantitative comparison of RF center sizes for noise vs. movies. We can see, that on movies, the RF centers are systematically larger for both midget cells (turquoise) and parasol cells (dark red). **C.** Surround strengths measured as the maximal (absolute minimal) value of the DoG paramtetrized RFs with a negavite (positive) center Gaussian on noise vs. movies. We can see that on noise, the surround is basically non-existent whereas on movies, all cells have developed a surround.

generalization from movies to noise somewhat, but still a fairly large gap remained (Fig. 4A). In contrast, the model trained on noise generalized fairly well to movies when allowing the readout to adapt (Fig. 4B), almost closing the gap for the LN model. This latter result suggests that while the linear filters change across stimulus statistics, the basis set learned from white noise spans the space required to predict movie data.

We now ask *how* the computations of the RCGs change with stimulus statistics. For this analysis, we used a subset of the cells which we could classify as OFF midget and OFF parasol cells. We synthesized most exciting inputs (MEIs) using the three-layer CNNs trained on noise and movies, respectively (Fig. 5). MEIs characterize the receptive field of non-linear neurons by revealing the stimulus to which they are most sensitive (Walker et al., 2019; Hoefling et al., 2022). MEIs were approximately space-time separable, and we focused on their spatial component (Fig. 5A). The spatial component of the MEI changed in two ways: First, the receptive field centers were systematically larger for movies than for noise (Fig. 5B). Second, the surround was much more pronounced under the movie stimulus than for noise (Fig. 5C). Parasol cells exhibited no surround at all under noise stimulation. Midget cells did show a weak surround, but it was approximately one order of magnitude smaller than for movies.

Discussion

In this paper we showed that a range of carefully optimized models ranging from LN to CNN models give rise to different response functions depending on which stimulus ensemble – white noise or natural scenes – they are trained on.

A number of previous studies have argued for using natural scenes to study visual processing, because models trained on white noise stimuli do not generalize well to natural movie inputs (McIntosh et al., 2017; Sinz et al., 2018; Heitman et al., 2016) and do not predict crucial non-linear features of retinal processing, such as contrast adaptation, motion extrapolation and omitted stimulus response (Maheswaranathan et al., 2023). Our results differ from and add to these earlier results in several ways:

First, also in our hands, models trained on noise do not perform as well on movies as models

trained directly on movies. However, unlike in the work by Heitman et al. (2016) the gap is fairly small for LN models. In contrast, for the more non-linear CNNs the gap widens, suggesting that whatever non-linear processing the CNNs capture is not shared between noise and movie stimulus context.

Second, the converse effect is also true: Training only on movies does not allow to accurately predict how RGCs respond to noise. This finding is in contrast to work in primary visual cortex, where the generalization from movies to noise was substantially better than from noise to movies (Sinz et al., 2018). However, we note that one should be careful with interpreting such numbers. Overall, responses to movies were better predicted in our data than responses to noise, presumably because movies contain stronger local luminance fluctuations, which are "easy" to predict and account for a larger fraction of the total variance than for noise stimuli. Thus, absolute differences in predictive performance may not be very meaningful.

Generally, metrics based on explained variance are somewhat difficult to interpret. While they are 1 if and only if one has a perfect model, differences in explained variance do not necessarily tell us how interesting this explained variance is. For instance, non-linear features of retinal processing, such as contrast adaptation, motion extrapolation and omitted stimulus response (Schwartz, Harris, et al., 2007; Schwartz, Okawa, et al., 2012; Kastner and Baccus, 2014; Maheswaranathan et al., 2023) are clearly important, but since they do not occur frequently in natural scenes, correctly accounting for them will provide only small improvements in terms of explained variance. Despite this, quantitative metrics are still important: Maheswaranathan et al. (2023) showed that models trained on white noise do not capture several interesting non-linear response properties, while models trained on movies do. It would be tempting to conclude that using movie stimuli is somehow superior. What quantitative evaluation shows us is that generalization does not work in either way. Simply because we may not have identified any interesting properties of neuronal responses that are not predicted from training on movies, this does not mean they do not exist. As of now, it appears that neither stimuli is "superior" – processing simply adapts to the stimulus statistics and we do not yet have a constructive, low-parametric or even parameter-free model of this adaptation.

Our results point towards the possibility that the efficient coding hypothesis may at least in part account for this adaptation. The lack of a surround in the receptive fields during stimulation with white noise is consistent with there not being a necessity for whitening, because the stimulus is already white. During stimulation with natural movies, in contrast, we observed a more pronounced surround, consistent with the whitening filter changing based on the input statistics.

The interpretation of the increased center size is less clear. While the natural movie stimuli are around 25% darker on average than the white noise, this decrease in luminance is unlikely to fully account for the lower cutoff frequency of the lowpass component of the filter predicted by Atick and Redlich (1992).

