bioRxiv preprint doi: https://doi.org/10.1101/2021.02.14.431169; this version posted February 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Individual variability of neural computations in the primate retina

Nishal Shah[1,2]*, Nora Brackbill[2,3], Ryan Samarakoon[2,4], Colleen Rhoades[2,5], Alexandra Kling[2,4], Alexander Sher[6], Alan Litke[6], Yoram Singer[7], Jonathon Shlens[8], and E.J. Chichilnisky [2,4]

1 Stanford University, Department of Electrical Engineering
2 Hansen Experimental Physics Lab, Stanford University,
3 Department of Physics, Stanford University
4 Department of Neurosurgery, Stanford University
5 Department of Bioengineering, Stanford University
6 University of California, Santa Cruz
7 WorldQuant, LLC
8 Google Brain

Abstract

Variation in the neural code between individuals contributes to making each person unique. Using ~100 neural population recordings from major ganglion cell types in the macaque retina, we develop an interpretable computational representation of individual variability using machine learning. This representation preserves invariances, such as asymmetries between ON and OFF cells, while capturing individual variation and covariation in properties such as nonlinearity, temporal dynamics, and spatial receptive field size. The similarity of these properties across cell types was dependent on the similarity of their synaptic connections. Surprisingly, male retinas exhibited higher firing rates and faster temporal integration than female retinas. By exploiting data from previously recorded macaque retinas, a new macaque retina (and crucially, a human retina) could be efficiently characterized. Simulations indicated that combining a vast dataset of healthy macaque recordings with behavioral feedback could be used to identify the neural code and improve retinal implants for treating blindness.

Main

An emerging frontier in biomedicine is understanding variability between individuals, with implications ranging from the mathematical modeling of living systems to ethics and personalized medicine. In neuroscience, differences in mental function between individuals are substantial, yet little is known about the underlying variation in the information processing performed by neural circuits. Neuroimaging, gross structural, and behavioral measurements in humans cannot reveal the neural code at circuit resolution, and physiological measurements in invertebrates or rodents have uncertain applicability to humans. These limitations have led to a large gap in our understanding of variability in the neural code and its implications for translational medicine and neuroengineering. Two technical challenges have limited our understanding of this variability in higher animal models: high-resolution, large-scale physiological recordings from many animals, and methods for deciphering variability in complex circuit level computations.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.14.431169; this version posted February 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Modeling the shared and individual components of neural coding variability

To overcome these limitations, we exploited large-scale multi-electrode recordings from ~21626 retinal ganglion cells (RGCs) aggregated over 112 recordings from 66 isolated primate retinas, in which the functional properties of diverse cell types have been extensively studied[1–5]. These data, gathered over a decade of experimentation, exhibited significant neural coding variability across recordings. As a baseline, response properties in each recording were summarized by the parameters of a linear-nonlinear-Poisson (LNP) encoding model. This widely used model[6] captures light-evoked responses in RGCs using a spatiotemporal linear filter applied to the stimulus, followed by an output nonlinearity and stochastic spike generation. Receptive field sizes of two known RGC types -- ON and OFF parasol -- exhibited substantial variation across recordings. This variability was evident across retinal eccentricities, but was also present at a given eccentricity (Figure 1A, first column, Figure 1G). Diversity was also seen in the kinetics of light response (Figure 1A, second column) and in the form of the output nonlinearity (Figure 1A, third column). When the LNP model was extended to capture an additional computational motif, nonlinear integration of spatial subunits within the receptive field[7], the change in prediction accuracy was also variable across recordings (Figure 1A, last column).

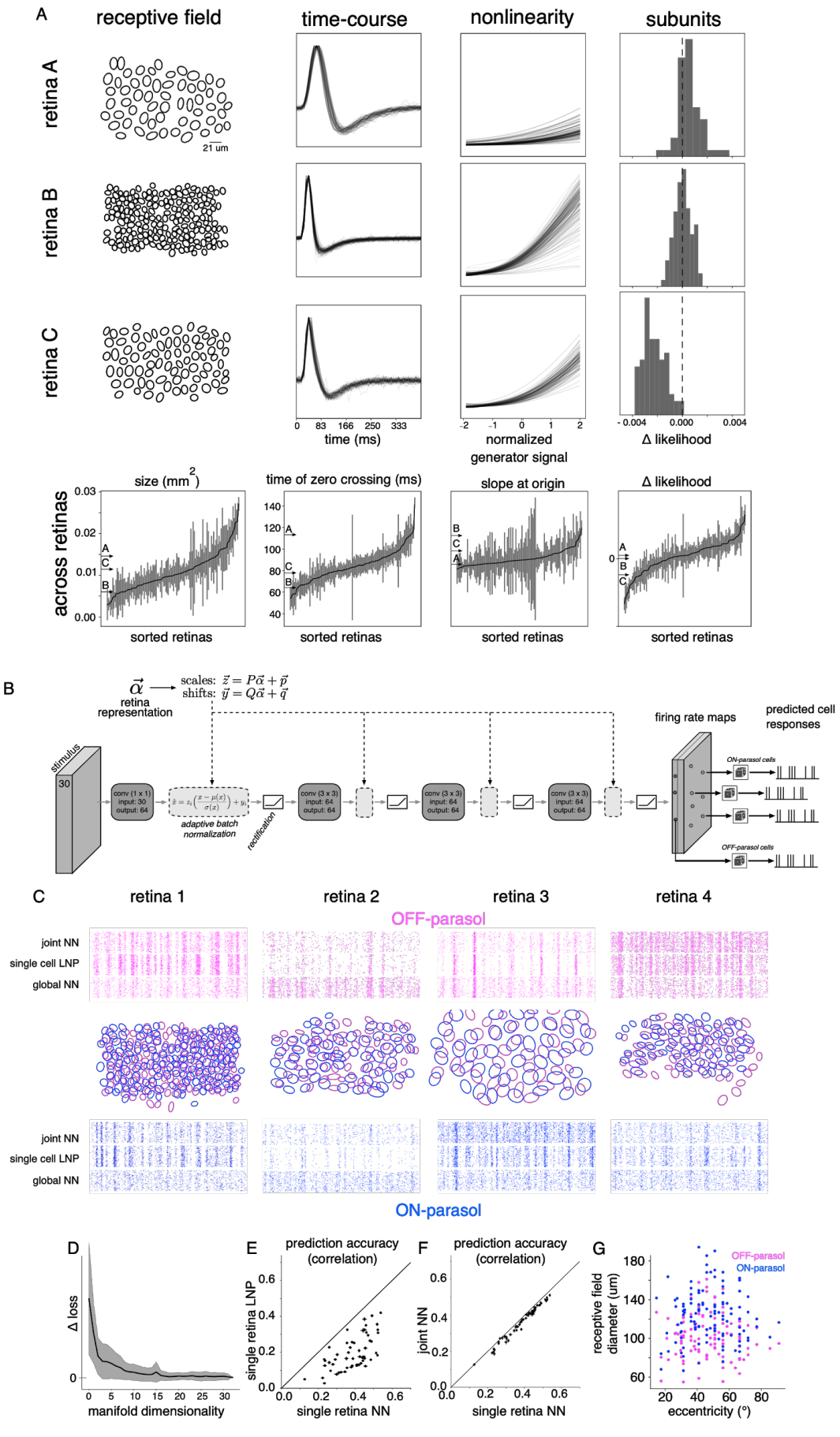
The structure and covariation of these response properties was therefore explored using a flexible model that captured nonlinear spatial integration, and combined shared and recording-specific parameters. The shared component was a multilayered convolutional neural network (CNN), an extension of the LNP model, consisting of multiple alternating stages of spatio-temporal filtering, normalization and rectification. The rectification captured nonlinear spatial integration, and the convolutional structure captured the known translational invariance of visual signals in each cell type (cells of the same type at different locations have very similar response properties[8]). The model output consisted of multiple firing rate maps in each retina, one for each cell type. To predict a given cell's responses, the model-predicted firing rate was read off from the map at the cell's location (Figure 1B). Due to the translational invariance constraint, the proposed model cannot capture differences between cells belonging to the same cell type, resulting in poorer prediction accuracy compared to models that allow for cell-specific parameters such as single-cell LNP[6] or other state-of-the-art models[9,10]. However, this constraint enabled the proposed neural network architecture to predict responses across different recordings with different numbers of cells, while focusing on variation across recordings. When trained using ON and OFF parasol cell responses in each retina separately, the CNN model exhibited performance superior to the single-retina LNP model (LNP model with common parameters for all cells of a given type), as expected given its more flexible structure (Figure 1E). However, when the CNN model was trained on multiple retinas together, it failed to capture the responses of retinas with low firing rates, highly modulated responses, or other features that varied between recordings (Figure 1C, rasters).

