Supporting Information

Palmer et al. 10.1073/pnas.1506855112

Bound Calculation

Solving Eq. **6** in *Materials and Methods* is difficult, in general. However, in the present case the underlying signal x_t is Gaussian, and so analytic approaches are possible, following ref. 1. If we consider the past history of the trajectory to be a vector $\mathbf{X}_p$, and the future trajectory to be a vector $\mathbf{X}_f$, then we can define two probability distributions, $P(\mathbf{X}_f)$ and $P(\mathbf{X}_f|\mathbf{X}_p)$. Both of these are Gaussian, so they are described completely by the means and covariance matrices. Let us call the covariance matrices of the two distributions Σ and Σ_c, respectively. Then, as explained in ref. 1, a crucial role is played by the matrix $\mathcal{M} \equiv \Sigma_c \Sigma^{-1}$ and its eigenvalues $\lambda_1, \lambda_2, \cdots$ (in decreasing order). The underlying parameters of the stimulus—Γ, ω, and D in Eqs. **4** and **5**—determine these eigenvalues, but the functional relationship is complicated and (for us) not very illuminating; one can also estimate the matrix $\mathcal{M}$ numerically from a long simulation of the trajectory x_t.

We are trying to calculate the bounding curve in Fig. 2*A*, which determines the maximum possible value of I_{future} given a fixed value of I_{past}. Adapting the results of ref. 1 to this case, we can write the following:

$$I_{\text{past}} - I^*_{\text{future}} = \frac{n_I}{2} \log \left(\prod_{i=1}^{n_I} (1-\lambda_i)^{\frac{1}{n_I}} + e^{\frac{2I_{\text{past}}}{n_I}} \prod_{i=1}^{n_I} \lambda_i^{\frac{1}{n_I}} \right), \qquad \textbf{[S1]}$$

where the index n_I defines the cutoff on the number of eigenvalues used to compute the bound segment. The bound curve is composed of several segments with increasing numbers of eigenvalues added as our information about the past trajectory increases. This bound is continuous, smooth, and concave. For the particular dynamics defined by Eqs. **4** and **5** in the main text, the bound curve was obtained for each Δt, by computing the covariance of the position and velocity in a long trajectory generated by these dynamics.

Linear–Nonlinear Model

To test whether simple receptive field properties of retinal ganglion cells can account for the saturation of the bound on the predictive information, we constructed linear–nonlinear (LN) model neurons based on our data. In LN models, the probability of spiking is an instantaneous, nonlinear function of a linearly filtered version of the sensory input. In the case of retinal ganglion cells that we study here, the inputs are the image or contrast as a function of space and time, $s(\vec{x}, t)$. Thus, if we write the probability per unit time of a spike (the firing rate), we have the following:

$$r_{\text{LN}}(t) = r_0 g(z), \qquad \textbf{[S2]}$$

where r_0 sets the scale of firing rates, $g(z)$ is a dimensionless nonlinear function, and

$$z(t) = \int_0^t d\tau \int d^2x \, f(\vec{x}, \tau) s(\vec{x}, t-\tau); \qquad \textbf{[S3]}$$

the function $f(\vec{x}, \tau)$ is the receptive field.

It is a theorem that, if we deliver stimuli that are drawn from a Gaussian white noise ensemble, then

$$f(\vec{x}, \tau) \propto \left\langle s(\vec{x}, t-\tau) \delta(t - t_{\text{spike}}) \right\rangle, \qquad \textbf{[S4]}$$

where t_{spike} is the time of a spike and $\langle \cdots \rangle$ denotes an average over a long movie. As described in Fig. 4*D* ("checker"), we have done experiments with randomly flickering checkerboards that approximate Gaussian white noise down to the frame time of 1/30 s and the pixel size of 40 × 40 μm. We used these data to estimate receptive fields by reverse correlation (Eq. **S4**) and used cubic spline interpolation to extend these receptive fields down to a resolution of $\Delta\tau = 1/60$ s.