The larger center size could be caused by other factors, including the inductive bias of the models favoring lower-frequency solutions given that the natural movies have little high-frequency content in comparison with white noise. One inductive bias of the CNN-based models is that their filters are often regularized to be smooth by penalizing nearby filter value differences with a Laplace filter. However, turning off this regularization did not change the results, suggesting that it is not the cause for the larger filters under natural movie stimulation. There could be implicit biases imposed by factors such as model architecture (convolution) or learning algorithm (stochastic gradient descent) that may favor convergence to lower-frequency solutions in a non-convex optimization problem. However, we also observed the increased size of the center in single-cell LN models, which represent a convex optimization problem, suggesting that implicit inductive biases of complex CNN-based models are unlikely to be a sufficient explanation.

One potentially relevant aspect of the dataset is that the natural movies were jittered from frame to frame to simulate fixational eye movements. As this jitter is independent from frame to frame, it may act as an effective lowpass filter on the stimulus if the neurons cannot resolve it

temporally. The standard deviation of this fixational noise was 15 µm. This factor could contribute to further removing high-frequency content from the natural movie stimulus which could account for smoothing the receptive fields.

In conclusion, our results show that retinal processing adapts between the stimulus statistics of natural scenes and white noise. While some of this adaptation is at least qualitatively consistent with efficient coding, it cannot fully account for all results. Making the link between theory and data more quantitative by combining normative modeling with parameter fitting as well as developing new theories that capture a broader range of phenomenology would be an exciting avenue for future research.

Methods

Data

Salamander Data

We used five datasets from adult axolotl salamander retinas, each comprising the responses of retinal ganglion cells to spatiotemporal white-noise stimuli measured using microelectrode arrays (MEAs). The stimuli were structured into trials, with each trial consisting of a 50-second non-repeating training segment followed by a 10-second test segment that was repeated across trials. A total of 39 to 248 trials were recorded per dataset. The acquisition of these five datasets has been previously described in Liu et al., 2017 and they are publicly available (Gollisch and Liu, 2018).

Marmoset data

We also used three datasets recorded using microelectrode arrays from the retinas of adult common marmosets (*Callithrix Jacchus*). Each dataset consisted of responses of retinal ganglion cells under white-noise and naturalistic movie stimulation. Each trial of the white noise stimulus consisted of 150 seconds of non-repeating frames and 30 seconds of repeating frames, whereas each trial of the naturalistic movie stimulus consisted of 300 seconds of repeating frames and 60 seconds of non-repeating frames. The stimuli were shown at a frame rate of 85 Hz. All three datasets contained responses to the same 10 trials of the white-noise stimulus. *Dataset 1* and *Dataset 3* contained 20 trials of the naturalistic movie stimulus, whereas *Dataset 2* contained 10 trials. The aquisition and preprocessing of all datasets was done in the same way, as described for *Dataset 1* in Sridhar et al. (2024).

Cell selection

Only cells with reliable responses to the stimuli were used for the analysis. The reliability of the cells was determined based on the fraction of explainable variance (Heitman et al., 2016; Cadena et al., 2019) in each cell's responses to the repeating segments of the stimuli across trials. All cells that exceeded a threshold of 0.15 were used for further analysis. In the case of the marmoset datasets, we retained all cells which exceeded this threshold for white-noise and naturalistic movie stimuli. This left us with 267 reliable cells from the salamander retinas and 236 reliable cells across *Dataset 1* and *Dataset 2* of the marmoset retinas. All cells in *Dataset 3* were deemed unreliable on the white-noise stimulus. Therefore, we used 91 cells that responded reliably to the naturalistic movie stimulus and used this dataset only for experiments on naturalistic movies.

Cell classification

In order to study specific differences in the characteristics of receptive fields learned from different models, we used the cell-type classification provided along with the marmoset *Dataset 1* ((Sridhar et al., 2024)). In particular, we used the OFF midget and OFF parasol cell clusters. The clusters were obtained by grouping cells based on their receptive field and response characteristics using a semi-automated approach.

LN models

White noise LN model filters

Filters for each cell for the white-noise model were estimated using spike triggered averaging (STA). We calculated the STA as

$$\text{STA} = \frac{1}{N} \sum x_{t-\delta:t} \cdot y_t$$

where x_t is the frame presented at time t, y_t is the spike-count recorded at time-point t and N is the recorded number of spikes. To this baseline filter, we then apply the following inductive biases:

- *Spatial* and *Temporal Cropping* around receptive field – The original STA has size 60x60 pixels spatially and 25 frames temporally for salamander and 40x40 pixels spatially and 30 frames temporally for marmoset. These size are able to capture the complete receptive field both spatially and temporally. However, such large crops mean a lot of parameters need to be fit and lead to overfitting. Therefore, we cropped more tightly around the receptive field to reduce the number of parameters. Specifically, we used crop sizes 20x20 pixels and 15 frames for salamander and 15x15 pixels and 20 frames for marmoset.
- *Space-time separation* – To get a space-time separation we calculate the STA as above. We then computed the singular value decomposition (SVD) and used the first resulting decomposition kernels as the spatial and temporal filters.
- *Gaussian fit* – Another way to reduce the number of parameters and prevent over-fitting is fitting a Gaussian to the spatial filter of the LN model. We estimated the spatial filter as described in space-time separation and then fitted a 2D Gaussian to it. The values that are further away from the mean than two standard deviations were set to zero. The temporal filter remains the same as in the space-time separated version.
- *parameterized non-linearity* – As the default output non-linearity, we used the softplus function. This can be parameterized by adding learnable parameters α, β and γ as follows:

$$softplus(x) = \alpha \cdot \log(1 + \exp(\gamma \cdot x + \beta)) \tag{1}$$

  where x is the input. We optimized these parameters using Poisson loss Eq. 2 between predicted firing rate p and recorded firing rate r using stochastic gradient descent.

Naturalistic movie LN model filters

Natural movie stimulus is not space and time independent. We therefore estimated the whole LN model filter learning it end-to-end, minimizing the Poisson loss (Eq. 2) between recorded firing rate r and the predicted firing rate p using stochastic gradient descent. We applied the same biases as for the white noise model filters with the following differences:

- *Spatial* and *Temporal Cropping* – The original STA had size 40x40 pixels spatially and 30 frames temporally. We cropped tightly around the receptive field. Specifically, we used 15x15 pixels and 20 frames, same as crops for marmoset noise models.
- *Gaussian fit* – We randomly initialized the temporal filter and the mean and covariance matrix of a non-isotropic Gaussian. The Gaussian then parameterized the spatial component of the filter. We trained all the learnable parameters using stochastic gradient descent.

$$L(r, p) = \sum_{c=0}^{N-1} p_c - r_c \cdot \log(p_c + \varepsilon) \tag{2}$$

Training

For all parameters that were optimized with stochastic gradient descent in LN models, we used the following training scheme. We split the trials in our datasets into 80% train and 20% validation splits. We randomly initialized the learnable parameters, and updated them by minimizing the Poisson loss (Eq. 2) between predicted p and recorded r response values. We trained on the training trials

Salamander on WN	Spat. input kernel	Temp. input kernel	Spat. hidden kernel	Temp. hidden kernel	Spat. dilations	Channels
Population LN	17	25	-	-	1	16
CNN 1	17	25	-	-	1	64
CNN 3	11	25	5	5	1, 2, 2	16, 32, 64
CNN 4	11	25	9	5	1, 1, 1, 1	8, 16, 32, 64

Table 1. Best model hyperparameters for models trained on salamander RGC responses to noise

Marmoset on WN	Spat. input kernel	Temp. input kernel	Spat. hidden kernel	Temp. hidden kernel	Spat. dilations	Channels
Population LN	27	35	-	-	1	32
CNN 1	27	35	-	-	1	32
CNN 3	21	35	3	5	1, 1, 1	16, 32, 64
CNN 4	25	25	3	5	1, 1, 1, 1	8, 16, 32, 64

Table 2. Best model hyperparameters for models trained on marmoset RGC responses to noise

for up to 1000 epochs, with early stopping when validation correlation (Eq. 3) has not increased for 30 epochs. The initial learning rate was 0.003 and we employed the Pytorch `ReduceLROnPlateau` learning rate scheduler with a patience of 15 on the validation correlation with minimal learning rate of 0.000001. The optimizer we used was `Adam` (Kingma and Ba, 2017).

Evaluation

We report the final performances of the best models based on validation correlation by calculating the correlation coefficient for a given RGC c (Eq. 3) between the predictions p_c and the trial averaged firing rates $\langle r_c \rangle$ on the held-out test sequence with length T.

$$CC_c = \frac{\sum_{t=0}^{T}(\langle r_c^{(t)} \rangle - \langle \bar{r} \rangle)(p_c^{(t)} - \bar{p})}{\sqrt{\sum_{t=0}^{T-1}(\langle r_c^{(t)} \rangle - \langle \bar{r} \rangle)^2}\sqrt{\sum_{t=0}^{T-1}(p_c^{(t)} - \bar{p})^2}} \tag{3}$$

CNN models

Architecture

We trained multiple models of various depths. All of them had a similar core + readout structure as in Hoefling et al. (2022). The core consisted of a varying numbers of layers of space-time separated convolutions. After each spatio-temporal convolution we applied an Exponential Linear Unit (ELU) non-linearity (Clevert, Unterthiner, and Hochreiter, 2015) and a batch normalization (Ioffe and Szegedy, 2015). We used different optimal hyper-parameters for each of the CNN depth, species and stimulus type combination. The best hyperparameter values are in Tab. 1 – Tab. 3. The kernels of the first layer are smoothed adding a 2D (3x3 pixels) Laplace filter in the spatial dimensions and a (1x3 pixels) 1D Laplace in the temporal dimension. In contrast to (Hoefling et al., 2022) temporal kernels were not parameterized as Fourier series but optimized directly. In the readout, each cell's RF was modeled as an isotropic Gaussian from which we sample at training time and take the mean at inference time. The response function was modeled as an affine function of the core's weighted feature maps at the RF positions, followed by a parameterized softplus (Eq. 1). The feature map weight vector was regularized using L1-norm .

