Supporting Information

Sederberg et al. 10.1073/pnas.1710779115

SI Materials and Methods

Dynamics of the Moving Bar. The dynamics for the position (x_t^b) and velocity (v_t^b) of the moving-bar stimulus follow the equations for a stochastic damped harmonic oscillator in the overdamped regime:

$$x_{t+\Delta t}^b = x_t^b + v_t^b \Delta t$$
$$v_{t+\Delta t}^b = [1 - \Gamma \Delta t] v_t^b - \omega^2 x_t^b \Delta t + \xi_t \sqrt{D \Delta t}.$$

The parameters are $\Gamma = 20\ \mathrm{s}^{-1}$ and $\omega = 2\pi \times (1.5\ \mathrm{s}^{-1})$, generating dynamics that are slightly overdamped: Without the stochastic kicks ξ_t, bar position decays back to the center position. The time step $\Delta t = 1/60$ s, matching the frame rate of the movie. The parameter $D = 2.7 \times 10^6 \mathrm{pixel}^2/\mathrm{s}^3$ was set so that bar position ranges across the screen extent.

Information Calculation Methods. Word–word internal predictive information is the information the binary word $\boldsymbol{x}_t$ at time t provides about the word $\boldsymbol{x}_{t'}$ at time $t' = t + dt$ for some temporal offset dt (1–3):

$$I(\boldsymbol{X}_t; \boldsymbol{X}_{t'}) = \sum_x P_X(\boldsymbol{x}_t)\, P_X(\boldsymbol{x}_{t'} | \boldsymbol{x}_t) \log_2 \frac{P_X(\boldsymbol{x}_{t'} | \boldsymbol{x}_t)}{P_X(\boldsymbol{x}_{t'})}.$$

Readout predictive information is the mutual information of the perceptron activity y_t at time t and the word $\boldsymbol{x}_{t+dt}$ at time $t' = t + dt$: $I(Y_t; \boldsymbol{X}_{t'})$. Word–word information is symmetric with dt [$I(\boldsymbol{X}_t; \boldsymbol{X}_{t+dt}) = I(\boldsymbol{X}_t; \boldsymbol{X}_{t-dt})$], but perceptron-word information is not [$I(Y_t; \boldsymbol{X}_{t+dt}) \neq I(Y_t; \boldsymbol{X}_{t-dt})$]. The shorthand "predictive information" is the perceptron-word information for $dt = 1/60$ s (the temporal bin size). Information was computed using CDMEntropy, a Bayesian entropy estimator for binary vector data (4). Information estimated through this method was consistent (within 1%) with the information estimated through finite-size scaling methods (5, 6), with computation run times that were significantly less. This was verified for all calculation types (word and perceptron readouts of internal predictive information driven by natural-movie stimuli, word and perceptron readouts of internal predictive information driven by moving-bar stimulus, and word and perceptron readouts of stimulus information from $t = -100$ ms to $t = +50$ ms in 33-ms increments) on a subset of sampled sets, in which we also estimated the uncertainty in the information calculation as the SD of the bootstrapped samples of half the data divided by $\sqrt{2}$, following ref. 5. Readout information estimate error bars were less than 0.001 bits per time bin (equivalently, 0.06 bits per s) and are generally smaller than marker size in the figures. Word information estimates were less than 0.003 bits per time bin (0.18 bits per s) are also smaller than marker size in the figures.

Definition of Similarity Metric Between Readout Rules. The similarity between two readouts is the fraction of time bins with one or more input spikes for which the readout functions produced the same output. Each readout corresponds to a set of output rules ($L_i = L(\boldsymbol{x}^{(i)})$), representing the readout response (0, 1) to each of the 16 possible input words ($\boldsymbol{x}^{(1)} = 0000; \boldsymbol{x}^{(1)} = 0001; \ldots . \boldsymbol{x}^{(16)} = 1111$) (Fig. 3*D*). For each learned readout, similarity between the learned rules L_i and the optimal rules O_i is quantified as

$$\frac{1}{1 - p(1)} \sum_{i \neq 1} p(i) \delta_{L_i, O_i},$$

where δ_{L_i, O_i} is 1 if $L_i = O_i$ and otherwise 0 and $p(i)$ is the probability of observing word i ($\boldsymbol{x}^{(i)}$).