If we choose the scale of firing rates to match the size of the time bins, $r_0 = 1/\Delta\tau$, then the function $g(z)$ is exactly the probability of a spike in a bin given that the output of the filter is z, that is, $g(z) = p(\text{spike}|z)$. Experimentally, we can sample the value of z in all of the bins with spikes, which allows us to estimate $p(z|\text{spike})$, and then Bayes' rule tells us that

$$g(z) \equiv p(\text{spike}|z) = \frac{p(z|\text{spike}) \cdot p(\text{spike})}{p(z)}, \qquad \textbf{[S5]}$$

where $p(z)$ is the distribution of z across the whole experiment. Nonlinearities derived in this way from the checkerboard experiments are very well fit by logistic functions,

$$g(z) = \frac{g_0}{1 + e^{-\gamma(z-\theta)}}. \qquad \textbf{[S6]}$$

Note that g_0 is the maximum spike probability, and hence is bounded by 1; γ defines a gain, and θ, a threshold for the responses.

If we take the LN model derived from the random checkerboard stimuli and use it to produce neural responses to the moving-bar stimulus, the predictive information carried by the neurons is drastically wrong. However, this is not surprising, because even the mean firing rates are wrong. This is because retinal ganglion cells adapt to match the scale of their nonlinear input/output relations [summarized here as $g(z)$] to the dynamic range of inputs. To give our model a chance of working, then, we should let the parameters in Eq. **S6** be adjusted to match some average properties of the neural response to the moving bar. We chose to match the mean spike rate,

$$\bar{r} = \frac{1}{T} \int_0^T dt \, r_{LN}(t), \qquad \textbf{[S7]}$$

where T is the duration of the stimulus movie, and the information that individual spikes provide about the (past) stimulus (2),

$$I_1 = \frac{1}{T} \int_0^T dt \, \frac{r_{LN}(t)}{\bar{r}} \log_2 \left[\frac{r_{LN}(t)}{\bar{r}} \right]. \qquad \textbf{[S8]}$$

To match the data, we found in all cases that we need to set $g_0 = 1$, its maximum possible value, and then matching I_1 and $\bar{r}$ fixed the values of the gain γ and the threshold θ.

Fig. S1*A* shows an example of the LN model for one neuron. In this cell, as in most, we found that the receptive field $f(\vec{x}, \tau)$ was separable into spatial and temporal components, as shown. Fig. S1*B* shows that, for all of the cells in our sample, we have been able to match the values of $\bar{r}$ (left) and I_1 (right). Having

built a population of model neurons, we can now perform the same analysis that we did for the real neurons: select groups of cells, compute the information that patterns of spiking and silence provide about both the past and future of the stimulus in the "common future" experiment, and then plot the results for the best of these groups in the information plane, as in Fig. 3*D* of the main text. Results are summarized in Fig. S1*C*.

Results shown in Fig. S1*C* reveal that the LN model fails to recapitulate the near-optimal behavior of the real data. All groups fall away from the bound determined by $\Delta t = 1/60$ s, the delay between the current response and the onset of the common future. Importantly, when we compute information about the future, we assume that the future starts now (as in real life!) and do not make any allowances for processing delays. We could, instead, compare the performance of the LN model with bounds calculated assuming that there is a delay between past and future, so that $\Delta t^* = \Delta t + t_{\mathrm{delay}}$. The bound for Δt^* is shown by the dashed curve in Fig. S1, where we have chosen $t_{\mathrm{delay}} = 117$ ms, comparable to the delay one might estimate from the peak of the information about position in Fig. 1*B*, or from the structure of the receptive fields themselves in Fig. S1*A*. Interestingly, the model neurons do come close to this less restrictive bound.