To capture the variation of light response properties across recordings in a compact and tractable way, a small number of recording-specific parameters were used to reweight the activations of different filters at each layer of the shared CNN. Data from all recordings were used to learn the ~100K parameters of the shared CNN, while data from each recording were

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.14.431169; this version posted February 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

used to obtain a small number recording-specific parameters. The collection of these recording-specific parameters were interpreted as a *manifold of neural coding variability*. When learned using 71 recordings, the low-dimensional manifold captured variations in background firing rate, sustained vs. transient dynamics, and response nonlinearities (Figure 1C, rasters). The ability to simultaneously predict responses across multiple retinas saturated at ~15 dimensions of the learned manifold (Figure 1D), much lower than the total number of CNN parameters, and the performance of this joint model based on the manifold was only slightly lower than a CNN model trained for each retina separately (Figure 1F). Thus, a simple, low-dimensional representation can efficiently and accurately capture the diversity of retinal computations.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.14.431169; this version posted February 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



bioRxiv preprint doi: https://doi.org/10.1101/2021.02.14.431169; this version posted February 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 1. Modeling variability in the neural code (A) Variability of response properties across preparations. Spatial receptive fields (first column), time of zero-crossing of temporal filter (second column), nonlinearity (third column) and change in response likelihood with five nonlinear subunits (fourth column) for OFF parasol cells in three representative recordings (rows, same y-axis for each response property across recordings). The last row represents the range of response parameters across 122 preparations, sorted according to their population means, and error bars corresponding to the robust standard deviation. Arrows indicate the values of the chosen retinas in the first three rows. (B) Architecture of the neural network for capturing response variation. The visual stimulus is passed through multiple layers of convolution with spatial filters, with adaptive batch normalization and rectification at each layer, producing two firing rate maps (one each for ON and OFF parasol cell types). The Poisson firing rate for each cell is read off from the value at the cell's location in the firing rate map of its cell type. Retina-specific tuning of responses is performed by adjusting the mean and standard deviation of the activation values at each layer, determined by a linear transformation of the retina's location in the low-dimensional manifold. (C) Response prediction across 4 representative training retinas (columns). The receptive field mosaics are shown for each retina (middle row), along with response predictions for a randomly selected OFF parasol (top row) and ON parasol cell (bottom row). Rasters (60 trials) for predicted responses to a 3 sec long white noise stimulus using the LNP model; neural network model, trained jointly on multiple retinas, with retina-specific parameters (15 dimensional manifold, joint NN) or without them (no manifold, global NN). (D) Model error (log likelihood) on test stimuli with varying manifold dimensionality; dimensions=0 indicates no retina-specific adaptation (global NN). Error bars indicate the standard deviation across retinas. (E) Prediction accuracy, averaged across cells, for different retinas (points), using a neural network trained on data from each individual retina (x-axis) and an LNP model with shared parameters across cells of a given type in each retina (y-axis). Prediction accuracy is measured as correlation between predicted firing rate and recorded responses smoothed with a Gaussian filter (σ: 11ms). (F) Similar to (E), comparing predictions from the neural network with a 15-dimensional manifold, trained jointly (y-axis) vs. trained on each retina separately (x-axis). (G) Range of eccentricities (x-axis) and the average receptive field sizes (y-axis) for ON parasol (blue) and OFF parasol (magenta) cells across 122 preparations used in this study.

Neural coding manifold smoothly captures systematic variation across recordings

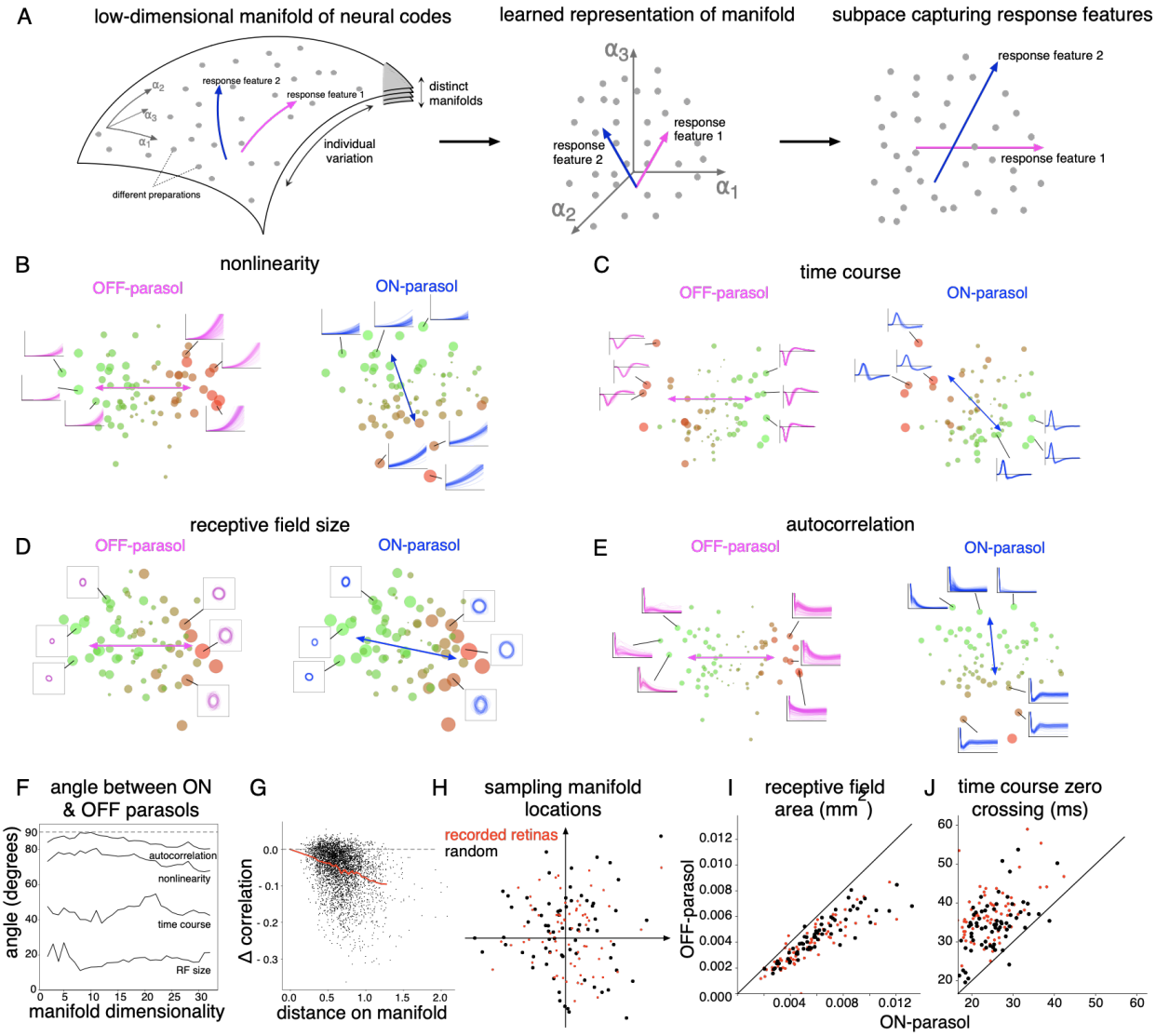
The learned manifold smoothly captured changes in neural code with a greater perturbation in manifold location leading to a greater decrease in response prediction accuracy (Figure 2G). The manifold geometrically represented variation in several light response properties, including receptive field size, time course, output nonlinearity, and spike train autocorrelation. This was observed by projecting the average response property for each retina onto its principal component across all recordings, and then identifying the manifold direction with maximum correlation to this projection. The Spearman rank correlation between these projections and the projections along the identified manifold direction was significantly higher than the value observed in random permutations of the data ($p < 0.001$ for all response properties) (Figure 2B, C, D, E, F), indicating that the geometry of the manifold is well-suited to representing the response properties linearly.