Sampling the Space of Possible Readouts of Internal Predictive Information (Fig. S1). To judge the efficiency of a learned perceptron for a particular set of cells we need to compare the predictive information encoded by the learned readout function to that of the set of possible single-bit readout functions. For a subset of groups of four cells, internal predictive information during the fish movie was computed exhaustively for each of the 2^{15} possible (binary) readout functions (for examples see Fig. S1 *A* and *B*). For these groups of four cells we computed the maximum readout predictive information as a function of readout firing rate for the exhaustively sampled set of readout functions ("exhaustive sampling," Fig. S1*B*, purple bound) and for the perceptron readouts alone ("perceptron hull," Fig. S1*B*, red bound). We noted that while there were places that the hull computed from the exhaustive sample was higher than the perceptron hull, over most of the firing rate the two hulls are indistinguishable, despite there being >32,000 readouts sampled for one and <150 for the other. We tested how well the bound could be estimated using samples of randomly drawn readouts, and we found that it is possible to estimate the upper bound on readouts with relatively small samples of all possible readouts (Fig. S1*C*). The estimate of the bound on predictive information was better for middle to high firing rates, which is the range over which the most informative readouts are typically found (see Fig. S1 *A*, *B*, *D*, and *E* for examples). At midrange firing rates (half of the maximum firing rate), a subsample containing a randomly selected fraction (1%, or 328 of 32,768; Fig. S1*C*, red) of all possible readouts produced an estimate of the predictive information bound that was 98% of the true bound calculated from the set of exhaustively sampled readouts (Fig. S1*C*). (Across sets of four cells, the average SE of the hull estimate efficiency over repeatedly drawn subsamples was 3%.) Sampling fewer random readouts (than 1%, or ∼300) did not lead to a good estimate of the bound (Fig. S1*C*, blue), although sampling only 149 linear readouts (perceptrons, positive weights only) rather than random ones did give a good estimate of the bound (Fig. S1*B*, red vs. purple).

For sets larger than four cells it is not feasible to sample even 1% of readout functions, and it is possible that a significant gap between the predictive information of the sampled subset of readout functions and the optimal readout of predictive information grows rapidly as the number of cells in the group increases. Thus, another method is needed to estimate the bound on predictive information of readouts at a given firing rate. An efficient way to estimate this bound for larger sets is to permit probabilistic readout functions and optimize over the probability of responding to a particular word. This reduces the search space for the optimal rule for a set of N cells from possible readout functions to parameters that can be learned through gradient descent on the output-firing-rate-constrained mutual information. The results of comparing the sampled subset of readout functions to the optimal hull are shown in Fig. S1*B* (set of four) and Fig. S1 *D*–*F*. For the set of four cells, the probabilistic hull is an upper bound on the exhaustively sampled readouts' predictive information (Fig. S1*B*). For large sets, we reasoned that, because sampling perceptrons was more efficient than randomly sampled

readouts at estimating the readout information hull for sets of four, perceptrons may get closer to the bound for larger sets as well. Thus, predictive information was computed for a small sample of linear readout (perceptron) functions (~1,000 readouts; black points). From this sample, the perceptron hull (orange) was calculated. Finally, the probabilistic hull (blue) was optimized (Fig. S1*D*, an example set of 7; Fig. S1*E*, an example set of 10). The perceptron hull approaches the probabilistic hull over most of the firing rate regime. We compared the perceptron hull value to the probabilistic value across the range of firing rates, normalizing firing rates by the maximum for each set before averaging across sets. Across sets, the estimated perceptron hull was typically greater than 95% of the optimized perceptron hull. Note that the probabilistic readout, which has variable probabilities for each possible input word, is not easily mapped to a one-step biologically plausible readout function, which was the focus of our current work. Therefore, to calculate the efficiency of learned readouts we compare them to the sampled perceptron hull, which represents the best possible single-bit readouts and is close (within 5–10%) to the optimized probabilistic readout function.

SI Results

Learning Under Other Spike-Timing–Dependent Rules (Fig. S2). We simulated several variations of STDP: pair (see main text) and triplet-spike rules, as well as homeostatic variations of each rule.

