nature neuroscience

ARTICLES

https://doi.org/10.1038/s41593-020-0699-2

Systematic errors in connectivity inferred from activity in strongly recurrent networks

Abhranil Das[1,2] and Ila R. Fiete[1,2,3]

Understanding the mechanisms of neural computation and learning will require knowledge of the underlying circuitry. Because it is difficult to directly measure the wiring diagrams of neural circuits, there has long been an interest in estimating them algorithmically from multicell activity recordings. We show that even sophisticated methods, applied to unlimited data from every cell in the circuit, are biased toward inferring connections between unconnected but highly correlated neurons. This failure to 'explain away' connections occurs when there is a mismatch between the true network dynamics and the model used for inference, which is inevitable when modeling the real world. Thus, causal inference suffers when variables are highly correlated, and activity-based estimates of connectivity should be treated with special caution in strongly connected networks. Finally, performing inference on the activity of circuits pushed far out of equilibrium by a simple low-dimensional suppressive drive might ameliorate inference bias.

Fully understanding the mechanisms of computation and plasticity in neural circuits requires knowledge of how those neurons are connected. Despite groundbreaking developments in direct circuit tracing[1–5], obtaining connectivity data is still difficult and expensive. Hence, there is a strong interest in statistical methods to estimate connectivity[6–9] from simultaneous circuit-wide neural recordings.

Activity correlations between neurons offer a crude estimate of their coupling, but it is widely acknowledged that such correlations can arise either from a direct synaptic connection or from a common input. Statistically sophisticated inference techniques, such as maximum entropy-based inverse Ising inference[7], l_1-regularized logistic regression[10] and generalized linear models[6,11,12], aim to 'explain away' correlations that arise from a common observed input. Successful explaining away prevents the inference of (nonexistent) direct connections based on correlation and yields connectivity graphs that are sparser than the raw correlation graph (Fig. 1a).

Such methods, applied to recordings of low-level sensory circuits in the brain, produce greatly improved predictions of neural responses because they account not only for the influence of the stimulus but also for other cells in the network[6,7,13,14]. With the success of these models in activity prediction, it is tempting to interpret the inferred connectivity as biological connections[6,15,16], an inference step that remains untested.

There are two key requirements for these methods to succeed in explaining away, neither of which tends to hold in reality: all nodes must be observed, and the inference model used to link connectivity and activity must very closely approximate the real dynamical system that generates the data. Thus, there is reason to suspect that the inferred connectivity might deviate substantially from structural connectivity, at least under some conditions. Problems of circuit inference because of unobserved neurons are relatively well recognized[14,17–23]; therefore, we focused here on the problems of inference even in the fully observed setting, which arise from mismatches between the generative system and the inference model (model mis-specification; see Methods and Extended Data Fig. 1 for our results on model-matched inference).

We hypothesized that inference models that are slightly mismatched to the generative system cannot exactly capture (and thus explain away) all observed correlations derived from multi-hop interactions, and the residual unexplained correlations are then interpreted as excess direct connections. Thus, we expect inference to be poor when weakly connected, or unconnected neurons exhibit strong activity correlations. This is common in strongly recurrent networks that perform amplification or generate self-sustaining memory states through emergent phenomena such as pattern formation. We expect these networks to present a fundamental challenge in circuit inference. More generally, our findings imply that causal inference will likely be problematic in any system with strongly interacting variables.

To study the relationship between circuit inference and network correlations, we constructed recurrent networks where we vary absolute recurrent weight strengths (to manipulate correlation strengths) while keeping the network architecture (connectivity and relative weights) and feed-forward drive fixed. Turning the absolute recurrent weight dial moves the circuit across weak ('sensory'), medium ('amplifying sensory') and strong ('memory') recurrent regimes, producing, in the last case, large-scale emergent activity patterns with strong correlations between unconnected neurons. To illustrate the patterns of explaining away errors, we first present results on a highly structured ring network. We demonstrate later that these results hold more generally. Finally, we show that pushing networks out of equilibrium through simple global perturbations might provide a solution to the inference problem even in the strong weight regime.

Results

A simple recurrent network and the sensory–memory continuum. We first considered a network of threshold-crossing spiking neurons arranged on a one-dimensional ring. Neurons interact through rotation-invariant recurrent connections **W** with a local Mexican

[1]Department of Physics, The University of Texas, Austin, TX, USA. [2]Center for Learning and Memory, The University of Texas, Austin, TX, USA. [3]Department of Brain and Cognitive Sciences and McGovern Institute, Massachusetts Institute of Technology, Cambridge, MA, USA. e-mail: fiete@mit.edu

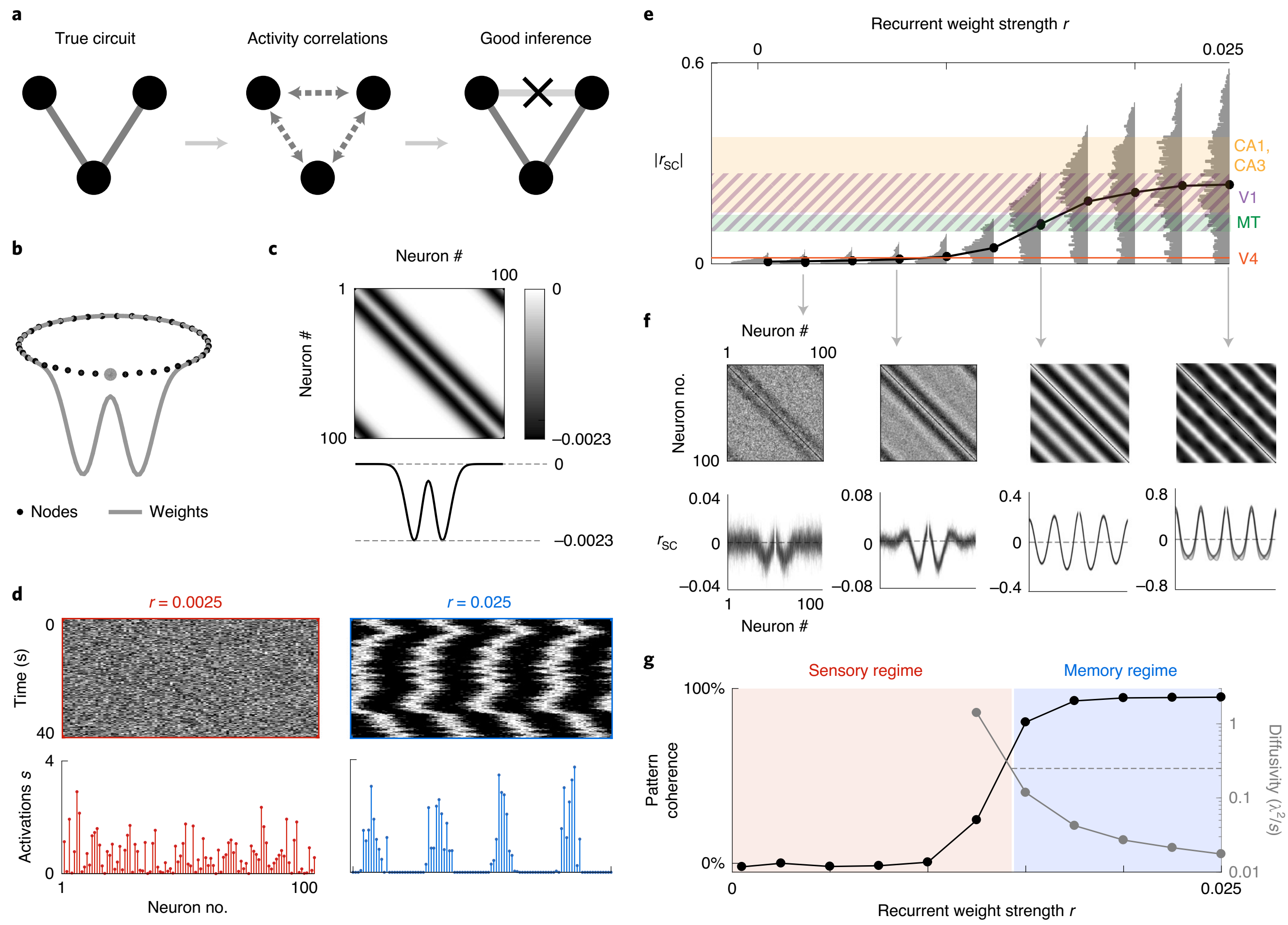


Fig. 1 | Structure and dynamics of the generative network. a, Left: schematic three-neuron circuit with two connections. Center: all neurons are correlated. Right: circuit inference algorithms should 'explain away' correlations between the top neurons based on their common input, ruling out a direct connection. **b**, Neurons (black dots) in a ring network. Gray curve, Mexican hat-shaped weights from an example node (gray dot) to the rest; all other neurons have the same weight profile. **c**, The resulting weight matrix **W** (top) is circulant, comprising rotations of the same row (bottom). **d**, Scalar parameter r modulates the strength of all recurrent weights. Spike raster plots (top) and snapshots of synaptic activity (bottom) of the network at weak (red) and strong (blue) weights (small and large r). **e**, Absolute noise correlations between neurons (spikes binned at 10 ms; gray, histogram; black, means) increases with weight strength. Horizontal color bands indicate experimentally measured ranges of noise correlation in different brain circuits. **f**, Noise correlation matrix at several weight strengths. Below each matrix, superposition of the rows, rotated so they align. **g**, Coherence (left ordinate, black) and diffusivity (right ordinate, gray) of the network pattern as a function of r. Horizontal line marks a diffusivity of 0.25 λ^2/s, where pattern phase information (memory) is typically lost after 0.5 s. The corresponding r value marks a working boundary between what we call 'sensory' (red) and 'memory' (blue) regimes of the network.

hat profile (Fig. 1b–c and Methods). All neurons also receive a shared feed-forward excitatory drive and independent identically sampled noisy drives (Methods). Cells generate spikes, which drive the dynamics at future times.

The ratio of recurrent to feed-forward input is set by the scalar recurrent weight strength parameter r, which multiplies the recurrent weight matrix **W**, whereas the average network firing rate is maintained by adjusting the firing threshold (Extended Data Fig. 2). When r is small, the feed-forward noise dominates, and network activity is relatively uncorrelated; when r is large, the network exhibits a global pattern of periodically spaced activity bumps[24] (Fig. 1d). The pattern undergoes noise-driven drift. Unconnected neurons in different co-active bumps exhibit strong correlations (Fig. 1d).

The structure of the noise correlation matrix changes with r (and depends on the time scale of binning; Extended Data Fig. 3). At small r, the matrix shows high variance with weak signatures of the true connectivity (Fig. 1e, first panel: variance appears as grains in the top plot and noisy superposition in the bottom). At intermediate r, the influence of noise decreases, and correlations better reflect connectivity (Fig. 1e, second panel). As r increases, however, pattern formation sets in, and noise correlations instead reflect the correlations induced by the periodic activity pattern (Fig. 1e, last two panels).

Before proceeding further, note that ranges of the parameter r over which the correlation structure changes can be substantially related to different operating regimes of circuits in the brain. We do so by characterizing noise correlations, coherence and temporal stability of network activity states as a function of r. Increasing r strengthens noise correlations (see Methods and Fig. 1f). These values can be compared with the mean pairwise noise correlation magnitudes measured in various sensory and non-sensory brain areas, such as V1[25–27], V4[28,29], MT[30–33] and hippocampus[34] (Fig. 1e, colored bands). As seen, medium to medium-large values in our range of r produce noise correlations consistent with primary and non-primary sensory processing stages in cortex. Large values correspond to noise correlations from CA1 and CA3, areas that are associated with memory. Medium-low and low values of r produce weaker correlations than found even in primary sensory areas such as V1. This comparison does not take into account the possibility of correlated noise inputs into primary sensory cortex; thus, we also use more qualitative measures—the coherence and diffusivity of activity states (see Methods)—to relate values of r to different operating regimes in the brain.

Coherence measures the fidelity of the activity pattern over time (regardless of where the pattern is centered). As r is increased, the network moves from a regime with zero pattern coherence to one with maximal coherence (Fig. 1g). The pattern drifts (Fig. 1d) in a non-restorative random walk (Ornstein–Uhlenbeck process), quantified by a diffusion coefficient (see Methods). As weights strengthen, the pattern becomes less diffusive (Fig. 1g). When diffusivity is low, the initial pattern phase is not rapidly lost and can be used as a memory state[35]. In Fig. 1g, the horizontal line marks the diffusivity value at which the expected root mean square spread of the pattern phase after 0.5 s equals half the pattern wavelength; thus, the starting phase is completely forgotten on this time scale. The weight strength at this point is $r = 0.0125$. Because both coherence and diffusivity of the pattern change sharply around this value, we take it as the working boundary between the 'sensory' and the 'memory' regimes of this network. The sensory regime close to the memory boundary is strongly amplifying and exhibits slow dynamics.

Inference at different recurrent weight strengths. We can now measure the quality of circuit inference along the sensory–memory continuum defined above. At each r, we generate 10^8 total spikes from the network (see Methods). We fit the data with a generalized linear model (GLM) because of its good performance in circuit activity prediction, at least at the sensory periphery[6]. The inputs to neurons in this model are the weighted sums of spikes from other neurons filtered by a temporal kernel, which are then exponentiated to produce Poisson spiking rates; parameters are estimated via maximum likelihood (see Methods). We extract the inferred weights $\hat{\mathbf{W}}$ from this model and define the normalized l_2-distance between the ground truth and optimally rescaled inferred weights to be the inference error (see Methods). Across weight strengths, inference is performed on the same number of total spikes; thus, differences in inference quality cannot be attributed to differences in data volume.

When the recurrent weights are weak, they have a small effect on neural activity relative to the ongoing noise; thus, the signal-to-noise ratio (SNR) is low, and the inferred connectivity exhibits uncorrelated errors (variance) (Fig. 2a, first panel), similar to the noise correlation matrix. As the weights become stronger, the SNR improves and the inference error decreases, but only up to a point (Fig. 2a, second panel). As the weights become stronger still, inferred weights begin to exhibit a new kind of error, visible as side bumps in the weight profile or as paradiagonal stripes in the weight matrix (last two panels of Fig. 2a).

Consistent with these results, the distribution of individual weight errors is Gaussian at small r. With increasing r, the width of the Gaussian initially shrinks to reflect improving SNR, and then the distribution becomes increasingly non-Gaussian (Methods and Extended Data Fig. 4).

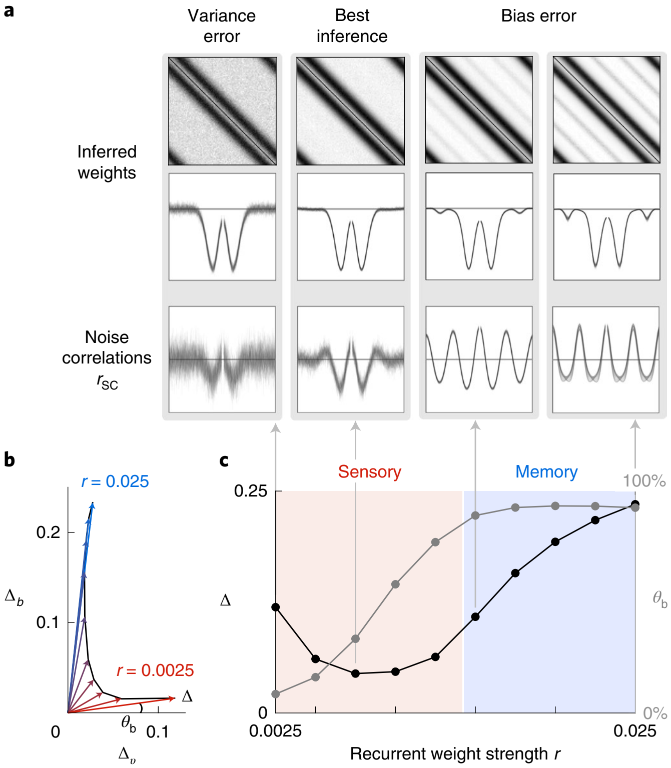


Fig. 2 | Quality of circuit inference using 10^8 spikes from a fully observed ring network, as a function of weight strength. a, Inferred weight matrices $\hat{\mathbf{W}}$ (top row), superposed rows of the weight matrix (middle row; line marks zero) and raw noise correlations (bottom row), at different weight strengths. **b**, Total inference error (arrows) as a vector of independent variance (Δ_v) and bias (Δ_b) error components (see Methods), at different weights. The vector magnitude Δ is the total inference error (as a fraction of the magnitude of the true weights vector), and the normalized angle θ_b is the fraction of bias error in this total error. **c**, Total inference error Δ and fraction of bias θ_b, against weight strength.

Thus, we see that the errors at high r are systematic (biased) overestimates of the existence and magnitudes of connections. They result from a partial failure to explain away strong correlations, induced by the emergence of the global activity pattern. The GLM estimate is a better approximation to the true weight matrix than is the raw noise correlation matrix, but it still exhibits a similar qualitative pattern of errors (compare the bottom two rows of Fig. 2a).

The total inference error Δ can be cast as the magnitude of a vector composed of two orthogonal components: the variance error with magnitude Δ_v, and the bias error with magnitude Δ_b—thus, $\Delta^2 = \Delta_v^2 + \Delta_b^2$ (Methods). The relative contribution of these two components is then simply visualized as the orientation of the total error vector relative to the variance and bias components (Fig. 2b). For small r, the variance error dominates, and the total error vector is near the variance axis. With increasing r, the error vector rotates from the variance toward the bias axis (Fig. 2b). Meanwhile, its magnitude first drops, then rises, and is smallest at intermediate weights when variance and bias contributions are both relatively low (Fig. 2c).

The point of best inference is far to the left of the sensory–memory boundary, in the very low weight regime (Fig. 2c). When the inference model is not exactly matched to the data-generating model, as is typical, this point is not an invariant that reflects a critical point of the network dynamics or of the inference process. Rather, it depends on the data volume used for inference, as we show next.