Stimulus Information in σ^{out} for One Group

To find the optimal downstream predictor neuron, we exhaustively sampled all possible Boolean transforms of the input. All partitions of four-cell input patterns into spike and no-spike responses (excluding the one-half that transform no input, $0000\ldots0$, into spiking output yielding high firing rates), and their resulting predictive information about the future input are shown in Fig. S2*A*. The density of a scatter plot of the 65,536 points representing a particular predictor neuron's output firing rate and predictive information are shown. Each point was convolved with a Gaussian and summed with other points. The plot is normalized to have a peak of 1. Not surprisingly, predictive information increases with output firing rate. These rates, however, remain within a biologically plausible range.

In Fig. 5*C* of the main text, we plotted the average stimulus information as a function of predictive information about the future inputs for 200 downstream cells. In Fig. S2*B*, we plot the same information for one group of four retinal input cells and all possible binary output rules that govern predictor neuron firing (density is represented in the same way as in *A*). The rate measured here is the firing rate of a predictor neuron with a particular rule, given the observed sequence of input spikes. This shows that capturing more of the predictive information in the patterns of retinal ganglion cell activity also allows the hypothetical predictor neuron to convey greater information about the visual stimulus: building better local predictions leads to better stimulus coding.

Feature Selectivity in Predictor Neurons

In Fig. 5*D* and Fig. S2 *C*–*E*, we show four kinds of stimulus feature selectivity that emerged in our analysis of optimized predictor neurons, constant velocity detection, velocity detection (regardless of direction of motion), position refinement, and time shift of the best position estimate toward the future. In Fig. S3, we show two more examples for each of these features.

We see that the predictor neurons respond to certain aspects of stimulus motion that might be useful for prediction—motion at constant speed but either direction (Fig. S2*C*) and long epochs of constant velocity (Fig. 5*D* and Fig. S3 *B* and *F*), followed by a reversal. These long constant-velocity epochs are predictive of reversals, as dictated by the equation of motion we defined for the bar's trajectory. After long excursions in one direction, the spring constant coupling the bar to the center of the visual world is engaged and pulls the bar back toward center. When a predictor neuron fires in response to this constant motion, its spiking could be used downstream to predict reversal.

The estimate of the bar position in the predictor neurons is better (lower variance) than in any one of its inputs (Fig. S2*D*), showing that optimizing for predicting inputs leads to a refinement in the stimulus estimate. Also, these downstream cells have interesting spike-triggered average stimuli when they are optimized (for the same inputs) to make predictions farther into the future (Fig. S2*E*): the time of sharpest stimulus discrimination moves closer to the time of a spike in the downstream cell when it is more predictive of its inputs farther in the future. This again shows that predictable components of the retina's firing map back to predictable components of the stimulus, but also that processing lags can be circumvented by coding for predictable firing in response to a moving stimulus.

Thus, searching for efficient representations of the predictive information in the state of the retina itself drives the emergence of motion estimation.

1. Chechik G, Globerson A, Tishby N, Weiss Y (2005) Information bottleneck for Gaussian variables. *JMLR* 6:165–188.

2. Brenner N, Strong SP, Koberle R, Bialek W, de Ruyter van Steveninck RR (2000) Synergy in a neural code. *Neural Comp* 12(7):1531–1552.

