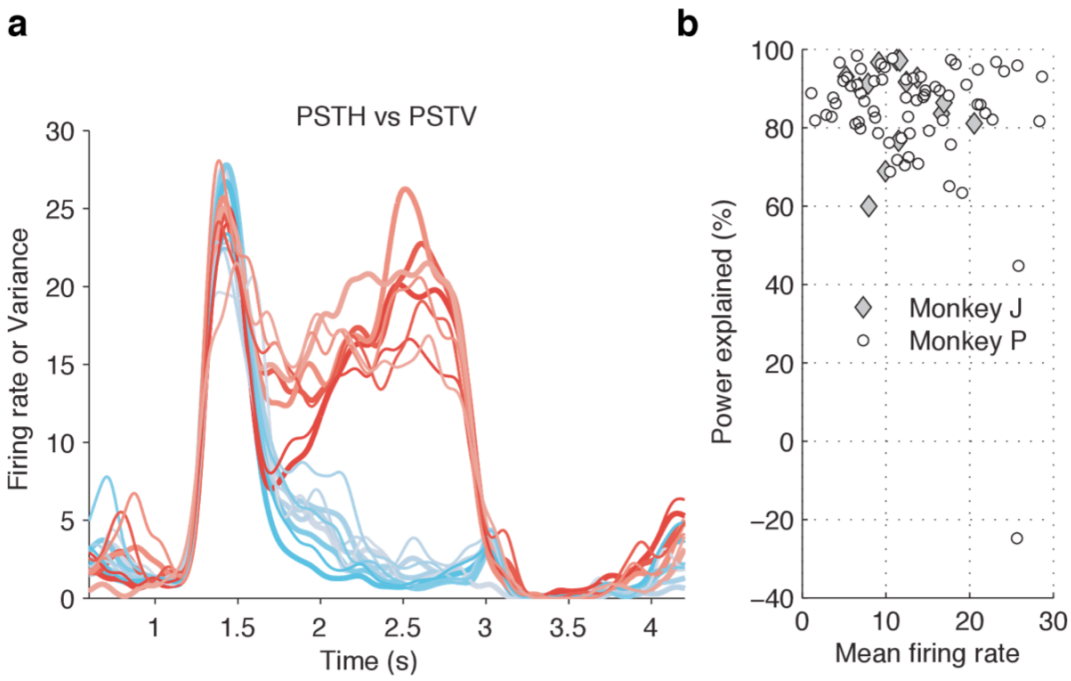
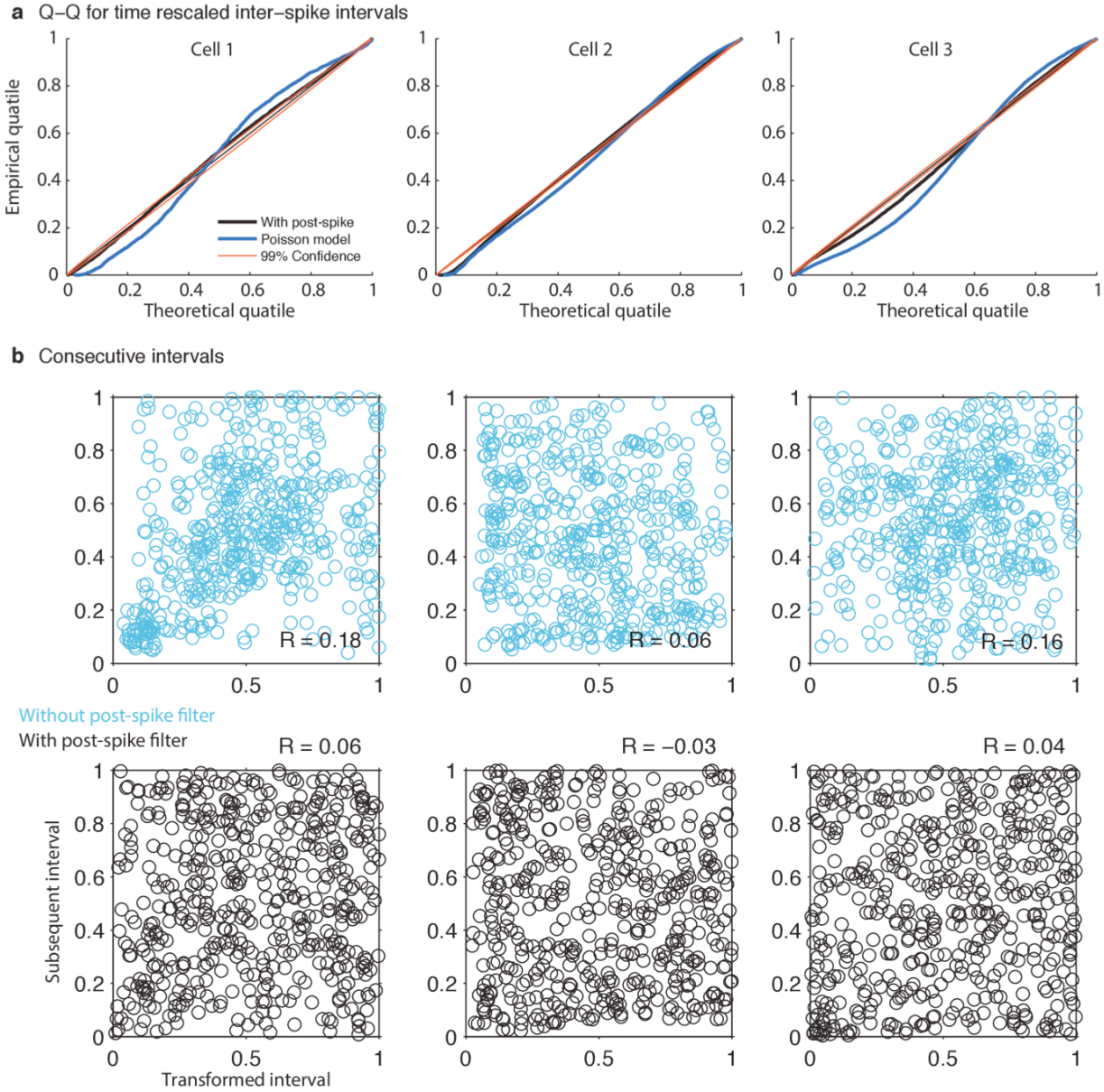