Variance, but not bias errors, decline with data volume. Data volume is the total number of spikes used for circuit inference. We hold

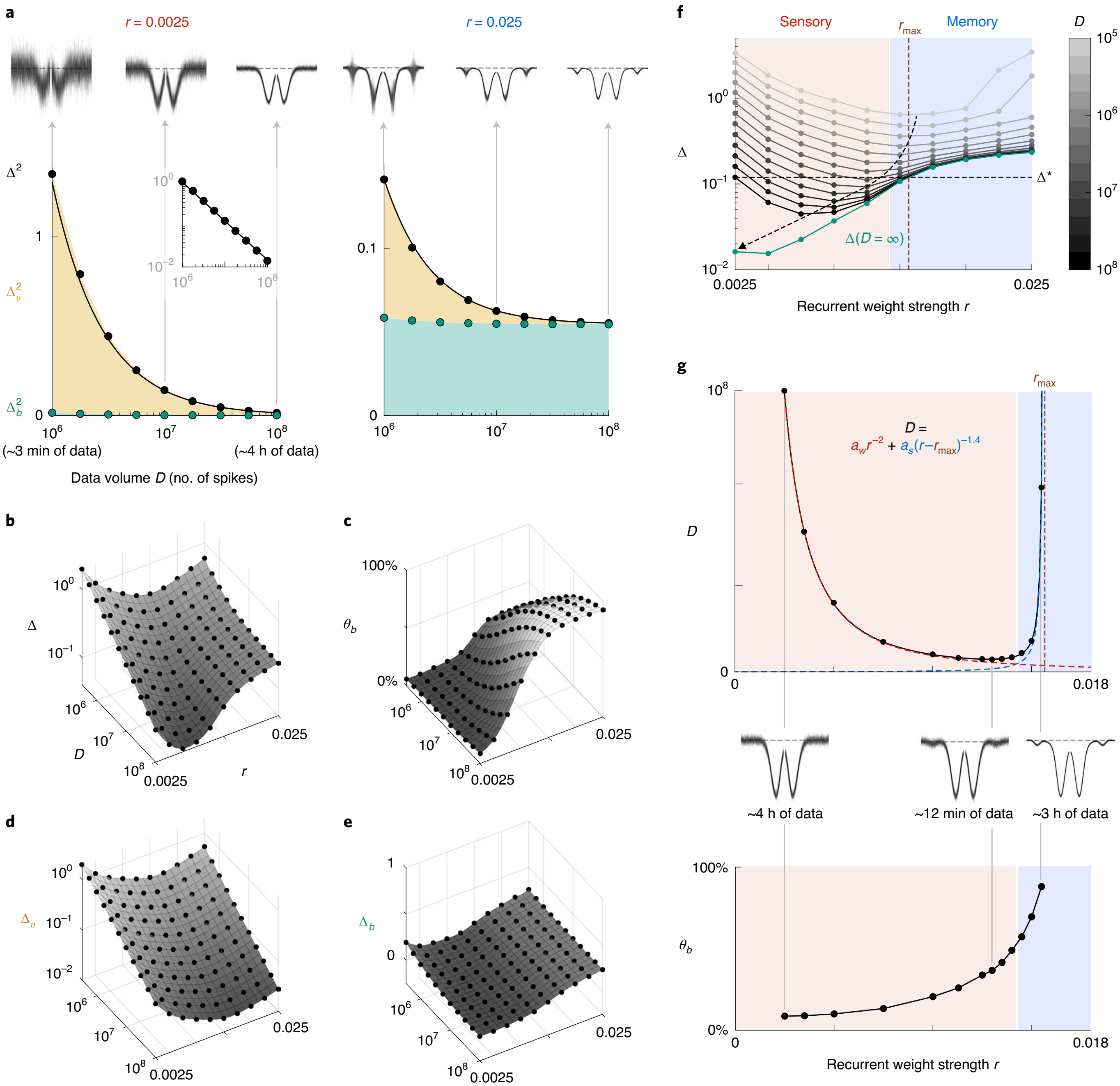


Fig. 3 | Inference quality as a function of data volume. a, Top: superposed rows of the weight matrix inferred with increasing data volume, at weak and strong weights. Bottom: squared inference error Δ^2, split into squared variance error Δ_v^2 and squared bias error Δ_b^2. Black curve is a fit ($\Delta^2 \sim 1/D$), confirmed by inset log–log plot at weak weight. **b–e**, Full surfaces of Δ, Δ_v, Δ_b and bias error fraction θ_b, versus weight strength and data volume. Black dots are data points, and surface is an interpolation. $\Delta_v^2 \sim 1/D$ at all weight strengths, but Δ_b is unchanged by data. **f**, Total error versus weight strength for different data volumes and in the limit of infinite data (green). Optimal inference point moves toward $r = 0$ (dashed arrow) with increasing data. Given some specified accuracy Δ^* (horizontal line), a network with weight exceeding r_{max} (brown) cannot achieve this accuracy with any data volume. **g**, Top: data volume required to meet the specified accuracy Δ^* at different weights. Fitted black curve is the sum of two power laws diverging at $r = 0$ (red) and $r = r_{max}$ (blue). Middle: superposed rows of the weight matrix at the criterion accuracy Δ^*. Bottom: corresponding bias error fractions.

the average firing rate in the generative model fixed as r is varied, so that data volume is proportional to data collection time. Variance and bias errors scale differently with data volume (Fig. 3a). Variance errors decrease inversely with data, as expected: $\Delta_v^2 \sim 1/D$ (Fig. 3a, inset at top right; error is well fit by a line of slope −1 on a log–log plot; see Extended Data Fig. 5a with fits and confidence intervals). By contrast, bias errors persist because they arise from correlated neural activities that are not averaged away with the addition of data (Fig. 3a, green areas, and Fig. 3b–e).

Because increasing data volumes erode variance but not bias errors, which dominate at weak and strong weights, respectively, it follows that, with increasing data volume, the total error will continue dropping at weak weights but remain relatively stable at strong weights. Thus, the curve of total error versus r will change shape with data volume, dropping progressively lower on the left, and the best weight strength for inference will shift progressively leftward (Fig. 3f). In the limit of infinite amounts of data, the weakest weights are optimal for inference (Fig. 3f, green).

By contrast, in the empirically implausible case where the generative and inference models exactly match, which we illustrate with inverse Ising inference applied to data generated from an Ising model, there are no bias errors, and errors decay according to a fixed power law at all weights. There is an intrinsic critical point for best inference (see Methods section on matched inference), independent of data volume. The critical point occurs at intermediate weight strength and corresponds to maximal magnetic susceptibility in the Ising network[36], which, in turn, is directly related to the maximization of Fisher information (FI).

Infeasible data volume for accurate inference in memory networks. To estimate the data volume required for circuit inference as a function of recurrent weight strength, we fix the desired inference error to a value Δ^* (marked by the horizontal line in Fig. 3f) and slice the surface of Fig. 3b within 1% of this level. The required data volume grows as r decreases, diverging at $r \to 0$ because the influence of recurrent weights on activity, and thus SNR in activity about weights, vanishes in that limit (Fig. 3g). At the opposite end, the required volume diverges again, at a finite value of the recurrent weight strength. We fit the curve with a sum of two power laws (Fig. 3g, black curve), fitting all five parameters simultaneously. The first power law has a form $a_w r^{-n_w}$, diverging at $r = 0$ (red dashed curve) and with a fitted exponent $n_w \approx 2$. The second has the form $a_s(r - r_{\max})^{-n_s}$, diverging at a fitted value of $r_{\max} = 0.0157$, with a fitted exponent $n_s \approx 1.4$ (blue dashed curve).

We can understand this divergence in required data volume at finite strong weight by plotting this fitted $r_{\max}$ on Fig. 3f: it corresponds to the point at which the infinite data inference error (green curve) exceeds the criterion error Δ^* (due to larger data-intractable bias errors at stronger weights). Thus, a network with stronger weights than $r_{\max}$ cannot be inferred to the desired accuracy with any volume of data.

Results generalize across inference methods. To probe the generality of our result on inference errors, we first applied alternative inference methods to the same spike data.

We estimate weights with the Ising model (that is, find the maximum entropy model that fits the data means and covariances under the assumption of binary responses; see Methods for binarization procedures). Exact inference under an Ising model with general all-to-all weights is NP-hard[37], which has led to the development of approximate algorithms that involve different simplicity–accuracy tradeoffs. We used the minimum probability flow (MPF) algorithm[38], which provides good solutions at intermediate computational complexity and guarantees convergence to the correct (maximum likelihood) Ising parameters in the asymptotic data limit (Fig. 4a). We alternatively used mean-field approximations to the Ising model, including the naive mean field and the Thouless–Anderson–Palmer (TAP)[39] and Sessak and Monasson (SM)[40,41] approximations (see Methods): although less accurate, they are simple and fast (Fig. 4a). The naive mean-field connectivity estimate is simply the negative inverse of the activity covariance matrix, so we call it the ‘raw correlations’-based estimate. This estimate using binarized spike data is our low benchmark (expected lower bound) for the quality of binary data-based inference methods, such as Ising inference (Fig. 4a), whereas a raw correlations-based estimate using spike count data is the low benchmark for count-based methods such as the GLM and logistic regression, described next.

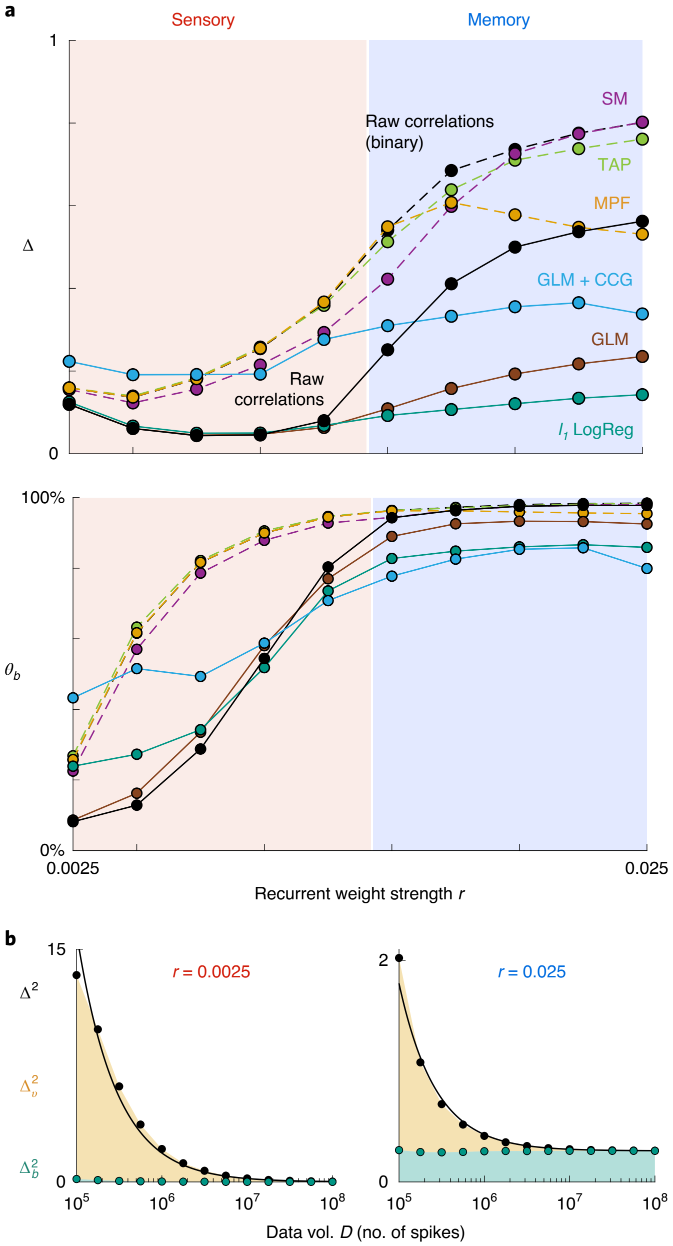


Fig. 4 | Results extend to different inference methods. a, Inference error (top) and bias fraction (bottom) versus weight strength, on 10^8 spikes from the ring network, using different methods. Dashed lines are for inference on binarized spikes. **b**, Inference error as a function of data volume when performing inverse Ising inference with minimum probability flow (as in Fig. 3a).

We next perform inference using logistic regression[10,42], in which the response of each neuron is regressed onto the activity of the rest. The response variable must be binary, but the predictor values need not be; thus, we binarize only the spike data of the response neuron. To try to reduce explaining-away errors by favoring a sparser matrix, we apply an l_1 penalty on the inferred coefficients[10], tuning the regularization parameter at each r to optimize inference accuracy (see Methods). The l_1 penalty barely improves inference (Extended Data Fig. 6): it truncates the flanks of peaks in the inferred weight matrix without removing side peaks. Nevertheless, logistic regression achieves slightly better performance than the GLM (Fig. 4a and Extended Data Fig. 6).

Finally, we consider whether gating the inferred connectivity matrices by the existence of short-latency peaks in spike train cross-correlations improves inference (see Methods). If there is a

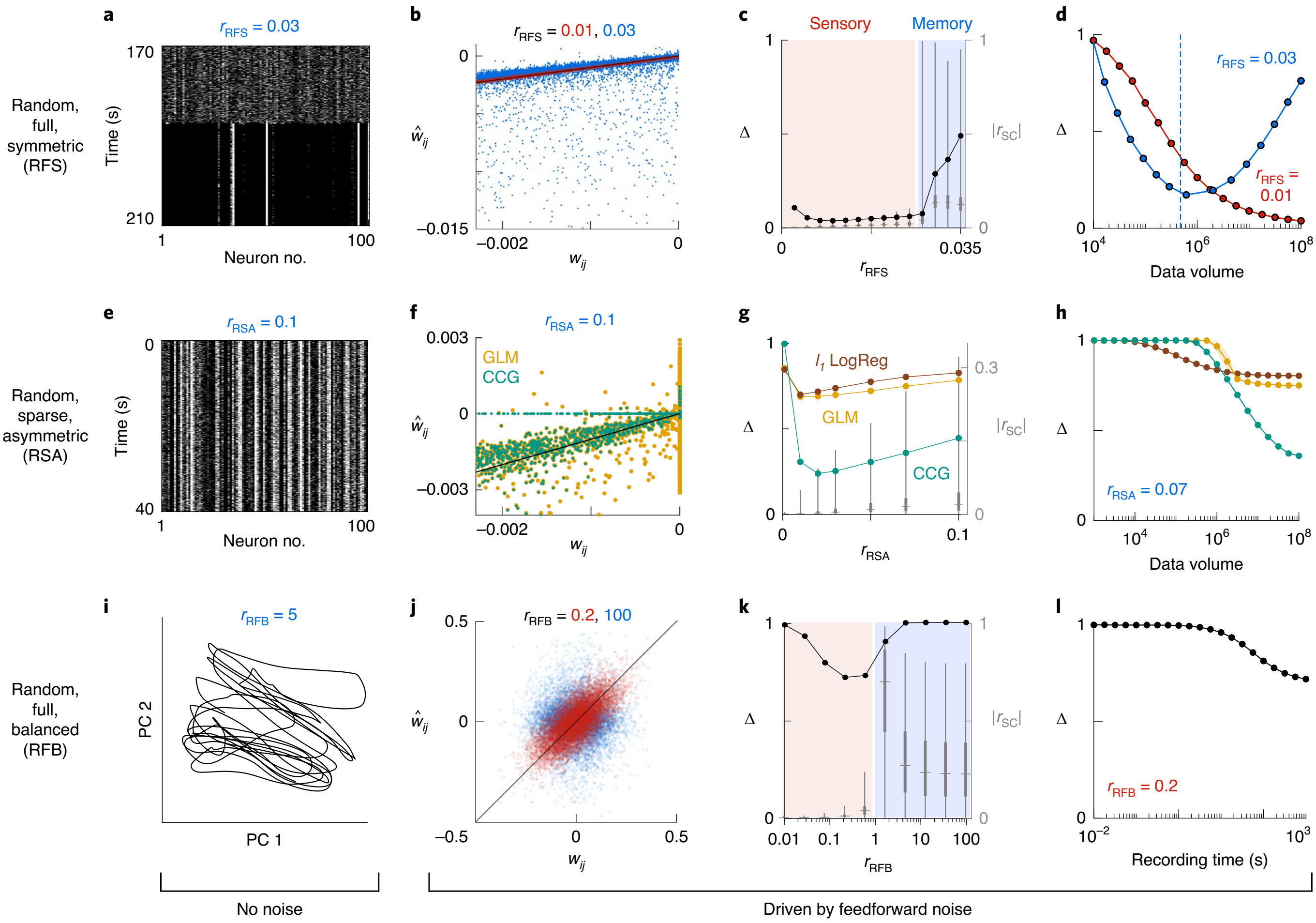


Fig. 5 | Results extend to different networks. a, Spike raster from a random, fully connected, symmetric recurrent network with strong weights, initialized randomly. Initially disperse activity settles into a stable activity pattern. **b**, Inferred versus true weights in this condition (blue) and at a weak weight where inference is best (red). Weak weight inference (red) falls along $y = x$ (black line; ground truth). **c**, Inference error (black) and absolute noise correlations (gray) of the $\binom{N}{2} = 4,950$ neuron pairs (box plot shows median, first to third quartile and range) as a function of weight strength. Blue indicates the strong weight regime that supports patterned activity attractor states. **d**, Blue curve, inference error on cumulative data collected from the initial condition in **a** (dashed blue line indicates the point of pattern onset). Red curve, cumulative data inference at a weak weight with no pattern onset. **e**–**h**, As in **a**–**d**, for a random, sparsely (10%) connected asymmetric recurrent network. **e**, Spike raster plot at strong weight. **f**, Inferred versus true weights in this condition based on a GLM (yellow) and a GLM with connections gated by spike CCG information (green). **g**, Inference errors using GLM alone, CCG-augmented GLM, l_1-regularized logistic regression and absolute noise correlations, for the 4,950 neuron pairs (box plot shows median, first to third quartile and range) versus weight strength. **h**, Inference error versus data volume for strong weights. Error band is s.e.m. over up to five different data subsets. **i**–**l**, Corresponding plots for a random, fully connected, asymmetric balanced recurrent network. **i**, Chaotic trajectory of first two principal components of neural fields for strong weights, when there is no noisy feed-forward drive. **j**–**l**, Inference results on this network when it is driven with feed-forward noise. **j**, Inferred versus true weights at strong weights (blue) and at weaker weights where inference is best (red). **k**, Inference error and absolute noise correlations of the 4,950 neuron pairs (box plot shows median, first to third quartile and range) versus weight strength. Blue indicates chaotic regime. **l**, Inference error versus data volume at the weak, optimal weight.

positive or negative peak in the cross-correlogram (CCG) of a pair of neurons within a lag τ, we posit a connection and use the weight estimate from the GLM inference; otherwise, connections are set to zero. The CCGs in this symmetric network exhibit only broad symmetric peaks, which are not ideal for connectivity inference (Extended Data Fig. 7). The resulting matrix is sparser but also contains false negatives; overall, CCG gating leads to larger errors in this circuit. We later use asymmetric CCG features in a sparse asymmetric network with better success.