The geometry of the manifold also captured co-variation in response properties of different cell types across recordings. For both spiking autocorrelation and response nonlinearity, the large angles between the manifold directions for ON and OFF parasol cells (86° and 76° respectively) were consistent with low Spearman rank correlation in these response properties (0.28 and 0.12 respectively). Conversely, for response time course and receptive field size, the small angles between the directions for ON and OFF cells (45° and 16° respectively) were consistent with larger Spearman rank correlation (0.56 and 0.95 respectively). Although high correlation is expected for receptive field size based on variation in the eccentricity of different recordings, the moderate correlation observed for response time course and the low correlation observed for

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.14.431169; this version posted February 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

nonlinearity and autocorrelation have not been previously reported. While this analysis could be done directly in the space of LNP model parameters, the manifold presents a more flexible way to study these response properties for a wide range of encoding models.

The manifold also captured known *invariances* in retinal encoding, i.e. properties that were consistent across retinas. To assess whether these invariances were present at many intermediate manifold locations not directly sampled in the experiments, the manifold locations of recorded retinas were perturbed using Gaussian noise with standard deviation equal to the median nearest neighbor separation (Figure 2H). Light responses were then generated using these randomly sampled manifold locations, and the encoding properties were summarized by fitting a LNP model. For the simulated responses, OFF parasol cells had consistently smaller RFs (Figure 2I) and slower time courses (Figure 2J) than ON parasol cells, as expected based on previously reported asymmetries[8].



bioRxiv preprint doi: https://doi.org/10.1101/2021.02.14.431169; this version posted February 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 2. **Geometry of manifold.** (A) Summary of the steps in subsequent analysis. Left: Schematic of the manifold representation of variability. Each recording is summarized by its neural encoding function, indicated by a point (gray) in space of all possible neural encoding functions. The observed neural encoding functions lie in a low-dimensional manifold (depicted as curved surface, but could be more than two dimensional) within this space. Different manifolds (other surfaces) would potentially correspond to different properties that are invariant across recordings. The training procedure learns a coordinate system (α)within the manifold. Middle: Directions corresponding to response features can be identified in the learned coordinate system for representing the manifold. Right: Geometry of the subspace corresponding to the identified directions lead to interpretation of the variations. (B) Manifold locations for 95 preparations (points), projected onto the 2D subspace given by the first principal components of variation in the output nonlinearity for OFF and ON parasol cells. Size and color of dots indicate deviation from the mean. Colored lines indicate the direction of maximum nonlinearity variation for ON parasol (blue) and OFF parasol (magenta) cells. Insets: Lines show output nonlinearities for all cells in representative retinas. (C, D, E) Similar to (B), for time course, receptive field size and autocorrelation, respectively. (F) Angle between ON and OFF parasol directions for particular response properties, as a function of manifold dimension. (G) Change in response prediction accuracy (y-axis) as the manifold location is perturbed from the learned location. Each black dot represents a different perturbation, the red line is the average. (H) Random manifold locations (red) were sampled by adding noise to the learned retina-specific locations (black). Responses to a white noise stimulus of 100 cells of each type (ON and OFF parasol) were sampled from random locations in the visual field using the firing rate maps for the two cell types associated with these sampled manifold locations. These simulated responses were then used to fit a LNP model. (I, J) Relationship between ON and OFF parasol cells for receptive field area and zero crossing of response time course, respectively, for recorded (red) and randomly sampled (black) retinas

Manifold reveals covariation associated with connectivity, and male-female differences

The manifold of variability revealed two novel properties of retinal encoding. First, the RGC types that receive synaptic input from bipolar cells at similar depths in the inner plexiform layer (IPL) showed greater covariation in their response properties across recordings. To examine covariation between cell types, the ON and OFF midget cell types were included with the ON and OFF parasol cell types considered thus far. In 85 recordings (53 macaques), the similarity of three response properties – nonlinearity, time course and autocorrelation – across different pairs of cell types was measured either directly, or in the manifold (Figure 3A-C). Using both methods, the highest correlation in these physiological properties was observed between cell-type pairs with the same polarity (ON or OFF), consistent with the lamination of ON and OFF cells in the inner and outer IPL, respectively (Figure 3D). Moreover, for cell type pairs with opposite polarities, a higher correlation in physiological properties was observed for the ON-parasol/OFF-parasol pair than for the ON-midget/OFF-midget pair, consistent with the lamination of parasol cells closer to the middle of IPL (Figure 3A-D). These observations support the approach of studying the response properties of newer cell types (such as ON and OFF smooth monostratified cells[4]) after normalizing to the properties of more commonly studied cells with similar synaptic inputs, thereby minimizing the effects of inter-retina variability.

Second, the recordings from male and female retinas were separated in the manifold (d'=1.8 for 15 dimensional manifold, Figure 3E), in a way that could not be explained by variations in experimental factors (Figure 3I). For both ON and OFF parasol cells, the differences between male and female retinas were partially due to differences in firing rate and speed of temporal filtering. This was determined by computing the direction separating the means of male and female retinas in the manifold, and determining the angle between this direction and the directions most aligned with variation in firing rate and response time course (cosine(angle) ~ 0.5 for both) (Figure 3H). Note that unlike gender, animal age did not separate different

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.14.431169; this version posted February 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

recordings on the manifold. In principle, the observed gender differences could be confounded by variation in experimental methods such as dissection procedure (isolation from the retinal pigment epithelium, or RPE) or temperature (31°-36°C across recordings). Because higher recording temperatures were associated with isolation from the RPE for technical reasons, the directions in the manifold representing dissection method and temperature variation were aligned (cosine(angle) ~ 0.57, Figure 3H). Compared to firing rate and response time course, these experimental factors were less aligned to the direction of gender separation (cosine(angle) ~ 0.26) (Figure 3H), suggesting that gender differences were probably not due to differences in these experimental methods. To eliminate experimental factors more rigorously, the separation of males and females in the manifold was measured after conditioning the data in several ways. For each condition, a bootstrap rank test was performed to verify if the mean locations of male and female recordings differed (see Methods). Significant separation ($p<0.05$) was observed for the recordings with RPE-isolated dissections and high temperature (≥33.5°C), whereas the separation was not significant ($p>0.1$) for RPE-attached dissections and lower temperatures (<33.5°C) (Figure 3I). For the RPE-isolated retinas, the male recordings exhibited higher firing rates (Figure 3F) and faster temporal integration (Figure 3G) ($p<0.01$, see Methods for details).

Although this conditioning on specific experimental conditions revealed statistically significant differences between males and females, the level of significance was lower when compared to all the recordings (Figure 3I), potentially due to a reduction in the number of samples when analysis is restricted to a particular set of experimental conditions. The manifold made it possible to separate experimental variations more efficiently, without reducing the number of data points. To accomplish this, the data were projected onto axes in the manifold orthogonal to the two identified directions of experimental variability. This projection increased the statistical separation between the male and female retinas ($p<10^{-6}$) (Figure 3I). Thus, the geometry of the manifold, which smoothly organizes the neural computation across retinas, makes it possible to examine statistical trends in the data efficiently in spite of potential experimental confounds.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.14.431169; this version posted February 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