The triplet rule depends on the timing of one presynaptic and two postsynaptic spikes, such that potentiation is modulated by the postsynaptic interspike interval Δ_{ISI}:

$$\Delta w_t^{(i)} = \varepsilon\left(y_t x_t^{(i)} \exp\left(-\Delta_{ISI}/\tau_y\right) - \alpha_{LTD} y_{t-1} x_t^{(i)}\right).$$

In the triplet rule, potentiation only occurs if there was recently a postsynaptic event at the time of the prepost pairing. We simulated learning under the triplet rule for $\tau_y = 115$ ms and $\tau_y = 167$ ms. This rule was defined following Gjorgjieva et al. (7), in which equivalent parameters $\tau_y = 114$ ms and $\alpha_{LTD} = 0.92$. Here $\alpha_{LTD} = 0.9$, which we found to be where simulation results usually generated weight vectors with both zero and nonzero weights.

For homeostatic learning rules, the sum of weights was constrained to equal 0.75N, where N is the size of the cell group, at each learning time step.

For each set of cells we generated a random set of initial conditions, and for each of those initial conditions four separate learning simulations were carried out, employing each learning rule in turn. We then directly compared the readouts learned under each rule. We visualize this in a matrix, showing for each initial condition of each cell set which of the four learning rules led to the readout with the highest predictive information (Fig. S2*A*). Rows are ordered by the predictive information of the full cell set. Most of the matrix is blue, indicating that the pair rule generally led to the highest predictive information in the final readout. For 80% of tested cell sets the pair rule was optimal for 80% (or more) of initial conditions (Fig. S2*B*). After the pair rule, the next most common optimal rule was the triplet rule with homeostasis (purple, Fig. S2 *A* and *B*). We next examined the fraction of recovered predictive information for each simulation under an optimal learning rule. For each cell set there is a particular pattern of recovered information across all tested initial weight patterns. One difficulty in analyzing these patterns is that the labeling of the four input cells is arbitrary, making comparisons across groups difficult. We noted that the pattern of each set was often most strongly coupled to one of the initial input weights. For each set we identify the dominant initial input weight based on the correlation of final predictive information with the initial weight strength, and we then order the initial conditions for that set by the initial value of the dominant weight (Fig. S2*C*). (The same ordering was adopted in Fig. S2*A*.) The strong patterning in Fig. S2*C* induced by this ordering shows that having high predictive information in the final learned readout is often dependent on starting with a particular pattern of input weights, specifically with a strong weight in the "dominant" input. However, there is still variability across the cell sets: For many other sets, with both small (0.04) and large (0.14) predictive information, nearly all initial conditions reach the same final state, rather than being strongly determined by initial condition. Finally, we examined the distribution of predictive information learned when each rule was optimal (Fig. S2*D*). Readouts learned under the pair rule have higher median predictive information, but the difference is small: less than 0.01 bits per 16 ms. In summary, the optimal learning rule depends on details of the cell sets and initial conditions, but most often the simple, pair-spike learning rule produced the most efficient readouts.

We observe that a range of reasonable rules find near-optimal readouts. Interestingly, a simple pair-STDP rule outperforms a triplet-STDP rule for the majority of cell sets sampled. Triplet-STDP, implemented as a prepost-post rule (7, 8), only potentiates input synapses in a window following a postsynaptic spike, which means that the first spiking pattern after a long pattern of silence will not potentiate the input weights. The observation that the pair-based rule finds more predictive readouts than the triplet-based rule suggests that the first spiking pattern after a long pattern of silence is important for prediction in cells receiving visual inputs from the retina. This aligns with the observation that population silent periods in retinal recordings carry significant amounts of information about the stimulus (9).

Learned Readout and Optimal Readouts: Relationship Between Structural and Information Efficiency (Fig. S3). In the main text we show the overall distribution of learned readout efficiency (Fig. 3*B*) and of the similarity of the structures of learned readouts and optimal readouts (Fig. 3*D*). Here we show the relationship between learned readout efficiency and structure for sets of 4, 7, and 10 (Fig. S4) cells. While there is a relationship between structural similarity to the optimal rule and efficiency of the readout, many readouts have a high degree of predictive information efficiency with low structural similarity to the optimal rule, especially for sets of 7 and 10 cells (Fig. S4 *B* and *C*). In Fig. S4 we look at properties of the readout space that explain why finding the exact structure of the optimal readout was not necessary for finding highly efficient readouts of predictive information.