All these inference methods, regardless of statistical sophistication, perform equally well—and no better than the corresponding (binary or non-binary) raw activity correlation matrix—when the recurrent weights are small, and the limiting factor is noise. The primary difference between methods in this regime is the (expected) better performance of models that use count rather than binary spike data. Across weights, all methods replicate the qualitative U-shaped curve of inference error versus r and produce data impervious bias errors (Fig. 4b). Thus, when there is inevitably some mismatch between the inference model and real system dynamics, inference methods will tend to overestimate connectivity.

Results generalize to different networks. To test whether our results generalize to other neural network models, we first switch the neural dynamics (to a linear–nonlinear Poisson model; see Methods) while keeping network architecture fixed. We apply GLM inference with an exponential nonlinearity, as before, and obtain the same qualitative inference results (Extended Data Fig. 8).

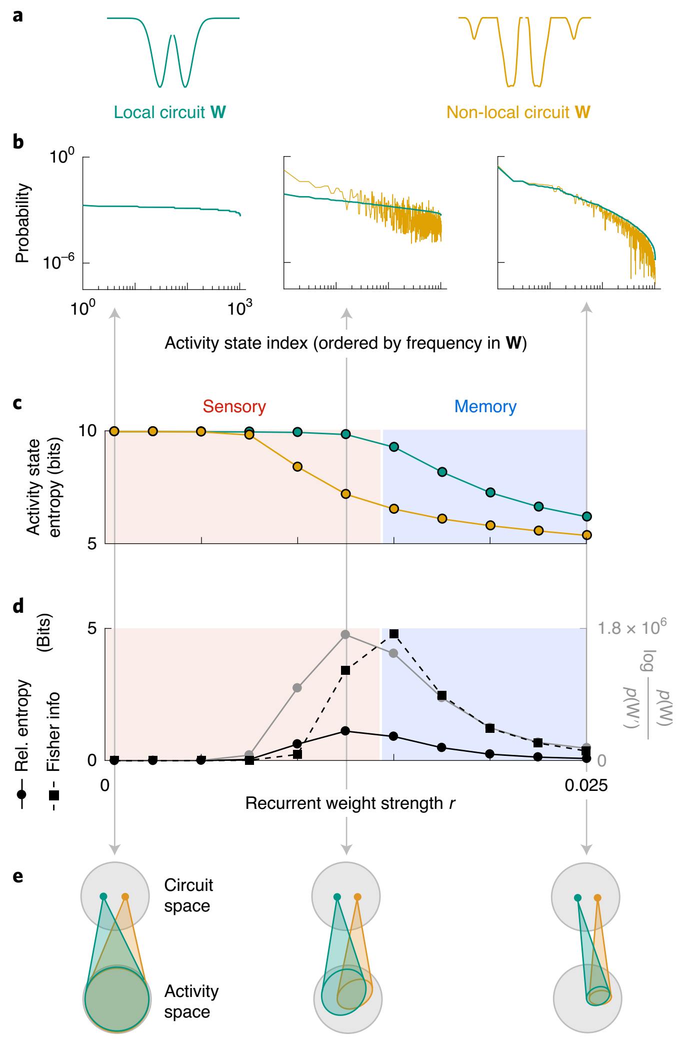


Fig. 6 | Circuit-to-activity map is inherently less invertible when correlations are strong. a, True weight profile for the ring circuit (with only local weights) and an alternative circuit (with a different weight profile and extra non-local weights). **b**, Probabilities of activity states (binary ten-neuron spike words) in each circuit, at weak, intermediate and strong *r*. States are ordered according to their probability in **W**. **c**, Entropies of the state distributions against *r*. **d**, Relative entropy (KL divergence) between the distributions, FI and the log likelihood ratio of the true versus alternative circuit given the true circuit activity, against *r*. **d**, Schematic of the mapping from the space of circuits to the space of their activity states, at different weight strengths.

Second, we change the circuit itself: instead of a structured, locally connected, low-dimensional ring, we consider a network that lies on the opposite extreme of symmetric recurrent architectures, with random all-to-all connectivity as in Hopfield networks (Methods). At weak weights, this network exhibits high-dimensional, statistically stationary network-wide firing (due to injected noise). At strong weights, initially high-dimensional activity states collapse to a stable pattern of activity, with a sparse set of firing neurons (Fig. 5a).

At each weight strength, we collect 10^8 spikes from the network and perform circuit inference with a GLM. Inference in the strong weight network (Fig. 5b, blue scatter) substantially overestimates overall weight strengths compared to in the weak weight condition (Fig. 5b, red). Inference error increases as a function of weight strength (Fig. 5c, black), with an abrupt transition near the weight strength that supports stable pattern formation (the pattern formation transition can be detected independently by the sudden growth in the noise correlation magnitudes; Fig. 5c, gray).

We inspect how the onset of pattern formation affects inference. Using only the sparse patterned state, inferred connectivity is extremely poor, with strong excitatory coupling between the sparsely active set and strong inhibitory coupling between them and all the rest. Instead, we initialize the network in an unpatterned state, collect data over time as the network settles into a patterned state and perform inference repeatedly on the accumulating data. At first, while the network activity is high dimensional, inference improves with increasing data volume (Fig. 5d, blue curve to the left of the dashed line). Inference actually improves more rapidly in the pre-pattern-onset strong weight network than in the unpatterned weak weight network (red curve). This observation leads us to a proposal that we explore later, that far-out-of-equilibrium activity might ameliorate inference bias. Once patterning sets in, however, including post-pattern-onset data degrades inference relative to the smaller subset of pre-pattern-onset data.

Third, we construct a recurrent network with non-symmetric sparse random connections (10% connection probability) and non-symmetric random weights (see Methods). To maintain stability with non-saturating neurons as the weight strengths are increased, and retain the same statistical distribution of weights as the ring network, recurrent connections are inhibitory (excitatory weights are used next). At strong weights, this network also exhibits patterned stable states (Fig. 5e). The GLM infers a broad distribution of weights, both positive and negative, for both connected and unconnected neuron pairs (Fig. 5f, large scatter of yellow dots away from black line and vertical yellow streak near the zero weight point, respectively).

In this circuit, augmenting GLM inference with CCG analysis produces a more pruned connectivity (Fig. 5f, green dots; note the reduced spread at the zero weight point) but also introduces false negatives (horizontal streak of green dots in Fig. 5f and Extended Data Fig. 7). This method outperforms GLM alone (Fig. 5g). On the whole, however, our earlier qualitative findings persist: inference error is higher in networks with stronger weights and even with increasing data volumes, asymptotes toward a sizeable non-zero level due to a persistent estimation bias from a failure to explain away (Fig. 5h). Despite the fact that connectivity in the network is sparse, l_1-regularized logistic regression provides at most marginal gains that do not approach the improvements from CCG analysis, outperforming it only at high noise (weak weights and little data) (Fig. 5g,h). This is arguably because the l_1 penalty suppresses weights based blindly on their magnitude, whereas the CCG method exploits insight into neuronal interactions.

It is unclear if CCG pruning will be fruitful in real circuits, even if they are sparse and non-symmetric. Here, CCG peaks are narrow and sharp (see Methods and Extended Data Fig. 7) because each neuron receives only ~10 inputs. (In a larger network, eg, of 10^4–10^5 neurons, the combined sparseness of activity and connectivity would have to be as small as 10^{-3}–10^{-4} to obtain only ~10 simultaneously active inputs to each neuron, or the inputs would have to be highly correlated, to avoid loss of information about precise temporal relationships between pairs of neurons.)

Fourth, we consider asymmetric balanced networks with mixed excitatory and inhibitory connections (using rate-based, saturating neurons and non-noisy inputs; see Methods) that exhibit chaotic rather than fixed point dynamics at sufficiently strong weights[43] ($r_{\mathrm{RFB}} > 1 + \epsilon$ for $\epsilon \to 0$ with network size; below these weight strengths, the dynamics admit a single stable fixed point at 0) (Fig. 5i).

Circuit inference in this network (by logistic regression on the rates) is ineffective at all weights, with poor inference performance even in the chaotic regime (see Extended Data Fig. 9; to stimulate activity when below the chaotic regime, we provide a brief input pulse to all neurons at the start of the simulation). Although chaotic dynamics appear noisy, they are nevertheless quite low dimensional; the deterministic balanced network is thus plagued by even worse inference problems than the pattern-forming networks explored above when they are noise driven. We therefore add high-dimensional noise to the network in the form of a fluctuating feed-forward drive, as before; Fig. 5j–l shows the results on this noise-driven network. Inference improves in the weak weight regime, where the noise input dominates; it also improves with increasing data volumes, but the improvement saturates. The strong weight regime remains dominated by the recurrent drive, and inference performance is poor.

Fifth, we studied a sparse version of this balanced network (10% connection probability) over a range of weight strengths r_{RSB}. Similarly to the fully connected balanced network, activity decays to a single fixed point at zero for sufficiently weak weights ($r_{\mathrm{RSB}} \lesssim 3.2$). At sufficiently strong weights ($r_{\mathrm{RSB}} \gtrsim 11$), the networks tend to exhibit limit cycle dynamics or chaotic dynamics. As with the non-sparse balanced network, inference is ineffective unless the network is driven with noise, and then it improves to similar levels as the non-sparse balanced network (see Extended Data Fig. 9). The cross-correlations of some connected pairs change most steeply around zero, reflected in short lag peaks in their time derivatives. Gating the connectivity matrix with this information reduces many false positives but also introduces false negatives; the resulting inference performance is not improved overall (Extended Data Fig. 9).

In sum, the problem of overestimation of connectivity in strongly recurrent networks generalizes across inference methods and across different circuits that exhibit disparate varieties of low-dimensional dynamics.

The circuit-to-activity map is inherently less invertible in strong weight networks. Given how general inference bias is at strong weights, the fault might lie less in the inference methods than in an inherent reduction of connectivity information in activity data at strong weights. Can the true activity of a circuit be distinguished from activity generated from a version of the circuit with the mis-inferred weights? If the two produce similar activity states, no inference method based only on activity could tell them apart in principle.

We consider two separate generative models: the true circuit **W** (our ring network with local Mexican hat matrix) and a related circuit **W**′ (the Mexican hat plus-side peaks matrix obtained from minimum probability flow-based Ising inference on the ring network in the strong weight regime) (Fig. 6a). The empirical activity state distributions of neural subsets (binary ten-neuron spike words; see Methods) in the weak weight regime are flat and identical for both models, as the states are driven by noise and not the weights (Fig. 6b). At intermediate weights, the distributions separate and then converge again at strong weights (Fig. 6b, center and right). With increasing *r*, the entropies of the distributions decrease monotonically from the maximum possible value of 10 bits (Fig. 6c). The Kullback–Leibler (KL) divergence (relative entropy), measuring the distinguishability between the distributions, is small at weak weights, peaks at intermediate weights and declines again at strong weights (Fig. 6d).

The FI measures the amount of information contained in the data to tell apart the circuit from similar ones. We can compute this by defining a family of intermediate circuits that linearly interpolate between **W** and **W**′ (see Methods). Similar to the KL divergence, the FI starts low at weak weights, peaks at intermediate weights and nearly vanishes at strong weights (Fig. 6d). Finally, we compute the closely related log likelihood ratio of the true versus alternative circuit model, given the activity from the true circuit (Fig. 6d; see Methods).

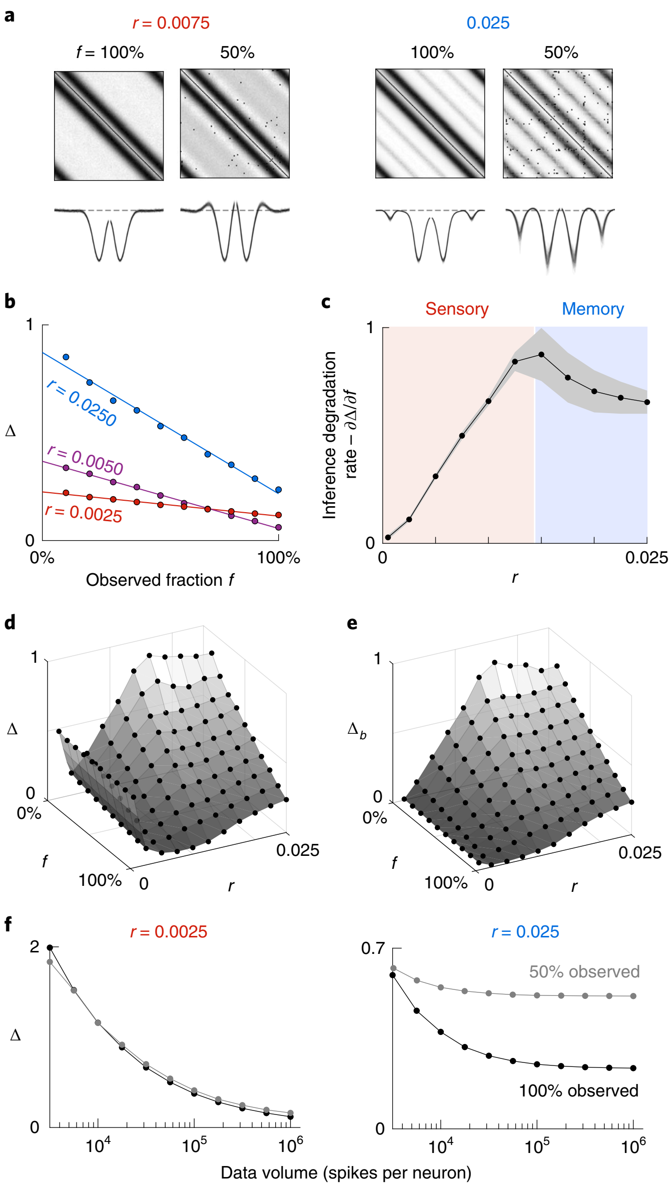


Fig. 7 | Inference bias due to unobserved neurons is exacerbated at strong weights. a, Weight matrices (top) and superposed rows (bottom) of merged inferred sub-circuits of the ring circuit, when all or 50% of the neurons are observed, at weak and strong weights. **b**, Average inference error as a function of the observed fraction, for different weight strengths. **c**, The rate at which inference degrades as the observed fraction shrinks, at different weight strengths (points in curve correspond to the slopes of the lines in **c**; gray region indicates 95% confidence interval). **d**,**e**, Full surfaces of average inference error and average bias error in the merged inferred circuit, against observed fraction and weight strength. Black dots are data points; surfaces are interpolations. **f**, Inference error against data volume for fully and 50% observed networks, at weak and strong weights.

Figure 6e shows a schematic mapping from the circuit to the state space that summarizes these results: with increasing *r*, as activity becomes more correlated and patterned, the entropy of states and thus the occupied state space shrinks for both models. Simultaneously, even though each model occupies a smaller region of state space, the states also become more distinct from each other, as they better reflect the respective weights and because the extra synapses in **W**′ induce an earlier pattern onset; thus, they have less overlap. At strong weights, patterning catches up in **W** and begins to saturate in both, so their state spaces continue to shrink toward a common point, and their overlap grows again.

Thus, the amount of information intrinsically available in activity about circuitry increases and then decreases again with increasing r, precisely mirroring the U-shaped inference error curves we have seen across inference methods.

Strong weights exacerbate inference errors due to unobserved neurons. To decouple known problems of inference in partially observed circuits[14,17–23] from the problems of inference in strongly recurrent circuits, we so far considered the fully observed setting. We now examine the combined effects of strong recurrence and partial observation on inference.

In the ring circuit, we use the activity data of a subset of neurons to infer connectivity within that sub-circuit. We repeat this process over multiple different subsets of the same size, and, to help visualization, 'merge' this patchwork of inferred sub-circuits together to obtain a complete $N \times N$ circuit (for each neuron pair i, j in the full circuit, we average the inferred weights $\hat{w}_{ij}$ from all sub-circuits that contained that pair).

Unsurprisingly, inference with partial observation results in worse performance at both weak and strong weights. Bias errors (side bands) emerge even at weak weights (Fig. 7a), because it is impossible to explain away the correlation between a pair of observed neurons if the correlation is being driven by an unobserved common input (eg, Fig. 1a).

The average inference error grows linearly as the observed fraction shrinks, but the growth is slow at weak weights (Fig. 7b, red and violet curves). As weights become stronger, degradation is more rapid (Fig. 7b–d). The speedup is specifically due to increasing bias errors (Fig. 7e)[22]. At the weakest weights, the quality of inference in a 50% observed circuit nearly matches that in the fully observed circuit (Fig. 7f, left), but, at strong weights, the gap between the two is large and does not decline with data volume (Fig. 7f, right). In sum, strong weights combined with partial observation produces potent data-impervious bias errors in inference. The flipside of our finding is that, with weak weights (and/or with large high-dimensional noise), the effects of partial observation can be minimized, and it should be possible to perform accurate inference even with many missing nodes.

Inference on far-out-of-equilibrium activity can mitigate bias errors. We have seen that, if network activity is high dimensional—that is, it traverses many relatively uncorrelated activity states—it leads to good inference even when recurrent weights are strong. How can one harness this observation in practice?

One way to reduce bias errors in strong weight networks is thus to inject high-dimensional noise into the neural circuit. This is a conceptually simple strategy but might be practically difficult to implement. Another strategy is to globally weaken the recurrent weights through a neuromodulator or other pharmacological agent capable of multiplicatively and equally affecting the strengths of all synapses without regard to their chemical and physical makeup.