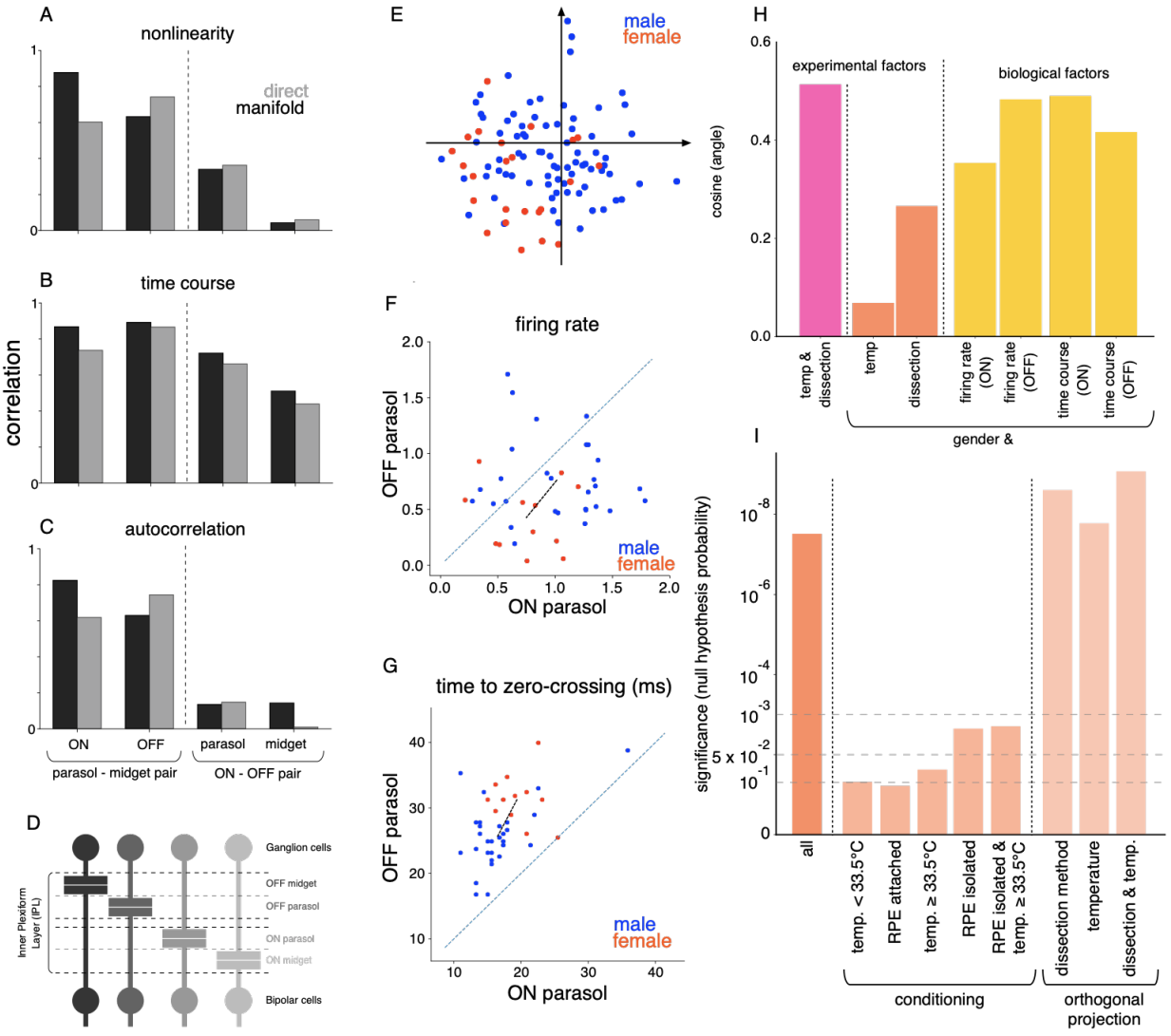


Figure 3 Biological factors underlying variability. Relation between the first principal component of (A) nonlinearity, (B) response time course and (C) spike train autocorrelation variation for different pairs of cell types. The relationship is either measured directly using Spearman rank correlation, or the cosine of the angle between corresponding directions in the 15-dimensional manifold. (D) Distinct lamination depths for the bipolar-ganglion cell synapse for different ganglion cell types[11]. (E) Two dimensional PCA projection of manifold locations for recordings from male (blue) and female (red) retinas. (F) The average firing rate for ON parasol (y-axis) and OFF parasol (x-axis) cells for preparations with isolated RPE. The mean manifold location of male (blue) and female (red) recordings were different ($p<0.01$ for bootstrap and $p<0.05$ for hierarchical bootstrap). Black line joins the mean male and female locations. (G) Similar to (F) for the time course of STA, with separation of male and female retinas ($p < 0.01$ for bootstrap and $p<0.05$ for hierarchical bootstrap[12]). (H) Cosine of the angle (y-axis) between the manifold directions corresponding to different pairs of factors, which are either biological (gender, firing rate, time of zero crossing of time course) and experimental (temperature, dissection - whether retinal pigment epithelium (RPE) was attached or isolated). (I) Degree of separation of male and female recordings measured using a resampling test, for all the recordings, conditioned on the subset of recordings with specific dissection procedure or temperature, or all recordings with locations projected orthogonal to directions for dissection procedure and temperature variation.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.14.431169; this version posted February 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Manifold generalizes to a novel recording

The manifold permitted efficient response modeling of a new, previously unseen retina by leveraging trends in the large data set of retinas used for training. Response modeling can be performed efficiently by identifying the manifold location of a new retina in several ways, using limited data. In the absence of any new data, the "typical" manifold location for a new retina can be obtained by merely *averaging* the locations of all training retinas. In the case of a degenerated retina with no light evoked response, partial information such as the autocorrelation function of the spiking of recorded neurons can still be identified from spontaneous activity. In this case, the manifold location can be *approximated* by averaging the locations of training retinas that have similar autocorrelation. Finally, in the presence of light evoked responses, gradient descent can be used to *optimize* the manifold location based on the likelihood of the data, leveraging the training retinas as prior information (Figure 4A).

These three approaches were examined using a model trained with 71 preparations and tested with 24 preparations. The proximity of manifold locations identified by approximation and optimization suggest that these approaches accurately capture properties of the new retina (Figure 4B). To examine the efficiency obtained by using the manifold, the optimization approach with limited data was studied by measuring convergence of the manifold location and of response prediction accuracy. First, the manifold location converged quickly as the recording duration increased, with ~3 minutes of light response data producing a location similar to that produced by ~30 minutes of data (Figure 4C). Second, whereas optimization of the manifold location performed similarly to training the full model (along with the CNN) when performed with a large amount of data, the manifold approach more accurately predicted light responses when the data were limited (~8 sec) (Figure 4D).

The efficiency of the low dimensional manifold could come at a cost of its generalization accuracy to new retinas. On varying the manifold dimensionality, the ability to predict responses to previously unseen retinas saturated at ~15 dimensions (Figure 4E, black line), a value that did not change with the number of retinas used for training (Figure 4E, colored lines). Hence, in addition to the previous observation on generalization to new stimuli within the collection of retinas used for learning (Figure 1D), the low dimensional manifold is also able to generalize to new, previously unseen retinas.

These findings suggest that the manifold may aid in translating our understanding of the macaque retina to the human retina, an important goal for biomedical research. Recent work[13,14] has shown that the receptive field properties of the four numerically dominant RGC types (ON and OFF parasol and midget) are similar to those of their macaque counterparts. To test whether light responses in the human retina fall within the range observed across many macaque retinas, the manifold location of a single human retina was identified using the three operations described above (averaging, approximation, optimization). For each method, the estimated manifold location of the human retina was well within the span of manifold locations of many macaque retinas (Figure 4B,C), and the responses of a new retina were predicted with similar accuracy and efficiency in the two species (Figure 4D). Hence, the manifold reveals similar light response properties of corresponding RGC types in macaques and humans.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.14.431169; this version posted February 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Given that the macaque retinal code translates accurately to humans, it may provide a valuable tool for the development of an advanced artificial retina for treating vision loss. However, a challenging first step in restoring vision with such a device is to identify how the neurons in a blind retina should encode visual stimuli with the implant. In this setting, the retina is no longer light-sensitive, but the blind person with an implanted device could potentially report the similarity of artificially induced images to a verbally described object. The correct neural encoding could be identified by estimating the location of the retina on the low-dimensional manifold, using a psychophysical discrimination task with the following iterative procedure: (i) sample a few of the plausible manifold locations , (ii) use each of these manifold locations to predict retinal responses for a particular visual stimulus, (iii) stimulate with the implanted device to produce each of these responses, (iv) ask the subject which stimulus produced a sensation that most closely matches a verbal description of it, and (iv) update the set of plausible manifold locations.

To illustrate the feasibility of this procedure, the above steps were simulated assuming (for simplicity) an artificial retina that has perfect cellular selectivity of stimulation. For updating the plausible locations, the perceptual accuracy was assumed to be governed by the Kullbeck-Leibler divergence between the distribution of neural responses associated with the tested manifold location and the distribution of responses associated with the true location. The set of plausible locations identified by the procedure converged to the true location (given by optimization; see above) in a small number of steps (see Methods, Figure 4F). Hence, the low-dimensional manifold could provide valuable efficiency in translational applications.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.14.431169; this version posted February 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

