Supporting Information

Roadmap

This supporting information document contains our theoretical and algorithmic developments, further inspections of our synthetic and real data analysis, and experimental procedures. Fig. S1 shows the organization of this document. Readers who are primarily interested in the theoretical and algorithmic developments may consider only the first six sections, whereas those who find further applications of AGC inference immediately more useful may start at Section 7.

Fig. S1. Table of contents of the supporting information.

1. The Main Theorem

In this section, we state the main theorem of this work, followed by its proof and discussion. We finally present an empirical validation of the result of our main theorem. Before presenting the main theorem, we make the following additional technical assumptions:

1) We consider a scaling of $\gamma = \mathcal{O}(\sqrt{(1-\beta)\log M})$, where M denotes the model order ($M^{(F)}$ or $M^{(R)}$). This assumption leads to asymptotic consistency of ℓ_1-regularized ML estimation (1, 2).

2) We assume that the stimuli $\{s_t\}_{t=1}^{T}$ form a Markovian random sequence. This assumption facilitates the limiting arguments used in our asymptotic analysis.

Our main theorem below extends the asymptotic inference results of the classical deviance difference to our adaptive de-biased variant:

Theorem S1 *Consider simultaneous spike train observations $\{\{n_t^{(c)}\}_{t=1}^{T}\}_{c=1}^{C}$ from an ensemble of C neurons. Let $\widehat{\boldsymbol{\omega}}_k^{(c)}$ and $\widehat{\boldsymbol{\omega}}_k^{(c\setminus\tilde{c})}$ denote the estimated sparse parameter vectors of neuron (c) at time window k in two nested logit-linked point process GLM models, where the contribution of neuron $(\tilde{c})$ is suppressed in the latter. Suppose that the adaptive estimation is carried out through solving the ℓ_1-regularized ML problem of Eq. 3 at time window k. Then,*

i) in the absence of a GC link from $(\tilde{c})$ to (c), we have $D_{k,\beta}^{(\tilde{c}\ \mapsto\ c)} \to \chi^2\big(M^{(d)}\big)$, and

ii) in the presence of a GC link from $(\tilde{c})$ to (c), and assuming that the cross-history coefficients from $(\tilde{c})$ to (c) scale at least as $\mathcal{O}\Big(\sqrt{\frac{1-\beta}{1+\beta}}\Big)$, then $D_{k,\beta}^{(\tilde{c}\ \mapsto\ c)} \to \chi^2\big(M^{(d)}, \nu_k^{(\tilde{c}\mapsto c)}\big)$,

as $\beta \to 1$, where $M^{(d)} := M^{(F)} - M^{(R)}$ is the dimensionality difference of the two nested models, and $\nu_k^{(\tilde{c}\ \mapsto\ c)} > 0$ is the corresponding non-centrality parameter and is only a function of the true model parameter of neuron (c) at time k (explicitly given in closed form in the proof).

Proof of the Main Theorem. Before presenting the proof, we introduce the following notation. For a log-likelihood function $\ell(\boldsymbol{\omega})$ with a parameter vector $\boldsymbol{\omega}$, we have:

$$\dot{\boldsymbol{\ell}}(\boldsymbol{\omega}) := \nabla_{\boldsymbol{\omega}}\ell(\boldsymbol{\omega}), \qquad [S1]$$

$$\ddot{\boldsymbol{\ell}}(\boldsymbol{\omega}) := \nabla^2_{\boldsymbol{\omega}}\ell(\boldsymbol{\omega}), \qquad [S2]$$

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\omega}) := \mathbb{E}\left\{\dot{\boldsymbol{\ell}}(\boldsymbol{\omega})\dot{\boldsymbol{\ell}}'(\boldsymbol{\omega})\right\}, \qquad [S3]$$

where $\dot{\boldsymbol{\ell}}(\cdot)$ is the gradient of the log-likelihood with respect to the parameter vector $\boldsymbol{\omega}$, known as the *score* statistic, $\ddot{\boldsymbol{\ell}}(\cdot)$ is the Hessian of the log-likelihood, and $\boldsymbol{\mathcal{I}}(\cdot)$ denotes the Fisher information matrix as the covariance of the score vector, where the expectation is over the realization of the process.

For simplicity of analysis, we consider a piece-wise constant model in which $\boldsymbol{\omega}_k$ is constant within observation windows indexed by $i = k-N, k-N+1, \ldots, k$ for some large $N = \mathcal{O}(\frac{1}{1-\beta})$, following the tradition of performance analysis of RLS-type algorithms (3). Recall that the exponentially weighted log-likelihood at window k is given by:

$$\ell_k^{\beta}(\boldsymbol{\omega}_k) := (1-\beta)\sum_{i=k-N}^{k}\beta^{k-i}\ell_i(\boldsymbol{\omega}_k). \qquad [S4]$$

Let $\boldsymbol{\omega}_k$ and $\widehat{\boldsymbol{\omega}}_k$ denote the true and estimated parameter vectors of length M associated with a unit at window k, where M can take any of the two values $M^{(F)}$ and $M^{(R)}$ corresponding to full and reduced models, respectively. Suppose that the inverse Hessian exists at $\boldsymbol{\omega}_k$ for each time k, which we denote by $\boldsymbol{\Theta}_k := (\ddot{\boldsymbol{\ell}}_k^{\beta}(\boldsymbol{\omega}_k))^{-1}$ for notational convenience. Throughout the proof, we make use of our earlier results on the consistency of the ℓ_1-regularized exponentially-weighted maximum likelihood (1, 4). These results imply that for β close enough to 1, we have $\|\widehat{\boldsymbol{\omega}}_k - \boldsymbol{\omega}_k\|_2 = \mathcal{O}(\sqrt{(1-\beta)s\log M})$, with a choice of $\gamma = \mathcal{O}(\sqrt{(1-\beta)\log M})$ for the regularization parameter.

The de-biased deviance $D_{k,\beta}(\widehat{\boldsymbol{\omega}}_k;\boldsymbol{\omega}_k)$ of Eq. 6 can be expressed in the following quadratic form:

$$D_{k,\beta}(\widehat{\boldsymbol{\omega}}_k;\boldsymbol{\omega}_k) = -\left(\frac{1+\beta}{1-\beta}\right)(\mathbf{w}_k-\boldsymbol{\omega}_k)'\,\ddot{\boldsymbol{\ell}}_k^{\beta}(\boldsymbol{\omega}_k)\,(\mathbf{w}_k-\boldsymbol{\omega}_k), \qquad [S5]$$

where

$$\mathbf{w}_k := \widehat{\boldsymbol{\omega}}_k - \boldsymbol{\Theta}_k\dot{\boldsymbol{\ell}}_k^{\beta}(\widehat{\boldsymbol{\omega}}_k). \qquad [S6]$$

By rearranging some terms, Eq. S5 can be expressed as:

$$\begin{aligned}\left(\frac{1-\beta}{1+\beta}\right)D_{k,\beta}(\widehat{\boldsymbol{\omega}}_k;\boldsymbol{\omega}_k) = {}& 2(\widehat{\boldsymbol{\omega}}_k-\boldsymbol{\omega}_k)'\dot{\boldsymbol{\ell}}_k^{\beta}(\widehat{\boldsymbol{\omega}}_k)\\ & -(\widehat{\boldsymbol{\omega}}_k-\boldsymbol{\omega}_k)'\ddot{\boldsymbol{\ell}}_k^{\beta}(\widehat{\boldsymbol{\omega}}_k)(\widehat{\boldsymbol{\omega}}_k-\boldsymbol{\omega}_k)\\ & -B_k+\Delta_1, \end{aligned} \qquad [S7]$$

where $B_k := \dot{\boldsymbol{\ell}}_k^{\beta}(\widehat{\boldsymbol{\omega}}_k)'\boldsymbol{\Theta}_k\dot{\boldsymbol{\ell}}_k^{\beta}(\widehat{\boldsymbol{\omega}}_k)$ denotes the bias term due to ℓ_1-regularization, and Δ_1 denotes a remainder term given by:

$$\Delta_1 := (\widehat{\boldsymbol{\omega}}_k-\boldsymbol{\omega}_k)'\big(\ddot{\boldsymbol{\ell}}_k^{\beta}(\widehat{\boldsymbol{\omega}}_k)-\ddot{\boldsymbol{\ell}}_k^{\beta}(\boldsymbol{\omega}_k)\big)(\widehat{\boldsymbol{\omega}}_k-\boldsymbol{\omega}_k). \qquad [S8]$$

Next, we use the Taylor's series expansion as follows:

$$\ell_k^\beta(\boldsymbol{\omega}_k) = \ell_k^\beta(\widehat{\boldsymbol{\omega}}_k) + (\boldsymbol{\omega}_k - \widehat{\boldsymbol{\omega}}_k)' \dot{\boldsymbol{\ell}}_k^\beta(\widehat{\boldsymbol{\omega}}_k) + \frac{1}{2}(\boldsymbol{\omega}_k - \widehat{\boldsymbol{\omega}}_k)' \ddot{\boldsymbol{\ell}}_k^\beta(\widetilde{\boldsymbol{\omega}}_k)(\boldsymbol{\omega}_k - \widehat{\boldsymbol{\omega}}_k), \quad \text{[S9]}$$

where $\widetilde{\boldsymbol{\omega}}_k := t\boldsymbol{\omega}_k + (1-t)\widehat{\boldsymbol{\omega}}_k$ is an intermediate vector for some $t \in (0,1)$, such that $\|\widetilde{\boldsymbol{\omega}}_k - \boldsymbol{\omega}_k\| < \|\widehat{\boldsymbol{\omega}}_k - \boldsymbol{\omega}_k\|$. Combining Eqs. S7 and S9, we get:

$$\left(\frac{1-\beta}{1+\beta}\right) D_{k,\beta}(\widehat{\boldsymbol{\omega}}_k; \boldsymbol{\omega}_k) = 2(\ell_k^\beta(\widehat{\boldsymbol{\omega}}_k) - \ell_k^\beta(\boldsymbol{\omega}_k)) - B_k + \Delta_2, \quad \text{[S10]}$$

where the remainder term Δ_2 takes a similar form to Eq. S8 with the Hessian evaluated at $\tilde{\boldsymbol{\omega}}_k$ instead. Using the Lipschitz property of the second-order derivative of the logistic function, boundedness assumption on the covariates ($\|\mathbf{X}\|_\infty = \mathcal{O}(K)$), and the consistency of $\widehat{\boldsymbol{\omega}}_k$, it can be proved that both remainder terms Δ_1 and Δ_2 are asymptotically negligible with a rate of $\|\widehat{\boldsymbol{\omega}}_k - \boldsymbol{\omega}_k\|^3 = o_{\mathbb{P}}((1-\beta)^{3/2})$ as $\beta \to 1$.

In order to adapt the treatment of Davidson and Lever (5) to our setting, we first consider a sequence of forgetting factors $\{\beta_j\}_{j=1}^\infty$ approaching unity, i.e., $\lim_{j\to\infty} \beta_j = 1$. Then, at window k, we test the null hypothesis $H_{0,k} : \boldsymbol{\omega}_k^0 = (\boldsymbol{\omega}_{0,k}, \mathbf{0})$ against a sequence of local alternatives $\{H_{1,k}^{\beta_j}\}_{j=1}^\infty = \{H_{1,k}^{\beta_j} : \boldsymbol{\omega}_k^{\beta_j} = (\boldsymbol{\omega}_{0,k}^*, \boldsymbol{\omega}_{1,k}^{\beta_j})\}$, where $\boldsymbol{\omega}_{1,k}^{\beta_j} = \sqrt{\frac{1-\beta_j}{1+\beta_j}}\boldsymbol{\delta}_k$ corresponds to the unspecified sub-vector excluded in the reduced model for some constant vector $\boldsymbol{\delta}_k$ of dimension $M^{(d)}$.

Statistical inference under the sequence of local alternatives $\{H_{1,k}^{\beta_j}\}$ is carried out through testing local departures from null hypothesis to the limiting true parameter $\boldsymbol{\omega}_k^*$ at the rate of $\mathcal{O}\left(\sqrt{\frac{1-\beta_j}{1+\beta_j}}\right)$ as $\beta_j \to 1$. For notational convenience, we drop the dependence of β_j on the index j. It is understood that expressions involving limits of β are interpreted as the sequential limit. From the definition of $\mathbf{w}_k$ in Eq. S6, it follows that:

$$\begin{aligned}\mathbf{w}_k - \boldsymbol{\omega}_k &= \widehat{\boldsymbol{\omega}}_k - \boldsymbol{\omega}_k - \boldsymbol{\Theta}_k \dot{\boldsymbol{\ell}}_k^\beta(\widehat{\boldsymbol{\omega}}_k) \\ &= -\boldsymbol{\Theta}_k \dot{\boldsymbol{\ell}}_k^\beta(\boldsymbol{\omega}_k) + \boldsymbol{\Delta}, \end{aligned} \quad \text{[S11]}$$

where $\boldsymbol{\Delta} := (\mathbf{I} - \boldsymbol{\Theta}_k \ddot{\boldsymbol{\ell}}_k^\beta(\widetilde{\boldsymbol{\omega}}_k))(\widehat{\boldsymbol{\omega}}_k - \boldsymbol{\omega}_k)$, and we used:

$$\dot{\boldsymbol{\ell}}_k^\beta(\widehat{\boldsymbol{\omega}}_k) = \dot{\boldsymbol{\ell}}_k^\beta(\boldsymbol{\omega}_k) + \ddot{\boldsymbol{\ell}}_k^\beta(\widetilde{\boldsymbol{\omega}}_k)(\widehat{\boldsymbol{\omega}}_k - \boldsymbol{\omega}_k), \quad \text{[S12]}$$

in Eq. S11 which holds for some intermediate vector $\widetilde{\boldsymbol{\omega}}_k = t\boldsymbol{\omega}_k + (1-t)\widehat{\boldsymbol{\omega}}_k$ for some $t \in (0,1)$. It can be shown that $\boldsymbol{\Delta} = o_{\mathbb{P}}(1-\beta)$ is asymptotically negligible, following the aforementioned argument used for Δ_1 and Δ_2.