Alternatively, given that the problem of biased inference arises from low-dimensional correlated activity states, these can be disrupted by globally pushing the system far from equilibrium by a simple suppressive input, which effectively reduces recurrent synaptic driven within the network. To this end, in the ring network, we briefly turn off the excitatory feed-forward input to push the network into a low-activity state (in experiments, a similar effect could be achieved by suppressing excitatory drive within or to the circuit or by driving inhibitory inputs to or within the circuit). After this low-dimensional perturbation, we collect data for approximately the synaptic time scale τ while the network recovers toward its equilibrium pattern and then reapply the perturbation and again collect data as the network relaxes, and so on. The network does not fully recover before the next pulse, so that all samples are out of equilibrium (Fig. 8a). We continue this until we collect enough out-of-equilibrium samples. When we perform inference on the out-of-equilibrium data with the same total data volume as before, inference performance is superior, and bias errors are abolished (Fig. 8b).

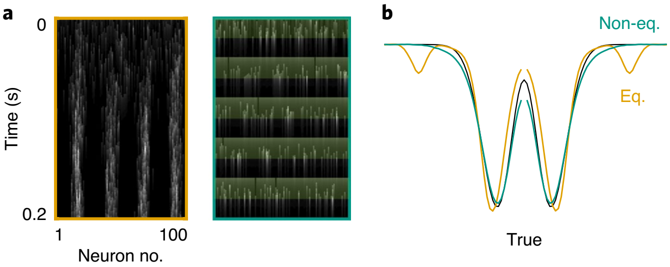


Fig. 8 | Sampling non-equilibrium data mitigates inference bias. a, Left: slow synaptic activations, s, of the ring network neurons at strong weight ($r = 0.025$). Right: the same when the feed-forward drive is pulsed on and off (green overlays mark 'on' periods). **b**, Average inferred weight profiles using data collected at equilibrium versus out of equilibrium.

Discussion

We have shown that activity-based connectivity estimates of strongly recurrent circuits can be incorrect in substantial and systemically biased ways in strongly recurrent circuits. Our definition of strong weights encompasses not only circuits that can hold states without external inputs, as required for memory, but also sensory circuits that moderately or strongly amplify their inputs.

The inference errors we identified in neural microcircuits might arise more generally, including when estimating inter-area connectivity from functional magnetic resonance imaging data. Our findings augment the refrain 'correlations are not causation' with the result that strong correlations obfuscate causation. Our results point to the need to be especially wary when assigning causal roles to variables based on statistical models applied to data, whenever variables are highly correlated with each other.

Despite some of our pessimistic results, the challenge of discovering the connectivity of recurrent neural networks is difficult but not unsurmountable. Rapid advances in connectomics[2–5] offer the promise of obtaining connectivity matrices of many complete circuits. However, the costs and technical challenges remain great, and even after obtaining a connectivity matrix, determining the link between structure and function requires another inference step, or model, of how activity emerges from connectivity.

More immediately accessible and interpretable are experiments that rely on perturbation of the system. Clearly, if each neuron could be perturbed, one at a time, and its effects monitored in all other available neurons, that would provide detailed causal connectivity information even in partially observed circuits. Alternatively, tracking the effects of targeted high-resolution and high-dimensional ('holographic') perturbations can also disentangle correlation from causation, but this approach requires inducing specific high-dimensional perturbation patterns and then associating them with their downstream effects to back out connectivity[44–47]. Alternatively, low-dimensional global perturbations have been shown to suffice in discriminating between a few specific candidate models in some highly structured neural circuits[48–50].

Our approach of far-out-of-equilibrium sampling is complementary to these strategies and combines their strengths. It allows for statistical inference of connectivity in arbitrary circuits without the construction of high-resolution or high-dimensional perturbation strategies and requires no tracking of the relationships between perturbation and effect. It requires only that the system be out of equilibrium while acquiring data. Thus, a simple, low-dimensional perturbation strategy, such as transient global silencing in the circuit, can be sufficient for more accurate circuit inference. Indeed, our results suggest that, with enough out-of-equilibrium data, performing even simple correlational inference could provide much better estimates of the true connectivity in strongly recurrent networks than using sophisticated inference algorithms on equilibrium data.

We have examined low-dimensional and locally connected versus high-dimensional and globally connected networks. We thus expect that intermediate cases, such as small-world networks with a mixture of dense local and sparse long-distance connectivity, will exhibit similar behavior, a potential direction for future investigation. The present study can be extended in many ways. On the generative side, it will be interesting to study more varied neural circuit architectures and richer temporal dynamics (such as neural and synaptic adaption).

On the inference side, when the inference model exactly matches the generative model, when all neurons are observed and when the mapping from circuits to activity is injective, one can exactly estimate connectivity from activity (eg, Ising-on-Ising inference for certain architectures; Fig. 1). Although it is impossible for any inference model to truly match actual neural circuit dynamics, improvements in this match can shrink the gap between effective and structural connectivity. Finally, it will be important to broadly characterize inference improvements from far-out-of-equilibrium sampling.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41593-020-0699-2.

Received: 1 February 2019; Accepted: 28 July 2020;
Published online: 07 September 2020

References

1. Burns, R., et al. The Open Connectome Project data cluster: scalable analysis and vision for high-throughput neuroscience. In *Proceedings of the 25th International Conference on Scientific and Statistical Database Management*, 1–11 (Association for Computing Machinery, 2013).
2. Xu, C. S. et al. A connectome of the adult *Drosophila* central brain. Preprint at *bioRxiv* https://doi.org/10.1101/2020.01.21.911859 (2020).
3. Helmstaedter, M. et al. Connectomic reconstruction of the inner plexiform layer in the mouse retina. *Nature* **500**, 168 (2013).
4. Takemura, S.-Y. et al. A visual motion detection circuit suggested by *Drosophila* connectomics. *Nature* **500**, 175–181 (2013).
5. Lee, W.-C. A. et al. Anatomy and function of an excitatory network in the visual cortex. *Nature* **532**, 370–374 (2016).
6. Pillow, J. W. et al. Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature* **454**, 995–999 (2008).
7. Schneidman, E., Berry, M. J., Segev, R. & Bialek, W. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature* **440**, 1007–1012 (2006).
8. Friston, K. J. Functional and effective connectivity: a review. *Brain Connect.* **1**, 13–36 (2011).
9. Pakman, A., Huggins, J., Smith, C. & Paninski, L. Fast state-space methods for inferring dendritic synaptic connectivity. *J. Comput. Neurosci.* **36**, 415–443 (2014).
10. Ravikumar, P. et al. High-dimensional Ising model selection using l1-regularized logistic regression. *Ann. Stat.* **38**, 1287–1319 (2010).
11. Nelder, J. A. & Wedderburn, R. W. M. Generalized linear models. *J. R. Stat. Soc. Series A* **135**, 370–384 (1972).
12. Truccolo, W., Eden, U. T., Fellows, M. R., Donoghue, J. P. & Brown, E. N. A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects. *J. Neurophysiol.* **93**, 1074–1089 (2005).
13. Shlens, J. et al. The structure of multi-neuron firing patterns in primate retina. *J. Neurosci.* **26**, 8254–8266 (2006).
14. Pillow, J. W. and Latham, P. E. Neural characterization in partially observed populations of spiking neurons. In *Advances in Neural Information Processing Systems* 1161–1168 (NIPS, 2008).
15. Mishchencko, Y., Vogelstein, J. T. & Paninski, L. A Bayesian approach for inferring neuronal connectivity from calcium fluorescent imaging data. *Ann. Appl. Stat.* **5**, 1229–1261 (2011).
16. Ramirez, A. D. & Paninski, L. Fast inference in generalized linear models via expected log-likelihoods. *J. Comput. Neurosci.* **36**, 215–234 (2014).
17. Nykamp, D. Q. Revealing pairwise coupling in linear–nonlinear networks. *SIAM J. Appl. Math.* **65**, 2005–2032 (2005).
18. Soudry, D. et al. Efficient "shotgun" inference of neural connectivity from highly sub-sampled activity data. *PLoS Comput. Biol.* **11**, e1004464 (2015).
19. Kulkarni, J. E. & Paninski, L. Common-input models for multiple neural spike-train data. *Network* **18**, 375–407 (2007).
20. Vidne, M. et al. Modeling the impact of common noise inputs on the network activity of retinal ganglion cells. *J. Comput. Neurosci.* **33**, 97–121 (2012).
21. Brinkman, B. A. W., Rieke, F., Shea-Brown, E. & Buice, M. A. Predicting how and when hidden neurons skew measured synaptic interactions. *PLoS Comput. Biol.* **14**, e1006490 (2018).
22. Dunn, B. & Battistin, C. The appropriateness of ignorance in the inverse kinetic ising model. *J. Phys. A Math. Theor.* **50**, 124002 (2017).
23. Mehler, D. M. A. and Kording, K. P. The lure of causal statements: rampant mis-inference of causality in estimated connectivity. Preprint at *arXiv* https://arxiv.org/abs/1812.03363 (2018).
24. Burak, Y. & Fiete, I. R. Accurate path integration in continuous attractor network models of grid cells. *PLoS Comput. Biol.* **5**, e1000291 (2009).
25. Gutnisky, D. A. & Dragoi, V. Adaptive coding of visual information in neural populations. *Nature* **452**, 220–224 (2008).
26. Poort, J. & Roelfsema, P. R. Noise correlations have little influence on the coding of selective attention in area V1. *Cereb. Cortex* **19**, 543–553 (2009).
27. Samonds, J. M., Potetz, B. R. & Lee, T. S. Cooperative and competitive interactions facilitate stereo computations in macaque primary visual cortex. *J. Neurosci.* **29**, 15780–15795 (2009).
28. Cohen, M. R. & Maunsell, J. H. R. Attention improves performance primarily by reducing interneuronal correlations. *Nat. Neurosci.* **12**, 1594 (2009).
29. Mitchell, J. F., Sundberg, K. A. & Reynolds, J. H. Spatial attention decorrelates intrinsic activity fluctuations in macaque area v4. *Neuron* **63**, 879–888 (2009).
30. Huang, X. & Lisberger, S. G. Noise correlations in cortical area MT and their potential impact on trial-by-trial variation in the direction and speed of smooth-pursuit eye movements. *J. Neurophysiol.* **101**, 3012–3030 (2009).
31. Cohen, M. R. & Newsome, W. T. Context-dependent changes in functional circuitry in visual area MT. *Neuron* **60**, 162–173 (2008).
32. Zohary, E., Shadlen, M. N. & Newsome, W. T. Correlated neuronal discharge rate and its implications for psychophysical performance. *Nature* **370**, 140 (1994).
33. Bair, W., Zohary, E. & Newsome, W. T. Correlated firing in macaque visual area MT: time scales and relationship to behavior. *J. Neurosci.* **21**, 1676–1697 (2001).
34. Kemere, C., Carr, M. F., Karlsson, M. P. & Frank, L. M. Rapid and continuous modulation of hippocampal network state during exploration of new places. *PLoS ONE* **8**, e73114 (2013).
35. Burak, Y. & Fiete, I. R. Fundamental limits on persistent activity in networks of noisy neurons. *Proc. Natl Acad. Sci. USA* **109**, 17645–17650 (2012).
36. Mastromatteo, I. & Marsili, M. On the criticality of inferred models. *J. Stat. Mech.* **2011**, P10012 (2011).
37. Barahona, F. On the computational complexity of Ising spin glass models. *J. Phys. A Math. Gen.* **15**, 3241 (1982).
38. Sohl-Dickstein, J., Battaglino, P. B. & DeWeese, M. R. New method for parameter estimation in probabilistic models: minimum probability flow. *Phys. Rev. Lett.* **107**, 220601 (2011).
39. Thouless, D. J., Anderson, P. W. & Palmer, R. G. Solution of 'Solvable model of a spin glass'. *Philos. Mag.* **35**, 593–601 (1977).
40. Roudi, Y., Tyrcha, J. & Hertz, J. Ising model for neural data: model quality and approximate methods for extracting functional connectivity. *Phys. Rev. E* **79**, 051915 (2009).
41. Sessak, V. & Monasson, R. Small-correlation expansions for the inverse Ising problem. *J. Phys. A Math. Theor.* **42**, 055001 (2009).
42. Lee, S.-I., Lee, H., Abbeel, P. & Ng, A. Y. Efficient L1 regularized logistic regression. In The Twenty-First National Conference on Artificial Intelligence 6, 401–408 (Association for the Advancement of Artificial Intelligence, 2006).
43. Sompolinsky, H., Crisanti, A. & Sommers, H.-J. Chaos in random neural networks. *Phys. Rev. Lett.* **61**, 259 (1988).
44. Widrow, B. & Hoff, M. E. Adaptive switching circuits. Technical Report No. 1553-1. https://apps.dtic.mil/dtic/tr/fulltext/u2/241531.pdf (Office of Naval Research, 1960).

45. Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.* **8**, 229–256 (1992).
46. Fiete, I. R. & Seung, H. S. Gradient learning in spiking neural networks by dynamic perturbation of conductances. *Phys. Rev. Lett.* **97**, 048104 (2006).
47. Shababo, B., Paige, B., Pakman, A. & Paninski, L. Bayesian inference and online experimental design for mapping neural microcircuits. In *Advances in Neural Information Processing Systems* 1304–1312 (NIPS, 2013).
48. Aronov, D., Veit, L., Goldberg, J. H. & Fee, M. S. Two distinct modes of forebrain circuit dynamics underlie temporal patterning in the vocalizations of young songbirds. *J. Neurosci.* **31**, 16353–16368 (2011).
49. Casadiego, J., Nitzan, M., Hallerberg, S. & Timme, M. Model-free inference of direct network interactions from nonlinear collective dynamics. *Nat. Commun.* **8**, 2192 (2017).
50. Widloski, J., Marder, M. P. & Fiete, I. R. Inferring circuit mechanisms from sparse neural recording and global perturbation in grid cells. *eLife* **7**, e33503 (2018).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Generative network models. Here, we outline the network models that were used to generate the data.

Ring circuit. $n = 100$ neurons are arranged on a ring. The outgoing synaptic weights w_{ij} from each neuron to all of the others around the ring have a local Mexican hat (difference of Gaussians) shape:

$$w_{ij} = e^{-d_{ij}^2/2\sigma_1^2} - ae^{-d_{ij}^2/2\sigma_2^2}, \tag{1}$$

where d_{ij} is the distance (in neurons) between neurons i and j, $\sigma_1 = 6.98$ and $\sigma_2 =$ 7 (in units of neuron index). An excitatory bump at the Mexican hat center would require a second neural nonlinearity (saturation) to stabilize activity. Instead, in a small variation on typical Mexican hat architecture, the local excitation is replaced by a weak inhibition, by setting $a = 1.0005 > 1$. As a result, all recurrent weights are inhibitory; this permits both pattern formation (with a uniform feed-forward excitatory drive) and dynamical stability without need of a second (saturating) neural nonlinearity, similar to ref. [24].

Neural dynamics. Dynamics are updated in discrete time, with a time step Δt of size 0.1 ms. The summed input to each neuron at time step t is given by:

$$\boldsymbol{g}(t) = r\mathbf{W}\boldsymbol{s}(t) + \boldsymbol{b}(t), \tag{2}$$

where $s(t)$ is the $N \times 1$ vector of synaptic activations and W is the $N \times N$ matrix of recurrent connectivity defined above. The feed-forward inputs are $b(t) = b(1 + \xi(t))$, where $b = 0.001$ is a uniform excitatory drive and $\xi(t)$ is a multiplicative private Gaussian white noise per neuron, with zero mean and s.d. $\sigma_\xi = 0.3$, resulting in a Poisson-like variance proportional to the mean activation. This noise is only injected into each neuron with probability = .07 in each time step of the discrete-time equations, because, with more noise, the dynamics loses coherence. The relative influence of recurrent weights is scaled by the scalar weight strength parameter r.

If the input $g_i(t)$ to neuron i at time step t exceeds a threshold Θ, the neuron emits a spike. The binary vector of spikes from the network at t is $\sigma(t)$. The synaptic activation from neurons that just spiked is incremented by 1 and otherwise decays exponentially with time constant $\tau = 10$ ms according to the following equation:

$$\boldsymbol{s}(t + \Delta t) = \boldsymbol{s}(t)\left(1 - \frac{\Delta t}{\tau}\right) + \boldsymbol{\sigma}(t). \tag{3}$$

Generating spike data. To move from the weakly to the strongly coupled regime, we increase the weight strength r. The threshold Θ is adjusted at each r to hold the average inter-spike interval (ISI) of the network roughly fixed (within 16.0 ± 0.1 ms) (Extended Data Fig. 2). For each parameter setting, we initialize the network with random activations s and wait for it to equlibriate and then collect a total of 10^8 spikes from the network.

Linear–nonlinear Poisson model. A linear–nonlinear Poisson (LNP) network was used to generate data from the ring circuit $\mathbf{W}$, for inference with a GLM (Extended Data Fig. 8). Its neural inputs are the same as the original generative model (see Methods):

$$\boldsymbol{g}(t) = r_{\mathrm{LNP}}\mathbf{W}.\boldsymbol{s}(t) + \boldsymbol{b}, \tag{4}$$

except that the external input is uniform and constant ($b = 0.001$) without an additive noise component. The summed neural input is passed through a point-wise rectifying nonlinearity to obtain neural firing rates $\lambda(t)$. These firing rates determine the rate of an inhomogeneous Poisson process that generates $n_i(t)$ spikes in neuron i (iid per neuron) in time step t with probability:

$$\lambda_i(t) = \lambda_0\left[g_i(t) - \Theta_{\mathrm{LNP}}\right]_+, \quad n_i(t) \sim \mathrm{Pois}(\lambda_i(t)). \tag{5}$$

As before, the outgoing synaptic activations s follow:

$$\dot{\boldsymbol{s}}(t) + \frac{\boldsymbol{s}(t)}{\tau} = \boldsymbol{\sigma}(t), \quad \sigma_i(t) = \sum_{\mathrm{spikes}} \delta(t - t_i^{\mathrm{spike}}). \tag{6}$$

To maintain an average network ISI of 16.0 ± 0.1 ms at each weight strength r_{LNP}, λ_0 was fixed at 32, and the threshold Θ_{LNP} was adjusted for each weight strength (Extended Data Fig. 2).

Generalized linear model. To perform model-matched inference using GLMs, we need a generative GLM with the same exponential inverse link used for inference. This generative GLM is the same as the inference GLM (Methods), except for a slight parametric difference in the inverse link function:

$$\lambda_i(t) = \frac{1}{\alpha}e^{10^4 g_i(t)}. \tag{7}$$

The prefactor α was adjusted at each weight strength r_{GLM} to maintain the average network ISI at 16.0 ± 0.1 ms (Extended Data Fig. 2).