Importance of Perceptron Structure to Perceptron Efficiency (Fig. S4). Learning the precise structure of the optimal perceptron is only important if information readout depends on finding that optimal structure. Based on the learning results for sets of 4–10 cells, this may not be the case. To see how an efficient readout of predictive information can be found without exactly matching the optimal readout structure we examined the relationship between the range of readout predictive information and predictive information of the full cell set across many sets of cells (Fig. S4). The perceptron performance range (Fig. S4*A*) was defined as the maximum, over all firing rates, of the difference between the best readout at or below a firing rate and the worst readout at or above a firing rate. This reflected how important finding the optimal structure is: If almost all readouts were within a 20% performance range, then readout structure counted for at most 20% of the readout efficiency. Each set had a particular pattern of predictive information across readouts, but several trends emerged across sampled cell sets. The widest ranges of readout performance were observed when the internal predictive information of the full cell set was smaller (Fig. S4*B*). For some of these cell sets perceptron performance range was 90%: The difference between the optimal readout and the worst readout

was 90% of the total readout word–word information. The range of perceptron performance narrowed to ∼20% for word–word information of 0.4 bits per bin (24 bits per s). The overall efficiency of readouts was highest for the lowest internal information: Some sets with 0.1 bits per bin of internal information had readouts that captured 90% of that information. For sets with 0.4 bits per bin of internal information, optimal readouts captured 50% of the information (Fig. S4*C*). These are most frequently readouts of sets of 10 cells (blue dots). The difference between the best and worst perceptrons was smaller for larger sets, so many perceptrons will tend to have high efficiencies relative to the optimal readout. This is reflected in our observations of the learned perceptrons associated with sets of 7 and 10 cells: High readout efficiency occurs without high similarity to the optimal rule structure. In summary, the decrease in efficiency means that the total information in the group of 7 (or 10) cells could not be compressed to a single, perceptron-readable bit, and that either a nonlinear readout function (i.e., multilayer perceptrons) or multiple readout channels are required to transmit the total information in larger groups of cells.

On the Diversity of Stimulus and Internal Predictive Information Values Across Sampled Sets. The diversity of predictive information values encoded by random groups of input cells (and their subsequent readout perceptrons) reflects the fact that RGCs are responsive to different spatial locations in the visual scene as well as different classes of spatiotemporal stimulus features. Combining some of these cells results in highly predictive groups, while combining others does not. In the sampled population all cells participated in at least one group that was highly predictive, and nearly every cell participated in more than a third of the highly predictive groups (Fig. S5). Some cells were more predictive than others and participated in a larger fraction of highly predictive groups (Fig. S5 *A* and *B*), but no clear subset of cells emerged as the carrier of the majority of the predictive information; contributions to highly predictive groups were spread evenly across all cells. Participating in a large fraction of groups with high stimulus-predictive information correlated with participation in a large fraction of groups encoding a large amount of future activity (internal information) (Fig. S5*C*).

1. Shannon C (1948) A mathematical theory of communication. *Bell Syst Tech J* 27: 379–423, 623–656.
2. Cover TM, Thomas JA (2005) *Elements of Information Theory* (Wiley, Hoboken, NJ).
3. Rieke F, Warland D, De Ruyter Van Steveninck R, Bialek W (1997) *Spikes: Exploring the Neural Code* (MIT Press, Cambridge, MA).
4. Archer EW, Park IM, Pillow JW (2013) Bayesian entropy estimation for binary spike train data using parametric prior knowledge. *Adv Neural Inf Process Syst* 26:1700–1708.
5. Palmer SE, Marre O, Berry MJ, 2nd, Bialek W (2015) Predictive information in a sensory population. *Proc Natl Acad Sci USA* 112:6908–6913.
6. Strong SP, Köberle R, de Ruyter van Steveninck RR, Bialek W (1998) Entropy and information in neural spike trains. *Phys Rev Lett* 80:197–200.
7. Gjorgjieva J, Clopath C, Audet J, Pfister J-P (2011) A triplet spike-timing-dependent plasticity model generalizes the Bienenstock-Cooper-Munro rule to higher-order spatiotemporal correlations. *Proc Natl Acad Sci USA* 108:19383–19388.
8. Pfister J-P, Gerstner W (2006) Triplets of spikes in a model of spike timing-dependent plasticity. *J Neurosci* 26:9673–9682.
9. Schneidman E, et al. (2011) Synergy from silence in a combinatorial neural code. *J Neurosci* 31:15732–15741.