Next, we need to determine the asymptotic behavior of the Hessian $\ddot{\boldsymbol{\ell}}_k^\beta(\boldsymbol{\omega}_k)$ as $\beta \to 1$. Due to the dependencies of the covariates, the common law of large numbers (LLN) for i.i.d. random variables cannot be applied. Due to the logistic link used in defining the log-likelihood, the Hessian can be written as $(1-\beta)\mathbf{X}'\mathbf{W}\mathbf{D}\mathbf{X}$, where $\mathbf{W}$ is a diagonal bounded weighing matrix, $\mathbf{D}$ is a diagonal matrix containing the exponential weights, and $\mathbf{X}$ is the matrix of covariates (1). Also, for finite M, $\{n_i^{(c)}\}_{c=1}^C$ form a 2^C-state Markov chain with ϕ-mixing property. Hence, the version of LLN for bounded functions of ϕ-mixing random variables can be used to characterize the limit (e.g., (6) or Theorem 27.4 in (7)). Hence, as $\beta \to 1$:

$$\ddot{\boldsymbol{\ell}}_k^\beta(\boldsymbol{\omega}_k) \xrightarrow{p} \mathbb{E}[\ddot{\boldsymbol{\ell}}_i(\boldsymbol{\omega}_k)] = -\boldsymbol{\mathcal{I}}(\boldsymbol{\omega}_k), \quad \text{[S13]}$$

where the second equality is obtained using the Fisher information equality.

Similarly, in order to characterize the asymptotic behavior of the score statistic, a version of the Central Limit Theorem (CLT) for dependent random variables is required. Note that the Lindeberg CLT for i.i.d. random variables does not apply, since the covariates are highly dependent. In the absence of the stimuli in the logistic model, i.e., $\mathbf{s}_i = \mathbf{0}, \forall i$, by invoking the aforementioned ϕ-mixing property of the equivalent 2^C-state Markov chain $\{n_i^{(c)}\}_{c=1}^C$, we use a version of the martingale CLT (6). In the presence of stimuli, by the hypothesis that the stimuli are generated by a Markov process, we invoke stronger versions of the CLT for autoregressive models (8, 9). Hence, the score statistic at the true parameter converges in distribution to a Gaussian random vector with zero mean and covariance given by the Fisher information matrix:

$$\sqrt{\frac{1+\beta}{1-\beta}}\, \dot{\boldsymbol{\ell}}_k^\beta(\boldsymbol{\omega}_k) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\mathcal{I}}(\boldsymbol{\omega}_k)), \quad \text{[S14]}$$

as $\beta \to 1$. Note that this result holds both under $H_{0,k}$ when $\boldsymbol{\omega}_k = \boldsymbol{\omega}_k^0$ is the true parameter vector, and for the sequence of alternatives $H_{1,k}^\beta$, where $\boldsymbol{\omega}_k = \boldsymbol{\omega}_k^\beta$ is the sequence of true parameters.

The asymptotic normality of $\mathbf{w}_k$ under $H_{0,k}$ follows by invoking the Slutsky's theorem using Eqs. S13 and S14:

$$\sqrt{\frac{1+\beta}{1-\beta}}(\mathbf{w}_k - \boldsymbol{\omega}_k) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \boldsymbol{\mathcal{I}}(\boldsymbol{\omega}_k)^{-1}\right), \quad \text{[S15]}$$

as $\beta \to 1$. Hence, under H_0, combining the asymptotic result on the Hessian in Eq. S13, and the asymptotic normality of $\mathbf{w}_k$ in Eq. S15 leads to the weak convergence of the adaptive de-biased deviance to a central chi-squared distribution with M degrees of freedom:

$$\left[D_{k,\beta}(\widehat{\boldsymbol{\omega}}_k; \boldsymbol{\omega}_k) \,|\, H_{0,k}\right] \xrightarrow{d} \chi^2(M), \quad \text{[S16]}$$

as $\beta \to 1$. Following on the classical results (10, 11), it can be shown that the deviance difference of two nested full and reduced models asymptotically converges in distribution to a central chi-squared with $M^{(d)}$ degrees of freedom:

$$\left[D_{k,\beta}(\widehat{\boldsymbol{\omega}}_k^\beta; \widehat{\boldsymbol{\omega}}_k^0) \,\middle|\, H_{0,k}\right] \xrightarrow{d} \chi^2(M^{(d)}), \quad \text{[S17]}$$

as $\beta \to 1$, where $M^{(d)}$ is the dimension of the specified sub-vector $\boldsymbol{\omega}_{1,k} = \mathbf{0}$ under the null hypothesis, i.e, the dimensionality difference of the two nested models. This establishes part (i) of the statement of Theorem S1. ■

As for part (ii), such an asymptotic result under the sequence of local alternative hypotheses will be slightly different, as the limiting Gaussian distributions are non-zero mean. To see this, we define the de-biased vector $\mathbf{w}_k^\beta$ associated with each local alternative $H_{1,k}^\beta$ at time step k as:

$$\mathbf{w}_k^\beta := \widehat{\boldsymbol{\omega}}_k^\beta - \boldsymbol{\Theta}_k^* \dot{\boldsymbol{\ell}}_k^\beta(\widehat{\boldsymbol{\omega}}_k^\beta), \quad \text{[S18]}$$

where $\boldsymbol{\Theta}_k^* := \boldsymbol{\Theta}_k(\boldsymbol{\omega}_k^*)$. By similar arguments leading to Eq. S11, it follows that:

$$\begin{aligned}\mathbf{w}_k^\beta - \boldsymbol{\omega}_k^* &= \widehat{\boldsymbol{\omega}}_k^\beta - \boldsymbol{\omega}_k^* - \boldsymbol{\Theta}_k^* \dot{\boldsymbol{\ell}}_k^\beta(\widehat{\boldsymbol{\omega}}_k^\beta) \\ &= -\boldsymbol{\Theta}_k^* \dot{\boldsymbol{\ell}}_k^\beta(\boldsymbol{\omega}_k^*) + o_{\mathbb{P}}(1-\beta) \qquad \text{[S19]} \\ &= \boldsymbol{\omega}_k^\beta - \boldsymbol{\omega}_k^* - \boldsymbol{\Theta}_k^* \dot{\boldsymbol{\ell}}_k^\beta(\boldsymbol{\omega}_k^\beta) + o_{\mathbb{P}}(1-\beta), \qquad \text{[S20]}\end{aligned}$$

where we have respectively used the following linear expansions around $\boldsymbol{\omega}_k^*$ in Eq. S19 and S20:

$$\dot{\boldsymbol{\ell}}_k^\beta(\widehat{\boldsymbol{\omega}}_k^\beta) = \dot{\boldsymbol{\ell}}_k^\beta(\boldsymbol{\omega}_k^*) + \ddot{\boldsymbol{\ell}}_k^\beta(\boldsymbol{\omega}_k^*)(\widehat{\boldsymbol{\omega}}_k^\beta - \boldsymbol{\omega}_k^*) + o_{\mathbb{P}}(1-\beta), \qquad \text{[S21]}$$

$$\dot{\boldsymbol{\ell}}_k^\beta(\boldsymbol{\omega}_k^\beta) = \dot{\boldsymbol{\ell}}_k^\beta(\boldsymbol{\omega}_k^*) + \ddot{\boldsymbol{\ell}}_k^\beta(\boldsymbol{\omega}_k^*)(\boldsymbol{\omega}_k^\beta - \boldsymbol{\omega}_k^*) + o_{\mathbb{P}}(1-\beta). \qquad \text{[S22]}$$

Using similar arguments leading to Eqs. S13 and S14, the asymptotic form of the Hessian and the asymptotic normality of the score function at the true parameter vector $\boldsymbol{\omega}_k^\beta$ under the sequence of local alternatives $H_{1,k}^\beta$ will follow:

$$\ddot{\boldsymbol{\ell}}_k^\beta(\boldsymbol{\omega}_k^\beta) \xrightarrow{p} -\boldsymbol{\mathcal{I}}(\boldsymbol{\omega}_k^*), \qquad \text{[S23]}$$

$$\sqrt{\frac{1+\beta}{1-\beta}}\, \dot{\boldsymbol{\ell}}_k^\beta(\boldsymbol{\omega}_k^\beta) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\mathcal{I}}(\boldsymbol{\omega}_k^*)). \qquad \text{[S24]}$$

Hence, incorporating the asymptotics of Eqs. S23 and S24 into Eq. S20, the de-biased estimate $\mathbf{w}_k^\beta$ under the sequence of local alternatives $H_{1,k}^\beta$ converges in distribution to a multivariate normal distribution:

$$\sqrt{\frac{1+\beta}{1-\beta}}(\mathbf{w}_k^\beta - \boldsymbol{\omega}_k^*) \xrightarrow{d} \mathcal{N}(\bar{\boldsymbol{\delta}}_k, \boldsymbol{\mathcal{I}}(\boldsymbol{\omega}_k^*)^{-1}), \qquad \text{[S25]}$$

with non-zero asymptotic mean $\bar{\boldsymbol{\delta}}_k := [\mathbf{0}', \boldsymbol{\delta}_k']'$ as $\beta \to 1$. The asymptotic mean is obtained from the asymptotic rate of the *Pitman drift*, where the sequence of true local parameter vectors $\{\boldsymbol{\omega}_k^\beta\}$ approach the limit $\boldsymbol{\omega}_k^*$ at a rate of $\|\boldsymbol{\omega}_k^\beta - \boldsymbol{\omega}_k^*\| = \mathcal{O}\left(\sqrt{\frac{1-\beta}{1+\beta}}\right)$.

Next, consider the decomposition of $\mathcal{I}(\boldsymbol{\omega}_k^*)$ into blocks corresponding to $\boldsymbol{\omega}_{0,k}$ and $\boldsymbol{\omega}_{1,k}$:

$$\mathcal{I}(\boldsymbol{\omega}_k^*) = \begin{bmatrix} \mathcal{I}_{0,0}(\boldsymbol{\omega}_k^*) & \mathcal{I}_{0,1}(\boldsymbol{\omega}_k^*) \\ \mathcal{I}_{1,0}(\boldsymbol{\omega}_k^*) & \mathcal{I}_{1,1}(\boldsymbol{\omega}_k^*) \end{bmatrix}. \qquad \text{[S26]}$$

By invoking a similar treatment as in the proof of Theorem 1 of (5) via the extension of Cochran's theorem to non-central chi-squared distribution (12, 13), and using the asymptotic result of Eq. S25 in the quadratic forms of Eq. S5 for both the reduced and full model estimates $(\widehat{\boldsymbol{\omega}}_k^0, \widehat{\boldsymbol{\omega}}_k^\beta)$, it can be shown that the deviance difference of two nested models converges in distribution to a non-central chi-squared distribution under the sequence of local alternatives $H_{1,k}^\beta$ as $\beta \to 1$:

$$\left[D_{k,\beta}(\widehat{\boldsymbol{\omega}}_k^\beta; \widehat{\boldsymbol{\omega}}_k^0) \,\middle|\, H_{1,k}^\beta\right] \xrightarrow{d} \chi^2(M^{(d)}, \nu_k), \qquad \text{[S27]}$$

where $M^{(d)}$ is the dimensionality difference of two nested models as before, and $\nu_k := \boldsymbol{\delta}_k' \bar{\boldsymbol{\mathcal{I}}}_{1,1}(\boldsymbol{\omega}_k^*)\boldsymbol{\delta}_k$ is the non-centrality parameter with $\bar{\boldsymbol{\mathcal{I}}}_{1,1}(\boldsymbol{\omega}_k^*) := \boldsymbol{\mathcal{I}}_{1,1}(\boldsymbol{\omega}_k^*) - \boldsymbol{\mathcal{I}}_{1,0}(\boldsymbol{\omega}_k^*)\boldsymbol{\mathcal{I}}_{0,0}^{-1}(\boldsymbol{\omega}_k^*)\boldsymbol{\mathcal{I}}_{0,1}(\boldsymbol{\omega}_k^*)$. This establishes part (ii) of the statement of Theorem S1. ■

Discussion of the Result of Theorem S1. Two remarks regarding the bias correction and implications of the result of Theorem S1 are in order:

Remark 1. The bias term B_k that emerged in the derivation of $D_{k,\beta}$ in Eq. S7 can be estimated as $\widehat{B}_k = \dot{\boldsymbol{\ell}}_k^\beta(\widehat{\boldsymbol{\omega}}_k)' \widehat{\boldsymbol{\Theta}}_k \dot{\boldsymbol{\ell}}_k^\beta(\widehat{\boldsymbol{\omega}}_k)$, where $\widehat{\boldsymbol{\Theta}}_k = \left(\ddot{\boldsymbol{\ell}}_k^\beta(\widehat{\boldsymbol{\omega}}_k)\right)^{-1}$. Proof of the consistency of this estimate, i.e., $\widehat{B}_k \xrightarrow{p} B_k$ follows directly from the consistency of the inverse Hessian $\widehat{\boldsymbol{\Theta}}_k \xrightarrow{p} \boldsymbol{\Theta}_k$. Since we assumed that the Hessian is invertible at true parameter $\boldsymbol{\omega}_k$, there exists a subsequence of the estimators $\{\widehat{\boldsymbol{\omega}}_k^{(\beta_\ell)}\}_\ell$, at which the Hessians are invertible, and approach the true inverse Hessian $\boldsymbol{\Theta}_k$, given that M is fixed.

In the case that the Hessian $\ddot{\boldsymbol{\ell}}_k^\beta(\widehat{\boldsymbol{\omega}}_k)$ is not invertible, either due to the rank-deficiency at $\widehat{\boldsymbol{\omega}}_k$ for some k, or the case of infinitely growing dimensions $M^{(F)}$ and $M^{(R)}$ with *fixed* difference $M^{(d)}$, we adopt the approach taken in (2) and compute $\widehat{\boldsymbol{\Theta}}_k$ using the so-called *node-wise regression*, for which similar asymptotic results have been proven, implying that $\|\widehat{\boldsymbol{\Theta}}_k - \boldsymbol{\Theta}_k\|_\infty = o_{\mathbb{P}}(1)$.

Remark 2. In the conventional asymptotic analysis of deviance, the true parameters $\{\boldsymbol{\omega}^N\}_{N=1}^\infty$ associated with the sequence of local alternatives H_1^N approach the limiting true parameter $\boldsymbol{\omega}^*$ at the rate of $\mathcal{O}(1/\sqrt{N})$, where N is the number of observations. In our case, given a forgetting factor β, it follows from our asymptotic analysis that the true (cross-history) parameter $\boldsymbol{\omega}_{1,k}^\beta$ of order $\mathcal{O}\left(\sqrt{\frac{1-\beta}{1+\beta}}\right)$ associated with the alternative $H_{1,k}^\beta$ will lead to a non-trivial asymptotic distribution of the test statistic, i.e., a *non-central* chi-squared distribution. Hence, one expects that the underlying cross-history coefficients taking small values would still be detectable for β close enough to 1. In other words, the more number of observations we have for hypothesis testing, the easier it gets to distinguish between the null $H_0 : \boldsymbol{\omega}_{1,k} = \mathbf{0}$ and the alternative $H_1^\beta : \boldsymbol{\omega}_{1,k} = \boldsymbol{\omega}_1^\beta$. Therefore, we expect to detect causal links resulting from regression coefficients as small as $\mathcal{O}\left(\sqrt{\frac{1-\beta}{1+\beta}}\right)$, as stated in the theorem.