Ising model. To use an Ising model as a generative model, we set the Ising coupling matrix to $\mathbf{J} = r_{\mathrm{Ising}}\,\mathbf{W}$, where $\mathbf{W}$ is the ring circuit, and r_{Ising} was varied from 10 to 80 to be in the physical regime containing the inference optimum. The biases (external fields) of the Ising model are taken to be uniform: $h_i = 1$. Ising states were generated using a Gibbs sampling algorithm that updates the spins in a random sequence in each pass.

Random, fully connected, symmetric circuit. An all-to-all symmetric circuit of 100 neurons is constructed by drawing lower-triangular recurrent weights uniformly randomly from the same range as the Mexican hat weight profile of the ring circuit and reflecting these across the diagonal to generate a symmetric matrix (self-connections permitted). We generate spike data using the same neural dynamics as for the ring network (see Extended Data Fig. 2). Unlike in the ring network, we could not find a threshold to generate an ISI of 16 ms across weights; instead, we tuned the threshold to maintain an average ISI of 34.0 ± 0.1 ms across weak and moderate weight strengths. At strong weights, activity in the network abruptly switches to stable sparse patterns with much lower average ISI. At each weight strength, we collected an equal number of spikes for inference.

Random, sparse, asymmetric circuit. Recurrent weights are drawn uniformly randomly from the same range as that of the ring circuit for a network of 100 neurons, and then 90% of all weights are randomly selected and set to zero. We generate spikes from this network using the same neural dynamics as in the ring network, over a range of weight strengths r_{RSA} (setting the firing threshold to maintain an average ISI of 16.0 ± 0.1 ms at each weight; see Extended Data Fig. 2).

Random, fully connected, balanced circuit. We adapt a fully connected balanced network with asymmetric random weights[43]. The circuit comprises $n = 100$ neurons with weights $w_{ij} \sim N(\mu = 0, \sigma = 1/\sqrt{N})$. The governing dynamics of the neural fields (related to the membrane potentials) are given by:

$$\tau\dot{h}_i(t) = -h_i(t) + \sum_j r_{\mathrm{RFB}}\, w_{ij} \tanh(h_j(t)) + b_i(t), \tag{8}$$

with $\tau = 10$ ms. As the weight strength r_{RFB} is increased, this network transitions from fixed-point dynamics to limit cycles to chaos.

In the noisy version of this network, the feed-forward drive $b_i(t) = \xi_i(t)$, independent Gaussian white noise with zero mean and unit s.d. to each neuron.

Characterizing the dynamical regime of the ring network. *Noise correlations.* The spatially periodic activity pattern of the ring network drifts around the ring over time; this effectively induces a signal correlation between neurons, which we wish to exclude. Therefore, we silence the feed-forward input b to one of the neurons, which forces a dip in activity there, pinning the periodic pattern in place. We then generate spikes from the network, bin them at 10 ms and calculate the correlation coefficient between the spike count trains of neurons, excluding the suppressed neuron.

Activity pattern coherence. The degree of coherence of the activity pattern of the ring network is reflected in the periodicity of the pattern, which is captured by activity correlations regardless of the pattern phase. We compute the correlation coefficient matrix for all neuron pairs (with the diagonal zeroed out). Periodicity in the pattern is reflected as a periodicity in the rows of this matrix. The average autocorrelation of the rows is the mean pattern autocorrelation. We compute the average of this mean autocorrelation function at three shifts—at $N/4$, $N/2$ and $3N/4$—which are multiples of the four-bump pattern period. This measure is close to zero when there is weak or no pattern coherence and approaches 1 for a coherent, four-period pattern.

Diffusivity of activity pattern phase. A measure of the temporal stability of the pattern is the rate at which it random walks around the ring. We compute the location of the pattern as the time-varying phase $\phi(t)$ of the Fourier component of the neural input vector $g(t)$ at the spatial frequency of the activity pattern (four cycles around the ring). The slope of the linear fit of $\langle(\phi(t+\tau) - \phi(t))^2\rangle_t$ against τ then gives the diffusion coefficient of the random walk. We did not compute diffusivity for weights below $r = 0.0125$, because g was too noisy to extract a periodic pattern.

Circuit inference. *Generalized linear model.* Here, we outline the GLM used for many of our inference results. The model assumes neural inputs $\hat{\boldsymbol{g}}(t)$ at time t of the form:

$$\hat{\boldsymbol{g}}(t) = \hat{\mathbf{W}}((a \star \hat{\boldsymbol{\sigma}})(t)) + \hat{\boldsymbol{b}} \tag{9}$$

where the spike trains $\hat{\boldsymbol{\sigma}}$ are convolved with a decaying temporal kernel $a = \hat{a}e^{-t/\hat{\tau}}$, truncated at $T = 200$ time steps, with $\hat{\tau} = \tau$ (the decay time constant in the kernel is set to the biophysical synaptic time constant in the generative neural network model). The filtered spike trains are weighted by effective weights $\hat{\mathbf{W}}$ and summed together with an additive term $\hat{\boldsymbol{b}}$ and then passed through a nonlinearity (inverse link function) ϕ. The resulting rate drives a Poisson spike generation process:

$$\hat{\boldsymbol{\lambda}}(t) = \phi[\hat{\boldsymbol{g}}(t)], \quad \hat{\boldsymbol{\sigma}}(t) \sim \mathrm{Pois}\left(\hat{\boldsymbol{\lambda}}(t)\hat{\Delta t}\right). \tag{10}$$

We used $\phi(\,\cdot\,) = exp(\,\cdot)$, the canonical inverse link function for the Poisson distribution. $\hat{\Delta t}$ is the temporal discretization used for inference. We chose $\hat{\Delta t} = \Delta t$ —that is, the time discretization for the inference model matched the discretization for the generation model. These choices are conservative, in that they ensure the best possible match between the generative and inference models, aside from the mismatch in the inverse link function relative to the neural nonlinearity and the lack of an additive Gaussian noise in the input to each neuron. This model is like a dynamical GLM[51] but with parameters ($\hat{\mathbf{W}}$,$\hat{\boldsymbol{b}}$) tied across time. When T in the kernel function is much larger than the exponential decay time constant, the nonlinearity ϕ is the same as the neural nonlinearity in the generative model; and, if the additive term $\hat{\boldsymbol{b}}$ included an injected Gaussian noise term with statistics as in the generative model, this inference model would exactly match the generative model.

The parameters $\hat{\mathbf{W}}$, $\hat{\boldsymbol{b}}$ and $\hat{a}$ are inferred by simultaneous gradient descent on the log-likelihood of the model given the data.

CCG gating. Short latency peaks in the CCG of neuron pairs are often used to infer the existence of a direct synaptic connection. For the symmetric ring network, we compute the (Pearson's) cross-correlation of the spikes of neuron pairs and check if it peaks (positively or negatively) within a lag of τ, to posit a connection. For the weight estimate, we may use the value of the peak cross-correlation itself, or the weight inferred by some other inference method (eg, GLM) (Extended Data Fig. 7).

For our random, sparse and non-symmetric network, the CCG provides better discriminatory information. Broad symmetric peaks in the CCG can now be explained away to multiple indirect connections, whereas directly connected neurons will show sharp asymmetric short-lag features. Normally, one looks for a peak shifted by a few milliseconds as an indication of a direct connection. But the inhibitory connections of this network produce, instead, a sharp asymmetric dip at zero and then recovery (Extended Data Fig. 7). We detect these by computing the time derivatives (discrete difference) of the CCG at each shift, expressing them as z-scores and detecting any discontinuity at zero greater than a threshold z-score of 3. Once a connection is posited using this criterion, the weight estimate from an inference method, such as GLM, can be used.

Ising inference model. In the limit of infinite data generated from a matched Ising model, Boltzmann machines perform exact inference, but they are slow. At the other end of the speed–accuracy tradeoff for inverse Ising inference are mean-field methods, which generate fast approximations on the coupling inference problem.

Inverse Ising inference performs best when the spike data are binned at an appropriate time scale. The Ising model additionally accommodates only binary data, thus too-large bins result in information loss. We binarize spike counts (spike/no spike) in bins of various widths. We then apply the MPF[38] algorithm to solve the inverse Ising inference problem and determine inference error as a function of bin width. We select the bin width that yields minimum inference error for all comparisons (Extended Data Fig. 3b). We do the same for logistic regression.

Bin width optimization requires knowledge of the ground truth weight matrix. Thus, it is a supervised approach that provides an upper bound on inference quality when applied without knowledge of the true connectivity matrix. We find, however, that the optimal bin width is approximately τ, the biophysical time scale of single-neuron integration in the generative model network. This is reasonable because τ is the time scale over which a neuron integrates its inputs before responding with a spike and thus is the relevant window for revealing connectivity (Extended Data Fig. 3a). This suggests that, in general, when bin width cannot be directly optimized by comparison with the ground truth weight matrix, setting it to equal the neural integration time constant is a good choice.

With stronger weights, activity peaks become higher and narrower, resulting in a less equal distribution of spikes across bins, so more spikes must be discarded (Extended Data Fig. 3c). To implement inverse Ising inference on a total of 10^8 spikes, the actual number of spikes collected from the network before binarization was greater.

Mean-field Ising models. Mean-field Ising models are a fast way to infer weights from binarized spikes, but their simplifying assumptions produce different degrees of approximate solutions[52]. The following are expressions for the weights with the naive mean field, which we also call 'raw correlations' (C is the spike train covariance matrix, and m_i is the average state of the i^{th} node):

$$J_{ij}^{\mathrm{NMF}} = -(C^{-1})_{ij}, \tag{11}$$

the TAP model[39]:

$$J_{ij}^{\mathrm{TAP}} = \frac{\sqrt{1 - 8m_i m_j (C^{-1})_{ij}} - 1}{4m_i m_j}, \tag{12}$$

and the SM model[40,41]:

$$J_{ij}^{\mathrm{SM}} = \tfrac{1}{4}\ln\frac{[(1+\mathrm{m_i})(1+\mathrm{m_j})+\mathrm{C_{ij}}][(1-\mathrm{m_i})(1-\mathrm{m_j})+\mathrm{C_{ij}}]}{[(1+\mathrm{m_i})(1-\mathrm{m_j})-\mathrm{C_{ij}}][(1-\mathrm{m_i})(1+\mathrm{m_j})-\mathrm{C_{ij}}]} - \left(\mathrm{C}^{-1}\right)_{\mathrm{ij}} - \frac{\mathrm{C_{ij}}}{(1-\mathrm{m_i^2})(1-\mathrm{m_j^2})-(\mathrm{C_{ij}})^2}. \tag{13}$$

l_1-regularized logistic regression. When performing logistic regression, we first mean-subtract all the spike channels. Then, we binarize the spike counts of the dependent channel but leave the predictor channels unbinarized. The matrix of regression coefficients of each node against all others is taken to be the inferred weight matrix. At each weight strength, we find the regularization strength λ for the l_1-norm of the inferred coefficients, which minimizes the inference error (Extended Data Fig. 6). This is not possible without knowledge of the ground truth, but we use it to provide an upper bound of performance.

Inference with matched models. To contrast the real-world scenario in which the inference model is not exactly matched with the generative model, we consider the theoretically idealized case in which they are identical. We use the ring circuit and the Ising model for both generation and inference (Ising-on-Ising inference) and also the same GLM (GLM-on-GLM inference).

Comparing the Ising-on-Ising inference error curve with the pattern coherence curve (the Ising model does not have a notion of time, so diffusivity cannot be computed) reveals that optimal inference is now just inside the memory regime (Fig. 1). Unlike with mismatched models, and as theoretically expected, there is no bias error here; all inference errors are variance errors, which decay with data volume as $\Delta^2 \sim 1/D$ at all weights, as seen in the uniform drop in the log–log axes of Fig. 1a, secondary axis (see Extended Data Fig. 5 for power law fits). Consequently, the optimal inference point remains stationary with increasing data, reflecting an intrinsic critical point of the system.

These findings are qualitatively reproduced with GLM-on-GLM inference (see Methods). Like Ising-on-Ising inference, inference error is almost entirely due to variance at all weights. Optimal inference is again just inside the memory regime at the point of pattern onset, and the point of optimal inference remains stationary, because increasing data volume erodes variance error everywhere.

Unfortunately, when one is inferring neural connectivity from experimental recordings, there will always be a gap between the dynamical system generating the data and the inference model: the full biological dynamics that generate the observed activity are unknown and far more complex than theoretical models describe. Thus, data-immune bias errors become inevitable, at least when recurrent interactions are moderate or strong.

Measuring inference error. The true and inferred weights are the elements of $\mathbf{W}$ and $\hat{\mathbf{W}}$. Because the inference model is generally different from the generative model, $\hat{\mathbf{W}}$ might have some arbitrary overall scale factor with respect to $\mathbf{W}$, which we wish to ignore. So the first step is to scale $\hat{\mathbf{W}}$ to match the scale of $\mathbf{W}$, which we do in the following way.

When $\mathbf{W}$ is circulant, as in the ring circuit, each row is a rotation of a single weight-shape vector $\boldsymbol{\omega}$. The rows of $\hat{\mathbf{W}}$ are then noisy rotations of each other. Let $\hat{\boldsymbol{\omega}}_i$ represent row i in $\hat{\mathbf{W}}$ (re-aligned; that is, appropriately unrotated). Then, the average weight-shape estimate $\overline{\boldsymbol{\omega}} = \frac{1}{N}\sum_i \hat{\boldsymbol{\omega}}_i$. We re-scale $\hat{\mathbf{W}}$ to minimize the l_1 deviation between $\overline{\boldsymbol{\omega}}$ and the true shape $\boldsymbol{\omega}$. The averaging and the l_1 measure make this scaling tolerant to noise.

After $\hat{\mathbf{W}}$ has been scaled, the inference error is the l_2 distance between the vectors of ground truth and inferred weights, expressed as a fraction of the vector magnitude of the ground truth weights:

$$\Delta = \frac{\parallel \mathbf{W} - \hat{\mathbf{W}} \parallel}{\parallel \mathbf{W} \parallel}. \tag{14}$$

In the case where the weights are not translation invariant (eg, in the random circuits), we resort to the less noise-tolerant choice of scaling the vector of weights in $\hat{\mathbf{W}}$ to have the least l_2 distance from the weights in $\mathbf{W}$. The inference error is then the sine of the angle between the ground truth and inferred weight vectors in the space of weight parameters.

Variance and bias errors. For circulant $\mathbf{W}$, let us denote the $N - 1$ elements (ignoring the self-coupling term) of the shape vectors $\boldsymbol{\omega}$, $\hat{\boldsymbol{\omega}}_i$ by ω^α, $\hat{\omega}_i^\alpha$. The square of the inference error (Equation 14) can thus be expressed as:

$$\Delta^2 = \frac{\sum_{\alpha,i}\left(\omega^\alpha - \hat{\omega}_i^\alpha\right)^2}{\parallel \mathbf{W} \parallel^2}. \tag{15}$$

This can be rewritten in terms of $\overline{\omega}^\alpha = \frac{1}{N}\sum_i \hat{\omega}_i^\alpha$, the elements of the average estimated weight shape $\overline{\boldsymbol{\omega}}$:

$$\Delta^2 = \frac{\sum_{\alpha,i}\left[(\omega^\alpha - \overline{\varpi}^\alpha) + \left(\overline{\varpi}^\alpha - \hat{\omega}_i^\alpha\right)\right]^2}{\|\mathbf{W}\|^2} \tag{16}$$

$$= \frac{(N-1)\sum_\alpha(\omega^\alpha - \overline{\varpi}^\alpha)^2}{\|\mathbf{W}\|^2} + \frac{(N-1)\sum_\alpha \mathrm{var}_i\left(\hat{\omega}_i^\alpha\right)}{\|\mathbf{W}\|^2} \tag{17}$$

$$= \frac{\|\mathbf{W} - \overline{\mathbf{W}}\|^2}{\|\mathbf{W}\|^2} + \frac{\|\hat{\mathbf{W}} - \overline{\mathbf{W}}\|^2}{\|\mathbf{W}\|^2} \equiv \Delta_b^2 + \Delta_v^2. \tag{18}$$

where $\mathrm{var}_i\left(\hat{\omega}_i^\alpha\right) = \frac{1}{N-1}\sum_i\left(\hat{\omega}_i^\alpha - \overline{\varpi}^\alpha\right)^2$ is the variance of the estimates of ω^α and $\overline{\mathbf{W}}$ is a circulant matrix consisting of rotations of the mean estimate $\overline{\boldsymbol{\varpi}}$. We call $\Delta_v = \frac{\|\hat{\mathbf{W}} - \overline{\mathbf{W}}\|}{\|\mathbf{W}\|}$ the variance error, as it is the error due to the variance among the different noisy estimates ω_i. We call $\Delta_b = \frac{\|\mathbf{W} - \overline{\mathbf{W}}\|}{\|\mathbf{W}\|}$ the bias error, as it is the error due to deviation of the mean estimate $\overline{\boldsymbol{\varpi}}$ from the true weights ω.

Geometrically, Δ_v and Δ_b are the lengths, measured in units of $\|\mathbf{W}\|$, of two orthogonal vector steps going from the true to the inferred weights in the space of weight parameters. Their orthogonality can be seen through:

$$(\overline{\mathbf{W}} - \mathbf{W}).(\hat{\mathbf{W}} - \overline{\mathbf{W}}) = \sum_{\alpha,i}(\overline{\varpi}^\alpha - \omega^\alpha)(\hat{\omega}_i^\alpha - \overline{\varpi}^\alpha) \tag{19}$$

$$= \sum_\alpha\left[(\overline{\varpi}^\alpha - \omega^\alpha)\sum_i(\hat{\omega}_i^\alpha - \overline{\varpi}^\alpha)\right] = \sum_\alpha(\overline{\varpi}^\alpha - \omega^\alpha)(N-1)(\overline{\varpi}^\alpha - \overline{\varpi}^\alpha) = 0. \tag{20}$$

Hence, the total inference error is the magnitude of a vector given by the sum of the perpendicular vector components of variance and bias.