The architectural feature which seemed most significant with respect to good model performance was making the first layer spatial and temporal kernels larger relative to those in subsequent layers. Presumably because the large kernel allows the models to capture the whole RF, including the surround. Regularization on the other hand did not have a large effect on model performance.

Marmoset on NM	Spat. input kernel	Temp. input kernel	Spat. hidden kernel	Temp. hidden kernel	Spat. dilations	Channels
Population LN	27	35	-	-	1	32
CNN 1	25	35	-	-	1	32
CNN 3	21	35	5	3	1, 1, 1	16, 32, 64
CNN 4	17	27	3	5	1, 1, 2, 3	8, 16, 32, 64

Table 3. Best model hyperparameters for models trained on marmoset RGC responses to movies

Training

We split the trials in the datasets in the same way as for the LN models into 80% train and 20% validation splits. We trained the CNNs minimizing Poisson loss (Eq. 2) on the training trials for up to 1000 epochs, with early stopping when correlation on the validation trials has not increased for 50 steps. The initial learning rate was 0.008 for natural movies and 0.006 for white noise and we employed the `ReduceLROnPlateau` learning rate scheduler from Pytorch with a patience of 15 on the validation correlation and minimal learning rate of 0.000001. The optimizer was `Adam` (Kingma and Ba, 2017).

Evaluation

For each complexity level i.e. number of layers, we selected the best model based on the estimated validation correlation and report its final performances by calculating the correlation coefficient (Eq. 3) between the model predictions p and the trial averaged responses $\langle r \rangle$ of the held-out test sequence of the dataset.

Cross-stimulus evaluation

To evaluate cross stimulus generalization, we used the marmoset datasets *Dataset 1* and *Dataset 2*, which contain responses to both white noise and natural movie stimuli of the same cells. We used only 5 trials of natural movie data from both of the datasets when training models for this analysis to match the white noise dataset size. In the unadapted setting models, trained on one stimulus type, were directly applied to the other to assess correlation between predicted and recorded spike counts. In the adapted readout setting, the readout weights and final non-linearity parameters of models trained on one stimulus were adjusted by training on the other stimulus, following the same procedure as initial training.

MEI optimization

We optimized spatio-temporal MEIs based on (Walker et al., 2019) using the trained CNN models with frozen weights. For each modelled RGC c we synthesized an input x_c such that $x_c = argmax_x(r_c)$ where r_c is the predicted activation for cell c. Starting from a random Gaussian noise with mean 0 and standard deviation 0.1, we updated the input x using stochastic gradient descent. Throughout the optimization, x was kept under a total L2 norm budget of 5 and the single element values within a [-1, 1] range not to exceed the range of the original stimulus values.

RF size estimation

To get the center size and surround amplitude of the RF of a given cell, we took its MEI and space-time separated it using singular value decomposition. We then took the spatial part of the MEI and fitted a difference of Gaussians (DoG) using `scipy.optimize` with mean squared error loss function and an initial guess of $A_c = 1, \sigma_{cx}^2 = 1, \sigma_{cy}^2 = 1, A_s = 0.5, \sigma_{sx}^2 = 2, \sigma_{sy}^2 = 2$ where $A_c(A_s)$ is the center (surround) amplitude and $\sigma_{cx}^2(\sigma_{sx}^2)$ and $\sigma_{cy}^2(\sigma_{sx}^2)$ the variances of the center (surround) Gaussians in pixel space. We applied a positivity constraint on the variances of both Gaussians and enforced

that both have the same amplitude sign. We established the *center size* S as $S = \sqrt{\pi * \sigma_{cx} * \sigma_{cy}}$. The *surround strength* was taken from the DoG-parameterized spatial component of the MEI as the lowest negative value if center amplitude was positive and the highest positive value if the center amplitude was negative.

Acknowledgements

MFB thanks the International Max Planck Research School for Intelligent Systems (IMPRS-IS).