Empirical Validation of the Results of Theorem S1. In order to validate the results of Theorem S1 empirically, we examine the distributions of the adaptive de-biased deviance difference statistics $D_k^{(\tilde{c} \mapsto c)}$ for two representative links from Fig. 3: the GC link $(1 \mapsto 7)$ which was present in the first segment of the experiment and vanished in the last segment, and the GC link $(5 \mapsto 2)$ which did not exist in the first segment, but emerged in the last segment.

To this end, we consider 500 realizations of simulated spike trains corresponding to the network dynamics of Fig. 3 in order to construct the empirical distribution of $D_k^{(\tilde{c} \mapsto c)}$'s in the form of uniform histograms. Fig. S2 shows the resulting histograms and theoretical density fits (solid curves), as predicted by Theorem S1, at two selected time points of 40 s (endpoint of the first segment) and 120 s (endpoint of the third segment). The histograms are constructed using 15 uniform bins. The theoretical density $\chi^2(M_d)$ from part (i) of the theorem is plotted for $M_d = 10$. The theoretical density from part (ii), i.e., $\chi^2(M_d, \nu_k^{(\tilde{c} \mapsto c)})$, is plotted for $M_d = 10$ and the non-centrality parameter estimates $\widehat{\nu}_k^{(\tilde{c} \mapsto c)}$ obtained by subtracting M_d from the average deviance differences across the 500 realizations.

As it can be observed from Fig. S2, the theoretical predictions closely match the empirical estimates of the densities, even at a practical value of $\beta = 0.999$ (i.e., $N_{\text{eff}} = 10000$) close enough to unity. We confirmed that similar results hold for the rest of the links in the network, but have only plotted

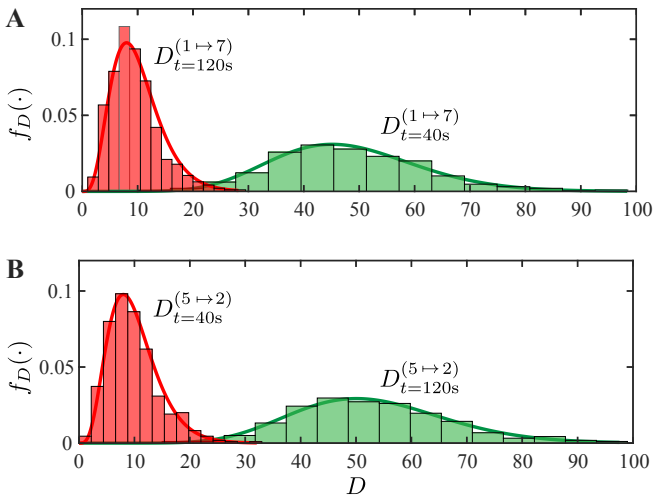

Fig. S2. Empirical and theoretical fits to the distributions of the adaptive de-biased deviance difference $D_k^{(\tilde{c}\mapsto c)}$ for two selected links from Fig. 3. The empirical densities are shown as histograms using 15 bins (colored bars) and the theoretical fits are plotted as solid curves. (A) Empirical and theoretical densities of $D_k^{(1\mapsto 7)}$ at $k = 40$ s (existing GC link) and $k = 120$ s (non-existing GC link), (B) Empirical and theoretical densities of $D_k^{(5\mapsto 2)}$ at $k = 40$ s (non-existing GC link) and $k = 120$ s (existing GC link).

those corresponding to the aforementioned representative links for the sake of brevity.

2. Derivations of the Statistical Inference Procedures

Given the deviance data and an estimate of the non-centrality parameters, the G-causal links can be detected at a fixed FDR according to Algorithm S1.

Algorithm S1 BY FDR Control and Characterizing the J-statistics

Input: $\left\{\{D_{k,\beta}^{(\tilde{c}\mapsto c)}, \widehat{\nu}_k^{(\tilde{c}\mapsto c)}\}_{=1}^K \,|\, (\tilde{c}\mapsto c)\in\mathcal{C}\right\}$, $M^{(d)}$, and α.

1: **for** $k = 1, 2, \ldots, K$ **do**
2: **for** $(\tilde{c}\mapsto c)\in\mathcal{C}$ **do**
3: Define p-values $p_k^{(\tilde{c}\mapsto c)} := 1 - \mathrm{F}_{\chi^2(M^{(d)})}\left(D_k^{(\tilde{c}\mapsto c)}\right)$
4: Sort the calculated p-values as $p_k^{(m_1)} \le p_k^{(m_2)} \le \cdots \le p_k^{(m_{|\mathcal{C}|})}$ where $\{m_1, \cdots, m_{|\mathcal{C}|}\}$ is a permutation of $\{1, \cdots, |\mathcal{C}|\}$
5: Find largest $i_{\max}$ for which $p_k^{(m_i)} \le \alpha_i := \frac{i\alpha}{|\mathcal{C}|\log\left(|\mathcal{C}|\right)}$
6: Reject all null hypotheses $\{H_0^{(m_i)} | i \le i_{\max}\}$ associated with the GC links $m = m_1, m_2, \ldots, m_{i_{\max}}$
7: $J_k^{(m_i)} = 0$ for $i = i_{\max}+1, \ldots, |\mathcal{C}|$
8: $J_k^{(m_i)} = 1 - \bar{\alpha} - \mathrm{F}_{\chi^2(M^{(d)},\widehat{\nu}_k^{(m_i)})}\left(\mathrm{F}^{-1}_{\chi^2(M^{(d)})}(1-\bar{\alpha})\right)$ for $i = 1, \ldots, i_{\max}$

Output: $\left\{\{J_k^{(\tilde{c}\mapsto c)}\}_{k=1}^K \,|\, (\tilde{c}\mapsto c)\in\mathcal{C}\right\}$

Fig. S3 summarizes the quantities involved in the FDR control procedure. Fig. S3-A illustrates the variables involved in hypothesis testing, and Fig. S3-B exhibits the receiver operating characteristic (ROC) curves for different values of $(M^{(d)}, \nu)$, as well as how the J-statistic is calculated for $\alpha = 0.05$.

In order to estimate the unknown non-centrality parameters $\nu_k^{(\tilde{c}\mapsto c)}$ given in Theorem S1, we make two additional assumptions. First, although the result of Theorem S1 establishes convergence in distribution as $\beta \to 1$, we make the assumption that $D_{k,\beta}^{(\tilde{c}\mapsto c)}$ is a sample drawn from a $\chi^2\left(M^{(d)}, \nu_k^{(\tilde{c}\mapsto c)}\right)$ density, when β is close to 1. This assumption is akin to the common adoption of a Gaussian density to parametrically describe uncertainties which are known to converge in distribution to a Gaussian random variable, thanks to the law of large numbers. In what follows, the dependence of $D_{k,\beta}^{(\tilde{c}\mapsto c)}$ and $\nu_k^{(\tilde{c}\mapsto c)}$ on c, $\tilde{c}$, and β will be suppressed for notational convenience.

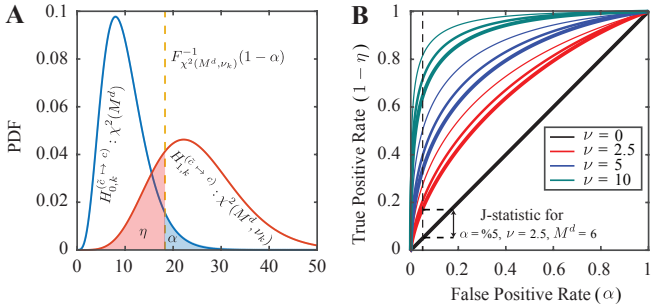

Fig. S3. A) PDFs of H_0 and H_1 for $M^{(d)} = 10$, $\nu = 15$, and $\alpha = 0.05$, B) ROC curves for different values of $M = \{2$ (narrow) , 4 (medium) , 6 (thick)$\}$ and $\nu = \{0$ (black) , 2.5 (red) , 5 (blue) , 10 (green)$\}$.

Second, we assume that ν_k changes smoothly in time. To this end, given that $\nu_k \ge 0$, we define the exponential link $\nu_k = \exp(z_k)$, for some random variable z_k in the range of $(-\infty, \infty)$ and impose first-order autoregressive dynamics of the form:

$$z_k = \rho z_{k-1} + e_k, \qquad \text{[S28]}$$

where $0 < \rho \le 1$ is a scaling factor, and $e_k \sim \mathcal{N}(0, \sigma_e^2)$ is a zero-mean i.i.d. Gaussian random variable with a variance of σ_e^2. Together with the assumption of $D_k \sim \chi^2\left(M^{(d)}, \nu_k\right)$, Eq. S28 forms a state-space model describing the dynamics of ν_k.

The parameters ρ and σ_e^2 are unknown, and need to be estimated. Assuming that the values of ρ and σ_e^2 are known, we can estimate $\{z_k\}_{k=1}^K$ given the sequence of deviance differences $\{D_k\}_{k=1}^K$ using approximate state-space smoothing (14). The resulting estimator consists of two steps: a forward filter, and a backward fixed interval smoother.

For the filtering algorithm, we exploit the unimodal property of non-central chi-squared distribution, and make a recursive Gaussian approximation to the posterior probability density function $p(z_k|D_{1:k}^{\beta})$, where the posterior modes and variances are computed recursively (14). Let $z_{k|l}$ and $\sigma_{k|l}^2$ denote the respective mode and variance of the state variable z_k, given the deviance samples up to and including time l, $\{D_{i,\beta}^{(\tilde{c}\mapsto c)}\}_{i=1}^l$. For notational convenience, we drop the dependence of $D_{i,\beta}^{(\tilde{c}\mapsto c)}$ on $\tilde{c}$, c and β, and denote it by D_i. Using the Bayes' rule and substituting the non-central chi-squared density function into the log-posterior, we get:

$$z_{k|k} := \underset{z_k}{\mathrm{argmax}} \left\{ -\frac{(D_k + \exp(z_k))}{2} + \frac{\xi}{2}(\log D_k - z_k) + \log I_\xi(\zeta_k) - \frac{(z_k - z_{k|k-1})^2}{2\sigma_{k|k-1}^2} \right\}, \qquad \text{[S29]}$$

where $\zeta_k := \sqrt{D_k \exp(z_k)}$, and $I_\xi(.)$ denotes the modified Bessel function of the first kind of order $\xi := M^{(d)}/2 - 1$. Note that in Eq. S29 a Gaussian approximation is applied

to the density $p(z_k|D^{\beta}_{1:k-1}) \sim \mathcal{N}(z_{k|k-1}, \sigma^2_{k|k-1})$, where the mode and variance are easily derived from Eq. S28 as $z_{k|k-1} = \rho z_{k-1|k-1}$ and $\sigma^2_{k|k-1} = \rho^2\sigma^2_{k-1|k-1} + \sigma^2_e$. From Eq. S29, the posterior mode $z_{k|k}$ can be computed as the solution to the following nonlinear equation:

$$z_k = z_{k|k-1} + \frac{\sigma^2_{k|k-1}}{2}\big(\zeta_k\, r_\xi(\zeta_k) - \exp(z_k)\big), \qquad \text{[S30]}$$

where the function $r_\xi(\zeta) := I_{\xi+1}(\zeta)/I_\xi(\zeta)$ is the ratio of modified Bessel functions of the first kind with order difference of one. This nonlinear equation can be solved numerically using iterative techniques such as Newton's method.

Given $z_{k|k}$, the posterior variance $\sigma^2_{k|k}$ can be computed as the negative inverse of the second order derivative of the log-posterior at $z_{k|k}$:

$$\sigma^2_{k|k} = \left((\sigma^2_{k|k-1})^{-1} + \frac{\exp(z_{k|k})}{2} - \frac{\zeta^2_{k|k}}{4}\left(1 - \frac{I_{\xi-1}(\zeta_{k|k})I_{\xi+1}(\zeta_{k|k})}{I_\xi(\zeta_{k|k})^2}\right)\right)^{-1}, \qquad \text{[S31]}$$

where $\zeta_{k|k} := \sqrt{D_k \exp(z_{k|k})}$, and we used the recurrence relation $I_{\xi-1}(\zeta) = I_{\xi+1}(\zeta) + (2\xi/\zeta)I_\xi(\zeta)$ to simplify the update rule. Unlike the ordinary Bessel functions, the modified Bessel functions of the first kind $I_\xi(\cdot)$ are exponentially growing. This could cause numerical stability issues for the recursive update rules of Eqs. S30 and S31, as the input ζ_k may take large values through recursion leading to extremely large values of the modified Bessel functions. To resolve potential numerical instability, we use the following sharp bounds on the ratio of Bessel function (15):

$$\sqrt{\zeta^2+(\xi+1)^2} - (\xi+1) \le \zeta\, r_\xi(\zeta) \le \sqrt{\zeta^2+(\xi+1/2)^2} - (\xi+1/2).$$

We select the upper-bound as the more accurate approximate of the ratio $\zeta\, r_\xi(\zeta)$ in Eq. S30 for large values of ζ. Moreover, the second order Bessel ratio in Eq. S31 can be replaced using a sharp upper bound on the Turánian of the modified Bessel functions of the first kind, $I_\xi(\zeta)^2 - I_{\xi-1}(\zeta)I_{\xi+1}(\zeta)$ (16):

$$\frac{I_\xi(\zeta)^2 - I_{\xi-1}(\zeta)I_{\xi+1}(\zeta)}{I_\xi(\zeta)^2} \le \frac{1}{\sqrt{\zeta^2+\xi^2-1/4}}. \qquad \text{[S32]}$$

Given filtered outputs $z_{k|k}$ and $\sigma^2_{k|k}$ obtained from the forward filtering algorithm, we next perform backward smoothing using the fixed interval smoothing algorithm (14), yielding the smoothed posterior modes $z_{k|K}$ and variances $\sigma^2_{k|K}$ for $k = K, K-1, \ldots, 1$ as follows:

$$\begin{cases} z_{k-1|K} &= z_{k-1|k-1} + s_k(z_{k|K} - z_{k|k-1}) \\ \sigma^2_{k-1|K} &= \sigma^2_{k-1|k-1} + s_k^2(\sigma^2_{k|K} - \sigma^2_{k|k-1}) \end{cases}, \qquad \text{[S33]}$$

where $s_k := \rho\sigma^2_{k-1|k-1}/\sigma^2_{k|k-1}$ is the backward smoothing gain. It should be noted that unlike the forward filtering, the backward smoothing step results in an overall batch-mode algorithm, as it refines the preceding filtered estimates $z_{k|k}$ using the deviance data D_i for $i > k$. Nevertheless, for real-time implementations one can always resort to the filtered estimates of the non-centrality parameters. Statistical confidence regions for both the filtered estimates $\widehat{z}^{\,\text{filtered}}_k \sim \mathcal{N}(z_{k|k}, \sigma^2_{k|k})$ and smoothed estimates $\widehat{z}^{\,\text{smoothed}}_k \sim \mathcal{N}(z_{k|K}, \sigma^2_{k|K})$ can be computed at each time step k and mapped to those of $\widehat{\nu}^{\,\text{filtered}}_k = \exp\big(\widehat{z}^{\,\text{filtered}}_k\big)$ and $\widehat{\nu}^{\,\text{smoothed}}_k = \exp\big(\widehat{z}^{\,\text{smoothed}}_k\big)$ in a straightforward fashion. Algorithm S2 summarizes the non-central χ^2 filtering and smoothing procedure.