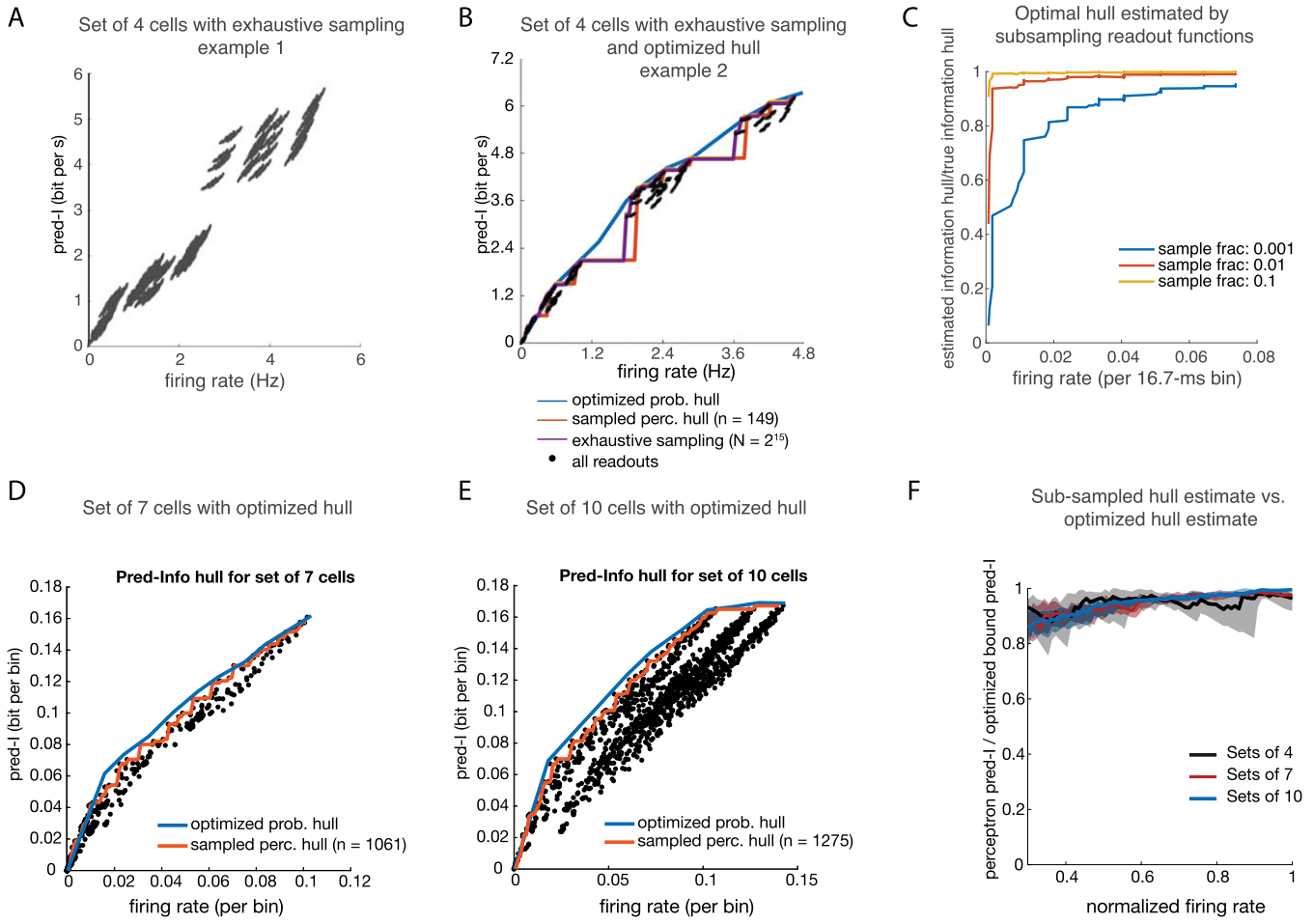

Fig. S1. Sampling the space of possible readouts. (*A*) An example of the exhaustive sampling of readout function for a single set of four cells. There are 32,768 readouts shown. (*B*) A second example of exhaustive sampling of readout functions for a set of four cells (black dots), with information bounds overlaid. The optimized probabilistic hull (blue) is the predictive information for a readout that responds probabilistically to each of the 16 possible words; this function was optimized at multiple firing rates to find the optimal bound on predictive information, a method that can be used for larger sets of cells as well. The sampled perceptron hull (red) is the maximum predictive information of any sampled perceptron readout at or a below a fixed firing rate. The exhaustive sampling hull (purple) is the maximum predictive information of any readout rule, including nonlinear (nonperceptron) readouts, and is occasionally higher than the perceptron readout hull. (*C*) Maximum readout information of randomly drawn subsamples vs. firing rate relative to the true maximum calculated for the exhaustively sampled set. Curves are the average over 200 cell sets from which the predictive information hull was estimated from subsets (0.1%, 1%, and 10%) of all possible readout functions (linear and nonlinear). Sampling 1% (328 readouts, red line) was sufficient to estimate the bound within 5% error across middle to high firing rates. (*D*) An example of the efficiency of the perceptron hull calculations from partial sampling of perceptron (linear) readouts for a single set of seven cells. The perceptron hull calculated from these points (orange) is close to the optimized probabilistic hull (blue). (*E*) Same as *D*, for a set of 10 cells. (*F*) Median (across randomly drawn sets of cells) efficiency of the estimate of the predictive information bound from sampled perceptron readouts relative to the optimized probabilistic hull for sets of 4 (blue), 7 (red), and 10 (yellow) cells for middle (30% of maximum) to high (100% of maximum) firing rate readout regime. Shaded region represents the 17th to 83rd percentile of cell sets sampled. The hull