The relative contribution of bias to the inference error can thus be measured by the angle between the vectors associated with the total inference error and the variance error, as a fraction of 90°:

$$\theta_b = \frac{\tan^{-1}\frac{\Delta_b}{\Delta_v}}{90^\circ}. \tag{21}$$

If the true matrix is non-circulant, these variance and bias errors cannot be computed for a single inferred weight matrix. In that case, it is possible to break the data into batches, compute separate inferred matrices for each batch and then compute the variance errors across batches instead of across rows of a single inferred matrix.

Negentropy of inference errors. Once we scale-match $\hat{\mathbf{W}}$ with $\mathbf{W}$, the elements of $\mathbf{W} - \mathbf{W}$ are the errors in individual inferred weights. Negentropy is a way to characterize how much the distribution of these errors differs from Gaussian (which would result from purely random errors). If the error distribution has variance σ^2, and p_i denotes the relative frequencies in the distribution histogram binned with bin size b, the negentropy is:

$$\log\left(2\pi e\sigma^2\right) + \sum_i p_i \log\frac{p_i}{b}. \tag{22}$$

This is the difference between the differential entropy of a normal distribution with the same variance compared to a continuous version of the discrete entropy of the error distribution.

Discriminating circuits using activity data. *Constructing the alternative circuit.* To examine the inherent indistinguishability of models at strong weights, we construct an alternative network weight matrix $\mathbf{W}'$ based on the connectivity inferred by inverse Ising inference on the original ring model in the strong weight regime. We compute a mean weight-shape vector $\overline{\boldsymbol{\varpi}}$ from the inferred weight matrix by averaging the rows (after appropriately unrotating). Next, we set the positive parts of the shape vector to zero (to ensure stability with threshold-linear neurons) and re-scale to match the minimum with that of the original, ground truth weight matrix. Finally, we create a circulant matrix from this weight shape.

As with the original circuit, firing thresholds Θ for this circuit are tuned at each weight strength to set the average network ISI at 16.0 ± 0.1 ms (Extended Data Fig. 2). We can now generate spike data from this circuit using the original dynamics.

Selecting a well-sampled subspace of neural activity. Even with binary spike data, there are 2^N possible instantaneous spike words (vector of spike counts at one time), or states, of an N-neuron network, prohibitively big for large N ($\approx 10^{30}$ for $N = 100$). The number of states obtained from 6 h of data binned at 10 ms is of the order of 10^6, a miniscule fraction of all states, and statistics to characterize the data distribution, such as entropy, will be biased[53,54].

To address this problem, we consider, instead, the distribution of binary spike words of segments of n adjacent neurons from our ring network (neurons 1 through n, 2 through $n+1$, etc, pooled together). Ideally, n should be as large as possible while making sure that the states are well enough sampled to accurately estimate the desired statistical measures. This value is dictated by the particular statistical measure, as explained in the following sections.

Information theoretic measures of neural activity distribution. **Entropy.** A distribution can be considered well sampled for entropy estimation if the entropies computed with increasing fractions of the data converge as we approach the total data volume. Our data volumes permitted convergence in the entropy estimates for spike words of length up to $n = 22$ in the ring network (see Extended Data Fig. 10). As described in the following sections, we were constrained to smaller n for other measures.

Relative entropy. The dissimilarity between the data distributions $p_\mathbf{W}$ and $p_{\mathbf{W}'}$, considering the former as the 'true' distribution, is measured by their relative entropy (KL divergence):

$$D_{\mathrm{KL}}(p_\mathbf{W}, p_{\mathbf{W}'}) = \sum_{\boldsymbol{\sigma} \in \mathrm{supp}(p_\mathbf{W})} p_\mathbf{W}(\boldsymbol{\sigma})\log\frac{p_\mathbf{W}(\boldsymbol{\sigma})}{p_{\mathbf{W}'}(\boldsymbol{\sigma})} \tag{23}$$

where σ runs over the n-neuron spike vectors that constitute the support of $p_\mathbf{W}$.

Fisher information. FI can be used to quantify the amount of information contained in data about the generative model. It is measured by the sensitivity of the data distribution to changes in the model parameters.

It is not feasible to compute the FI about each circuit parameter—that is, the $\binom{N}{2} = 4,950$ weights in $\mathbf{W}$. Instead, consider a single-parameter family of models that passes through the true circuit $\mathbf{W}$ and the circuit with non-local, correlation-based weights $\mathbf{W}'$:

$$\mathbf{W}(\theta) = (1-\theta)\mathbf{W} + \theta\mathbf{W}'. \tag{24}$$

with $\theta \in [0, 1]$. The FI $I(\theta = 0)$ about the true model $W = W(\theta = 0)$ can be written in terms of the KL divergence between the data distribution p_W ($= p_0$) that it produces and the distribution $p_{d\theta}$ that a neighboring model $\mathbf{W}(d\theta)$ produces[55]:

$$D_{\mathrm{KL}}(p_0, p_{d\theta}) \approx \frac{1}{2}d\theta^2 \underbrace{\sum_{\boldsymbol{\sigma}\in\mathrm{supp}(p_0)} p_0(\boldsymbol{\sigma})\left(\left.\frac{d\log \mathrm{p}_\theta(\boldsymbol{\sigma})}{d\theta}\right|_{\theta=0}\right)^2}_{I(\theta=0)}, \tag{25}$$

where σ again constitute n-length binary spike vectors.

We approximate this measure by constructing the circuit $\mathbf{W}(\theta = \frac{1}{3})$ as a neighboring model in the family. At each weight strength, we adjust its firing threshold to maintain an average network ISI of 16.0 ± 0.1 ms as before, generate spike data, binarize counts and compute the KL divergence with respect to the data distribution generated by $\mathbf{W}$. The FI is then approximated by:

$$I \approx 18 D_{\mathrm{KL}}(p_\mathbf{W}, p_{\mathbf{W}(\theta=\frac{1}{3})}). \tag{26}$$

Likelihood ratio of circuit models. The log-likelihood that the observed data distribution p_W was produced by the circuit $\mathbf{W}'$ is:

$$\log p(\mathbf{W}'|p_\mathbf{W}) = \sum_{\boldsymbol{\sigma} \in \mathrm{supp}(p_\mathbf{W})} s\, p_\mathbf{W}(\boldsymbol{\sigma})\log p'(\boldsymbol{\sigma}) \tag{27}$$

where s is the total number of samples in $p_\mathbf{W}$.

We can collect a second sample $\tilde{p}_\mathbf{W}$ from the local circuit to account for sample-to-sample variability and then calculate the log-likelihood ratio of $\mathbf{W}$ versus $\mathbf{W}'$, given data generated from $\mathbf{W}$:

$$\log\frac{p(\mathbf{W}|p_\mathbf{W})}{p(\mathbf{W}'|p_\mathbf{W})} = \sum_{\boldsymbol{\sigma} \in \mathrm{supp}(p_\mathbf{W})} s\, p_\mathbf{W}(\boldsymbol{\sigma})\log\frac{\tilde{p}_\mathbf{W}(\boldsymbol{\sigma})}{p_{\mathbf{W}'}(\boldsymbol{\sigma})} \tag{28}$$

For the relative entropy, FI and likelihood ratio to be defined, the distributions must be sampled well enough that each spike state σ that occurs in $p_\mathbf{W}$ occurs in $\tilde{p}_\mathbf{W}$, $p_{\mathbf{W}'}$ and $p_{\mathbf{W}(\theta=\frac{1}{3})}$. With the total volume of spike data that we collected from the network, the largest value of n for which this was possible is 10; at this n, the number of state samples in the data exceeded the total number of possible states by a factor of more than 100.

Inferring a partially observed network. For each of a range of sub-population sizes n, we randomly select multiple n-neuron sub-populations from the 100-neuron ring network. We use the spike data from each such sub-population to infer its corresponding $n \times n$ connectivity submatrix.

Each n-neuron sub-network covers a fraction $\frac{n^2-n}{N^2-N}$ of the full weight matrix (ignoring diagonal terms); thus, randomly selecting

$$s_n(p) = \left\lceil\frac{\log(1-\mathrm{p})}{\log\left(1-\frac{\mathrm{n}^2-\mathrm{n}}{\mathrm{N}^2-\mathrm{N}}\right)}\right\rceil \tag{29}$$

sub-networks ensures with probability p that all synapses of the full network have been sampled at least once. For each sub-population size n, we collect enough samples to have $p = 0.99$.

For more information on the methods, refer to the Life Sciences Reporting Summary associated with this paper.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data and code availability

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

References

51. West, M., Harrison, P. J. & Helio, S. M. Dynamic generalized linear models and bayesian forecasting. *J. Am. Stat. Assoc.* **80**, 73–83 (1985).
52. Federico, R.-T. The Bethe approximation for solving the inverse Ising problem: a comparison with other inference methods. *J. Stat. Mech.* **2012**, P08015 (2012).
53. Nemenman, I., Shafee, F. & Bialek, W. Entropy and inference, revisited. In *Advances in Neural Information Processing Systems* 471–478 (NIPS, 2002).
54. Paninski, L. Estimating entropy on m bins given fewer than m samples. *IEEE Trans. Inf. Theory* **50**, 2200–2203 (2004).
55. Jeffreys, H. An invariant form for the prior probability in estimation problems. *Proc. Roy. Soc. Lond. Math. Phys. Sci.* **186**, 453–461 (1946).

Acknowledgements

We thank R. Chaudhuri, B. Gerçek, I. Kanitscheider, T. Taillefumier, D. Schwab, M. Bethge and P. Dayan for helpful discussions. This project was funded in part by the ONR, the Howard Hughes Medical Institute through the Faculty Scholars Program and the Simons Collaboration on the Global Brain through the Simons Foundation.

Competing interests

The authors declare no competing interests.

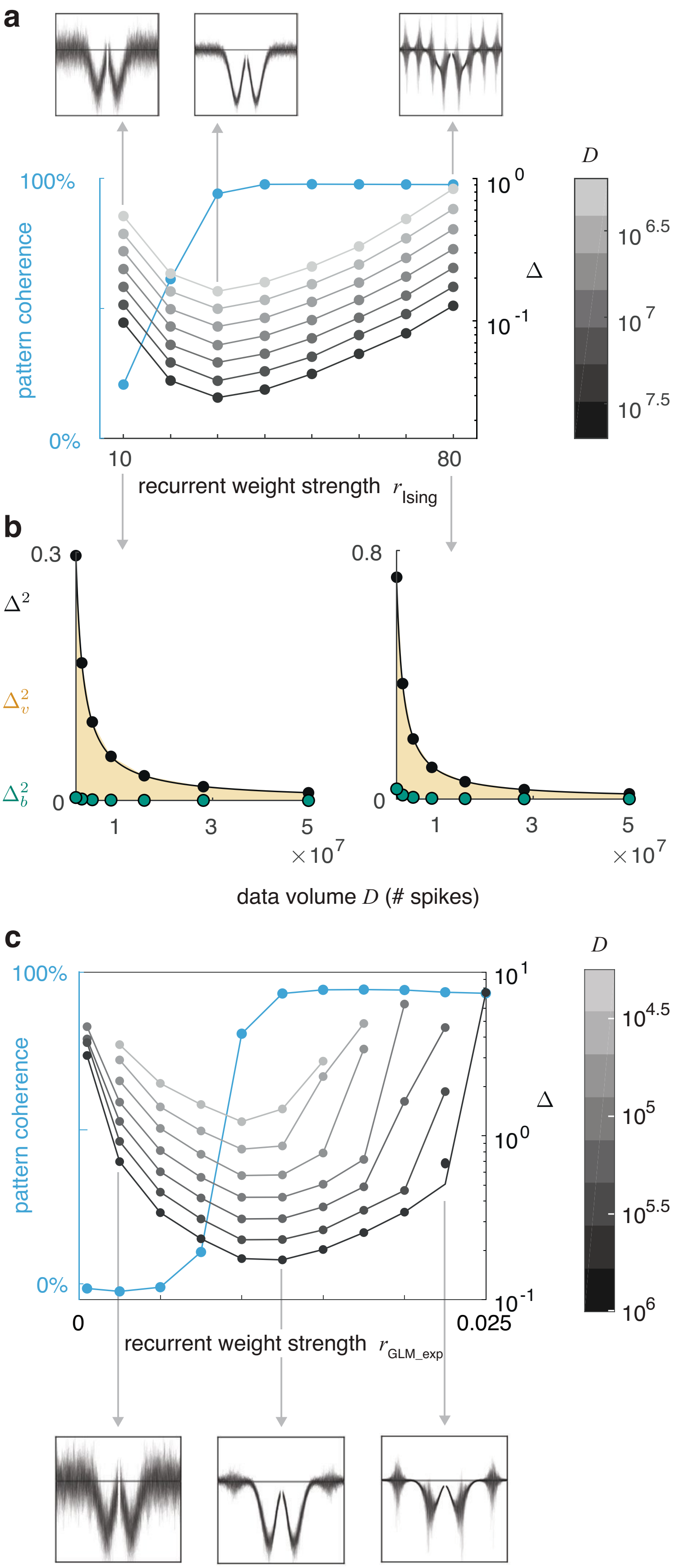
Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41593-020-0699-2.

Supplementary information is available for this paper at https://doi.org/10.1038/s41593-020-0699-2.

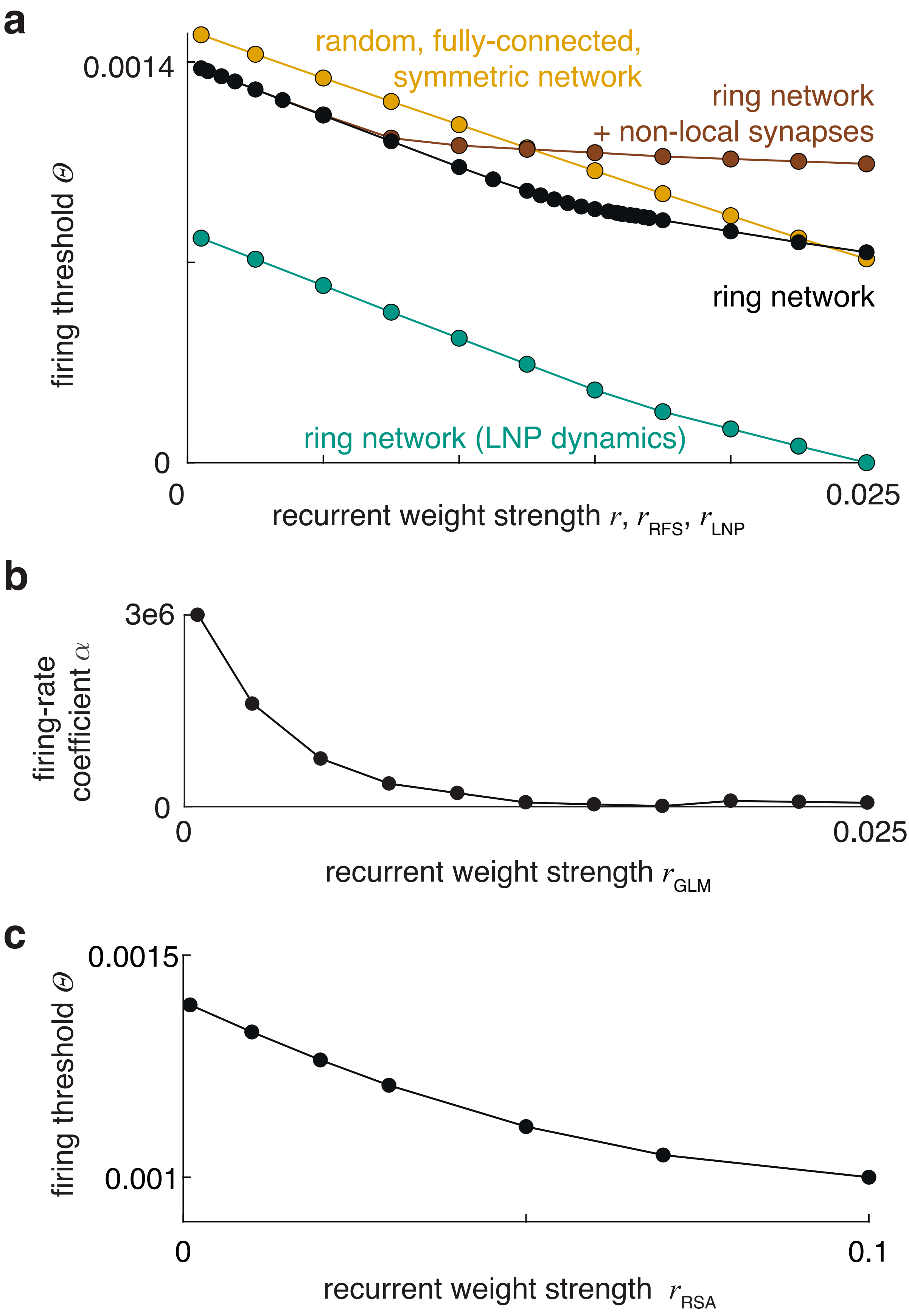
Correspondence and requests for materials should be addressed to I.R.F.

Reprints and permissions information is available at www.nature.com/reprints.