Algorithm S2 Non-central χ^2 Filtering and Smoothing Algorithm

Input: D_k, $M^{(d)}$, ρ, σ^2_e, $z_{0|0}$ and $\sigma^2_{0|0}$.

for $k = 1, 2, \ldots, K$ **do**

  Define $\xi := M^{(d)}/2 - 1$ and $\zeta_{k|k} := \sqrt{D_k \exp(z_{k|k})}$

  $z_{k|k-1} = \rho\, z_{k-1|k-1}$

  $\sigma^2_{k|k-1} = \rho^2\sigma^2_{k-1|k-1} + \sigma^2_e$

  $z_{k|k} = z_{k|k-1} + \frac{\sigma^2_{k|k-1}}{2}\left(\sqrt{\zeta^2_{k|k}+(\xi+1/2)^2} - (\xi+1/2) - \exp(z_{k|k})\right)$

  $\sigma^2_{k|k} = \left((\sigma^2_{k|k-1})^{-1} + \frac{\exp(z_{k|k})}{2} - \frac{\zeta^2_{k|k}}{4\sqrt{\zeta^2_{k|k}+\xi^2-1/4}}\right)^{-1}$

  $\widehat{\nu}^{\,\text{filtered}}_k = \exp(z_{k|k})$

  $\mathcal{CR}^{\,\text{filtered}}_k = \left[\exp\left(z_{k|k} \pm \Phi^{-1}(1-\epsilon/2)\sigma_{k|k}\right)\right]$

Given $\{z_{k|k}\}_{k=1}^K$ and $\{\sigma^2_{k|k}\}_{k=1}^K$

for $k = K, K-1, \ldots, 1$ **do**

  $z_{k-1|K} = z_{k-1|k-1} + s_k(z_{k|K} - z_{k|k-1})$

  $\sigma^2_{k-1|K} = \sigma^2_{k-1|k-1} + s_k^2(\sigma^2_{k|K} - \sigma^2_{k|k-1})$

  $\widehat{\nu}^{\,\text{smoothed}}_{k-1} = \exp(z_{k-1|K})$

  $\mathcal{CR}^{\,\text{smoothed}}_{k-1} = \left[\exp\left(z_{k-1|K} \pm \Phi^{-1}(1-\epsilon/2)\sigma_{k-1|K}\right)\right]$

Output: Filtered estimates $(\widehat{\nu}^{\,\text{filtered}}_{1:K}, \mathcal{CR}^{\,\text{filtered}}_{1:K})$, and smoothed estimates $(\widehat{\nu}^{\,\text{smoothed}}_{1:K}, \mathcal{CR}^{\,\text{smoothed}}_{1:K})$

In order to simultaneously smooth z_k's and estimate the unknown parameters ρ and σ^2_e, an Expectation-Maximization (EM) algorithm can be used (17). The details of EM algorithm are given in Section 4.

3. Derivation of the Recursive Computation of the AGC

In order to achieve recursive computation, we exploit the smoothness of the point process log-likelihood function, and approximate each scalar-valued log-likelihood function $\ell_i(\widehat{\boldsymbol{\omega}}_k)$ using a second order Taylor's series expansion around $\widehat{\boldsymbol{\omega}}_i$ for $i \le k$. Retaining the first three terms of the expansion yields:

$$\ell_i(\widehat{\boldsymbol{\omega}}_k) \approx \ell_i(\widehat{\boldsymbol{\omega}}_i) + (\widehat{\boldsymbol{\omega}}_k - \widehat{\boldsymbol{\omega}}_i)'\dot{\boldsymbol{\ell}}_i(\widehat{\boldsymbol{\omega}}_i) + \tfrac{1}{2}(\widehat{\boldsymbol{\omega}}_k - \widehat{\boldsymbol{\omega}}_i)'\ddot{\boldsymbol{\ell}}_i(\widehat{\boldsymbol{\omega}}_i)(\widehat{\boldsymbol{\omega}}_k - \widehat{\boldsymbol{\omega}}_i), \qquad \text{[S34]}$$

where $\dot{\boldsymbol{\ell}}_i(\cdot)$ and $\ddot{\boldsymbol{\ell}}_i(\cdot)$ denote the gradient vector and Hessian matrix with respect to $\boldsymbol{\omega}$, which can be computed from Eq. 2 for the logit-linked GLM model as follows:

$$\dot{\boldsymbol{\ell}}_i(\widehat{\boldsymbol{\omega}}_i) = \mathbf{X}_i'\boldsymbol{\varepsilon}_i, \qquad \text{[S35]}$$

$$\ddot{\boldsymbol{\ell}}_i(\widehat{\boldsymbol{\omega}}_i) = -\mathbf{X}_i'\boldsymbol{\Lambda}_i\mathbf{X}_i, \qquad \text{[S36]}$$

where $\boldsymbol{\varepsilon}_i := \mathbf{n}_i - \boldsymbol{\lambda}_i(\widehat{\boldsymbol{\omega}}_i)\Delta$ denotes the point process innovation vector at time window i, and $\boldsymbol{\Lambda}_i := \operatorname{diag}\left(\boldsymbol{\lambda}_i\Delta \odot (1 - \boldsymbol{\lambda}_i\Delta)\right)$ is a $W \times W$ diagonal matrix with $(\boldsymbol{\Lambda}_i)_{m,m} := \lambda_{(i-1)W+m}(\widehat{\boldsymbol{\omega}}_i)\Delta(1 - \lambda_{(i-1)W+m}(\widehat{\boldsymbol{\omega}}_i)\Delta)$ as the m-th diagonal element obtained from the second-order derivative of the logistic log-likelihood function. Substituting the quadratic Taylor's approximation of Eq. S34 into Eq. 2 and rearranging terms will lead to the following recursive update rule for the adaptive log-likelihoods at time step k:

$$\ell^\beta_k(\widehat{\boldsymbol{\omega}}_k) = a_k + \widehat{\boldsymbol{\omega}}_k'\mathbf{b}_k + \frac{1}{2}\widehat{\boldsymbol{\omega}}_k'\mathbf{B}_k\widehat{\boldsymbol{\omega}}_k, \qquad \text{[S37]}$$

where

$$a_k = \sum_{i=1}^{k} \beta^{k-i} \big(\mathbf{1}_W{}' \boldsymbol{\ell}_i(\widehat{\boldsymbol{\omega}}_i) - \widehat{\boldsymbol{\omega}}_i' \mathbf{X}_i' \boldsymbol{\varepsilon}_i - \frac{1}{2} \widehat{\boldsymbol{\omega}}_i' \mathbf{X}_i' \boldsymbol{\Lambda}_i \mathbf{X}_i \widehat{\boldsymbol{\omega}}_i \big),$$

$$\mathbf{b}_k = \sum_{i=1}^{k} \beta^{k-i} \mathbf{X}_i' \big(\boldsymbol{\varepsilon}_i + \boldsymbol{\Lambda}_i \mathbf{X}_i \widehat{\boldsymbol{\omega}}_i \big),$$

$$\mathbf{B}_k = - \sum_{i=1}^{k} \beta^{k-i} \mathbf{X}_i' \boldsymbol{\Lambda}_i \mathbf{X}_i, \qquad \text{[S38]}$$

in which $\boldsymbol{\ell}_i(\widehat{\boldsymbol{\omega}}_i) := [\ell_{(i-1)W+1}(\widehat{\boldsymbol{\omega}}_i), \ldots, \ell_{iW}(\widehat{\boldsymbol{\omega}}_i)]'$ denotes the vector of log-likelihoods corresponding to the ith time window, and $\mathbf{1}_W := [1, \ldots, 1]'$ is the vector of all ones of length W. It is easy to see that a_k, $\mathbf{b}_k$ and $\mathbf{B}_k$ also admit recursive update rules at time step k:

$$a_k = \beta\, a_{k-1} + \mathbf{1}_W{}' \boldsymbol{\ell}_k(\widehat{\boldsymbol{\omega}}_k) - \widehat{\boldsymbol{\omega}}_k' \mathbf{X}_k' \boldsymbol{\varepsilon}_k - \frac{1}{2} \widehat{\boldsymbol{\omega}}_k' \mathbf{X}_k' \boldsymbol{\Lambda}_k \mathbf{X}_k \widehat{\boldsymbol{\omega}}_k,$$

$$\mathbf{b}_k = \beta\, \mathbf{b}_{k-1} + \mathbf{X}_k' \left(\boldsymbol{\varepsilon}_k + \boldsymbol{\Lambda}_k \mathbf{X}_k \widehat{\boldsymbol{\omega}}_k \right),$$

$$\mathbf{B}_k = \beta\, \mathbf{B}_{k-1} - \mathbf{X}_k' \boldsymbol{\Lambda}_k \mathbf{X}_k. \qquad \text{[S39]}$$

By performing the recursive computation of Eq. S37 for both the full model and the reduced model, a fully recursive update procedure for the AGC measure of Eq. 5 is obtained, which enables us to track the G-causal interactions among the neurons in an online fashion. This fully recursive procedure can be further extended to our proposed statistical inference framework based on the de-biased deviance statistics. To this end, we obtain a recursive update rule for the quadratic bias terms in Eq. 6. The recursion for the score statistic evaluated at the current estimate, $\dot{\boldsymbol{\ell}}_k^{\beta}(\widehat{\boldsymbol{\omega}}_k)$, is readily available through a similar treatment using the Taylor's series expansion and is employed in the ℓ_1–PPF$_1$ filtering procedure for estimating the maximizers of ℓ_1-regularized ML problems recursively (1). This update rule simplifies to:

$$\dot{\boldsymbol{\ell}}_k^{\beta}(\widehat{\boldsymbol{\omega}}_k) = \mathbf{b}_k + \mathbf{B}_k \widehat{\boldsymbol{\omega}}_k. \qquad \text{[S40]}$$

The inverse Hessians $\ddot{\boldsymbol{\ell}}_k^{\beta}(\widehat{\boldsymbol{\omega}}_k)^{-1}$ can also be efficiently computed via the Woodbury matrix identity applied to the update rule of B_k. When the Hessians are not invertible, a recursive implementation of the node-wise regression procedure of (2) can be used, which is developed in (18) using the SPARLS iteration (19) for RLS-type exponentially weighted log-likelihoods. Algorithm S3 summarized the recursive computation of the exponentially-weighted log-likelihoods at window k.

Algorithm S3 Recursive update rule for $\ell_k^{\beta}(\widehat{\boldsymbol{\omega}}_k)$

Input: $\mathbf{n}_k$, $\mathbf{X}_k$, $\widehat{\boldsymbol{\omega}}_k$, a_{k-1}, $\mathbf{b}_{k-1}$, and $\mathbf{B}_{k-1}$.

$\mathbf{y}_k = \mathbf{X}_k \widehat{\boldsymbol{\omega}}_k$

$\boldsymbol{\lambda}_k \Delta = \text{logit}^{-1}(\mathbf{y}_k)$

$\boldsymbol{\varepsilon}_k = \mathbf{n}_k - \boldsymbol{\lambda}_k \Delta$

$\boldsymbol{\Lambda}_k = \text{diag}\left(\boldsymbol{\lambda}_k \Delta \odot (1 - \boldsymbol{\lambda}_k \Delta)\right)$

$a_k = \beta\, a_{k-1} + \mathbf{1}_W{}' \boldsymbol{\ell}_k(\widehat{\boldsymbol{\omega}}_k) - \mathbf{y}_k' \boldsymbol{\varepsilon}_k - \frac{1}{2} \mathbf{y}_k' \boldsymbol{\Lambda}_k \mathbf{y}_k$

$\mathbf{b}_k = \beta\, \mathbf{b}_{k-1} + \mathbf{X}_k' \left(\boldsymbol{\varepsilon}_k + \boldsymbol{\Lambda}_k \mathbf{y}_k \right)$

$\mathbf{B}_k = \beta\, \mathbf{B}_{k-1} - \mathbf{X}_k' \boldsymbol{\Lambda}_k \mathbf{X}_k$

Output: $\ell_k^{\beta}(\widehat{\boldsymbol{\omega}}_k) = a_k + \widehat{\boldsymbol{\omega}}_k' \mathbf{b}_k + \frac{1}{2} \widehat{\boldsymbol{\omega}}_k' \mathbf{B}_k \widehat{\boldsymbol{\omega}}_k$

4. Parameter Selection

In this section, we describe how the various parameters in our proposed AGC estimation procedure are selected, and discuss the underlying trade-offs thereof.

Forgetting Factor. In the adaptive filtering setting with a forgetting factor mechanism β and window size W, the effective block length of the filter is determined by $N_{\text{eff}} = \frac{W}{1-\beta}$. It was shown in (1) that the estimation error scales as $\mathcal{O}(\sqrt{s \log M / N_{\text{eff}}})$ in the ℓ_2 sense, where s denotes the sparsity level. Thus, the forgetting factor β controls the trade-off between the estimation and tracking performance of the filter. That is, a choice of β close to 1 corresponds to a large effective block length N_{eff}, which in turn results in a more accurate estimation of the modulation parameters $\widehat{\boldsymbol{\omega}}_k$, and consequently the AGC, at the cost of losing the trackability of the underlying dynamics. On the other extreme, a choice of β far from 1 reduces the effective block length, and thereby results in capturing the fast dynamics of the underlying time-varying process, although the estimation accuracy degrades. As discussed in the remark following the proof of Theorem 1, the proposed statistical testing procedure enables us to detect G-causal links associated with true cross-history components of the order of $\boldsymbol{\omega}_k^{\beta} = \mathcal{O}(\sqrt{1-\beta})$. Hence, a choice of β close to 1 will increase the test strengths. In the applications of interest in this paper, the underlying dynamics are slower than the sampling rate, which allows us to choose forgetting factor values sufficiently close to 1. While it may be beneficial to tune β via cross-validation, our numerical experiments show that the resulting values of β turn to be close to 1 (i.e., $1 - \beta \in [10^{-4}, 10^{-2}]$). Therefore, in order to simplify the cross-validation procedure, we fixed the value of β close to 1 in our analysis. It is noteworthy that the usage of the forgetting factor mechanism mitigates the problem of choosing a window size faced by GC inference methods based on sliding-window processing.