Supplementary Figure 1

Rate model captures variance in spike counts.

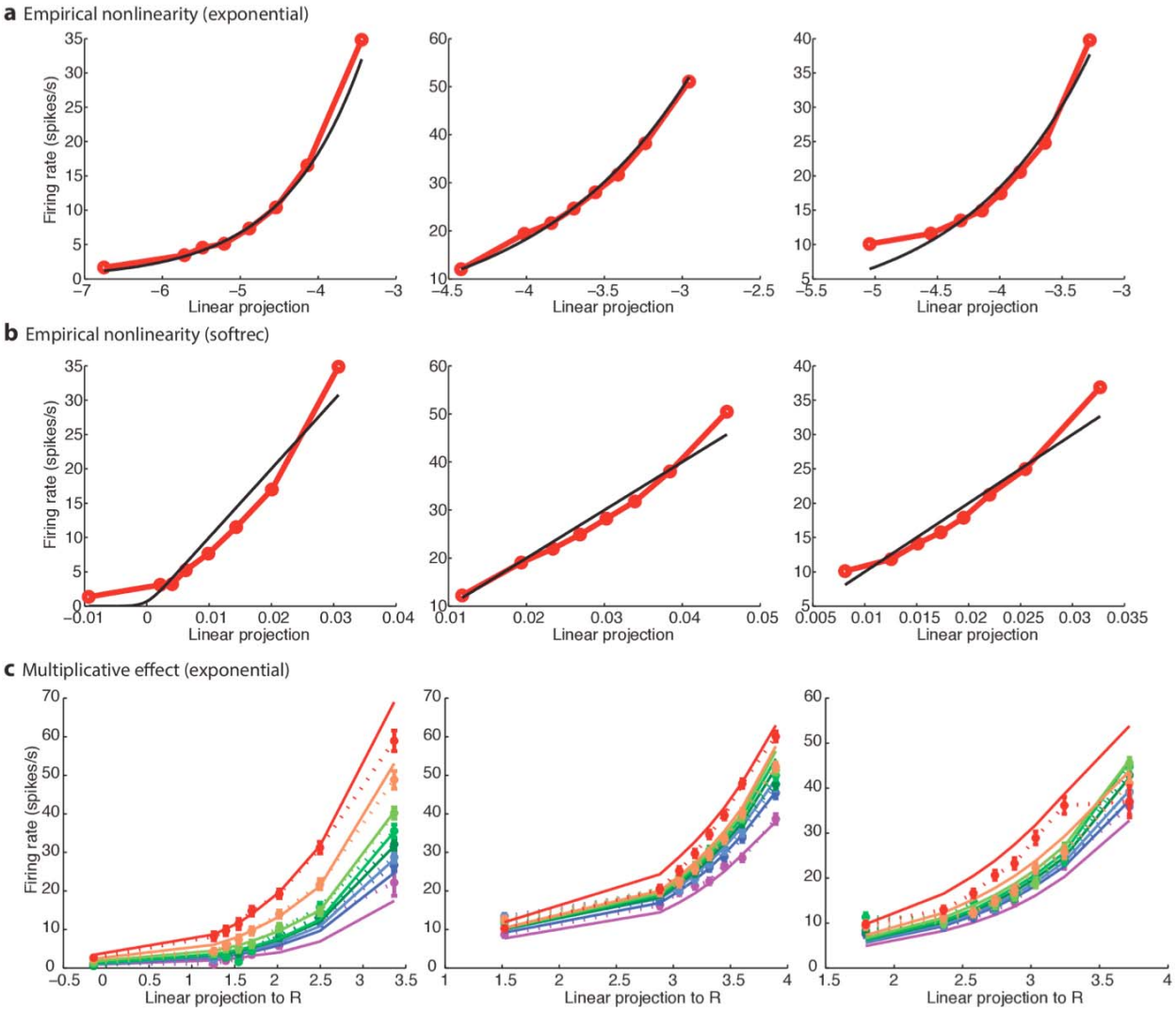
(a) Peri-stimulus time variance (PSTV) histogram (thin line), and PSTH (think line) plotted simultaneously. These are targets-ON trials conditioned on coherence and decision. Both traces are computed with 50 ms boxcar moving average and smoothed with a Gaussian of 50 ms standard deviation. Note that the estimated mean and variance are close to each other, consistent with Poisson variability. **(b)** Population summary of PSTV power explained analogous to **Fig. 3b**. On average, 84.1% of the power is captured by the rate model.