estimated from a limited sample of perceptrons is close to the hull defined by the optimized probabilistic rule, with the values of the firing rate averaged 17th/50th/83rd percentile across sets for each set size as follows: [0.88, 0.95, 0.98], sets of 4; [0.92, 0.95, 0.97], sets of 7; and [0.93, 0.95, 0.97], sets of 10. The jumps in the curve for sets of four cells come from the "clumpier" structure of readout predictive information-firing rates (*A* and *B*), which tend to smear out for larger sets of cells (*D* and *E*). frac, fraction; info, information; perc, perceptron; pred, predictive; pred-I, predictive information; prob, probability.

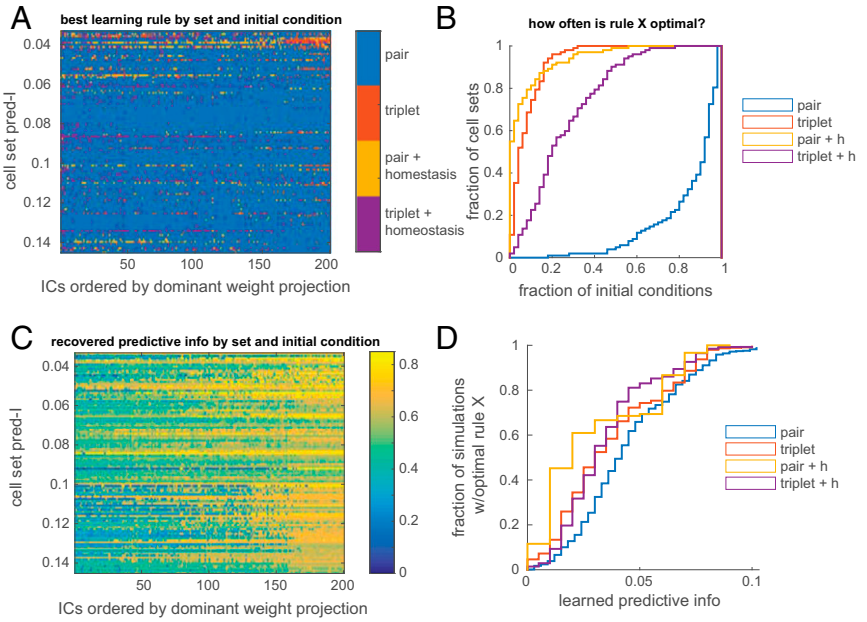

Fig. S2. Best rules for learning efficient readouts. (*A*) For each set of cells (rows) and each initial condition (columns) we show the rule that led to the readout function with the highest predictive information, indicated by color. The pair rule without homeostasis (blue) is most often the best rule and was used in the main text. Rows are ordered by the information of the full cell set, and columns are ordered as described in *Supporting Information*. (*B*) For each rule, we compute the fraction of cell sets for which that rule was optimal for fewer than *x*% of initial conditions. For example, the pair rule was optimal for fewer than 80% of initial conditions in 20% of cell groups; in the remaining 80% of groups, the pair rule was optimal for 80% or more of the sampled initial conditions. The triplet rule with homeostasis was optimal for more than 50% of initial conditions in fewer than 10% of cell groups. (*C*) For each set of cells (rows) and each initial condition (columns), squares are colored by the internal predictive information of the learned readout as a fraction of total cell set internal predictive information. (*D*) Cumulative distribution of the internal predictive information of learned readouts, separated by which learning rule was optimal for that cell set and initial condition. Predictive information of learned readouts was highest for readouts learned under the pair rule (main text) and next-highest for the triplet rules. info, information; pred-I, predictive information.