Model Order Selection. Our model selection procedure is grounded in the compressed sensing theory. In contrast to classical model order selection procedures (e.g., AIC), compressed sensing suggests choosing large model orders followed by sparse regularization to avoid overfitting. Indeed, our recent results on extending the theoretical guarantees of compressed sensing to processes with non-i.i.d. and history dependent covariates (1, 4), show that recovery of sparse history kernels with large ambient dimensions M is possible from a limited number of observations N, in which N may be comparable or smaller than M, as long as the sparsity level s is small enough. In more precise terms, long kernels of self-history can be robustly estimated given an effective number of observations N_{eff} scaling sub-linearly with M and s.

The benefit of employing such models with long self-history kernels is two-fold: first, long self-history kernels M_H^{self} enable us to maximally capture the intrinsic spiking statistics of a unit. Second, due to the autoregressive nature of these models, long self-history kernels allow for estimation, and thereby correcting for the effects of latent confounding variables, which cannot be explained by the cross-history influences from other units. Thus, we choose $M_H^{\text{self}} > M_H^{\text{cross}}$ to maximally capture the aforementioned intrinsic and latent confounding effects. At the same time, smaller values of M_H^{cross} are beneficial in increasing the statistical test strengths, as they directly set the statistical thresholds for multiple hypothesis testing. In Section 9, we present two illustrative numerical experiments that corroborate our choices for these parameters.

Adaptive Filtering Parameters. In order to achieve an estimation performance with high accuracy, we select the effective

block length $N_{\text{eff}} \gg M^{(F)}$ to be larger than the kernel length. We use non-overlapping spike counting windows of length W_H for parameterizing the self- and cross-history kernels, where W_H is often chosen to be comparable to the data window length W.

For the adaptive filtering setting, we first standardize the matrix of covariates (i.e., zero-mean columns with unit norm), and then apply the ℓ_1–PPF$_1$ adaptive filter, with a step size of $\varsigma := \frac{1-\beta}{cW}$, where c is a constant often chosen in the range $c \in [1, 10]$, to achieve different levels of smoothing (1).

The regularization parameter γ for the ℓ_1–PPF$_1$ is chosen as $\gamma = \mathcal{O}\big(\sqrt{\log M/N_{\text{eff}}}\big)$ based on the results of Theorem 1 in (1), and the asymptotic scaling requirement in (2), to obtain consistent ℓ_1-regularized ML estimates. In order to adapt this parameter to different neurons, we choose $\gamma^{(c)} = \bar{\gamma}^{(c)}\sqrt{\bar{\kappa}^{(c)} \log M/N_{\text{eff}}}$ for neuron (c), where $\bar{\kappa}^{(c)} := \text{var}(\mathbf{n}^{(c)}) = \bar{\lambda}\Delta^{(c)}(1 - \bar{\lambda}\Delta^{(c)})$, followed by tuning the normalized regularization parameter $\bar{\gamma}^{(c)}$ for each neuron in a data-driven fashion via the even-odd two-fold cross validation procedure discussed in (1).

Note that when the underlying functional network is fully connected, the cross-validation procedure for tuning the regularization parameter γ is expected to choose values near zero (i.e., no sparsity in the parameter vectors), and hence our methodology can adapt to non-sparse network connectivity as well. For the applications of interest in this work, the cross-validation procedure consistently resulted in sparse functional networks.

Finally, it is possible to generalize the ℓ_1-regularization scheme to have a different regularization parameter $\gamma^{(c,\tilde{c})}$ for the cross-history parameters of units (c) and $(\tilde{c})$. Theoretical analysis, however, suggest that there is little benefit in terms of estimation accuracy in doing so, which comes at the cost of higher computational complexity in the cross-validation and bias correction stages. More precisely, separate regularization of each of the cross-history parameters may result in better constants in the error rate, but the asymptotic scaling of the rate remains unchanged. For instance, as mentioned earlier the results of (1) show that the estimation error scales as $\mathcal{O}(\sqrt{s \log M/N_{\text{eff}}})$, which is optimal modulo the logarithmic factor. By viewing the concatenation of several sparse vectors as another sparse vector, we use a single regularization parameter that is tuned appropriately via cross-validation in order to select a sparse model at a near optimal error rate. Nevertheless, the ℓ_1-PPF$_1$ procedure can be generalized in a straightforward fashion to accommodate multiple regularization parameters, thanks to the separable nature of the ℓ_1 norm and the underlying proximal algorithms (See (1) for details).

Parameters of the Non-central χ^2 Filtering and Smoothing. For the non-central χ^2 filtering and smoothing algorithm, we select the scaling factor $\rho \in [0.999, 1]$ close (or equal) to one to promote temporal continuity. The state variance σ_e^2 plays the role of a smoothing factor for non-centrality parameter estimates $\widehat{\nu}_k$, and can be determined in two ways. First, we can choose a small value in the range $[10^{-7}, 10^{-4}]$ suggested by our numerical experiments, which results in smooth estimates of $\widehat{\nu}_k$ for a wide range of settings. Second, σ_e^2 can be systematically estimated via the expectation maximization (EM) algorithm (20) in a data-driven fashion using the observed deviance data $D_{1:K} := \{D_k\}_{k=1}^K$. We take $z_{1:K} := \{z_k\}_{k=1}^K$ as the set of latent variables for the EM algorithm. Given an estimate $\widehat{\sigma}_e^{2,(\ell)}$ at the ℓ-th iteration, the E-step at the $(\ell+1)$-st iteration computes:

$$\mathbb{E}_{\mathbf{z}}\Big[\log p\big(D_{1:K}, z_{1:K}|\sigma_e^2\big)\big|D_{1:K}, \widehat{\sigma}_e^{2,(\ell)}\Big] = -\frac{K}{2}\log(\sigma_e^2) - \frac{1}{2\sigma_e^2}\sum_{k=1}^{K}\Big\{\big(\sigma_{k|K}^2+z_{k|K}^2\big)+\rho^2\big(\sigma_{k-1|K}^2+z_{k-1|K}^2\big) - 2\rho\big(\sigma_{k-1,k|K}^2 + z_{k-1|K}z_{k|K}\big)\Big\} + \textsf{cnst.}, \tag{S41}$$

where $\mathbb{E}_{\mathbf{z}}\big[.|D_{1:K}, \widehat{\sigma}_e^{2,(\ell)}\big]$ denotes the expectation operator with respect to the latent variables given the complete set of deviance data $D_{1:K}$ and the current estimate of the parameter $\widehat{\sigma}_e^{2,(\ell)}$, and cnst. denotes all terms not dependent on σ_e^2. It is noteworthy that calculation of the E-step involves computation of the smoothed means and variances $\mathbb{E}_{\mathbf{z}}\big[z_k^2\,|\,D_{1:K}, \widehat{\sigma}_e^{2,(\ell)}\big] = \sigma_{k|K}^2 + z_{k|K}^2$, which are readily available from the non-central chi-squared smoothing given by Eq. S33, and the covariance terms $\mathbb{E}_{\mathbf{z}}\big[z_{k-1}z_k|D_{1:K}, \widehat{\sigma}_e^{2,(\ell)}\big] = \sigma_{k-1,k|K}^2 + z_{k-1|K}z_{k|K}$, which can be computed using a state-space covariance smoothing algorithm (17) as $\sigma_{k-1,k|K}^2 = s_k\sigma_{k|K}^2$. The M-step gives the update for $\widehat{\sigma}_e^{2,(\ell+1)}$ by maximizing S41 as follows:

$$\widehat{\sigma}_e^{2,(\ell+1)} = \frac{1}{K}\sum_{k=1}^{K}\Big\{\big(\sigma_{k|K}^2 + z_{k|K}^2\big)+\rho^2\big(\sigma_{k-1|K}^2 + z_{k-1|K}^2\big) - 2\rho\big(\sigma_{k-1,k|K}^2 + z_{k-1|K}z_{k|K}\big)\Big\}. \tag{S42}$$

5. Numerical Choices of the Parameters Used in the Applications Section

Parameters for the Simulated Example. We selected the modulation parameter vectors to be the same for all the G-causal interactions, and set to $\boldsymbol{\omega}_{\text{exc.}} = [1, 0, 0, 2, 0, 0, 0, 0, 0, 1]$ for excitatory links and $\boldsymbol{\omega}_{\text{inh.}} = -\boldsymbol{\omega}_{\text{exc.}}$ for the inhibitory links, where each component corresponds to a uniform non-overlapping spike counting window of length 10 bins (or 10 ms). The modulation parameter vector associated with the non-existing G-causal links (such as $(8 \mapsto 2)$) is set to all zeros. The self-history dependence for all neurons is chosen to be of inhibitory and static nature to maintain stable behavior for simulation purposes. The norm of all non-zero parameter vectors is normalized to 1. The average spiking probability is set to $\bar{\lambda}\Delta \approx 0.07 \ll 1$ by choosing the baseline firing parameter $\mu_k = -2.597$ to be the same for all neurons.

To model the dynamics of the G-causal links in the second segment of the simulation, we enforce a linear time evolution for all the coefficients of underlying parameter vector, with a respective decay and growth for the links associated with neurons (1) and (5). For estimation of G-causal interactions, we select the sparse parameter vector associated with the full GLM model of neuron (c) to be in form of $\boldsymbol{\omega}_k^{(c)} = [\mu_k^{(c)}, \boldsymbol{\omega}_k^{(c,1)\prime}, \boldsymbol{\omega}_k^{(c,2)\prime}, \ldots, \boldsymbol{\omega}_k^{(c,C)\prime}]'$ of length $M^{(F)}$, composed of the scalar baseline parameter $\mu_k^{(c)}$, and sub-vectors $\boldsymbol{\omega}_k^{(c,c)}$ of length M_H^{self}, and $\boldsymbol{\omega}_k^{(c,\tilde{c})}$ of length M_H^{cross} for $\tilde{c} \neq c$, denoting the respective parameter vectors tuning the self-history dependence and the cross-history effects from neuron $(\tilde{c})$. We select $M_H^{\text{cross}} = M_H^{\text{self}} = 10$ history components associated with

the respective kernel lengths of $L_H^{\text{cross}} = L_H^{\text{self}} = 100$ ms, obtained by non-overlapping windows of length $W_H = 10$ bins. Note that $M^{(F)} = 81$, and $M^{(R)} = M^{(F)} - M_H^{\text{cross}} = 71$.

We employ the sparse adaptive filter ℓ_1–PPF$_1$ to estimate the sparse parameter vectors $\widehat{\boldsymbol{\omega}}_k$ at every time step k for both the full and reduced models. For the ℓ_1–PPF$_1$ filtering algorithm, an effective block length of $N_{\text{eff}} = 10k$ is selected with a window size of $W = 20$, forgetting factor of $\beta = 0.998$ chosen sufficiently close to one, step size of $\varsigma := \frac{1-\beta}{W}$, and $L = 1$ number of iterations. The regularization parameter is tuned for each cell separately $\bar{\gamma}^{(c)} \in [0.3, 0.5]$, via the two-fold even-odd cross validation (1). For the χ^2 filtering and smoothing algorithm, the smoothing and scaling factors are selected as $\sigma_e^2 = 5 \times 10^{-6}$ and $\rho = 1$, respectively, using an initialization of $z_{0|0} = 0$ and $\sigma_{0|0}^2 = 1$ in the EM algorithm.

For the performance comparison in Fig. 3-E, we have adapted the methods in (21) and (22), which are designed for static connectivity inference, to the time-varying setting in full fairness. First, due to the batch-mode computation of these static methods, we divided the total $K = 120k$ observed bins to non-overlapping window segments of length $W^{\text{ML}} = 10k$, matching the effective block length N_{eff} of our dynamic method. Both methods compute the ML estimates of the network parameters for each segment. We have therefore selected the true model orders $M^{\text{ML}} = 10$ for both methods, matching the selected model order for our AGC inference method, so as to have a fair statistical comparison and to ensure that they operate at their optimal performance (note that the dimensionality difference M_d has a particularly pivotal role in the inference procedure).

The method in (21) computes a static GC connectivity map obtained from nested full and reduced ML estimates, followed by an FDR control procedure for correction of multiple comparison errors. The method in (22) performs a likelihood-ratio test to assess the significance of each pair-wise interaction. The same significance levels are chosen for the statistical tests in both methods, to match our FDR rate of $\alpha = 0.1$. Finally, both methods have been modified to the logit-linked GLM setting, in order to ensure their consistency with the generative model used for simulating the spike trains.

Parameters for the Analysis of Spontaneous Activity in the Mouse Auditory Cortex. We used time bins of length $\Delta = 33$ ms, equal to the sampling interval. For the GLM models, we chose $M_H^{\text{cross}} = 3$ cross-history components associated with a block length of $L_H^{\text{cross}} = 10$ samples, obtained by non-overlapping windows of length $[2, 4, 4]$ samples. To correct for possible latent confounding effects, we select a larger self-history kernel of $L_H^{\text{self}} = 30$ samples segmented using windows of $[2, 4, 4, \ldots, 4]$ samples, giving a total of $M_H^{\text{self}} = 8$ parameters. We corrected for the clustered spike detection effect of the constrained-foopsi method, using a masking window for rejecting multiple consecutive spikes. We selected an optimal data-driven masking window of size $W^{\text{mask}} = 8$ samples, obtained by computing minimum rise-time of the calcium peaks inferred from the smoothed fluorescence traces of all cells. Then, the spikes detected within an interval of length W^{mask} are rejected.

We employ ℓ_1–PPF$_1$ algorithm for estimating $\widehat{\boldsymbol{\omega}}_k$ with a forgetting factor of $\beta = 0.999$, a window size of $W = 10$ bins, and $L = 1$ number of iterations. The regularization parameter was tuned for each cell via two-fold even-odd cross validation. The χ^2 filtering and smoothing algorithm parameters are chosen as $\rho = 0.999$, and $\sigma_e^2 = 10^{-3}$. The J-statistics are evaluated at the mean FDR for the detected GC links.

Parameters for the Analysis of the Ferret A1-PFC Activity. We discretized the total duration of $\mathcal{T} = 420$ s using bins of length $\Delta = 1$ ms. The GLM modulation parameter $\boldsymbol{\omega}_k := [\mu_k; \boldsymbol{\omega}_k^{\text{hist}}; \boldsymbol{\omega}_k^{\text{STRF}}]$ at time k consists of the baseline firing parameter μ_k, the history dependence vector $\boldsymbol{\omega}_k^{\text{hist}}$, as well as the STRF vector denoted by $\boldsymbol{\omega}_k^{\text{STRF}}$. For the history dependence parameters, we selected $M_H^{\text{cross}} = 3$ cross-history and $M_H^{\text{self}} = 21$ self-history components associated with respective history block lengths of $L_H^{\text{cross}} = 100$ ms and $L_H^{\text{self}} = 1$ s, using non-overlapping windows of $W_H^{\text{cross}} = [20, 30, 50]$ and $W_H^{\text{self}} = [20, 30, 50, \ldots, 50]$ bins, respectively.