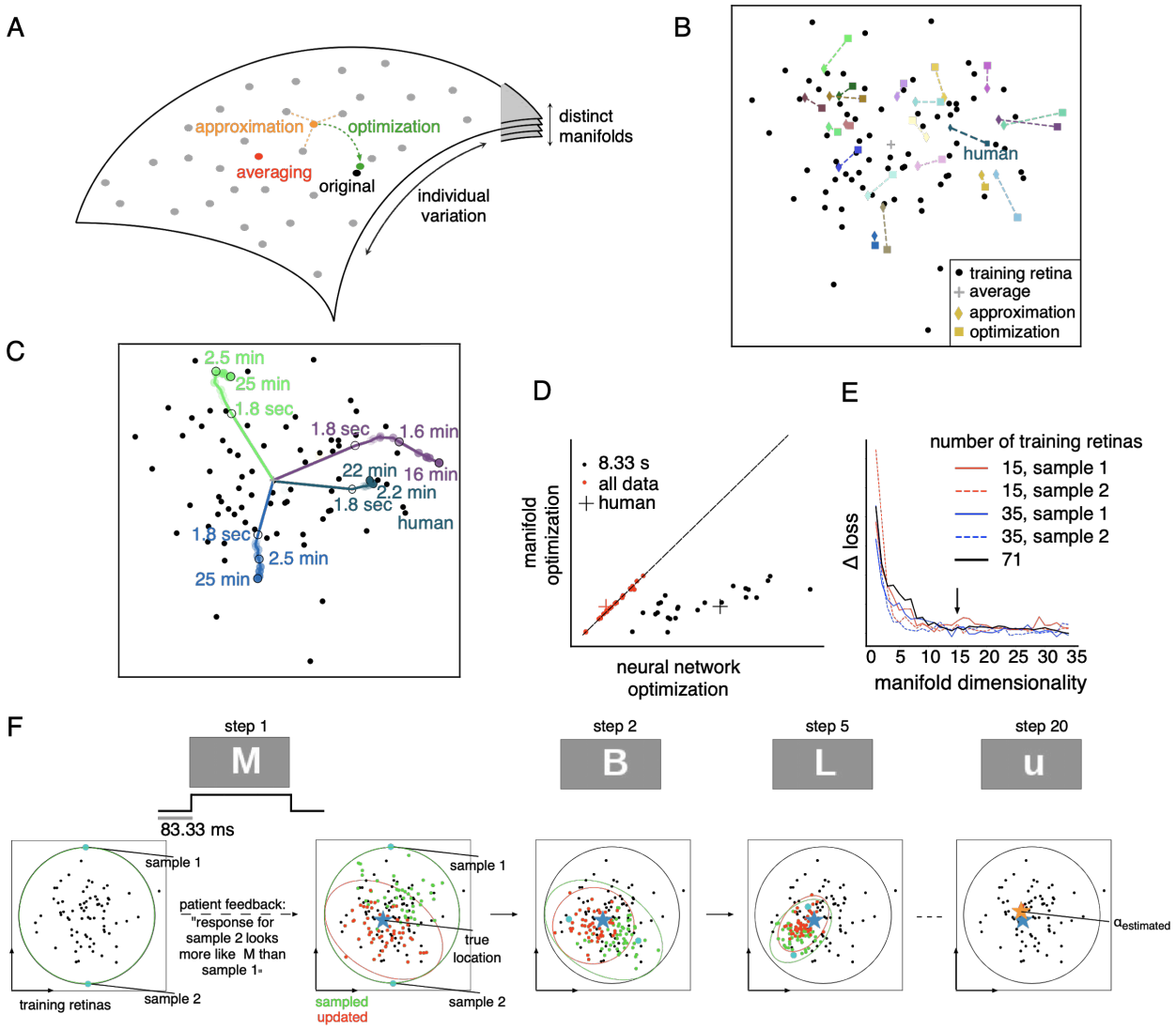


Figure 4. **Generalization to new, previously unseen retinas**. (A) Efficient identification of manifold location. Training retinas (gray points) are used to define the manifold. Then, the manifold location of a new retina is determined in one of several ways - averaging (red): compute the mean location of all training retinas; approximation (orange): compute the mean location of a subset of training retinas with specific features similar to the new retina; optimization (green): gradient descent on the manifold location to maximize the prediction accuracy for measured light responses of the new retina. (B) Identified manifold locations using averaging (+; same for all retinas), approximation (◊) and optimization (□) for testing retinas (colors, each pair joined with a line). Black points correspond to the locations of training retinas. (C) Optimized manifold location for three retinas (colored lines), with varying duration of recorded responses. Training retina locations (black points) and locations identified by averaging (+) are shown. For (B, C), the 15 dimensional manifold is projected into two dimensions that capture autocorrelation variation (same as Figure 2E) (D) Response prediction loss with optimization of manifold location (y-axis) vs. loss with learning all neural network parameters (x-axis), using either 8.33ms (red) or 15-30 min (black) of data. (E) Convergence of prediction loss with optimized manifold location, averaged across 24 testing retinas (y-axis) as a function of the number of manifold dimensions (x-axis). Loss measured as negative log-likelihood, averaged across cells. Colored lines indicate loss obtained using fewer training retinas. (F) Simulation of the discrimination task used to identify manifold location in the retina of a blind person. At each step, the visual stimulus is a letter from English alphabet (top row). Two dimensional projection of the manifold with locations of training retinas (black dots) is shown (bottom row). Posterior over the set of feasible manifold locations approximated using a Gaussian (red circle). At each step, random manifold locations are sampled from the posterior, and corresponding responses are reproduced

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.14.431169; this version posted February 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

using the artificial retina, and feedback from the subject is used to update the posterior using Monte Carlo methods (all samples in green, accepted samples in red). In ~20 steps, the estimated manifold location (orange) converges to the true underlying location (blue).

Conclusions and Discussion

Recent studies have shown that neural responses in the retina, which have often been described using pseudo-linear models[6,15], can be modeled more accurately using artificial neural networks[9,10]. However, the function of these complex models can be difficult to interpret. While using machine learning models to more accurately capture shared (and complex) aspects of neural response, the present work crucially maintained the information about variability between recordings in an interpretable low-dimensional manifold. [16]. As a result, the manifold revealed novel structure in inter-retina variation, such as gender-based differences and covariation of several response properties across cell types. In principle, such an approach could be used for other applications in which the varying neural computation of interest is captured using a simple low-dimensional manifold, while machine learning models improve accuracy by capturing complex but unchanging aspects of the computation that are of less immediate interest.

At least three sources of biological variability in the neural code may play a role in the present data: variation between animals, differences between the two eyes, and variation across retinal locations in a given eye. However, in the present data, these sources of variability are confounded with variation in experimental procedures for euthanasia, for eye removal, and for *ex vivo* recording, respectively. Thus, it is not possible to definitively isolate biological and experimental variation in the present data, though the observed male-female differences were separated from the clear influences of experimental procedures and therefore most likely reflect true biological differences. The tools developed here make it possible to capture and analyze both biological and experimental variability in a single framework, an asset for understanding the neural code more completely using imperfect experimental data. However, future work will be needed to parse the mechanisms of both biological and experimental variability, potentially exploiting the geometry of the manifold. For example, the low-dimensional manifold made it possible to project the data into a subspace orthogonal to known confounds, and thus to control for them without losing data by conditioning on specific experimental variables.

The present findings on neural coding differences between male and female retinas add to a large literature on gender-based differences in brain structure and function[17–19]. In the retina, genetic differences between male and female primates (including humans) produce different variation in cone photopigment spectral sensitivities, and thus different color vision, across the population[20]. However, to our knowledge, differences in neural coding between males and females have not been reported, perhaps due to the lack of appropriate physiological recordings and/or analysis tools. In the present work, the large data set and the geometry of the manifold representation made it possible to establish the observed male-female differences in spite of other potential confounds.

The manifold of neural coding variability may also be useful in other neuroengineering applications, such as motor prostheses. The goal in motor prosthesis is usually to read out the neural activity in a paralyzed person to control a computer cursor or a robotic limb[21]. Similar to

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.14.431169; this version posted February 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the problem of identifying the neural encoding in blind person, identifying the neural mapping in a paralyzed person is limited by the absence of simultaneous neural recordings and limb trajectory measurements, and thus may benefit from leveraging existing data that captures the diversity of neural coding across individuals. Specifically, a manifold of inter-individual variation may be useful for identifying neural decoding in a person with an implant, perhaps using a task involving imagined movements to identify the manifold location[22]. The manifold may also be useful for adjusting to the challenge of variability over time in chronic recordings[23].

The neural coding manifold may also be useful for harnessing brain plasticity, which could improve vision with an artificial retina. Indeed, present-day retinal implants make little attempt to reproduce the neural code of the retina, and thus implicitly rely heavily on plasticity to compensate for device limitations[24]. In motor prostheses, it has been shown that the brain can more easily adjust its activity to accommodate perturbations in the artificial neural decoder if these perturbations lie in a specific low-dimensional manifold[25]. In the case of an artificial retina, we hypothesize that the brain may more readily learn to interpret the neural activity produced by the implanted device if the visual encoding that it uses lies within the manifold of retinal coding variability.