Appendix

LN model version					Per cell avg	Avg	R1	R2	R3	R4	R5
SC	**TC**	**S-T**	**PNL**	**GF**							
X	X	X	X	X	0.1	0.09	0.17	0.08	0.07	0.06	0.08
✓	X	X	X	X	0.22	0.21	0.3	0.2	0.19	0.18	0.18
X	✓	X	X	X	0.12	0.11	0.2	0.09	0.09	0.07	0.08
✓	✓	X	X	X	0.23	0.23	0.32	0.22	0.21	0.19	0.19
X	X	✓	X	X	0.2	0.19	0.32	0.19	0.16	0.14	0.15
X	X	✓	✓	X	0.23	0.21	0.36	0.2	0.16	0.12	0.19
✓	X	✓	X	X	0.3	0.29	0.4	0.3	0.26	0.25	0.25
X	✓	✓	X	X	0.2	0.19	0.32	0.18	0.17	0.15	0.14
✓	X	✓	✓	X	0.36	0.35	0.47	0.36	0.29	0.25	0.3
X	✓	✓	✓	X	0.24	0.21	0.37	0.21	0.16	0.14	0.17
✓	✓	✓	X	X	0.3	0.29	0.4	0.3	0.28	0.25	0.25
✓	✓	✓	✓	X	0.37	0.34	0.48	0.36	0.3	0.26	0.27
X	X	✓	✓	✓	0.38	0.36	0.48	0.37	0.32	0.3	0.34
✓	X	✓	✓	✓	0.38	0.36	0.48	0.36	0.32	0.3	0.33
X	✓	✓	✓	✓	**0.39**	0.37	0.49	0.37	0.34	0.31	0.33
✓	✓	✓	✓	✓	**0.39**	0.36	0.48	0.37	0.34	0.31	0.32

Table 4. LN model for white noise salamander. Mean LN model performance on white-noise salamander data with different combinations of inductive biases. SC stands for spatial-cropping, TC for temporal crop, S-T for space-time separation, PNL for parameterized non-linearity and GF for Gaussian fit. X means inductive bias was not applied, ✓means it was applied. Columns R1 – R5 refer to the five different datasets. Per cell avg is the average over all cells in all five datasets. Avg is the average over datasets.

LN model version					Per cell avg	Avg	Dataset 1	Dataset 2
SC	**TC**	**S-T**	**NL**	**GF**				
X	X	X	X	X	0.18	0.17	0.19	0.15
✓	X	X	X	X	0.31	0.3	0.33	0.27
X	✓	X	X	X	0.2	0.19	0.21	0.17
✓	✓	X	X	X	0.33	0.32	0.34	0.29
X	X	✓	X	X	0.4	0.33	0.37	0.28
X	X	✓	✓	X	0.36	0.33	0.4	0.27
✓	X	✓	X	X	0.44	0.37	0.46	0.28
X	✓	✓	X	X	0.34	0.32	0.36	0.28
✓	X	✓	✓	X	0.49	0.47	0.54	0.39
X	✓	✓	✓	X	0.35	0.33	0.39	0.27
✓	✓	✓	X	X	0.43	0.42	0.45	0.38
✓	✓	✓	✓	X	0.48	0.46	0.52	0.39
X	X	✓	✓	✓	**0.54**	0.51	0.59	0.43
✓	X	✓	✓	✓	**0.54**	0.51	0.59	0.43
X	✓	✓	✓	✓	0.53	0.5	0.58	0.42
✓	✓	✓	✓	✓	0.53	0.5	0.58	0.42

Table 5. LN model for white noise marmoset. Mean LN model performance on white-noise marmoset data with different combinations of inductive biases. SC stands for spatial-cropping, TC for temporal crop, S-T for space-time separation, PNL for parameterized non-linearity and GF for gaussian fit. X means inductive bias was not applied, ✓means it was applied. Per cell avg is the average over all cells in the two datasets. Avg is the average over the datasets.

LN model version					Per cell avg	Dataset 1	Dataset 2	Dataset 3
SC	**TC**	**S-T**	**PNL**	**GF**				
X	X	X	X	X	0.37	0.41	0.26	0.38
✓	X	X	X	X	0.46	0.46	0.38	0.53
X	✓	X	X	X	0.41	0.46	0.29	0.43
✓	✓	X	X	X	0.47	0.44	0.51	0.5
X	X	✓	X	X	0.42	0.53	0.17	0.39
✓	X	✓	X	X	0.48	0.59	0.25	0.45
X	✓	✓	X	X	0.44	0.57	0.16	0.4
✓	✓	✓	X	X	0.48	0.62	0.24	0.41
X	X	✓	✓	X	0.5	0.53	0.3	0.61
✓	X	✓	✓	X	0.46	0.41	0.4	0.59
X	✓	✓	✓	X	0.55	0.64	0.32	0.54
✓	✓	✓	✓	X	0.57	0.65	0.42	0.54
X	X	✓	X	✓	0.43	0.41	0.34	0.56
✓	X	✓	X	✓	0.44	0.41	0.37	0.56
X	✓	✓	X	✓	0.4	0.38	0.3	0.52
✓	✓	✓	X	✓	0.42	0.39	0.36	0.53
X	X	✓	✓	✓	0.59	0.59	0.54	0.7
✓	X	✓	✓	✓	**0.62**	0.62	0.5	0.72
X	✓	✓	✓	✓	0.6	0.6	0.5	0.68
✓	✓	✓	✓	✓	0.61	0.6	0.51	0.69