For the STRF parameters, we use a vectorized array of size $I \times J$, with $I = 50$ time lag bins, and $J = 50$ frequency bins in logarithmic scale, uniformly spanning time lags in the range of $[0, 50]$ ms, and frequencies in the range of $f \in [500, 16k]$ Hz, respectively. To capture the inherent sparsity of the STRFs in the time-frequency domain, we use a representation $\boldsymbol{\theta}_k^{\text{STRF}} = \mathbf{G}\boldsymbol{\omega}_k^{\text{STRF}}$, where $\mathbf{G}$ is a Gaussian time-frequency dictionary of 49 Gaussian atoms (1), and $\boldsymbol{\omega}_k^{\text{STRF}}$ and $\boldsymbol{\theta}_k^{\text{STRF}}$ denote the sparse representation of the STRF (with 49 parameters) and the vectorized STRF at time k, respectively. We used two-dimensional symmetric Gaussian kernels with a variance of $D^2/4$ as Gaussian atoms in time-frequency plane, where atoms are distributed on a grid of size 7×7 with a spacing of $D = 7$ bins. The vectorized array of the TORC sequence spectrograms with J frequency bins and I time lags is considered as the stimulus sequence $\mathbf{s}_k$ in the GLM model.

We used the ℓ_1–PPF$_1$ filter to estimate the sparse parameter vectors $\widehat{\boldsymbol{\omega}}_k$ associated with the reduced and full GLMs for each neuron in a dynamic fashion. We selected a forgetting factor of $\beta = 0.9998$, a window size of $W = 8$, a step size $\varsigma = \frac{1-\beta}{5W}$, $L = 20$ number of iterations per step, and regularization parameters $\gamma^{(c)}$ tuned for each unit separately via two-fold even-odd cross validation. We chose the scaling factor $\rho = \beta$, and the smoothing factor $\sigma_e^2 = 5 \times 10^{-6}$ for the χ^2 filtering and smoothing algorithm. The FDR is controlled at the rate $\alpha = 0.1$, and the J-statistics computed at mean FDR $\bar{\alpha} = 0.0119$, testing for $|\mathcal{C}| = 9 \times 8 = 72$ possible GC links among the units.

6. Computational Complexity Considerations

The computational complexity of Algorithm 1 (per cross-validation iteration) is linear in the total data length T and quadratic in the network size C and parameter orders M, due to the RLS-type adaptive filtering procedure used (1). However, the high number of cross-validation iterations required to tune the regularization parameters increases the overall runtime of the algorithm. Substantial reduction of the runtime can be achieved by parallel implementation: the cross-validation steps for each unit can be done independently of the others, and therefore using a natural parallel implementation, the runtime would reduce by $1/C$. We have not used this parallel scheme in our current implementation deposited on GitHub (https://github.com/Arsha89/AGC_Analysis), as we deemed it beyond the scope of this work. In order to efficiently analyze data from high-density neuronal recordings, we suggest the use of a parallel implementation and view it as a future work.

7. Robustness to the Choice of Parameters

In this section, we inspect the robustness of the proposed AGC inference with respect to the choice of three major parameters: the dimensionality difference M_d, the regularization parameter γ, and the effective block length of the adaptive filter $T_{\text{eff}} := N_{\text{eff}}\Delta = \frac{W\Delta}{1-\beta}$. Before doing so, we describe the computation and statistical assessment of the TDR and FAR performance metrics in the analysis of Figs. 3 and 4, as well as the current and forthcoming section.

The TDR and FAR Performance Metrics. In the Applications section of the main manuscript, we compared the performance of the proposed AGC inference with the methods of (22) and (21) in terms of TDR and FAR performance metrics. Given the continuous nature of the AGC links, as opposed to the binary connectivity measures of the other two methods, we binarize the resulting J-statistics for fairness of comparison. To this end, let $A_R^{(\tilde{c}\mapsto c)}$ be the fraction of times within a time window where the AGC link $(\tilde{c} \mapsto c)$ is identified with high statistical significance $J_k^{(\tilde{c}\mapsto c)} > J_{\text{th}}$. We call an AGC link active within a given time window if $A_R^{(\tilde{c}\mapsto c)} > A_{\text{th}}$, and inactive otherwise. We selected the thresholds to be $A_{\text{th}} = \frac{1}{3}$ and $J_{\text{th}} = \frac{1-\bar{\alpha}}{3}$.

The TDR at each time window is computed as the ratio of the correctly identified links to the total number of existing GC links. The FAR at each time k is computed as the ratio of spuriously detected links to the total number of non-existent links. Given the ground truth GC map shown in Fig. 3-E, these performance metrics can be computed for the static (first and last) segments of the experiment in a straightforward fashion. For the middle segment, where the GC influences undergo dynamic changes, we define the ground truth as follows: a threshold of $G_{\text{th}} = \frac{1}{4}$ is used to binarize the ground truth GC links, which linearly ascend from 0 to 1 (emerging link) or descend from 1 to 0 (vanishing link) in the middle segment. For each repetition of the simulation, the FAR and TDR metrics are computed for each of the three segments by averaging over the time windows within, resulting in two summary statistics. Boxes indicate the mean values as well as the 90% confidence intervals pooled across all repetitions, and are plotted in green and red, respectively.

Due to the highly non-Gaussian nature of the empirical distributions of the paired difference metrics, we have used the non-parametric Wilcoxon signed-rank test for comparison, both in the main manuscript (Fig. 4) and the forthcoming section. The corresponding effect sizes are computed in the form of rank correlation $r := \mathscr{W}/\mathscr{S}$, where $\mathscr{W}$ is the Wilcoxon signed-rank statistic and $\mathscr{S}$ is the total sum of ranks (23).

The AUC values reported in the main manuscript (Fig. 4) are computed as the area under the ROC curves. These curves are obtained by varying the values of mean FDR $\bar{\alpha} \in [0, 1]$ for AGC and the statistical thresholds of (21) and (22) and plotting the corresponding (TDR,FAR) pairs averaged across repetitions. Figure S4 shows the ROC curves of the three methods for the three segments of simulation. While the methods of (21) and (22) exhibit similar ROC performances, the AGC achieves higher AUC values, particularly in the middle segment. We expect the performance gap between the AGC inference and the other two methods to increase for larger networks with higher sparsity.

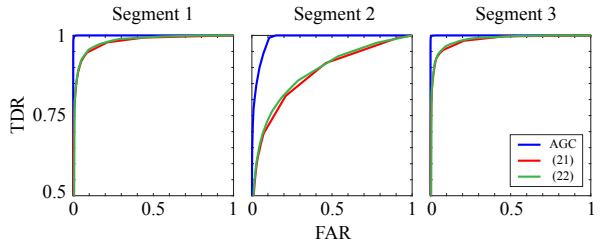

Fig. S4. ROC performance curves of the AGC inference (blue) and the methods of (21) (red) and (22) (green) for the three segments of the simulation setting of the main manuscript (Fig. 4).

Assessing the Robustness of AGC Inference to the Choice of Parameters. We consider three different choices for each parameter $\{M_d, c_\gamma, T_{\text{eff}}\}$, and for each choice, we run the simulation of Fig. 3 for $R = 100$ repetitions, where a random sequence of spike trains are generated at each repetition based on the network dynamics of Fig. 3-A. In each setting, the rest of the parameters are chosen as described in section 5.

Robustness to the choice of M_d**.** For the dimensionality difference M_d, we consider three settings of $M_d \in \{10, 15, 20\}$. Fig. S5 shows the performance results for different choices of M_d. While the FAR values remain consistently low (i.e., < 0.01, on average), as expected the larger choices of M_d would impose stricter statistical thresholds on the hypothesis tests (See Fig. S3-B), leading to slight degradation of the TDR performance.

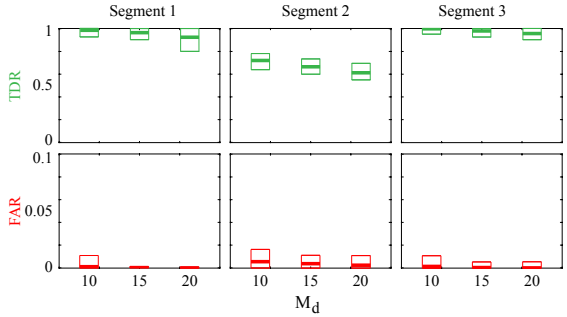

Fig. S5. Performance of the AGC inference for three different values of $M_d \in \{10, 15, 20\}$, in terms of TDR (top row) and FAR (bottom row).

Robustness to the choice of γ**.** For the choice of regularization parameter, we consider three different settings for $\boldsymbol{\gamma} = c_\gamma \boldsymbol{\gamma}^*, c_\gamma \in \{0.1, 1, 10\}$, where c_γ denotes a scaling factor and $\boldsymbol{\gamma}^*$ represents the optimally tuned regularization parameter vector obtained from cell-by-cell two-fold even-odd cross-validation. Fig. S6 reveals the robustness of the AGC inference with respect to the choice of the regularization parameter. It can be observed that the resulting performance metrics show resilience to under-regularization ($c_\gamma = 0.1$), while the TDR performance notably degrades due to over-regularization ($c_\gamma = 10$). This is due to the fact that larger choices of γ would shrink the inferred cross-history coefficients and thereby remove weaker GC effects, which would lead to reduced TDR (and FAR) performance for

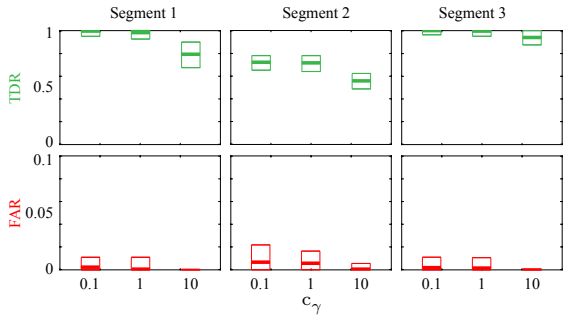

Fig. S6. Performance of the AGC inference for three different scalings γ for $c_\gamma \in \{0.1, 1, 10\}$, in terms of TDR (top row) and FAR (bottom row).

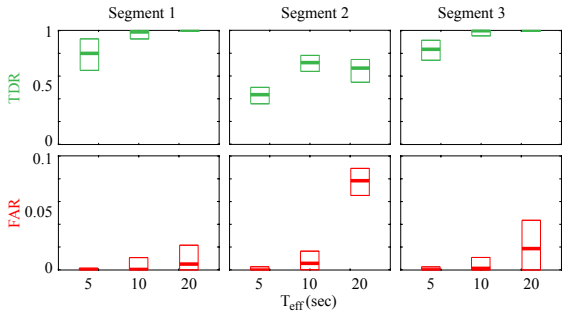

Fig. S7. Performance of the AGC inference for three different values of $T_{\text{eff}} \in \{5, 10, 20\}$ s, in terms of TDR (top row) and FAR (bottom row).

all segments. The optimally-tuned choice of the regularization parameter $\gamma = \gamma^*$ obtained via cross-validation achieves a favorable TDR-FAR performance trade-off.

Robustness to the choice of T_{eff}**.** For the effective filtering length, we select three different settings of $T_{\text{eff}} \in \{5, 10, 20\}$ s. Fig. S7 exhibits the significant influence of effective filtering length T_{eff} on the performance of AGC inference, where as expected larger choices of T_{eff} would increase both the TDR and FAR metrics. In other words, larger effective number of samples for GC inference at each time step would increase both the capability of correct identification (due to increased estimation accuracy of the existing links) and the risk of false detection (due to increased effective observation noise).

8. Inspecting the Roles of Adaptive Sparse Estimation and Bias Correction in AGC Inference

In this section, we inspect the roles of the bias correction procedure as well as sparse estimation in our proposed AGC inference method using an illustrative simulation study. We examine how these features affect the performance in terms of correct identification of the GC links and avoiding false positives. To this end, we compare the performance of our AGC inference method with a variant in which the bias correction stage is removed, as well as the conventional static ML-based GC inference (21), in which the dynamics and sparsity are not taken into account.

We consider $R = 100$ realizations of a random network configuration comprising $N_c = 10$ neurons, causally interacting through $N_{\text{Links}} = 10$ randomly selected directional links. For each repetition and given the network configuration, a sequence of spike trains with a duration of $\mathcal{T} = 30$ s is generated with a bin size of $\Delta = 1$ ms. The average baseline spiking probability is set to $\bar{\lambda}\Delta = 0.05$. For spike generation, we use a logit-linked GLM model with a static block-sparse parameter vector $\boldsymbol{\omega}^{(\tilde{c},c)}$ with a support set of $S = \{1, 5, 10\}$, and respective values of $(\boldsymbol{\omega}^{(\tilde{c},c)})_S = \{2, -1, 1\}$ to model the self- and cross-history dependencies among neurons. Each history component is associated with a non-overlapping history window of $W_H = 5$ time bins. The sign of the history kernel determines the aggregate excitatory or inhibitory effect of a causal link. We assume self-excitatory behavior for all neurons. For GLM estimation, we select $M_H^{\text{cross}} = M_H^{\text{self}} = 10$ history components, associated with spike counting windows of length 5 time bins. For the ℓ_1–PPF$_1$ algorithm, we select the forgetting factor $\beta = 0.999$, and a filtering window size of $W = 5$ bins (corresponding to an effective window length of 5 s), $c_w = 1$ and $L = 1$ number of iterations. The regularization parameter γ is tuned for each neuron via two-fold even-odd cross-validation. The χ^2 filtering and smoothing algorithm parameters are chosen as $\sigma_e^2 = 5 \times 10^{-6}$ and $\rho = 1$, and the FDR is controlled at a significance level of $\alpha = 0.1$.