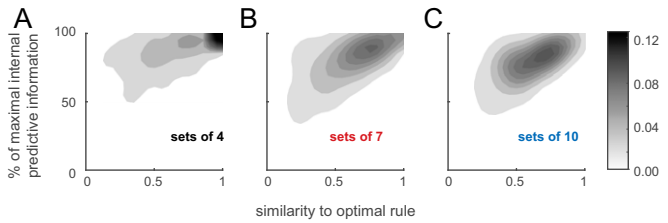

Fig. S3. Relationship between the efficiency of learned readouts and the similarity of learned readout rule to the optimal perceptron readout. (*A*–*C*) Density plot across sampled cell sets and initial conditions of the efficiency of learned readout internal information relative to the optimal readout vs. the similarity of learned readouts to optimal readout rules for sets of 4 (*A*), 7 (*B*), and 10 (*C*) cells. Many readouts have a high degree of predictive information efficiency with low structural similarity to the optimal rule, particularly for sets of 7 and 10 cells.

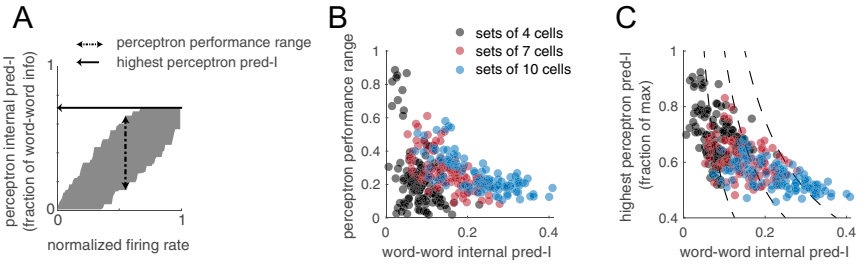

Fig. S4. For larger groups, more readouts are near-optimal, but the total readout efficiency decreases. (*A*) Quantification of the predictive information–firing rate landscape. We characterize each cell group by the highest perceptron predictive information (the maximum fraction of total internal predictive information of any readout, solid black line) and by the perceptron performance range (the maximum range of readout predictive information at fixed firing rate, dashed black line). (*B*) The largest perceptron performance range decreases as total predictive information of the group increases. This means that for cell groups with the largest total predictive information readouts are within 0.2 (as a fraction of the group information) of each other. (*C*) The highest perceptron predictive information is plotted against the total internal information of the group. Dashed lines show curves of constant readout predictive information. Internal predictive information of larger groups and of highly informative groups is less compressible onto a single-bit readout. info, information; pred-I, predictive information.

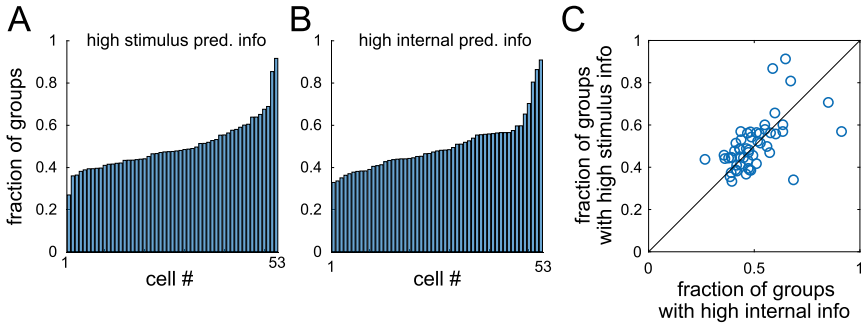

Fig. S5. Every cell is a member of several highly informative cell groups. (*A*) For each of the recorded cells (1–53) we plot the inclusion fraction: the fraction of the randomly sampled 10-cell groups that included this cell that have stimulus-predictive information higher than the median measured over all 10-cell groups. Bars are ordered by inclusion fraction. (*B*) Same as *A*, but for internal predictive information. Bars are ordered by inclusion fraction for internal predictive information. (*C*) Cells that contribute to many high-stimulus-information groups are also likely participate in many high-internal-information readouts. info, information; pred, predictive.