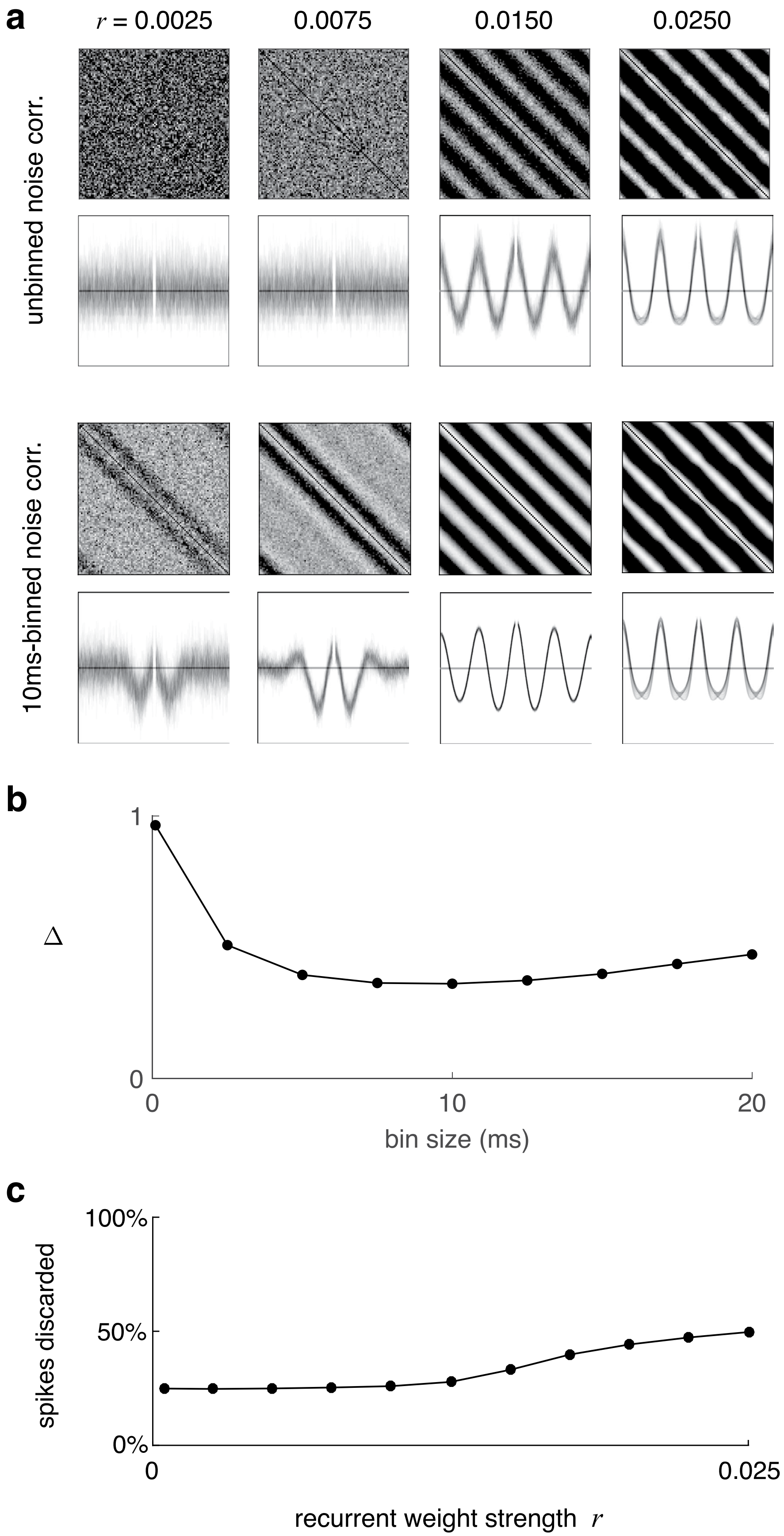
Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Inference with matched models. a, Using an Ising model for both generation and inference. Top: superposed inferred weights from each node to the rest (line marks zero). Bottom: pattern coherence, and inference error with different data volumes, against weight strength. **b**, Squared total, variance and bias errors against data volume at weak and strong weights. **c**, Using a generalized linear model with an exponential nonlinearity for both generation and inference (see Methods). Top: pattern coherence, and inference error with different data volumes, against weight strength. Bottom: Superposed inferred weights from each node to the rest.



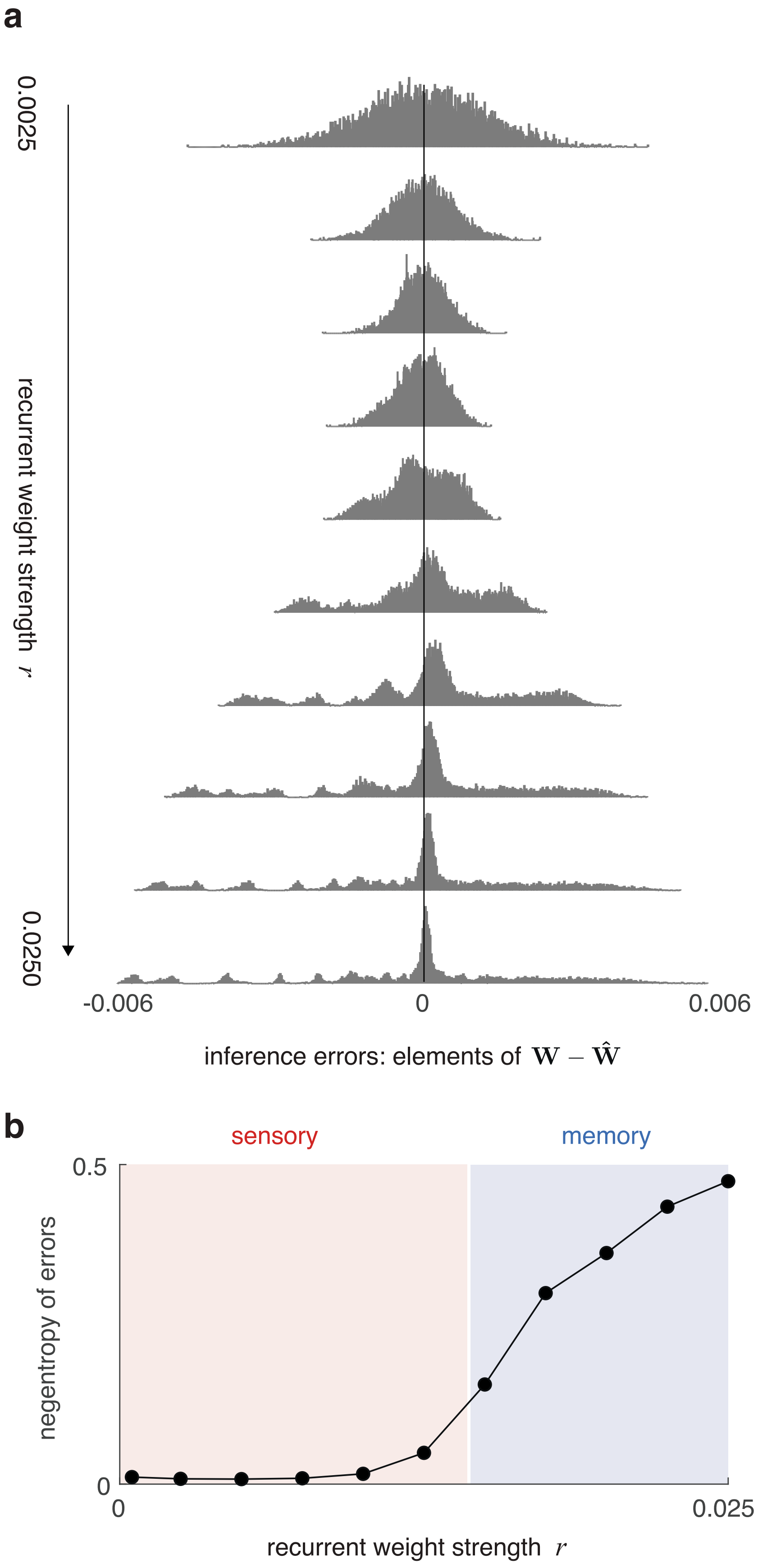
Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Tuning firing thresholds of different generative networks to control the average network inter-spike-interval as recurrent weight strength is varied. a, Firing thresholds to hold the average ISI of the ring network with local and non-local synapses, and the local ring network with rectifying LNP dynamics, at 16.0 ± 0.1 ms, and that of the random, fully-connected, symmetric network at 34.0 ± 0.1 ms. **b**, Firing rate coefficients α (see eq. (7)) to hold the average ISI of the ring network with GLM dynamics (logarithmic link function) at 16.0 ± 0.1 ms. **c**, Firing thresholds to hold the average ISI of the sparse, non-symmetric random network at 16.0 ± 0.1 ms.



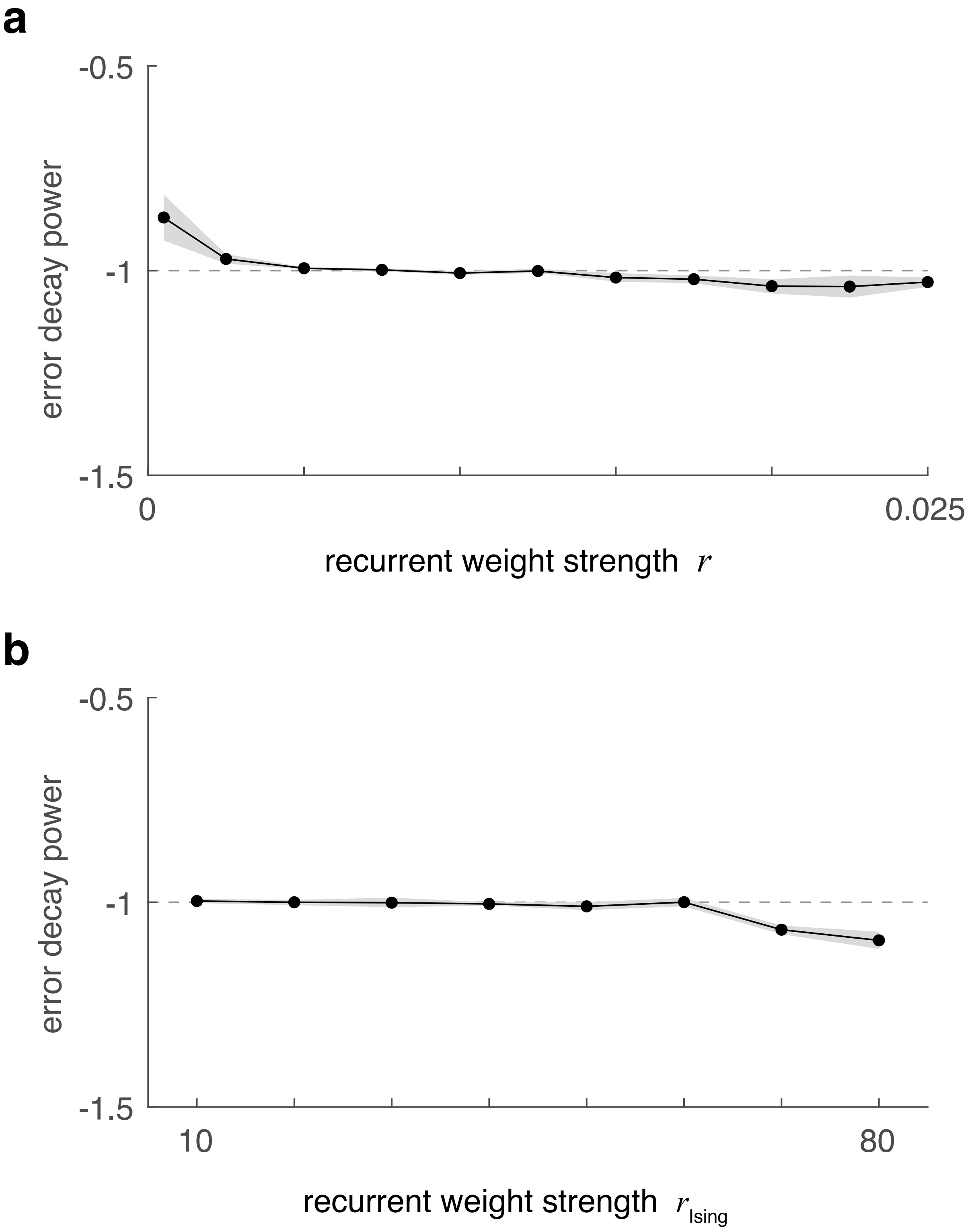
Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Binning affects noise correlations and inference. a, Noise correlations between neuron pairs (top: full matrix, bottom: superposed rows) for binned vs unbinned spikes from the ring network. The optimal bin-width groups causally related spikes together, and noise correlations at an intermediate *r* then reflect underlying weights. **b**, Inference error using inverse Ising with MPF on spike data binned at different widths. **c**, Fraction of spikes discarded when binarizing binned spike data for Inverse Ising inference.

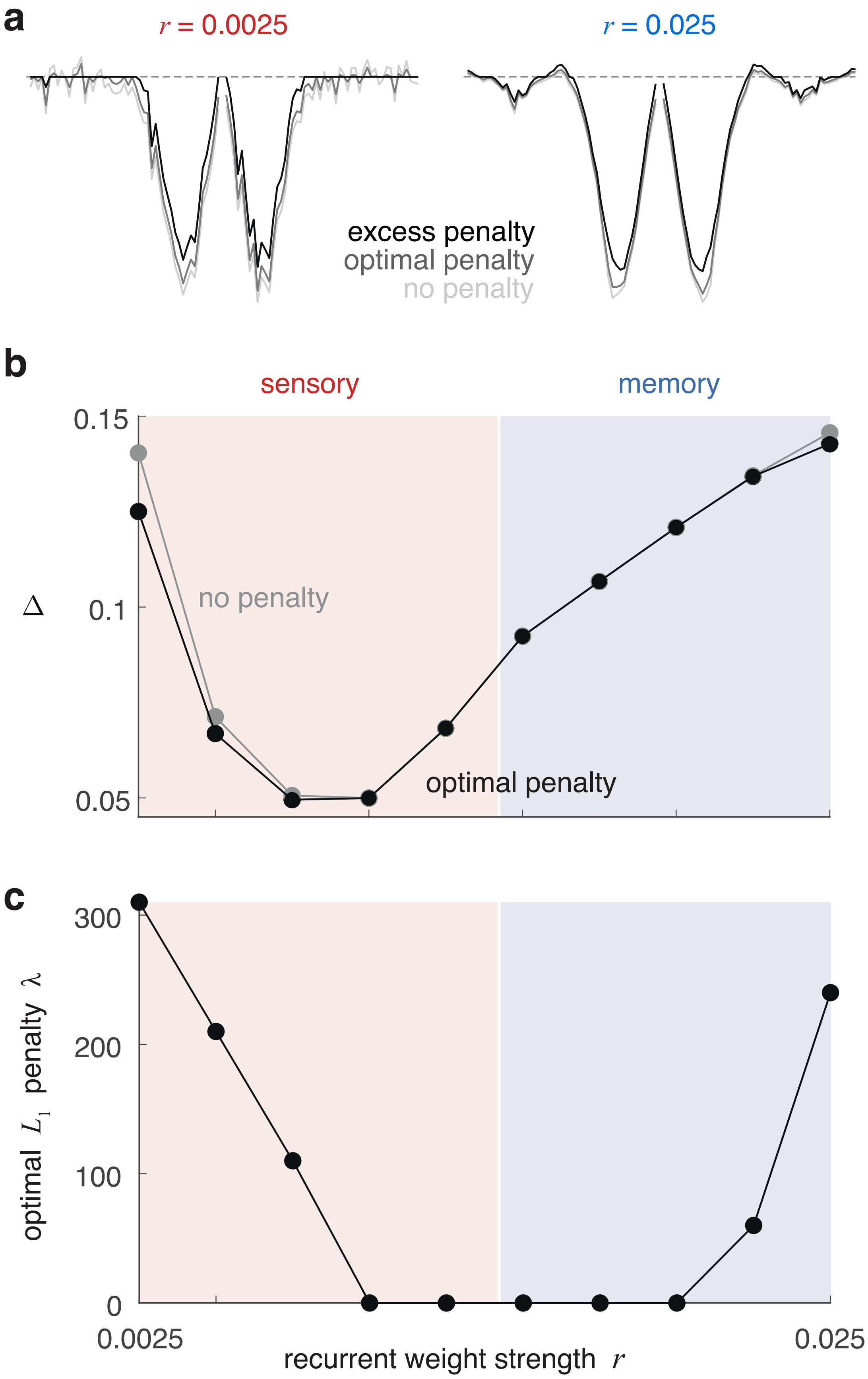


Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Distribution of inference errors of individual weights in the ring circuit, at different recurrent weight strengths. **a**, Histograms of the inference errors (relative to the length of the ground-truth weight vector). At weak weights, errors are random and normally distributed. As the weight increases, errors first shrink as noise weakens and SNR grows, then they become increasingly non-normal due to bias. **b**, Negentropy of the error distribution (see Methods) against weight strength.

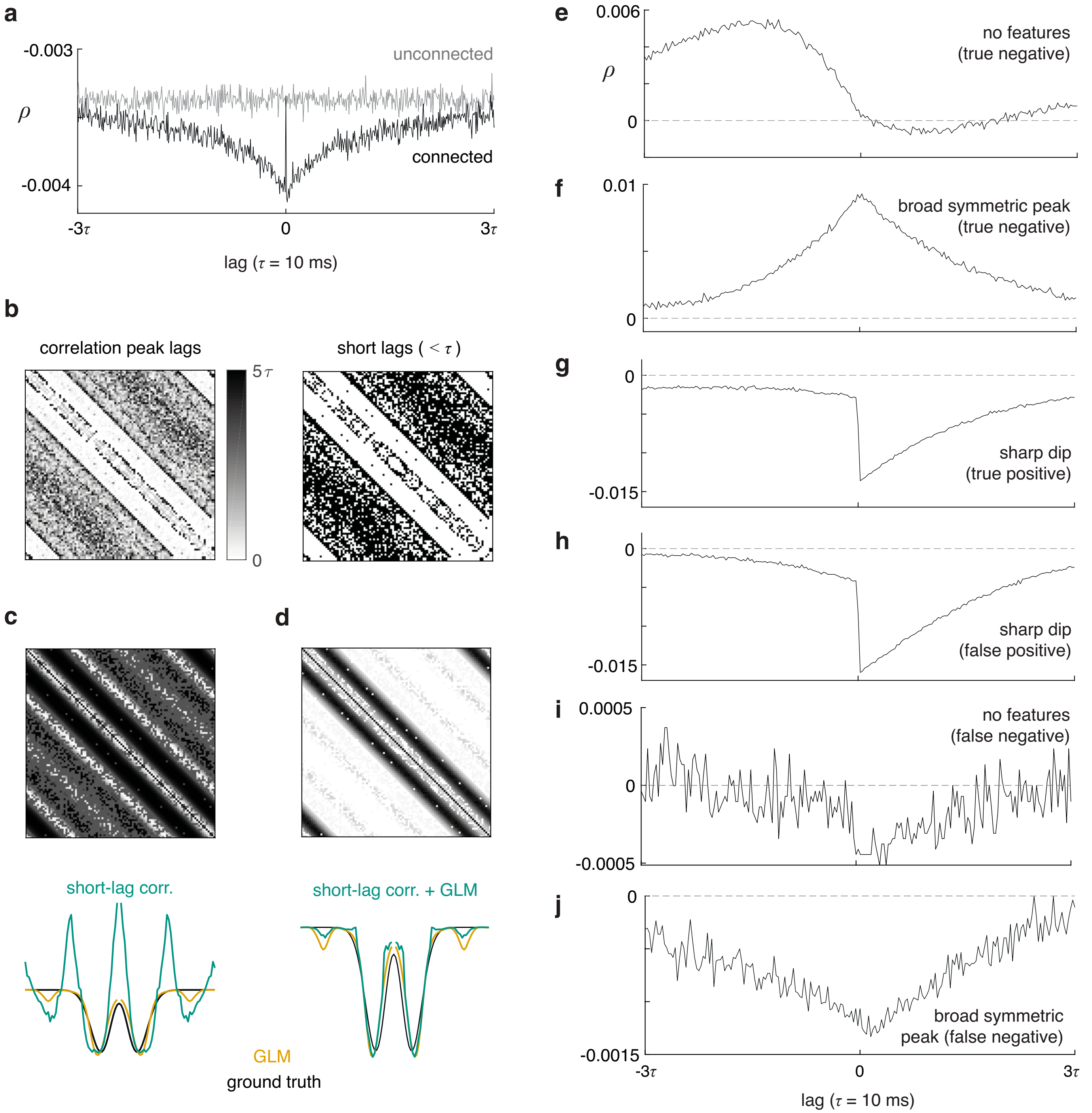


Extended Data Fig. 5 | Power-law decay of variance error of inferring the ring circuit, with increasing data volume. a, The fitted exponents α of $\Delta_v^2 \sim D^\alpha$ when using a generalized linear model for inference. Error-bands are 95% confidence intervals using 20 data points. The theoretical exponent is -1. **b**, The exponent α of the decay of total inference error $\Delta^2 \sim D^\alpha$ when using the Ising model for both data generation and inference. Error-bands are 95% confidence intervals using 20 data points. Here inference error is almost entirely due to variance, thus decays as the power law.