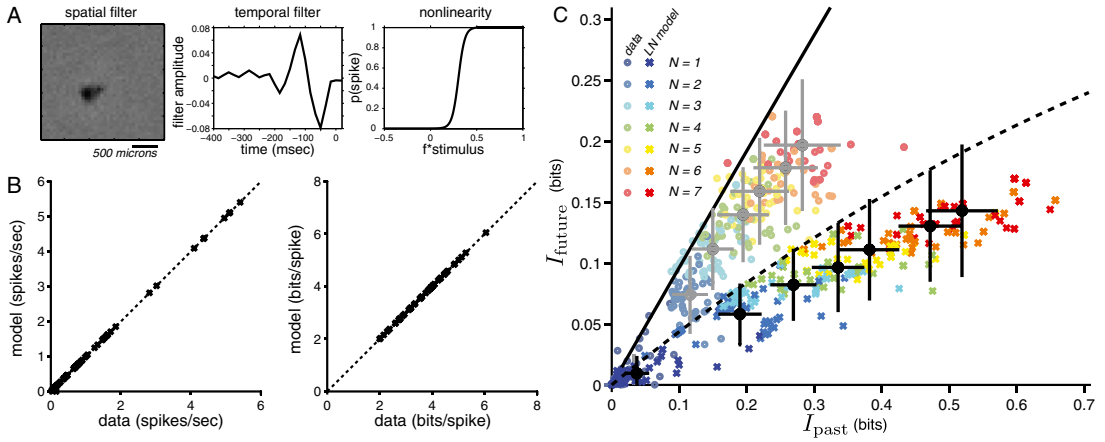

Fig. S1. An LN model cannot reproduce the saturation of the bound on predictive information observed in retinal data. (*A*) An example fit to a cell in our dataset. (*B*) Adjusting parameters of the nonlinearity reproduces both the mean firing rate and single spike information present in all cells in our dataset. (*C*) Populations of LN neurons (x's) modeled in this way do not saturate the bound on predictive information (solid curve; Δt = 1/60 s) as seen in the real data (small circles; less saturated coloring; same data points as shown in Fig. 3 in the main text). The dashed curve represents the bound on predictive information about the future possible with $\Delta t = 8\Delta\tau \sim 133$ ms. Larger populations of model neurons fall away from this curve.

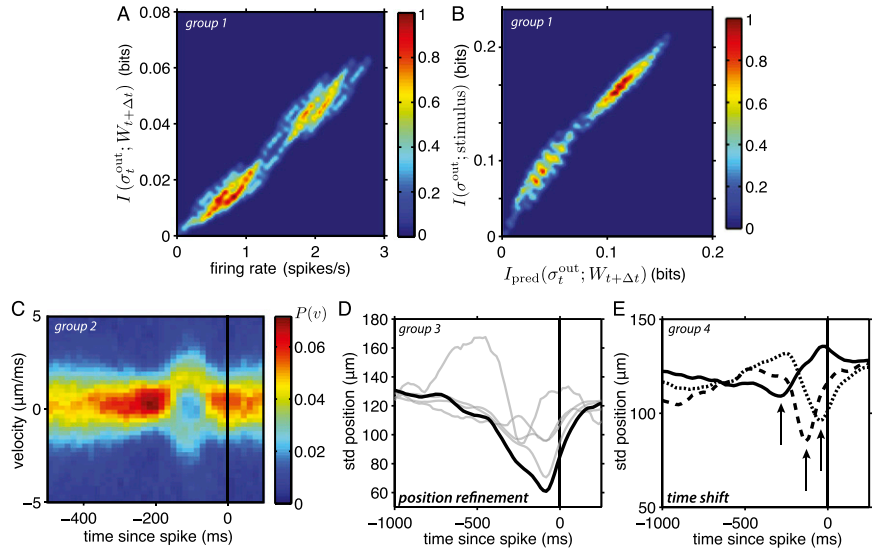

Fig. S2. Increasing word–word predictive information enhances stimulus coding for predictor cells. (*A*) Predictive information, $I(\sigma_t^{\mathrm{out}}; W_{t+\Delta t})$, captured by all possible mappings $w_t \to \sigma^{\mathrm{out}}$, as a function of the average firing rate of σ^{out}, for one particular four-cell input group. A scatter plot of the data are represented as a density plot. The scale bar on the *Right* indicates density, normalized to a peak of 1. (*B*) The stimulus information for all downstream rules for the group in *A*, also plotted as a density plot. (*C*) Distribution of stimuli that give rise to a spike in an optimized predictor neuron, for a second particular group of four cells in response to the moving bar stimulus ensemble in Fig. 1 of the main text. (*D*) For a third group of inputs, the SD of bar positions triggered on a spike in the predictor neuron (black) or on spikes in the individual input neurons (gray). Δt = 1/60 s in *A*–*D*. (*E*) For a fourth group of cells, the SD of bar positions conditional on a predictor neuron spike varies as we optimize for predictions with delays of Δt = 1/30 s (solid curve), Δt = 1/15 s (dashed curve), and Δt = 1/10 s (dotted curve).