Supplementary Figure 2

Post-spike filter corrects for non-Poisson interval statistics.

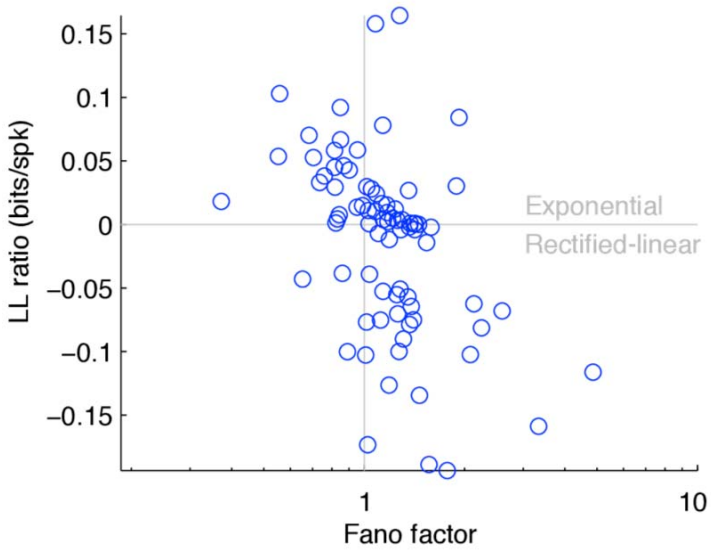
We use the time-rescaling theorem to transform the observed spike trains and check if the transformed interval statistics resulting from post-spike filter are consistent with the Poisson assumption. **(a)** Quantile-quantile plot of time-rescaled interval distribution. Diagonal represents Poisson process model assumptions. **(b)** Subsamples of adjacent time-rescaled inter-spike intervals. Note the uniformity of the intervals resulting from the model with the post-spike filter. The post-spike filter decorrelates the consecutive intervals, which is consistent with the Poisson assumption, indicated by the reduction of correlation coefficient (R) values.



Supplementary Figure 3

Empirical quantification of multiplicative effects.