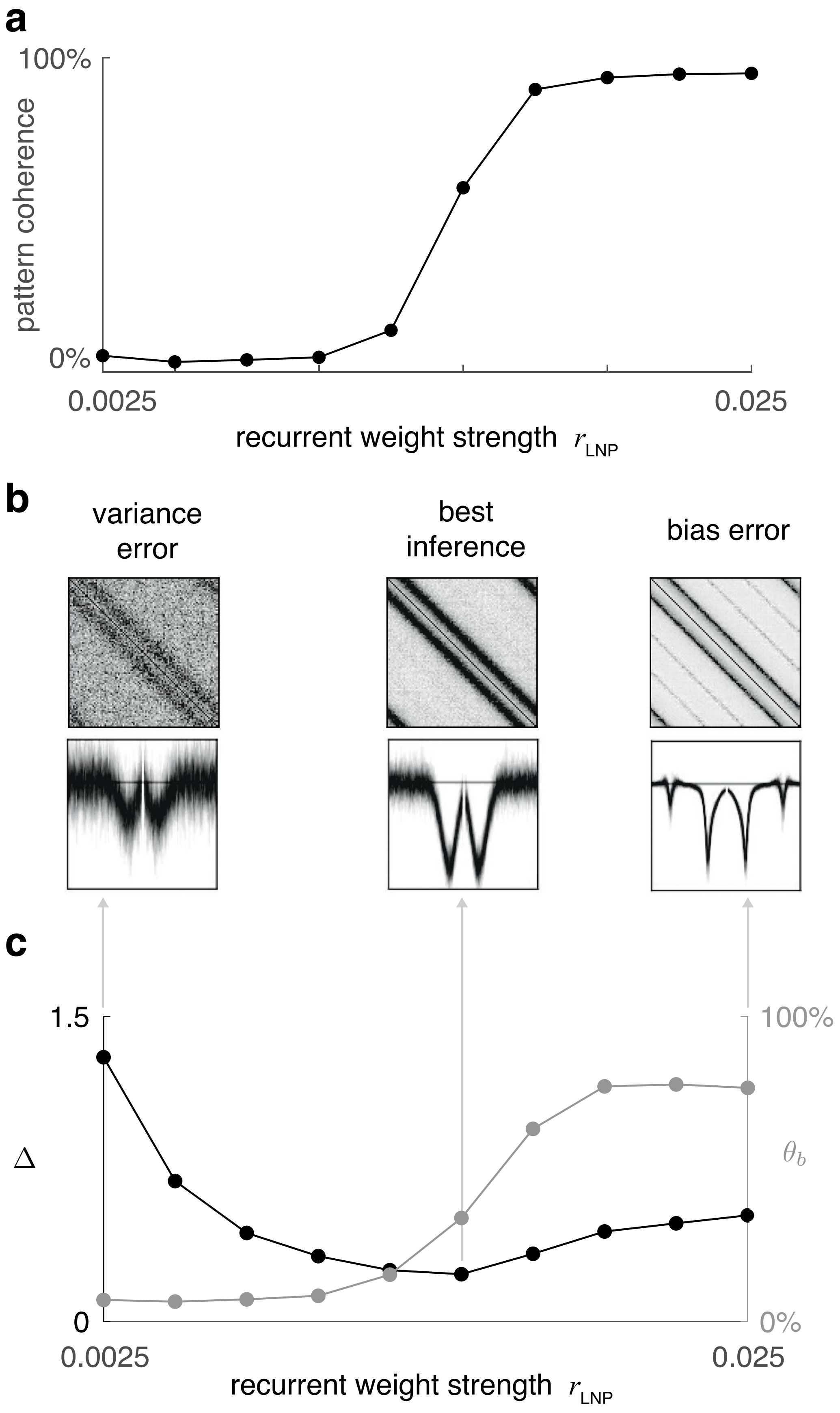


Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Circuit inference using logistic regression is not improved by l_1 regularization. a, Example ring network weight profiles inferred using logistic regression with zero, optimal and excessive regularization penalties. When weights are weak, regularization reduces some noise and marginally improves inference. At high weights, regularization suppresses both the spurious off-diagonal stripes and the true coupling shape, so is not helpful. **b**, Inference error vs weight strength using logistic regression with and without l_1 regularization. **c**, Optimal l_1 penalties (that produce the lowest inference errors) at each weight. Regularization improves inference in the strong and weak weight regimes, but barely.

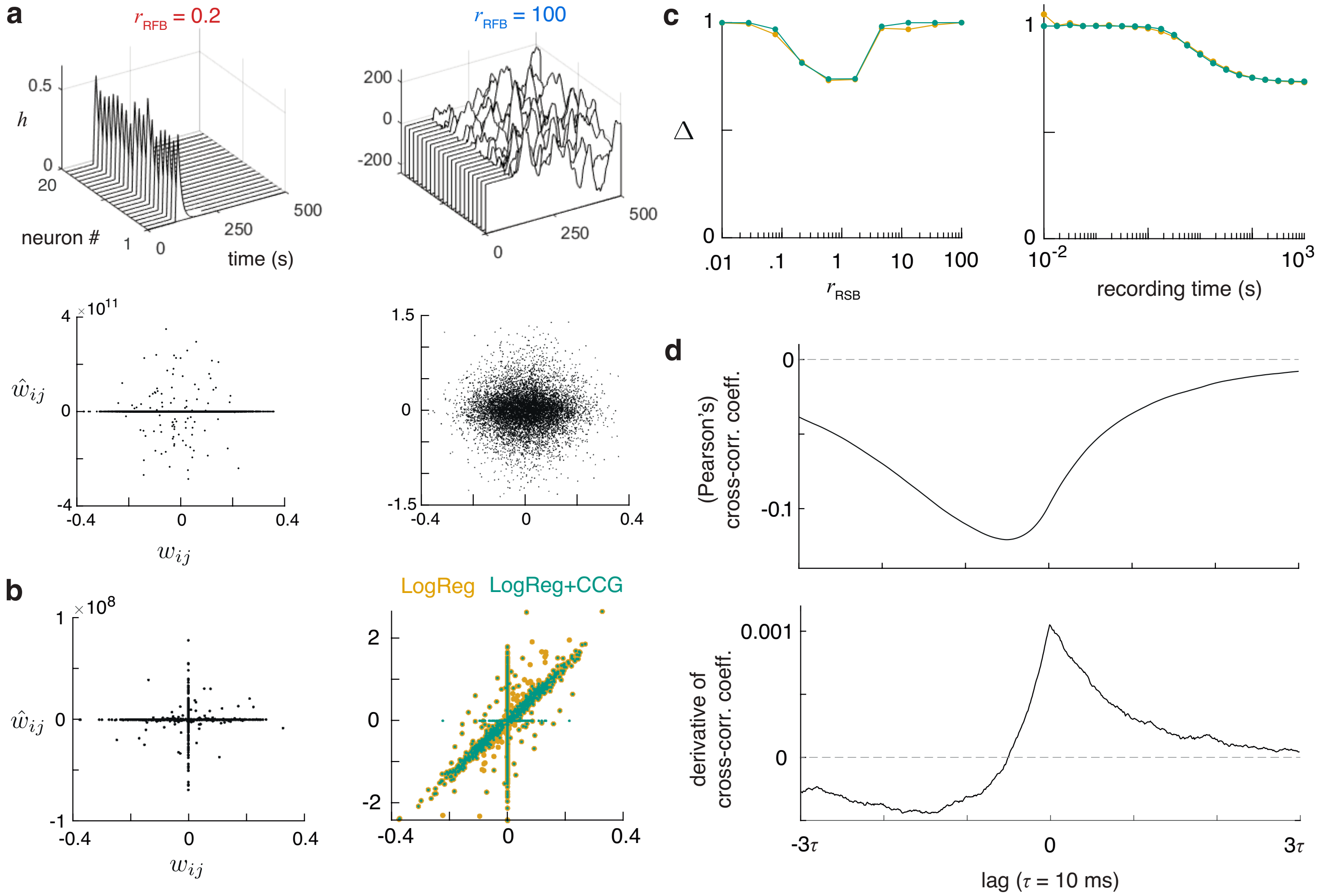


Extended Data Fig. 7 | Circuit inference using neural CCGs. (**a-d**) Inferring the strong weight ($r = 0.025$) ring circuit using short-lag peaks in neural CCGs. **a**, Pearson's cross-correlation of a connected and an unconnected neuron pair. **b**, Left: matrix of absolute lags of the CCG peaks. This partly reveals the circuit: directly connected neurons exhibit short lags. Right: binary matrix connecting neuron pairs with short lags. **c**, Top: CCG-based weight matrix: weight is set to the peak cross-correlation if the neurons are connected (lag $< \tau$), zero otherwise. Bottom: avg. weight profile. GLM fares better. **d**, Combining CCG and GLM. Top: matrix of GLM-inferred weights if neuron pairs have short CCG lag, zero otherwise. Bottom: avg. weight profile. Relative to pure GLM ($\Delta = 0.23$), this method ($\Delta = 0.32$) removes some biases but introduces others. (**e-j**) Spike CCGs from the sparse, non-symmetric, strong weight ($r_{\mathrm{RSA}} = 0.1$) random network. **e**, Unconnected pair. CCG has no sharp features around 0, indicating no direct connection. **f**, Unconnected pair. Broad symmetric peak, indicating multiple indirect influences through mutual connections, is discounted. **g**, Connected pair. Sharp asymmetric dip at 0 reveals direct (inhibitory) connection. **h**, Unconnected pair. CCG is indistinguishable from the previous, and passes the criterion. **i**, Connected pair, but no CCG features. **j**, Connected pair, but broad symmetric peak is discounted.

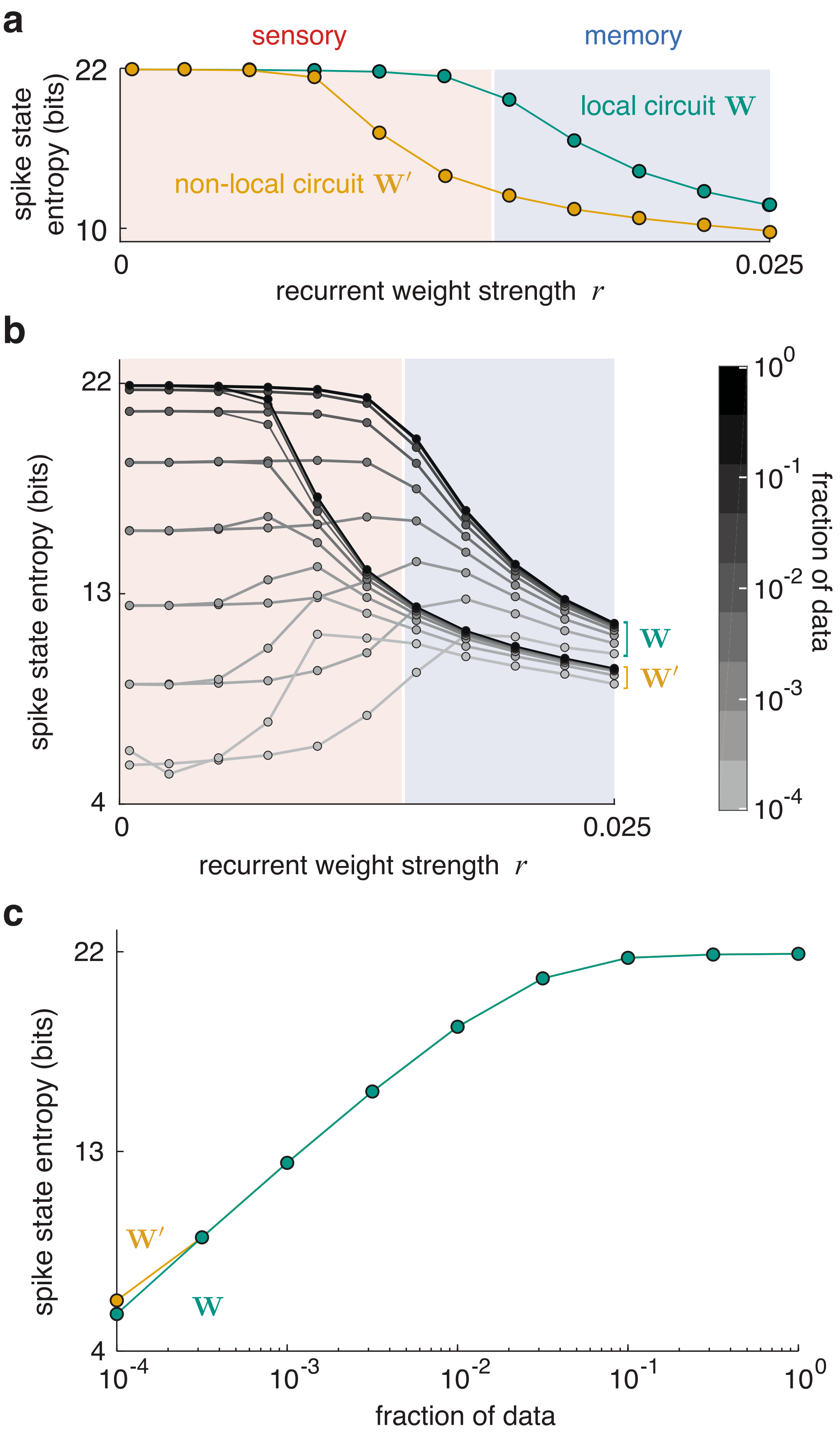


Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Results of inference using a GLM on data generated by a linear-nonlinear-Poisson model with a rectifying linear response (see Methods). **a**, Pattern coherence against weight strength for the generative LNP network. **b**, Inferred weight matrices (top) and superposition of rows (bottom, line marks zero), at several weight strengths. **c**, Inference error and bias fraction against weight strength. Optimal inference is at the point of pattern onset.



Extended Data Fig. 9 | Activity and inference in the random balanced networks that are fully or sparsely connected. a, Top: waterfall plots of neural fields of the full network at weak weights (when activity decays) and strong weights (when activity is chaotic), in response to a brief uniform feed-forward pulse. Bottom: corresponding inferences. (**b-d**) Inference on the sparse network. **b**, Left: true and inferred weights (using logistic regression) for the network with no noisy drive, at r_{RSB} = 0.6. Some zero weights are inferred to be non-zero, while some nonzero weights are underestimated. Right: true and inferred weights (using only logistic regression, and augmented with CCG information) for the same condition, but when the network is noise-driven. **c**, Inference error (using the two methods) vs recurrent weight strength, and data volume, on the noise-driven network. **d**, Example CCG (top) and its time-derivative (bottom) of a connected neuron pair in the noise-driven network.



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | Entropy of spiking activity states of neural circuits. a, Entropy of the distributions of 22-neuron spike sub-states from the true local ring circuit **W** and the non-local circuit **W**′. **b**, Entropies of the spike sub-states of the two circuits computed with different data fractions across weight strengths. At all weights, the computed entropies converge as the data approaches the total volume. **c:** Example slice of plot b at the weakest weights, where entropy convergence takes the longest.

nature research

Corresponding author(s):	Ila R. Fiete
Last updated by author(s):	Apr 23, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	□	A description of all covariates tested
☒	□	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	□	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	□	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Custom MATLAB (R2019b) code written during the current study is available from the corresponding author upon reasonable request.
Data analysis	Custom MATLAB (R2019b) code written during the current study is available from the corresponding author upon reasonable request.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The datasets generated and/or analysed during the current study are available from the corresponding author upon reasonable request.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences □ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	This work is theoretical, and data is synthesized from artificial neural networks. Thus, predominantly, sample sizes are not an issue. Where the analysis is limited by sample size, adequate tests have been described in the text to establish statistical significance. For example, in the information-theoretic analysis of neural activity, we show that the entropy measure is reliable by checking its convergence with sample size.
Data exclusions	No data generated were excluded from analysis.
Replication	This work is theoretical and computational. Thus, if the data are generated and analyzed as described in our methods, it should be fully reproducible.
Randomization	This is not applicable to our theoretical work using data synthesized from artificial neural networks.
Blinding	This is not applicable to our theoretical work using data synthesized from artificial neural networks.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	□ Antibodies
☒	□ Eukaryotic cell lines
☒	□ Palaeontology
☒	□ Animals and other organisms
☒	□ Human research participants
☒	□ Clinical data

Methods

n/a	Involved in the study
☒	□ ChIP-seq
☒	□ Flow cytometry
☒	□ MRI-based neuroimaging