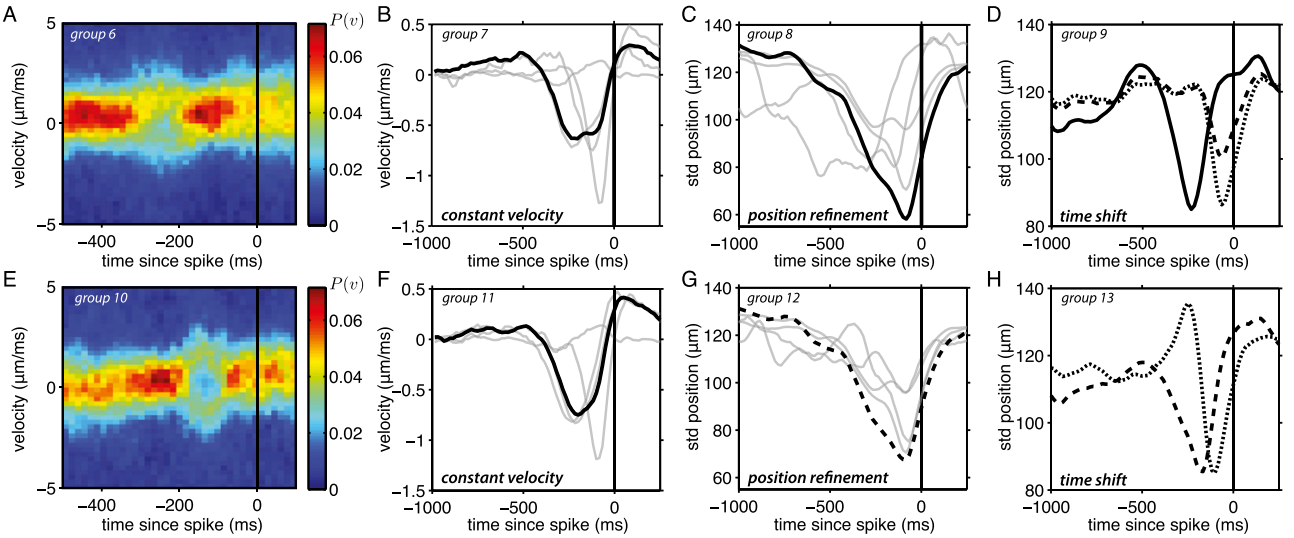

Fig. S3. Spike-triggered average stimuli for firing in optimized predictor neurons. (*A* and *E*) Distribution of stimuli that give rise to a spike in an optimized predictor neuron, for two particular groups of four cells in response to the moving-bar stimulus ensemble in Fig. 1 of the main text; the predictor neuron selects for motion at constant speed, with relatively little direction selectivity. $\Delta t = 1/60$ s. (*B* and *F*) The average velocity triggered on a spike of the predictor neuron; the predictor neurons select for a long epoch of constant velocity. Light gray lines show the spike-triggered average stimuli for the input cells. (*C* and *G*) The SD of bar positions triggered on a spike in the predictor neuron (black) or on spikes in the individual input neurons (gray); predictor neurons provide a more refined position estimate. (*D* and *H*) The SD of bar positions conditioned on a predictor neuron spike varies as we optimize for predictions with delays of $\Delta t = 1/30$ s (solid curve), $\Delta t = 1/15$ s (dashed curve), and $\Delta t = 1/10$ s (dotted curve); optimizing predictions can compensate for latencies. Not all groups were sampled exhaustively at every Δt; the results corresponding to the delay denoted by what would be a solid curve in *H* was not computed.