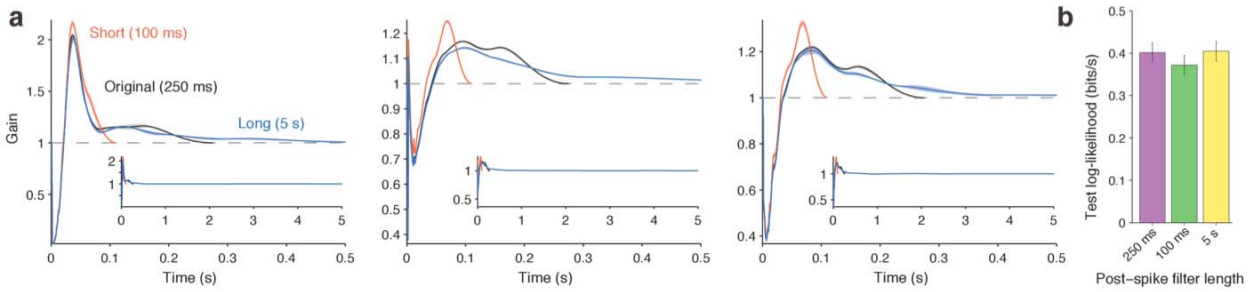
We compare an alternative nonlinearity with $g(x) = 1000 \times \log\left(1 + \exp\left(\frac{x}{1000}\right)\right)$ (soft rectification/threshold linear function). Each column represents a neuron corresponding to **Fig. 5**. **(a)** Exponential function (black) and empirical predicted nonlinearity on test-data (red). Empirical nonlinearities are estimated by first partitioning the net linear output for each time bin into eight equal-quantile buckets, then estimating the expected spike count conditioned on each bucket. The expected means are plotted at the mean net linear drive within each bucket. **(b)** Threshold linear function (black) and empirical predicted nonlinearity on test-data (red). Note the systematic deviation towards an exponential nonlinearity. **(c)** To check for multiplicative interactions, we divided the net linear input into two components, one from saccade-locked kernels and the rest. Then, we compared model prediction to that expected from multiplicative interaction between those two components. Each colored trace represents conditioning on different amounts of contribution from the saccade-locked kernels. Solid line is the true multiplicative effect predicted from the model (with exponential nonlinearity), while dotted lines are empirical rate prediction from the data.



Supplementary Figure 4

Nonlinearity comparison with cross-validated log-likelihoods for exponential and linear rectification.

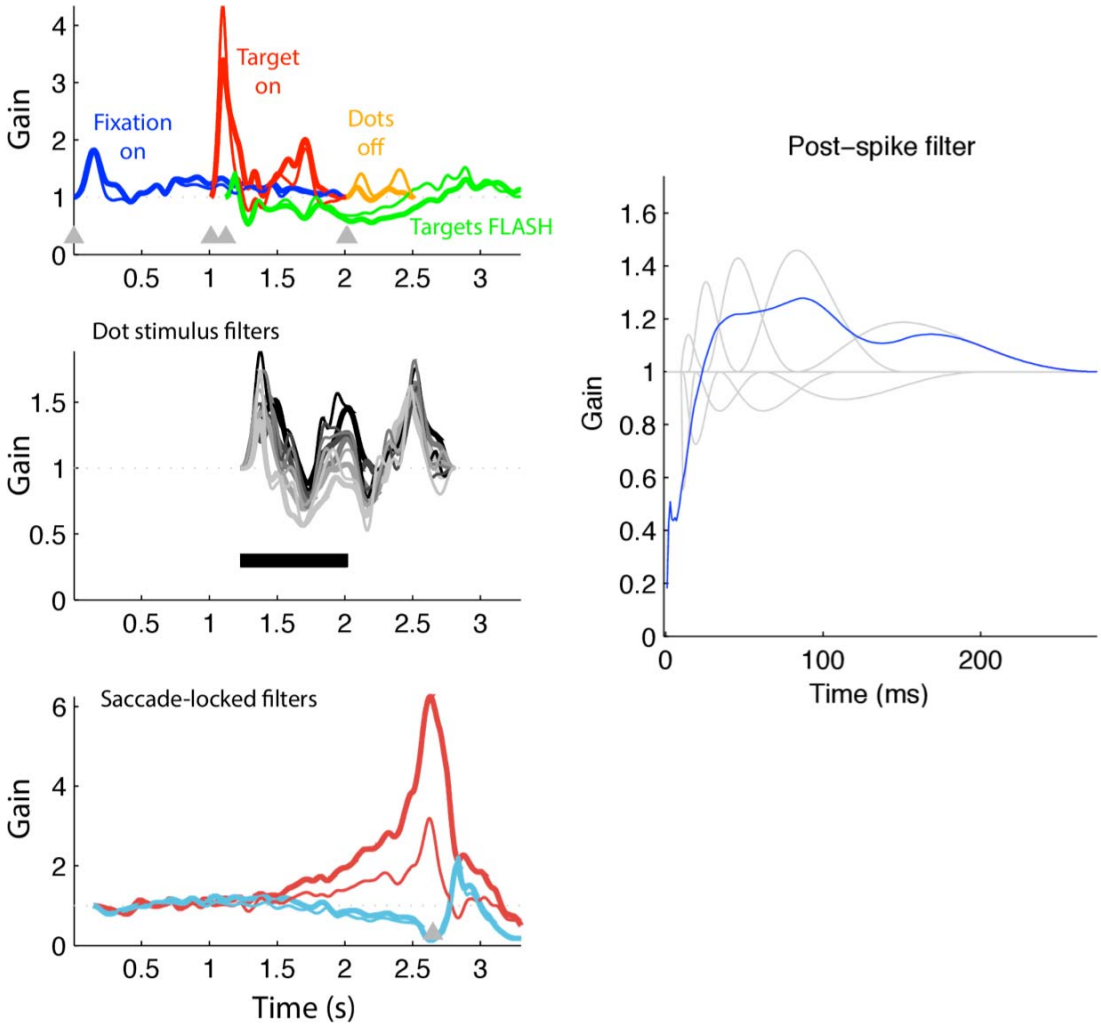
Positive cross-validated log-likelihood difference implies exponential model better fits the data. We used Fano factor as a measure of neural variability, to test for a relation between variability and the quality of the fit with an exponential nonlinearity. There is a noticeable trend (correlation coefficient -0.44) indicating that underdispersed neurons (Fano factor less than 1) are better modeled with an exponential nonlinearity, while overdispersed neurons are often better modeled with linear rectification.