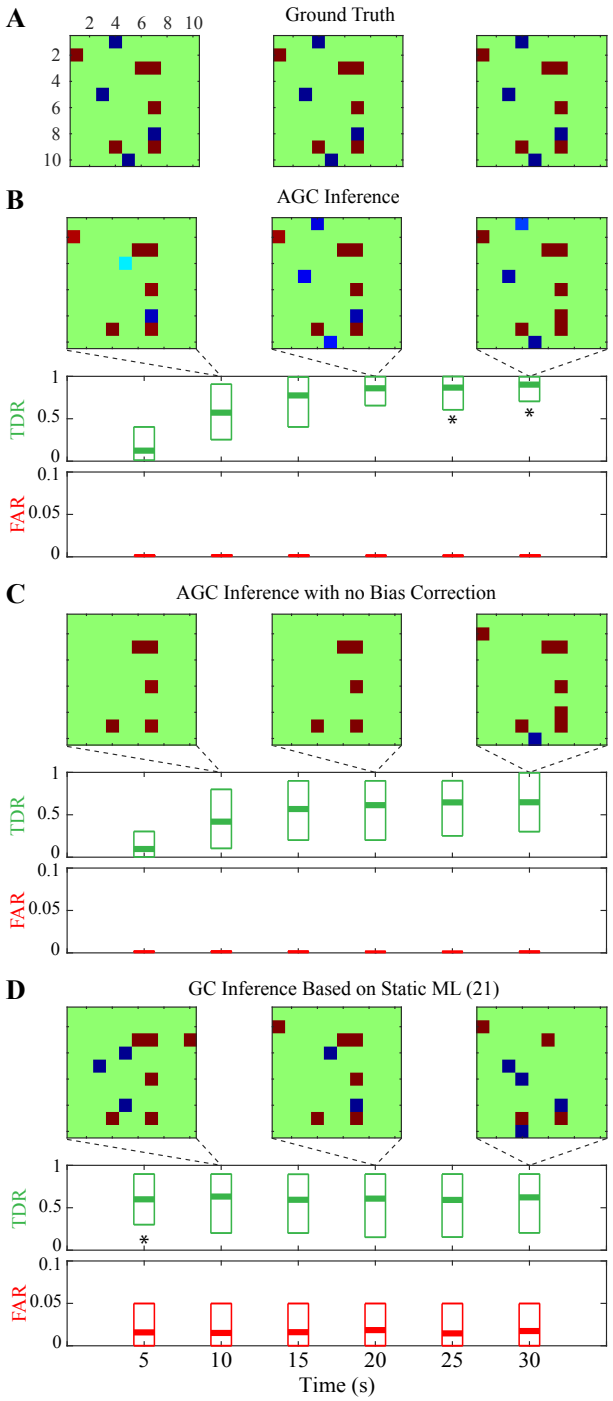

Fig. S8. Performance comparison of AGC inference to its variant without bias correction and GC inference based on static ML. (A) ground truth GC maps in a network of 20 neurons, (B) AGC inference, (C) AGC inference without bias correction, and (D) GC inference based on static ML (21). For the subplots B, C and D, the top rows correspond to three snapshots of the network inference result for a given realization, and the bottom rows show the TDR and FAR curves computed based on 100 realizations, for the six non-overlapping segments. Stars indicate significant differences between the AGC and static ML, with effect sizes of $r \geq 0.8$ (Wilcoxon signed-rank test, $p < 0.001$).

Fig. S8-A shows the static ground truth GC maps for a selected realization, plotted at three time instances in the form of 10×10 matrices. Fig. S8-B, C and D show the results for the full AGC inference method, its variant with no bias

correction, and the ML-based static GC method, respectively. Each panel consists of three rows: snapshots of the detected GC maps for one realization (top row), and the TDR (second row) and FAR (third row) performance metrics for consecutive non-overlapping 5 s windows, pooled across the $R = 100$ repetitions. Boxes show the mean and 90% confidence regions. The static GC maps in Fig. S8-D are estimated using non-overlapping windows of length $W_{\text{eff}} = 5$ s, equal to the effective filtering block length of the AGC method.

Figs. S8-B and C reveal the favorable FAR performance of the AGC method as compared to the static ML, even in the absence of bias correction. In particular, based on the Wilcoxon signed-rank test with a p-value of $p < 0.001$, the FAR performance of AGC inference is significantly lower than that of the static ML for all segments (effect sizes of $r = 0.44, 0.38, 0.44, 0.52, 0.37$, and 0.37 in the six segments, respectively). However, the lack of bias correction (Fig. S8–C) results in lower TDR performance compared to the AGC method. The TDR performance of the AGC method is also significantly higher than that of the static ML for the last 4 segments, but is outperformed by the static ML in the first segment (effect sizes of $r = 1, 0.17, 0.73, 0.74, 0.88$ and 0.90, in the six segments, respectively), which is expected due to the initialization period of ~ 5 s for the AGC method.

This illustrative example shows that the static ML-based approach that does not account for sparsity overfits the parameters when applied to limited data, and hence results in low true positive performance and a high number of spurious link detections. In comparison, the AGC method provides favorable TDR and FAR performance, but only after the initialization period, which is of the order of the effective window length. In addition, this example highlights the crucial role of bias correction for the deviance difference statistics in our proposed statistical inference procedure.

9. Robustness Against Latent Confounding Causal Effects: Three Simulation Studies

In this section, we will inspect the robustness of our proposed AGC inference method with respect to the problem of *latent confounding causal effects*, which is one of the major challenges in GC inference. When the two data time series X_t and Y_t subject to G-causal inference are driven by a third latent common process Z_t, with possibly different latencies, i.e., $(X_t \leftarrow Z_t \rightarrow Y_t)$, the GC inference may lead to spurious detection of causal effects between X_t and Y_t. This is due to the fact that the common information from Z_t is pronounced into both time series X_t and Y_t, and cannot be captured by the conditional covariates due to the latent nature of Z_t, and may result in false positive errors, thereby limiting the reliability of GC inference.

Although the original form of the GC measure does not take into account the latent confounding causal effects, several solutions have been proposed in the literature to resolve this issue. As an example, a variant of GC called "partial G-causality" is introduced in (24), which shows superior performance in terms of removing the effects of unknown confounding influences compared to the conditional GC.

Our proposed method for AGC inference mitigates this issue through several mechanisms. First, the hypothesis of sparsity allows for stable estimation of high order GLM models, with large number of self-history components in both the reduced and full models used in the conditional GC measure. Hence, we expect that the latent effects are captured via the high order self-history parameters due to the autoregressive nature of the GLM models, which promotes the detection of the actual GC links between the units using the cross-history components. This feature is akin to estimating latent Moving Average (MA) components using autoregressive models in the ARMA modeling paradigm.

Second, by explicitly modeling the dynamics of the non-centrality parameters describing the deviance statistics, and thereby using the χ^2 filtering and smoothing algorithm to reliably estimate them, we expect that only the temporally-salient G-causal effects are captured, and the transient G-causal links possibly due to confounding influences manifested in the deviance statistics are suppressed.

Third, the non-centrality parameter estimates allow us to characterize the test strengths for the rejected nulls (i.e., detected GC interactions) obtained by the FDR-controlled multiple hypothesis testing framework, in a model-based fashion. The resulting J-statistics can be further used to reject the detected GC links with low test strength, which may be due to transient latent effects.

In order to demonstrate these features, we test the performance of our proposed method for GC inference in the presence of confounding causal effects under three scenarios: 1) confounding deterministic common input, 2) confounding stochastic common input, and 3) confounding effects due to network subsampling.

Scenario 1: Confounding Effects Due to Latent Deterministic Common Input. We first consider an illustrative two-neuron example. We consider a setting with a GC-link from neuron (2) to (1), and no GC link in the opposite direction. We also consider a hidden (confounding) source (H) affecting both neurons. We assess the robustness of our algorithm in terms of two performance metrics: the detected false alarm rate (FAR) corresponding to the link $(1 \mapsto 2)$, and the true detection rate (TDR) for correctly identifying the link $(2 \mapsto 1)$, all in the presence of the confounding source $(H \mapsto 1, 2)$ (Fig. S9–A).

We assume a stationary environment with static GC links, and we use the same spiking statistics as in the previous simulation setting based on a logit-linked GLM model. We consider the case with no self-history dependence in order to more specifically inspect the trade-off between the cross-history and the latent confounding influences on our GC inference procedure. To model the cross-history dependence associated with the GC link $(2 \mapsto 1)$, we select a uniform modulation vector $\boldsymbol{\omega}_k^{(1,2)} = \frac{1}{\sqrt{W^{(1,2)}}} \mathbf{1}_{W^{(1,2)}}$ covering a window of $W^{(1,2)}$ time bins, where $\mathbf{1}_{W^{(1,2)}}$ denotes the vector of all ones of length $W^{(1,2)}$. The effect of the latent hidden source is later added to the contributing effects in the GLM models for both neurons. For the estimation of the GLM models, a larger number of $M_H^{\text{self}} = 5 \times M_H^{\text{cross}}$ self-history components are considered compared to the cross-history in order to capture the effects of the latent confounding influences.

In this first scenario, a sinusoidal signal $x_k = A_H \sin(2\pi k/200)$ is afferent to neurons (1) and (2) as the latent common input with different delays. We consider a phase difference of $\pi/2$ between the latent inputs to neurons (1) and (2) to account for the delay. For simulation of this scenario, we select a cross-history window of $W^{(1,2)} = 100$, and a uniform non-overlapping spike counting window of length

$\bar{\lambda}\Delta$ \ $\mathcal{E}$		0.01	0.05
0.01	FAR	0.01 ± 0.05	0.13 ± 0.20
	TDR	0.66 ± 0.32	1
0.05	FAR	0.11 ± 0.11	0.15 ± 0.10
	TDR	0.89 ± 0.11	1
0.1	FAR	0.15 ± 0.09	0.12 ± 0.06
	TDR	0.90 ± 0.08	1

Table S1. Performance metrics of AGC inference in presence of a latent deterministic sinusoidal common input. Entries show mean +/- standard deviation.

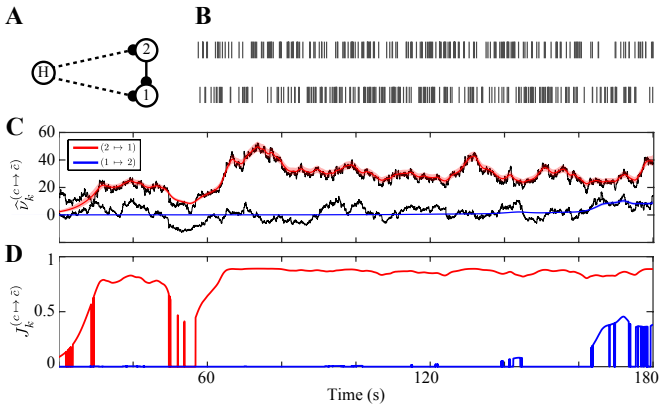

Fig. S9. The performance of the proposed GC inference method in presence of the latent confounding causal effects corresponding to the realization from Table S1 with median performance metric pair (FAR, TDR) at the setting $(\bar{\lambda}\Delta, \mathcal{E}_{CS}) = (0.05, 0.01)$. A) dual-neuron network model with hidden source, B) spike trains of both neurons within 4 s window, C) estimated non-centrality $\widehat{\nu}_k$ corresponding to GC links $(1 \mapsto 2)$ (blue) and $(2 \mapsto 1)$ (red) across time, along with the 95% confidence regions, and the shifted deviance differences $D_{k,\beta} - M^{(d)}$ (black traces), D) estimated J-statistics J_k for both GC links obtained via statistical inference procedure based on BY rejection.

$W_H = 50$ for parameterizing the history components, and $M_H^{\text{cross}} = 20$ number of cross-history components.

For fairness of comparison, we choose the mean power $\mathcal{E}_H$ of the latent confounding source to be equal to the mean power of the G-causal link $\mathcal{E}_{GC} := \text{var}(\boldsymbol{\omega}_k^{(1,2)T} \mathbf{x}_k^{(1,2)})$ for both scenarios, and denote them by $\mathcal{E}$. We run the simulation for $R = 50$ repetitions, where a spike train of $K = 180k$ samples covering a duration of $\mathcal{T} = 180$ s is generated with $\Delta = 1$ ms time bins for each realization. For the ℓ_1–PPF$_1$ sparse filtering setup, we chose an effective block length of $N_{\text{eff}} = \frac{W}{1-\beta}$ in the set $\{10k, 20k, 100k\}$, with respective average spiking probabilities of $\bar{\lambda}\Delta \in \{0.1, 0.05, 0.01\}$. The regularization parameter γ is tuned for both neurons via two-fold even-odd cross-validation. For the χ^2 filtering and smoothing setup, we selected a scaling factor $\rho = 1$, and a smoothing factor $\sigma_e^2 = 10^{-4}$. We infer the GC links for each run, and finally measure the mean FAR and TDR across all realizations.

Table S1 exhibits the (FAR, TDR) performance pairs for six different settings of $(\bar{\lambda}\Delta, \mathcal{E})$ for the sinusoidal latent source, pooled across the 50 repetitions. Each row and column correspond to specific choices of the average spiking probability $\bar{\lambda}\Delta$ and mean confounding power $\mathcal{E}_H$, respectively. The effective number of spikes per filtering window $n_{\text{eff}} := N_{\text{eff}}\bar{\lambda}\Delta$ is chosen to be the same across all rows. We selected two different values of $\mathcal{E} \in \{0.01, 0.05\}$. The FDR is controlled at the respective rates of $\alpha = 0.1$ and 0.05 for $\mathcal{E} = 0.01$ and 0.05. The respective small and large values of FARs and TDRs

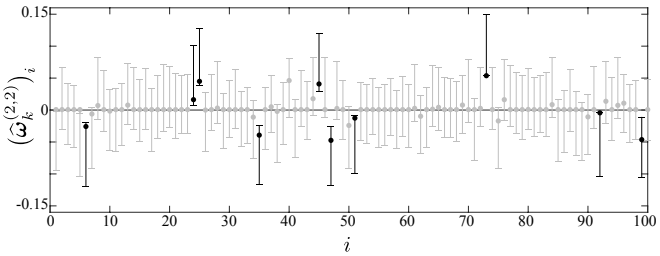

Fig. S10. A sample of the self-history coefficient of neuron (2) in the latent sinusoidal common input scenario for a generic trial and time window. The error bars show 90% confidence intervals. Coefficients that are significantly away from zero are highlighted in black.

in the entries of Table S1 reveal the utility of our proposed method in suppressing the effect of confounding latent causal influences, while identifying the true G-causal links between the two neurons with high sensitivity and specificity. In addition, they suggest that high-order self-history components are capable of capturing both deterministic and stochastic latent effects. It is worth mentioning that the six different settings in Table S1 are chosen to span the low-spiking ($\bar{\lambda}\Delta = 0.01$) and high-spiking ($\bar{\lambda}\Delta = 0.1$) regimes, both in presence of weak ($\mathcal{E} = 0.01$) and strong ($\mathcal{E} = 0.05$) confounding effects.