Table 6. LN model for natural mvoies marmoset. Mean LN model performance on natural movies marmoset data with different combinations of inductive biases. SC stands for spatial-cropping, TC for temporal crop, S-T for space-time separation, PNL for parameterized non-linearity and GF for Gaussian fit. X means inductive bias was not applied, ✓means it was applied. Per cell avg is the average over all cells in all three datasets. Avg is the average over the datasets.

References

Atick, Joseph J and A Norman Redlich (1992). "What does the retina know about natural scenes?" In: *Neural computation* 4.2, pp. 196–210.

Baccus, Stephen A. and Markus Meister (2002). "Fast and Slow Contrast Adaptation in Retinal Circuitry". In: *Neuron* 36.5, pp. 909–919. ISSN: 0896-6273. DOI: https://doi.org/10.1016/S0896-6273(02)01050-4. URL: https://www.sciencedirect.com/science/article/pii/S0896627302010504.

Baccus, Stephen A., Bence P. Ölveczky, Mihai Manu, and Markus Meister (2008). "A Retinal Circuit That Computes Object Motion". In: *The Journal of Neuroscience* 28, pp. 6807–6817. URL: https://api.semanticscholar.org/CorpusID:3470745.

Barlow, Horace (1961). "Possible Principles Underlying the Transformations of Sensory Messages". In: *Sensory Communication* 1. DOI: 10.7551/mitpress/9780262518420.003.0013.

Batty, Eleanor, Josh Merel, Nora Brackbill, Alexander Heitman, Alexander Sher, Alan M. Litke, E. J. Chichilnisky, and Liam Paninski (2016). "Multilayer Recurrent Network Models of Primate Retinal Ganglion Cell Responses". In: *International Conference on Learning Representations*. URL: https://api.semanticscholar.org/CorpusID:39002941.

Bonin, Vincent, Valerio Mante, and Matteo Carandini (2006). "The statistical computation underlying contrast gain control". In: *Journal of Neuroscience* 26.23, pp. 6346–6353.

Cadena, Santiago A, George H Denfield, Edgar Y Walker, Leon A Gatys, Andreas S Tolias, Matthias Bethge, and Alexander S Ecker (2019). "Deep convolutional models improve predictions of macaque V1 responses to natural images". In: *PLoS computational biology* 15.4, e1006897.

Chichilnisky, EJ (2001). "A simple white noise analysis of neuronal light responses". In: *Network (Bristol, England)* 12.2, pp. 199–213. ISSN: 0954-898X. DOI: 10.1080/713663221. URL: https://doi.org/10.1088/0954-898X/12/2/306.

Clevert, Djork-Arné, Thomas Unterthiner, and Sepp Hochreiter (2015). "Fast and accurate deep network learning by exponential linear units (elus)". In: *arXiv preprint arXiv:1511.07289*.

David, Stephen V, William E Vinje, and Jack L Gallant (2004). "Natural stimulus statistics alter the receptive field structure of v1 neurons". In: *Journal of Neuroscience* 24.31, pp. 6991–7006.

Dowling, John E (1963). "Neural and photochemical mechanisms of visual adaptation in the rat". In: *The Journal of general physiology* 46.6, pp. 1287–1301.

— (1960). "Chemistry of Visual Adaptation in the Rat". In: *Nature* 188, pp. 114–118. URL: https://api.semanticscholar.org/CorpusID:4150748.

Gollisch, Tim and Jian K. Liu (2018). *Data: Salamander retinal ganglion cells under finely structured spatio-temporal white noise*. DOI: 10.12751/G-NODE.62B65B. URL: https://doid.gin.g-node.org/62b65b1b4a6589c6b77c72e50b118c90.

Gollisch, Tim and Markus Meister (2010). "Eye smarter than scientists believed: Neural computations in circuits of the retina". In: *Neuron* 65.2, pp. 150–164. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2009.12.009. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3717333/.

Heitman, Alexander, Nora Brackbill, Martin Greschner, Alexander Sher, Alan M. Litke, and E. J. Chichilnisky (2016). *Testing pseudo-linear models of responses to natural scenes in primate retina*. Pages: 045336 Section: New Results. DOI: 10.1101/045336. URL: https://www.biorxiv.org/content/10.1101/045336v2.

Hoefling, Larissa, Klaudia P Szatko, Christian Behrens, Yongrong Qiu, David Alexander Klindt, Zachary Jessen, Gregory S Schwartz, Matthias Bethge, Philipp Berens, Katrin Franke, Alexander S Ecker, and Thomas Euler (2022). *A chromatic feature detector in the retina signals visual context changes*. preprint. Neuroscience. DOI: 10.1101/2022.11.30.518492. URL: http://biorxiv.org/lookup/doi/10.1101/2022.11.30.518492.