Supplementary Figure 5

Comparing temporal extent of post-spike filters.

Majority of neurons showed a long self-exciting tail in the post-spike filter (**Supplementary Fig. 7**). We quantify how much information the tail contributes to, and test if there is a trial length scale modulation of gain. The post-spike filter used throughout the paper consists of ten 1 ms bins to model fast response components, in addition to ten raised cosine bases stretched logarithmically over 250 ms; here we also consider a model with shorter post-spike filter which has 7 temporal bases over 100 ms, and a model with longer post-spike filter which has 25 temporal bases over 5 s, which is long enough to cover the entire trial. **(a)** Resulting fits for 3 example cells, comparing short/medium/long spans for post-spike filter basis functions. Inset shows 5 seconds scale. **(b)** Spike prediction accuracy comparison across 3 models averaged over 80 neurons. Errorbar indicates standard error. Our "original" model used throughout the paper explained 7.9% more information per spike compared to short filter model ($p < 10^{-10}$), and the long filter model explained 0.7% more information ($p < 0.001$). Thus, longer post-spike effect is negligible, and there is no systematic trial time scale gain modulation.

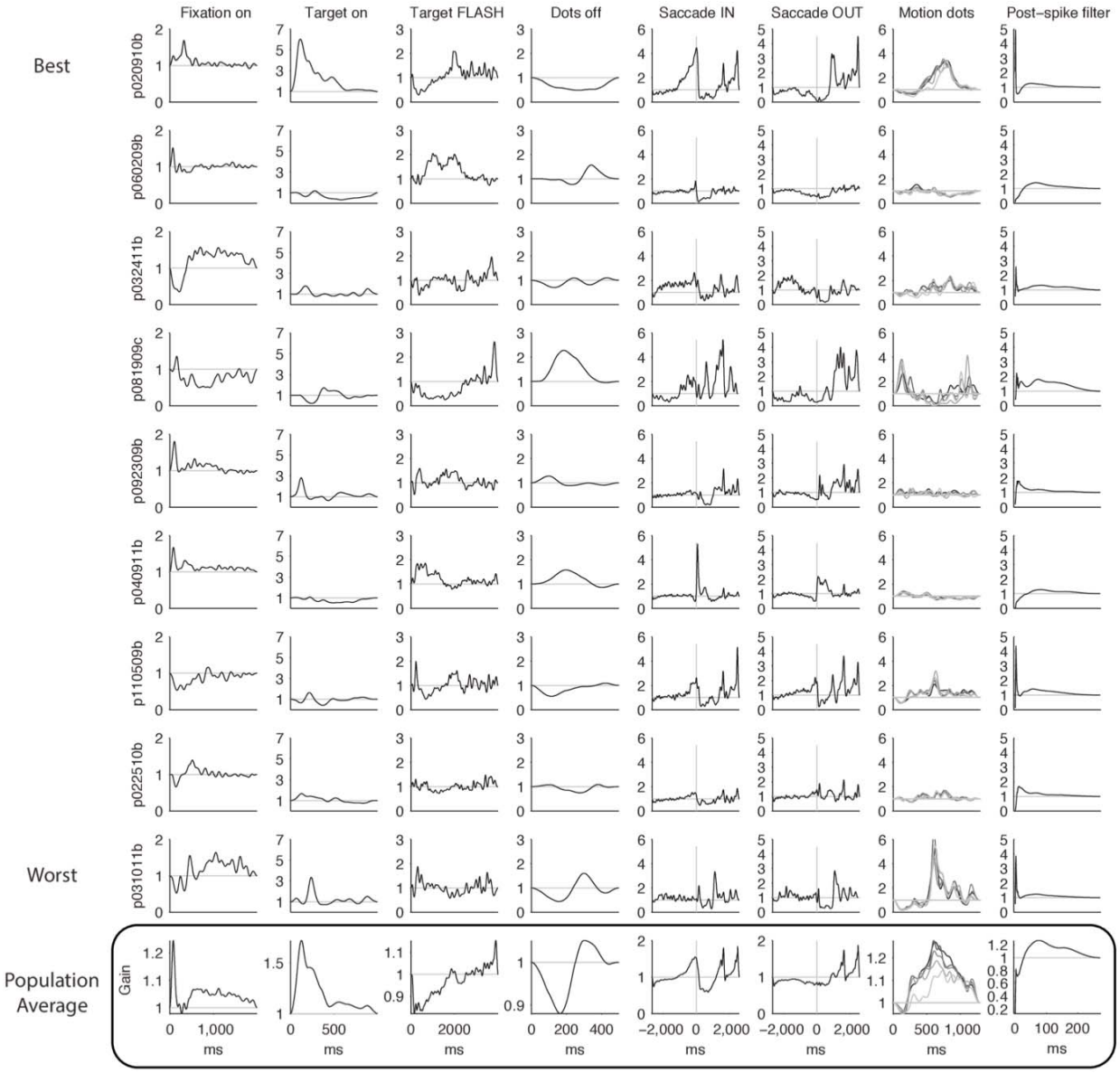


Supplementary Figure 6

Visualization of all kernels.

(Left) Kernels corresponding to each event. Thinner line shows the corresponding weight for the model with post-spike filter. Note that **Fig. 2** is fit with a rank-2 constraint on the motion coherence kernels. Here we show the unconstrained kernel in the middle panel. **(Right)** Post-spike filter (blue), which is represented as a sum of weighted temporal basis functions (gray).