Acknowledgements

We thank J. Carmena, K. Bankiewicz, T. Moore, W. Newsome, M. Taffe, T. Albright, E. Callaway, H. Fox, R. Krauzlis, S. Morairty, and the California National Primate Research Center for access to macaque retinas. We thank Kristy Berry, K. Williams, B. Morsey, J. Frohlich, and M. Kitano for accumulating and providing macaque retina metadata. The human eye was provided by Donor Network West (San Ramon, CA). We are thankful for the cooperation of Donor Network West and all of the organ and tissue donors and their families, for giving the gift of life and the gift of knowledge, by their generous donations. We thank E. Simoncelli, S. Mitra, V. Gupta, K. Talwar, V. Feldman, A. Gogliettino, E. Wu, S. Madugula and the entire Stanford Artificial Retina team for helpful discussions. We thank Google internship and Student Researcher programs (NPS), Research to Prevent Blindness Stein Innovation Award, Wu Tsai Neurosciences Institute Big Ideas, NIH NEI R01-EY021271, NIH NEI R01-EY029247, and NIH NEI P30-EY019005(EJC), NSF IGERT Grant 0801700 (NB) and NIH NEI F31EY027166 (CR), NSF GRFP DGE-114747 (NB and CR) for funding this work.

Methods

Recordings

Preparation and recording methods are described elsewhere[1,2,8]. Briefly, eyes were enucleated from terminally anesthetized macaque monkeys (*M. Mulatta* or *M. Fascicularis*) used by other experimenters in accordance with institutional guidelines for the care and use of animals. Immediately after enucleation, the anterior portion of the eye and the vitreous were removed in room light. The eye was stored in darkness in oxygenated Ames' solution (Sigma, St. Louis, MO) at 33°C pH 7.4. Segments of isolated or RPE-attached peripheral retina (approximately 3mm x 3mm, taken from 6-15mm temporal equivalent eccentricity[8]) were placed flat, RGC side down, on a planar array of 512 extracellular microelectrodes arranged in an isosceles triangular lattice. The electrode spacing was 60μm in each row, and the array covered a rectangular region measuring 1800 μm x 900 μm. While recording, the retina was perfused with Ames' solution (31-36°C; typically 32 °C for RPE attached and 34°C for RPE isolated dissections),

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.14.431169; this version posted February 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

bubbled with 95% O_2 and 5% CO_2, pH 7.4. Voltage signals on each electrode were bandpass filtered (80Hz - 2kHz), amplified, and digitized at 20 kHz [1].

A custom spike-sorting algorithm was used to identify and segregate spikes from distinct cells[1]. Briefly, candidate spike events were detected using a threshold on each electrode, and voltage waveforms on the electrode and nearby electrodes in the 4ms period surrounding the time of the spike were extracted. Candidate neurons were identified by clustering the waveforms using a Gaussian mixture model. Candidate neurons were retained only if the assigned spikes exhibited a 1 ms refractory period and had a stable firing rate for the entire duration of recording. Duplicate spike trains were identified by temporal cross-correlation and removed. For each cell, the autocorrelation of the recorded spike train was computed and normalized by its value at zero time lag.

Visual stimuli and cell type identification

Visual stimuli were delivered using the optically reduced image of a CRT monitor refreshing at 120 Hz and focused on the photoreceptor outer segments. The optical path passed through the transparent plug and Ames' solution or through the mostly transparent electrode array and the retina. The relative emission spectrum of each display primary was measured with a spectroradiometer (PR-701, PhotoResearch) after passing through the optical elements between the display and the retina. The total power of each display primary was measured with a calibrated photodiode (UDT Instruments). The mean photoisomerization rates for the cone photoreceptors were estimated by computing the inner product of the primary power spectra with the spectral sensitivity of each cone type, and multiplying by the effective collecting area of primate cones (~0.6 μm^2)[26,27], resulting in photoisomerization rates of approximately 800–2200, 800–2200, 400–900 for the long-, middle- and short-wavelength sensitive cones, respectively. The stimulus pixel size on the retina was either 41.6 microns (8 monitor pixels), 52 microns (10 monitor pixels) or 83.2 microns (16 monitor pixels). A new white noise frame was drawn at refresh rates of 60 Hz or 30 Hz. The pixel contrast (difference between the maximum and minimum intensities divided by the sum) was 96% for each display primary, with mean intensity of 50%. The white noise stimulus either modulated the three display primaries independently, or coherently, at each spatial location.

In each recording, RGCs were classified into distinct types using properties of the spatial receptive field and response time course obtained from the spike-triggered average (STA) stimulus [68,28]. A two-dimensional Gaussian fit to the spatial receptive field was used for determining the center location[8]. All analyses used recordings with stable firing rates and nearly complete tiling of ON and OFF parasol cell receptive field mosaics. For Figure 3, only recordings that also had nearly complete ON and OFF midget cell mosaics were used. For model fitting, both the visual stimulus and spike times were binned at 8.33ms (120Hz), and the visual stimulus was upsampled to 8 monitor pixels, resulting in a common 80x40 pixel grid across recordings.

Linear Nonlinear Poisson model

The Linear Nonlinear Poisson (LNP) model consists of a linear spatio-temporal filter followed by a point nonlinearity[6]. A filter that is separable in space and time was used[8], which is equivalent to a cascade of a spatial filter and a temporal filter. These filters were estimated for each cell in each recording as follows. First, the STA was computed by averaging the stimulus preceding spikes, over all pixel locations and 250ms (30 frames at 120Hz) prior to the spike. Next, the spatial filter was computed by choosing the STA frame with the single largest pixel magnitude. The spatial filter was restricted to a rectangular window around the receptive field. The receptive field was defined as the set of pixels with absolute magnitude greater than 2.5σ, contiguous with the strongest pixel, where σ is the robust standard deviation[29] of pixels

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.14.431169; this version posted February 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

in the STA, an estimate of the measurement noise. Next, the temporal filter was identified by averaging the time course of all pixels in the receptive field. Finally, the output nonlinearity was estimated by fitting a 5th order polynomial to the relationship between the observed responses and the generator signal, which was computed by filtering the stimulus with the estimated spatial and temporal filters[6].

For describing variation in spatial nonlinearities, subunit models were fitted using spike triggered clustering d[7]. Briefly, neural responses were modeled by passing the stimulus through multiple linear filters (subunits), followed by an exponential nonlinearity, and summation over the filter outputs. The temporal filtering of each subunit was assumed to be identical to the time course of the STA. The spatial filters were inferred by soft-clustering the collection of stimuli preceding a spike. To assess the degree of subunit nonlinearity, the log-likelihood of a model with five subunits was compared to that of a model with one subunit (which reduces to an LNP model).

Neural network model

A convolutional neural network was used to predict RGC responses across multiple recordings simultaneously. Below, the model architecture and the fitting procedure are described in detail.

Notation

The model $f(S, \alpha_i, C_i)$, takes as its input the visual stimulus S, recording-specific information about the collection of recorded cells C_i, and the recording-specific manifold location α_i, and yields as its output the predictions for recording-specific response R_i.

The recent history of the visual stimulus is given by $S \in R^{d_x \times d_y \times d_z}$, where $\boldsymbol{R}$ is the set of real numbers, $d_x \times d_y$ are the spatial dimensions (80 x 40), and $d_t$is the number of time bins (30). Stimuli presented at different spatial or temporal resolution were upsampled or downsampled to these dimensions.

The recording specific manifold location is given by $\alpha_i \in R^n$, where $n$is the manifold dimensionality.

The recording specific information about recorded cells is given by $C_i = \{x(c), y(c), t(c)\}_{c=1}^{c=|C_i|}$, where each cell c is described by its receptive location $(x(c), y(c))$ in the $d_x \times d_y$ visual space and its cell type $t(c)$. For models with only two cell types, $t(c) \in \{0, 1\}$, for ON and OFF parasols respectively. For models with four cell types, $t(c) \in \{0, 1, 2, 3\}$, corresponding to ON parasol, OFF parasol, ON midget, and OFF midget cell types.

The responses are given by $R_i \in {Z_+}^{|C_i|}$, where Z_+ denotes non-negative integers and $|C_i|$ is the collection of cells in recording i. Responses were binned at the same resolution as the stimulus (8.33ms).