Hosoya, Toshihiko, Stephen A. Baccus, and Markus Meister (2005). "Dynamic predictive coding by the retina". In: *Nature* 436, pp. 71–77. URL: https://api.semanticscholar.org/CorpusID:486803.

Ioffe, Sergey and Christian Szegedy (2015). "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift". In: *CoRR* abs/1502.03167. arXiv: 1502.03167. URL: http://arxiv.org/abs/1502.03167.

Karamanlis, Dimokratis and Tim Gollisch (2021). "Nonlinear Spatial Integration Underlies the Diversity of Retinal Ganglion Cell Responses to Natural Images". In: *Journal of Neuroscience* 41.15, pp. 3479–3498. ISSN: 0270-6474. DOI: 10.1523/JNEUROSCI.3075-20.2021. eprint: https://www.jneurosci.org/content/41/15/3479.full.pdf. URL: https://www.jneurosci.org/content/41/15/3479.

Kastner, David B and Stephen A Baccus (2014). "Insights from the retina into the diverse and general computations of adaptation, detection, and prediction". In: *Current Opinion in Neurobiology* 25. Theoretical and computational neuroscience, pp. 63–69. ISSN: 0959-4388. DOI: https://doi.org/10.1016/j.conb.2013.11.012. URL: https://www.sciencedirect.com/science/article/pii/S0959438813002237.

Kim, Kerry J. and Fred Rieke (2001). "Temporal Contrast Adaptation in the Input and Output Signals of Salamander Retinal Ganglion Cells". In: *The Journal of Neuroscience* 21, pp. 287–299. URL: https://api.semanticscholar.org/CorpusID:8866633.

Kingma, Diederik P. and Jimmy Ba (2017). *Adam: A Method for Stochastic Optimization*. arXiv: 1412.6980 [cs.LG].

Klindt, David, Alexander S Ecker, Thomas Euler, and Matthias Bethge (2017). "Neural system identification for large populations separating "what" and "where"". In: *Advances in neural information processing systems* 30.

Korenberg, Michael J and Ian W Hunter (1986). "The identification of nonlinear biological systems: LNL cascade models". In: *Biological cybernetics* 55.2-3, pp. 125–134.

Kuo, Sidney P., Gregory W. Schwartz, and Fred Rieke (2016). "Nonlinear Spatiotemporal Integration by Electrical and Chemical Synapses in the Retina". In: *Neuron* 90.2, pp. 320–332. ISSN: 0896-6273. DOI: https://doi.org/10.1016/j.neuron.2016.03.012. URL: https://www.sciencedirect.com/science/article/pii/S0896627316002129.

Linsker, R. (1988). "Self-organization in a perceptual network". In: *Computer* 21.3, pp. 105–117. DOI: 10.1109/2.36.

Liu, Jian K., Helene M. Schreyer, Arno Onken, Fernando Rozenblit, Mohammad H. Khani, Vidhyasankar Krishnamoorthy, Stefano Panzeri, and Tim Gollisch (2017). "Inference of neuronal functional circuitry with spike-triggered non-negative matrix factorization". In: *Nature Communications* 8.1, p. 149. ISSN: 2041-1723. DOI: 10.1038/s41467-017-00156-9.

Lurz, Konstantin-Klemens, Mohammad Bashiri, Konstantin Willeke, Akshay Jagadish, Eric Wang, Edgar Y. Walker, Santiago A Cadena, Taliah Muhammad, Erick Cobos, Andreas S. Tolias, Alexander S Ecker, and Fabian H. Sinz (2021). "Generalization in data-driven models of primary visual cortex". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=Tp7kI90Htd.

Maheswaranathan, Niru, Lane T. McIntosh, Hidenori Tanaka, Satchel Grant, David B. Kastner, Joshua B. Melander, Aran Nayebi, Luke E. Brezovec, Julia H. Wang, Surya Ganguli, and Stephen A. Baccus (2023). "Interpreting the retinal neural code for natural scenes: From computations to neurons". In: *Neuron* 111.17, 2742–2755.e4. ISSN: 0896-6273. DOI: https://doi.org/10.1016/j.neuron.2023.06.007. URL: https://www.sciencedirect.com/science/article/pii/S0896627323004671.

Marmarelis, Panos Z and Ken-Ichi Naka (1972). "White-noise analysis of a neuron chain: an application of the Wiener theory". In: *Science* 175.4027, pp. 1276–1278.