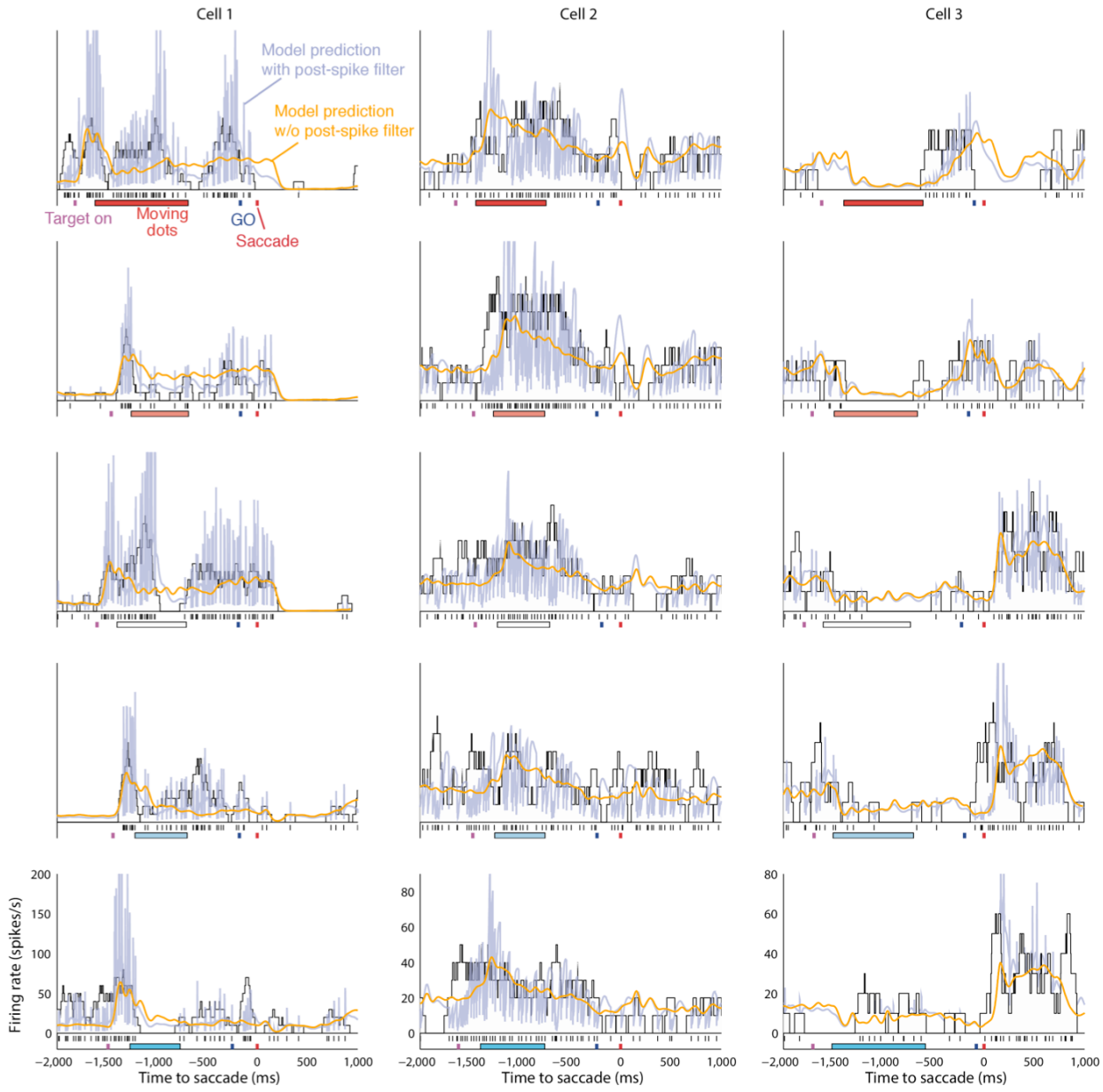
Nature Neuroscience: doi:10.1038/nn.3800



Supplementary Figure 7

Population diversity in all kernels.

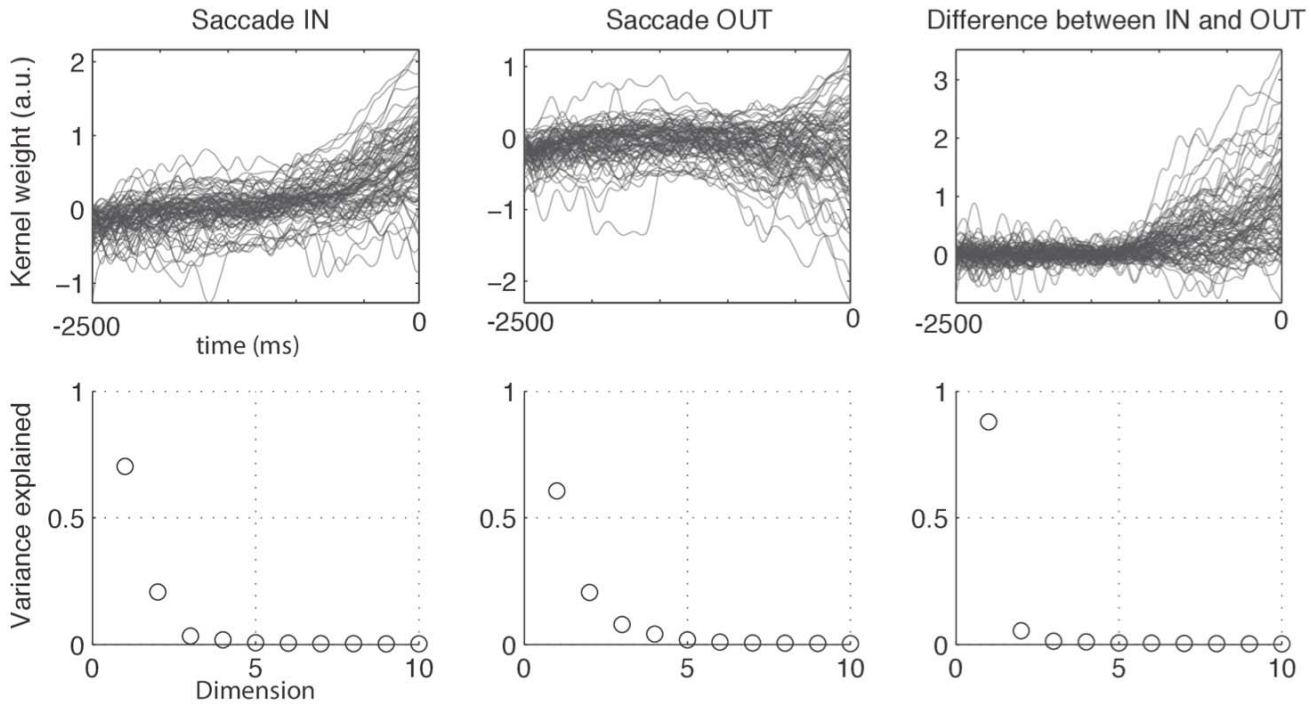
Each column represents a different kernel. The motion coherence kernels are convolved with a 500 ms boxcar for easier interpretation. **(Top 9 rows)** Each row shows one of the 9 example cells. They are sorted by decoding performance: The top cell is the best performing, followed by example neurons performing at 11%, 22%, …, 89% percentile. **(Bottom row)** Population average for all 80 neurons. Average kernel was transformed with exponential nonlinearity to represent gain. These fits are from models with the post-spike filter.