Model architecture

The model $f(S, \alpha_i, C_i)$ passes the visual stimulus S through a multilayered convolutional neural network, with each layer consisting of a convolution (stride 1), retina-specific normalization and a point-wise (softplus) nonlinearity (see Figure 1B). The model output is Poisson firing rate. This firing rate is used to predict the responses R_i. The number of channels and filter sizes are chosen by cross-validation, as described below. Recording-specific normalization and challenges associated with predicting responses for varying numbers of cells across recordings are also given below.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.14.431169; this version posted February 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Recording-specific normalization is inspired by previous work[30], in which a translationally-invariant affine transformation of the layer activations adapts the model to each recording. The scale and shift coefficients for this affine transform are determined linearly from the manifold locationα_i. Let $\hat{a}(x,y,l,t)$be the activation after convolution at location x,y in the channel l of layer t. First, the mean μ and standard deviation σ across samples in a batch are computed, and used to calculate normalized activations: $\tilde{a}(x,y,l,t) = \frac{\hat{a}(x,y,l,t)-\mu}{\sigma}$. Next, using the manifold location α_i, a learned affine transform determines the desired mean ($\tilde{\mu} = P\alpha_i + p$)and standard deviation ($\tilde{\sigma} = Q\alpha_i + q$)for each layer. Finally, the normalized activations are transformed to give recording-specific activations $a(x,y,l,t) \leftarrow \tilde{a}(x,y,l,t)\widetilde{\sigma_{l,t}} + \widetilde{\mu_{l,t}}$. Note that the retina-specific scales and shifts are the same for each location in visual space, preserving the translational invariance of convolutional networks and reflecting the homogenous response properties of the RGCs belonging to a single type.

A potential challenge is that the number of recorded neurons, and hence the number of outputs of $f(.)$, is variable across recordings. To address this issue, the model predicts multiple response maps, one for each cell-type, with the same spatial dimensions as the visual stimulus. The response for each cell is read off from its cell location in the response map of the corresponding cell type. Specifically, $f(.)$outputs $m_i(x,y)$, which corresponds to firing rate map of cell-type i, and for a cell with type $t(c)$and centered at $x(c), y(c)$, the Poisson firing rate is given by $m_{t(c)}(x(c), y(c))$.

Model fitting

Estimation of recording-specific parameters $(\alpha_i)$and the shared parameters are performed by maximizing the log-likelihood of observed responses, summed across all the cells, recordings and stimuli. This is performed by stochastic gradient descent, where at each step, a randomly sampled batch of stimuli and corresponding responses from a particular recording are used to update the shared and the corresponding recording-specific parameters. The batch size was 250 and the updates were performed using the Adam[31] update algorithm with learning rate of 0.1. For each training retina, the first 4 min of white noise data were used for testing and the remainder was used for training. The duration of the stimulus varied from 15-90 min (median 30 min) across experiments. A model with 4 layers, 3 x 3 or 1 x 1 filter size, 64 channels per layer and a 15 dimensional manifold was chosen based on cross validation and used for subsequent analysis (see Figure 1B for architecture).

Variation of neural coding on the manifold

The following steps were used to test if the manifold captured variations in neural response properties across recordings. First, the manifold direction that was maximally correlated to the variations of a particular response feature was identified by linear regression. Second, recordings were projected along this direction, and the Spearman rank correlation with the response property was measured. Statistical significance was measured with a permutation test, where the null distribution was generated by permuting the recordings with the manifold locations fixed. In Figure 2, the Spearman rank correlation and its statistical significance was measured for the first principal component projection of various response properties such as receptive field size (ON: 0.78, p<0.0001 ; OFF: 0.76, p<0.0001), time course (ON: 0.83, OFF: 0.85; p < 0.0001), output non-linearity of the LNP model (Spearman rank correlation for ON: 0.92; OFF: 0.92; p < 0.0001) and normalized auto-correlation (ON: 0.97, OFF: 0.94, p < 0.0001). The interdependence between response properties was either measured directly in raw data using Spearman rank correlation, or with the angle between the corresponding manifold directions.

Differences between male and female recordings

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.14.431169; this version posted February 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

For analysis of gender differences, only the subset of recordings (102) from one species (M. Mulatta) were used. First, the separation between male and female recordings was measured by computing the d' value of the projection of the two distributions onto the difference in the means. The d' value observed (~1.8) indicated that the gender based differences were not large on an individual basis. Second, a bootstrap test was performed to test whether the mean locations of the male and female recordings were statistically distinguishable. The distance between the mean manifold locations of male and female retinas was measured and compared to a null distribution of distances generated by resampling (with replacement) of the manifold locations. The null distribution was fitted with a normal distribution and the significance level was measured as the probability mass greater than the observed distance in data. Because multiple preparations were frequently recorded from the same animal, a hierarchical variant of this bootstrap test was performed, in which the resampling was performed according to the hierarchical structure of the data[12], by first sampling an animal and then sampling the manifold location of one of the recordings from that animal, both with replacement. Hierarchical bootstrap is more conservative and biased towards accepting the null hypothesis[12]. Mean manifold locations for male and female retinas were significantly different ($p<10^{-6}$ for bootstrap and $p<0.05$ for hierarchical bootstrap). Identical tests were applied for assessing male-female differences in firing rate and the speed of temporal filtering ($p<0.01$ for bootstrap and $p<0.05$ for hierarchical bootstrap for both quantities).

Invariance of neural coding on the manifold

The ability of the manifold to preserve previously reported invariances of the neural code was tested as follows. First, random manifold locations were sampled by perturbing the learned locations of training retinas with a Gaussian noise of standard deviation equal to their median nearest-neighbor distance. Second, a ~800 sec long white noise stimulus was sampled, and ON and OFF parasol firing rate maps were computed using the neural network, which was adapted using the manifold locations. Third, the Poisson firing rates for 200 cells (100 of each type) with random receptive field locations were read off from the firing rate maps. Finally, the cell responses were sampled, and used to estimate a Linear-Nonlinear Poisson model, which served as an interpretable summary of neural encoding captured by the manifold location. Comparison of average receptive field size and the zero crossing time of the temporal filter revealed known invariances between ON and OFF parasols (Figure 2J, K).

Estimation of the manifold location of a previously unseen retina

By fixing the shared parameters after learning, and estimating the recording-specific representation on the manifold, the trained model was adapted to predict responses in a new, previously unseen recording. Based on the amount of data available, several methods can be employed to identify the manifold location (Figure 4). These methods are described below in detail.

Averaging: When no data about the new retina are available, the simplest approach is to average the locations of all the retinas used for training.

Approximation: This is similar to averaging, but only using the subset of training retinas with similar response properties as the new retina. For Figure 4, locations of five training retinas with the most similar autocorrelation function were used for approximation.

Optimization: When light response data are available, the manifold location $\alpha_i$was determined by Bayesian inference. Bayesian inference combines a Gaussian prior $(P_{prior}(\alpha) \sim N(\mu_{prior}, \sigma_{prior}))$ over manifold locations determined from the training retinas and the likelihood $(P(R_i^t \mid S^t, \alpha_i))$ of stimulus-

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.14.431169; this version posted February 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

response data for the new retina. The posterior $P(\alpha_i|\ \{S^t, {R_i}^t\ \})$was maximized using gradient ascent (learning rate 0.1) :

$$\alpha_i^* = argmax_\alpha P(\alpha_i|\ \{S^t, {R_i}^t\ \}) = argmax_\alpha \sum_t log\ P({R_i}^t\ |\ S^t, \alpha_i) + log\ P_{prior}(\alpha_i)$$

Discrimination task: Simulation of manifold location estimation in a blind person implanted with a retinal prosthesis was performed using a discrimination task. For a given visual stimulus, the task involves using the implanted retinal prosthesis to reproduce responses corresponding to two manifold locations and the subject selects the response that yields perception most closely matching a verbally described stimulus. Multiple rounds of this task are used to update the posterior on manifold locations.

The discrimination task was simulated under the assumption that the perceptual difference of the responses generated by hypothetical retinas at two manifold locations $\alpha_1$and α_2 for a stimulus S is equal to the KL-divergence between the corresponding response distributions $P(R|S, \alpha_1)$ and $P(R|S, \alpha_2)$. Given α_{true}as the true underlying manifold location, the blind person's feedback$Y(\alpha_{true}, \alpha_1, \alpha_2) = 0$ if

$$D_{KL}(P(R|S, \alpha_1)\ ||\ P(R|S, \alpha_{true}) \leq D_{KL}(P(R|S, \alpha_2)\ ||\ P(R|S, \alpha_{true})$$

and $Y = 1$ otherwise. For simplicity, sampled responses were used to compute an unbiased estimate of the KL-divergence :

$$D_{KL}(P(R|S, \alpha_1)\ ||\ P(R|S, \alpha_{true}) \approx \sum_{R_i \sim P(R|S,\alpha_1)} log\ \left(\frac{P(R_i|S,\alpha_1)}{P(R_i|S,\alpha_{true})}\right).$$

Hence, the posterior over manifold location after t steps of the task is given by:

$$P_{posterior}(\alpha\ |\ \{S_t, \alpha_{1,t}, \alpha_{2,t}, Y_t\}_{t=1}^{t=T}) \propto \prod_{t=1}^{t=T} P(Y_t\ |\ S_t, \alpha_{1,t}, \alpha_{2,t})\ P_{prior}(\alpha)$$

where the prior is estimated from training retinas as a Gaussian distribution: $P_{prior}(\alpha) \sim N(\mu_{prior}, \sigma_{prior})$.