McFarland, James M., Yuwei Cui, and Daniel A. Butts (2013). "Inferring Nonlinear Neuronal Computation Based on Physiologically Plausible Inputs". In: *PLOS Computational Biology* 9.7, pp. 1–18. DOI: 10.1371/journal.pcbi.1003143. URL: https://doi.org/10.1371/journal.pcbi.1003143.

McIntosh, Lane T., Niru Maheswaranathan, Aran Nayebi, Surya Ganguli, and Stephen A. Baccus (2017). "Deep Learning Models of the Retinal Response to Natural Scenes". In: DOI: 10.48550/ARXIV.1702.01825. URL: https://arxiv.org/abs/1702.01825.

Ölveczky, Bence P, Stephen A Baccus, and Markus Meister (2007). "Retinal adaptation to object motion". In: *Neuron* 56.4, pp. 689–700.

Pillow, Jonathan W., Jonathon Shlens, Liam Paninski, Alexander Sher, Alan M. Litke, E. J. Chichilnisky, and Eero P. Simoncelli (2008). "Spatio-temporal correlations and visual signalling in a complete neuronal population". In: *Nature* 454, pp. 995–999. URL: https://api.semanticscholar.org/CorpusID:306143.

Ravello, César, Laurent Perrinet, Maria-Jose Escobar, and Adrian Palacios (2019). "Speed-Selectivity in Retinal Ganglion Cells is Sharpened by Broad Spatial Frequency, Naturalistic Stimuli". In: *Scientific Reports* 9. DOI: 10.1038/s41598-018-36861-8.

Sakai, Hiroko M, Naka Ken-Ichi, and Michael J Korenberg (1988). "White-noise analysis in visual neuroscience". In: *Visual neuroscience* 1.3, pp. 287–296.

Schwartz, Greg, Rob Harris, David Shrom, and Michael J Berry (2007). "Detection and prediction of periodic patterns by the retina". In: *Nature neuroscience* 10.5, pp. 552–554.

Schwartz, Greg, Haruhisa Okawa, Felice Dunn, Josh Morgan, Daniel Kerschensteiner, Rachel Wong, and Fred Rieke (2012). "The spatial structure of a nonlinear receptive field". In: *Nature Neuroscience* 15, pp. 1572–1580. DOI: 10.1038/nn.3225.

Shi, Qing, Pranjal Gupta, Alexandra Martinson, Joshua Singer, and Daniel Butts (2019). "Functional characterization of retinal ganglion cells using tailored nonlinear modeling". In: *Scientific Reports* 9. DOI: 10.1038/s41598-019-45048-8.

Sinz, Fabian, Alexander S Ecker, Paul Fahey, Edgar Walker, Erick Cobos, Emmanouil Froudarakis, Dimitri Yatsenko, Zachary Pitkow, Jacob Reimer, and Andreas Tolias (2018). "Stimulus domain transfer in recurrent models for large scale cortical population prediction on video". In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Vol. 31. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2018/file/9d684c589d67031a627ad33d59db65e5-Paper.pdf.

Smirnakis, Stelios M, Michael J Berry, David K Warland, William Bialek, and Markus Meister (1997). "Adaptation of retinal processing to image contrast and spatial scale". In: *Nature* 386.6620, pp. 69–73.

Sridhar, Shashwat, Michaela Vystrčilová, Mohammad H. Khani, Dimokratis Karamanlis, Helene M. Schreyer, Varsha Ramakrishna, Steffen Krüppel, Sören J. Zapp, Matthias Mietsch, Alexander Ecker, and Tim Gollisch (2024). *Modeling spatial contrast sensitivity in responses of primate retinal ganglion cells to natural movies*.

Tanaka, Hidenori, Aran Nayebi, Niru Maheswaranathan, Lane McIntosh, Stephen A. Baccus, and Surya Ganguli (2019). "From deep learning to mechanistic understanding in neuroscience: the structure of retinal prediction". In: DOI: 10.48550/ARXIV.1912.06207. URL: https://arxiv.org/abs/1912.06207.

Tkačik, Gašper, Anandamohan Ghosh, Elad Schneidman, and Ronen Segev (2014). "Adaptation to Changes in Higher-Order Stimulus Statistics in the Salamander Retina". In: *PLoS ONE* 9. URL: https://api.semanticscholar.org/CorpusID:14979163.

Turner, Maxwell H and Fred Rieke (2016). "Synaptic rectification controls nonlinear spatial integration of natural visual inputs". In: *Neuron* 90.6, pp. 1257–1271.

Walker, Edgar Y., Fabian H. Sinz, Erick Cobos, Taliah Muhammad, Emmanouil Froudarakis, Paul G. Fahey, Alexander S. Ecker, Jacob Reimer, Xaq Pitkow, and Andreas S. Tolias (2019). "Inception loops discover what excites neurons most using deep predictive models". In: *Nature Neuroscience* 22.12, pp. 2060–2065. ISSN: 1546-1726. DOI: 10.1038/s41593-019-0517-x.