Supplementary Figure 8

Single trial prediction of the model with and without post-spike filter.

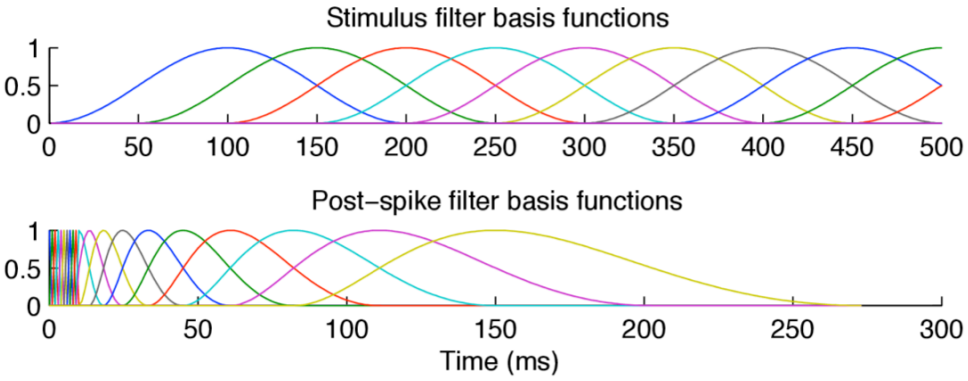
In **Fig. 4,** we only showed best and median predicted trials. Here are randomly selected trials per coherence strength for which the monkey was correct. Each column represents a cell corresponding to **Fig. 5** and each row corresponds to a coherence level. Light blue lines are predictions from the model with the post-spike filter. Although the fine time scale predictions are significantly better with the post-spike filter, we avoided showing the post-spike filter prediction in **Fig 4**, because it is difficult to visually assess the quality of fit. This is because a typical sharp self-excitatory gain appears after each spike, which also makes it difficult to smooth without inducing deceiving results.



Supplementary Figure 9

Dimensionality of saccade kernels compared to the decoding kernel.

(Top row) Individual traces of the kernel shapes of the population. Difference corresponds to **Fig. 7a**. **(Bottom row)** Dimensionality analysis corresponding to **Supplementary Table 1**. The first few dimensions explain most of the variance. Note that the difference of IN and OUT kernel has the tightest subspace.



Supplementary Figure 10

Visualization of temporal basis functions for the kernels.

(Top) Raised cosine basis functions spaced in 50 ms steps. **(Bottom)** 10 single bin boxcars followed by log-time scaled raised cosine basis functions. This is only used for the post-spike filter.

	1st	2nd	3rd	4th
IN	70.3%	91.0%	94.4%	
OUT	60.6%	81.2%	89.0%	93.1%
DIFF	87.9%	93.3%		

Supplementary Table 1

Dimensionality of saccade and optimal decoding kernels.

Cumulative variance explained by first 3 singular vectors for each saccade-locked component over the population of 80 neurons (see **Supplementary Fig. 9** for the corresponding data). The low dimensionality is quantified by the dimensionality of the subspace that explains most of the variance. For the saccade-locked kernels, IN and OUT kernels were less low-dimensional than the difference. Since we regularized IN and OUT kernels, and not on the difference, this is surprising.