In the simulations, the visual stimulus Sconsisted of letters of English alphabet, flashed for $100\ ms$and preceded and succeeded by $50\ ms$of gray screen. At each step, a Gaussian approximation of the posterior $P_{posterior}(\alpha) \sim N(\mu_{posterior}, \sigma_{posterior})$ was maintained, and updated using Monte-Carlo sampling. In summary, the steps for the t^{th}iteration of the algorithm are as follows :

1. Sample symmetric $\alpha_{1,t}$, $\alpha_{2,t}$ around posterior mean: $\alpha_{1,t} \sim P_{posterior}(\alpha)$; $\alpha_{2,t} = 2\mu_{posterior} - \alpha_{1,t}$.
2. Sample an English letter and a target stimulus S_t.
3. Sample responses $R_{1,t} \sim P(R|S_t, \alpha_{1,t})$; $R_{2,t} \sim P(R|S_t, \alpha_{2,t})$.
4. Get patient feedback $Y(\alpha_{true}, \alpha_{1,t}, \alpha_{2,t})$, based on an estimate of the KL divergence using sampled responses $R_{1,t}$, $R_{2,t}$.
5. Update the posterior of plausible manifold locations.
    a. Sample N retina locations $\alpha_i \sim P_{posterior}(\alpha)$ for $i \in [1, \cdots, N]$.
    b. For the set of sampled manifold locations, find the subset that matches user feedback, i.e., with $Y(\alpha_i, \alpha_{1,l}, \alpha_{2,l}) = Y(\alpha_{true}, \alpha_{1,l}, \alpha_{2,l})$for all $l\ (= 1, \ldots, t)$previous steps. Let this subset of be$\{\tilde{\alpha}_j\}$.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.14.431169; this version posted February 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

c. Update posterior distribution with $\mu_{posterior} = < \tilde{\alpha_j} >$and $\sigma^2{}_{posterior} = < \alpha_j \alpha_j^T > - \mu_{posterior} \mu_{posterior}^T$.

For results shown in Figure 4F, α_{true} was set as the result of optimizing the manifold location using light-evoked responses. In the simulations, the posterior distribution converged in ~20 steps, suggesting that the low dimensional manifold can be used for efficiently identifying the expected neural code in a blind person. However, the amount of noise in the simulation is probably lower compared to what would be encountered in practice, leading to a larger number of steps to identify the true manifold location and may perhaps require changes to the estimator of KL divergence and the method to update the posterior of α.

Data/Code Availability

The data/code that support the findings of this study are available from the corresponding author upon reasonable request.

References

1. Litke, A. M. *et al.* What does the eye tell the brain?: Development of a system for the large-scale recording of retinal output activity. *IEEE Trans. Nucl. Sci.* **51**, 1434–1440 (2004).
2. Frechette, E. S. *et al.* Fidelity of the ensemble code for visual motion in primate retina. *J. Neurophysiol.* **94**, 119–135 (2005).
3. Field, G. D. *et al.* Spatial properties and functional organization of small bistratified ganglion cells in primate retina. *J. Neurosci.* **27**, 13261–13272 (2007).
4. Rhoades, C. E. *et al.* Unusual Physiological Properties of Smooth Monostratified Ganglion Cell Types in Primate Retina. *Neuron* **103**, 658–672.e6 (2019).
5. Greschner, M. *et al.* A polyaxonal amacrine cell population in the primate retina. *J. Neurosci.* **34**, 3597–3606 (2014).
6. Chichilnisky, E. J. A simple white noise analysis of neuronal light responses. *Network* **12**, 199–213 (2001).
7. Shah, N. P. *et al.* Inference of nonlinear receptive field subunits with spike-triggered clustering. (2020) doi:10.7554/eLife.45743.
8. Chichilnisky, E. J. & Kalmar, R. S. Functional asymmetries in ON and OFF ganglion cells of primate retina. *J. Neurosci.* **22**, 2737–2747 (2002).

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.14.431169; this version posted February 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

9. McIntosh, L. T., Maheswaranathan, N., Nayebi, A., Ganguli, S. & Baccus, S. A. Deep Learning Models of the Retinal Response to Natural Scenes. *Adv. Neural Inf. Process. Syst.* **29**, 1369–1377 (2016).
10. Batty, E. *et al.* Multilayer Recurrent Network Models of Primate Retinal Ganglion Cell Responses. in *International Conference on Learning Representations* (2017).
11. Wassle, H. & Boycott, B. B. Functional architecture of the mammalian retina. *Physiological Reviews* vol. 71 447–480 (1991).
12. Saravanan, V., Berman, G. J. & Sober, S. J. Application of the hierarchical bootstrap to multi-level data in neuroscience. *bioRxiv* 819334 (2019) doi:10.1101/819334.
13. Soto, F. *et al.* Efficient Coding by Midget and Parasol Ganglion Cells in the Human Retina. *Neuron* **107**, 656–666.e5 (2020).
14. Kling, A. *et al.* Functional Organization of Midget and Parasol Ganglion Cells in the Human Retina. 2020.08.07.240762 (2020) doi:10.1101/2020.08.07.240762.
15. Pillow, J. W. *et al.* Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature* **454**, 995–999 (2008).
16. Schneidman, E., Brenner, N., Tishby, N., van Steveninck, R. R. de R. & Bialek, W. Universality and Individuality in a Neural Code. in *Advances in Neural Information Processing Systems 13* (eds. Leen, T. K., Dietterich, T. G. & Tresp, V.) 159–165 (MIT Press, 2001).
17. Cahill, L. Why sex matters for neuroscience. *Nat. Rev. Neurosci.* **7**, 477–484 (2006).
18. Poplin, R. *et al.* Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat Biomed Eng* **2**, 158–164 (2018).
19. Choleris, E., Galea, L. A. M., Sohrabji, F. & Frick, K. M. Sex differences in the brain: Implications for behavioral and biomedical research. *Neurosci. Biobehav. Rev.* **85**, 126–145 (2018).
20. Color Blindness | National Eye Institute. https://www.nei.nih.gov/learn-about-eye-

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.14.431169; this version posted February 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

health/eye-conditions-and-diseases/color-blindness.

21. Bensmaia, S. J. & Miller, L. E. Restoring sensorimotor function through intracortical interfaces: progress and looming challenges. *Nat. Rev. Neurosci.* **15**, 313–325 (2014).
22. Shenoy, K. V. & Carmena, J. M. Combining decoder design and neural adaptation in brain-machine interfaces. *Neuron* **84**, 665–680 (2014).
23. Chestek, C. A. *et al.* Long-term stability of neural prosthetic control signals from silicon cortical arrays in rhesus macaque motor cortex. *J. Neural Eng.* **8**, 045005 (2011).
24. Beyeler, M., Rokem, A., Boynton, G. M. & Fine, I. Learning to see again: Biological constraints on cortical plasticity and the implications for sight restoration technologies. *J. Neural Eng.* **14**, 051003 (2017).
25. Golub, M. D. *et al.* Learning by neural reassociation. *Nat. Neurosci.* **21**, 607 (2018).
26. Angueyra, J. M. & Rieke, F. Origin and effect of phototransduction noise in primate cone photoreceptors. *Nat. Neurosci.* **16**, 1692–1700 (2013).
27. Schnapf, J. L., Nunn, B. J., Meister, M. & Baylor, D. A. Visual transduction in cones of the monkey Macaca fascicularis. *J. Physiol.* **427**, 681–713 (1990).
28. Field, G. D. & Chichilnisky, E. J. Information processing in the primate retina: circuitry and coding. *Annu. Rev. Neurosci.* **30**, 1–30 (2007).
29. Rousseeuw, P. J. & Croux, C. Alternatives to the Median Absolute Deviation. *J. Am. Stat. Assoc.* **88**, 1273–1283 (1993).
30. Dumoulin, V., Shlens, J. & Kudlur, M. A Learned Representation For Artistic Style. *International Conference on Learning Representations* (2017).
31. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. (2014).