In order to illustrate the aforementioned features of our proposed method in detecting the salient effects, and characterizing the corresponding test powers, we have shown one realization from Table S1 in Fig. S9, corresponding to the setting $(\bar{\lambda}\Delta, \mathcal{E}) = (0.05, 0.01)$. The corresponding spike trains of neurons (1) and (2) within a small window of 4 s are shown in Fig. S9-B. Fig. S9-C shows the time-course of the estimated non-centrality parameter $\widehat{\nu}_k^{(1\mapsto 2)}$ associated with the false positive error $(1 \mapsto 2)$ (blue trace), and $\widehat{\nu}_k^{(2\mapsto 1)}$ associated with the true positive excitatory G-causal link $(2 \mapsto 1)$ (red trace). The black traces show the shifted deviance differences $D_{k,\beta} - M^{(d)}$. Fig. S9-D shows the time-course of the estimated J-statistics corresponding to the existing GC link $(2 \mapsto 1)$ (red trace) and the non-existing $(1 \mapsto 2)$ link (blue). As expected, the existing GC link is detected in a temporally-salient fashion with high test strength, whereas the non-existing link is overwhelmingly rejected, with low test strength otherwise. In order to highlight the effect of capturing the latent input using long self-history kernels, a sample of the self-history coefficients of neuron (2) in the sinusoidal latent input scenario are shown in Fig. S10. The coefficients that are away from zero at a significance level of 90% are highlighted in black. The estimated sparse high-order self-history components are able to capture the sinusoidal latent input based on the temporal correlations in the spiking history of the neuron.

Scenario 2: Confounding Effects Due to Latent Stochastic Common Input. For the second scenario, we consider a similar setting as the previous one, but generate a high-order AR process to model a general stochastic latent confounding effect. We use a block-sparse structure for the AR kernel with parameters $\boldsymbol{\omega}_H = [0.7, 0, 0, 0, -0.05, 0, 0, 0, 0.02]'$, where each coefficient is associated with a non-overlapping spike counting window of length $W_H = 25$ bins. The AR coefficients are normalized to result in a stable process. We selected an arbitrary delay of 40 bins between the common input to the two neurons. A cross-history window of length $W^{(1,2)} = 50$, and a spike counting window of length $W_H = 25$, and $M_H^{\text{cross}} = 10$ number of cross-history components are selected for this setting. All

$\bar{\lambda}\Delta$ \ $\mathcal{E}$		0.01	0.05
0.01	FAR	0.07 ± 0.18	0.05 ± 0.16
	TDR	0.81 ± 0.29	1
0.05	FAR	0.10 ± 0.09	0.06 ± 0.07
	TDR	0.92 ± 0.11	1
0.1	FAR	0.09 ± 0.06	0.06 ± 0.05
	TDR	0.93 ± 0.05	1

Table S2. Performance metrics of AGC inference in presence of a latent stochastic AR common input. Entries show mean +/- standard deviation.

$\bar{\lambda}\Delta$ \ $\mathcal{E}$		0.01	0.03
0.01	FAR	0.01 ± 0.03	0.04 ± 0.07
	TDR	0.74 ± 0.27	0.99 ± 0.01
0.05	FAR	0.01 ± 0.01	0.01 ± 0.02
	TDR	0.71 ± 0.14	1
0.1	FAR	0.01 ± 0.01	0.01 ± 0.01
	TDR	0.68 ± 0.11	1

Table S3. Performance metrics of AGC inference under network subsampling. Entries show mean +/- standard deviation.

the other parameters used for AGC inference are chosen the same as in the previous scenario.

In the same vein as Table S1, Table S2 exhibits the (FAR, TDR) performance pairs for six different settings of $(\bar{\lambda}\Delta, \mathcal{E})$ for the AR latent source. Similarly, the respective small and large values of FARs and TDRs in the entries of Table S2 confirm the utility of our proposed method in suppressing the effect of confounding latent causal influences.

Scenario 3: Confounding Effects due to Network Subsampling. In the third scenario, we evaluate the performance of our proposed AGC inference method in the context of the more general confounding setting of *network subsampling*. This scenario occurs when the observable neurons are subsampled from a large neuronal network, and are prone to significant confounding effects from the unobserved portion of the network (Fig. S11). This scenario often happens in the analysis of experimentally recorded data, in which the observable neuronal ensemble consists of a small subset of a larger latent network of neurons, due to the physical limitations of data acquisition.

In order to test the robustness of our method to the problem of network subsampling, we consider a network of $N_c = 20$ neurons, where the AGC inference is performed on a small subnetwork of $N_{cs} = 3$ observable neurons. We repeat the network subsampling simulation for $R = 100$ realizations, where a random network configuration consisting of $N_{Links} = 40$ links randomly selected out of 380 possible directional links is considered for each realization. To determine the observable ensemble for AGC inference, we randomly select a subset of N_{cs} neurons, such that there would be at least one direct latent common input to a pair of causally-linked observable neurons (e.g. neurons 1 and 2 in Fig. S11). For simulation, we consider a static setting for the GC links, where the underlying parameters remain constant throughout the entire duration. We use a block-sparse kernel of $\boldsymbol{\omega}_H = [2, -1, 0, 0, -0.5, 0, 0, 0, 0, -0.5]'$ with non-overlapping history windows of length $W_H = 5$ bins, to model the self- and cross-history dependence among the causally interacting neurons. The effective excitatory or inhibitory natures of the GC links are determined by positive ($\boldsymbol{\omega}^{(\tilde{c},c)} = +\boldsymbol{\omega}_H$) or negative polarity ($\boldsymbol{\omega}^{(\tilde{c},c)} = -\boldsymbol{\omega}_H$) of the kernel. We assume the self-history dependence to be of excitatory nature for all neurons, and the probability of excitatory or inhibitory cross-history dependence is set to 50% for all links.

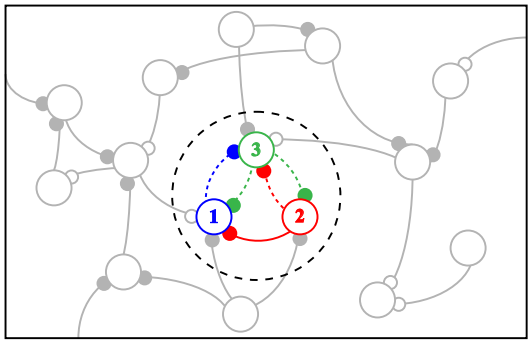

Fig. S11. A schematic depiction of the network subsampling scenario. A small observable subnetwork of three neurons (within the dashed circle) are sampled from a large latent neuronal network. The observable subnetwork and the interactions within are represented by blue, red and green colors, while the latent neurons and interactions are shown in gray.

For each realization, we generate spike trains with a total duration of $\mathcal{T} = 180$ s with $\Delta = 1$ ms time bins. For estimation of the GLM models, a total number of $M_H^{\text{cross}} = 10$ cross-history and $M_H^{\text{self}} = 20$ self-history components are considered with a spike counting window of length 5 for parameterizing the history components.

We repeat the network subsampling simulation and perform the AGC inference for six different settings of $(\bar{\lambda}\Delta, \mathcal{E})$ pairs, similar to the Tables S1 and S2, where two different values of the mean GC link power $\mathcal{E} \in \{0.01, 0.03\}$ and three different values of the average spiking probability $\bar{\lambda}\Delta \in \{0.01, 0.05, 0.1\}$ are selected. We use the same parameter settings for the ℓ_1–PPF$_1$ filter and the non-central χ^2 filtering and smoothing as in the previous two scenarios for the three different $\bar{\lambda}\Delta$ settings. The regularization parameter γ is tuned for each observable neuron separately via two-fold even-odd cross-validation. The FDR is controlled at a significance level of $\alpha = 0.1$.

As before, we evaluate the performance of AGC inference in terms of two performance metrics: FAR and TDR within the observable network across all realizations. Table S3 summarizes the performance results for the six different settings. The resulting metrics reveal the favorable performance of our proposed AGC inference in suppressing the false positives due to the latent confounding causal effects (low FAR rate of $\sim 1\%$, on average), while maintaining high true detection rates (high TDR rate of $\sim 70\%$, on average). Together with the results of the two foregoing scenarios, these results corroborate our earlier assessment of the AGC inference in maintaining a degree of immunity to latent confounding effects.

10. Cross-history Coefficient Dynamics of the Top-down and Bottom-up Links in the Ferret A1-PFC Analysis

In this section, we examine the dynamics of the cross-history coefficients involved in the extracted top-down and bottom-up GC links in the ferret A1-PFC interaction during active behavior (See Fig. 6). Recall that two of the major findings

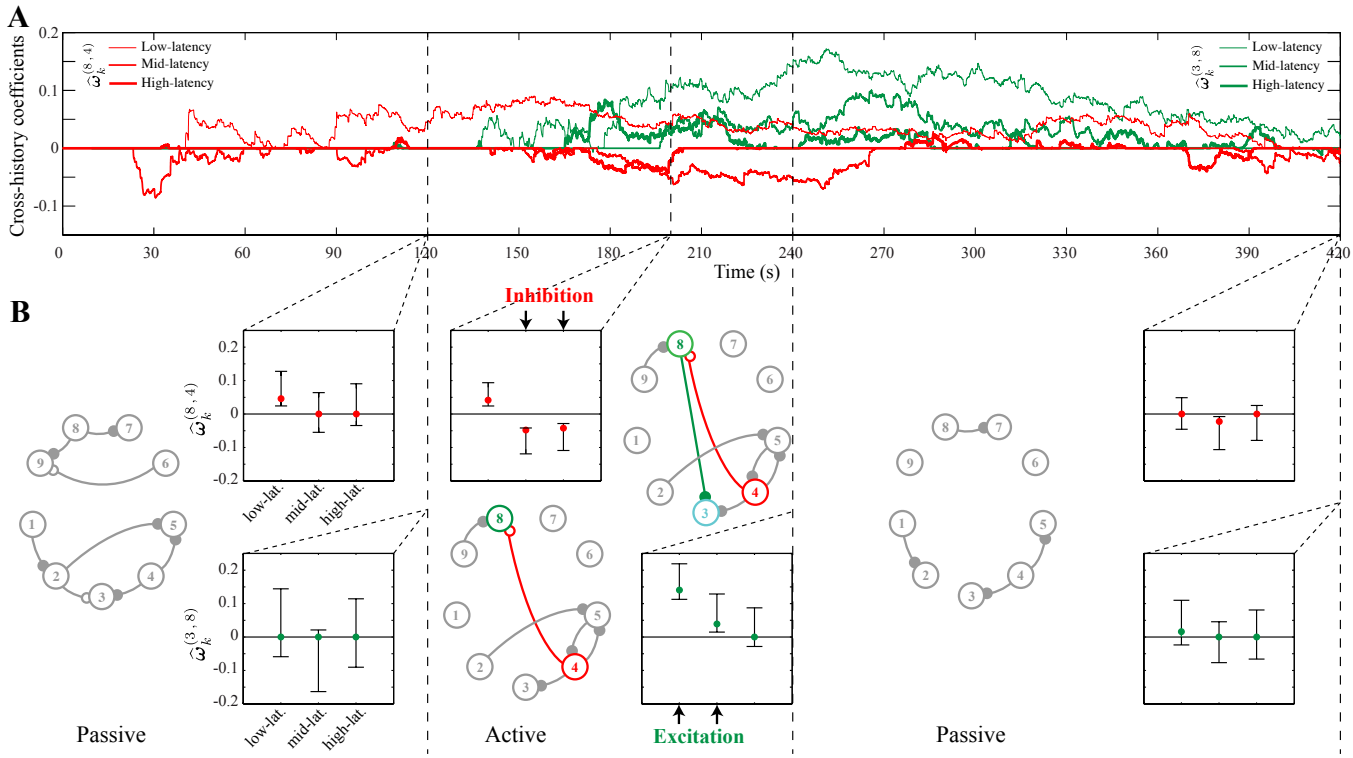

Fig. S12. Dynamics of the cross-history coefficients for the bottom-up ($4 \mapsto 8$) and the top-down ($8 \mapsto 3$) links from the simultaneous PFC and A1 recordings in the pure tone detection task with target 2.5 kHz (See Fig. 6). A) the low-, mid- and high-latency components of each set of coefficients are distinguished by their line widths, where red and green traces correspond to the bottom-up and top-down link, respectively. B) Bar plots of the cross-history coefficients and their 90% confidence intervals at times indicated by the dashed vertical lines. Each panel also depicts the inferred AGC network. Downward and upward vertical arrows in the middle panel highlight the significant changes in the coefficients.

of this analysis were: 1) emergence of a bottom-up inhibitory link from unit 4 in A1 to 8 in PFC, followed by 2) a top-down excitatory link from unit 8 in PFC to 3 in A1. The latter effect resulted in the disappearance of the frequency selectivity of unit 3 which was originally sharply tuned to $f = 8$ kHz. In addition, unit 4 which affects unit 8 is sharply tuned to the target frequency of $f = 2.5$ kHz.

In order to gain insight into the nature of these influences, we examine the time-course of the estimated underlying cross-history coefficients. Fig. S12-A shows the time-course of the cross-history coefficients $\widehat{\boldsymbol{\omega}}_k^{(8,4)}$ (red traces) and $\widehat{\boldsymbol{\omega}}_k^{(3,8)}$ (green traces) corresponding to the bottom-up and top-down links, respectively. As mentioned in Section 5, the cross-history coefficients consists of three components: a low-latency component corresponding to a cross-history window of 20 ms, a mid-latency component corresponding to a cross-history window of 30 ms, and a high-latency component corresponding to a cross-history window of 50 ms, which cover an overall cross-history window of 100 ms. The low-, mid- and high-latency components are distinguished by their line width in Fig. S12-A, as indicated in the figure legend.

Consistent with our AGC inference results of Fig. 6, these cross-history coefficients undergo major changes shortly after the onset of the active segment, some of which persist throughout a considerable portion of the post-active passive segment. Note that the observed delay of order ~ 40 s in adaptive parameter estimation is consistent with the choice of effective window length $\frac{W}{1-\beta}$ for $W = 8$ and $\beta = 0.9998$.

In order to dissect these dynamics more carefully, we have plotted three snapshots of these coefficients together with their 90% confidence intervals in Fig. S12-B. The confidence intervals are obtained based on the de-biasing procedure to account for the bias of the adaptive ℓ_1-regularized ML estimates (1, 2). Note that, unlike the conventional unbiased Gaussian case, the confidence intervals are not evenly centered around the estimates, which highlights the effect of bias correction. The confidence intervals are not shown in Fig. S12-A for graphical clarity. Each panel also shows the inferred AGC network from Fig. 6, in which the units not involved in the top-down and bottom-up GC links are grayed out for graphical simplicity.

The left panel shows that during the first passive task, most of the cross-history coefficients are insignificant, which is also reflected in the absence of any cross-region link in the inferred AGC network. The middle panel reveals the emergence of low-latency excitation together with strong mid- and high-latency inhibition from unit 4 to 8 (indicated by downward arrows), hence the overall inhibitory bottom-up GC link. Similarly, the strong low- and mid-latency excitation from unit 8 to 3 (indicated by upward arrows) results in the top-down excitatory GC link. The latter excitation locks the activity of unit 3 to that of unit 8, and as a result the high frequency responsiveness of unit 3 is suppressed. Finally, the right panel shows that the cross-history coefficients return to the original setting of the pre-active condition.

As mentioned in the discussion following Fig. 6, the fluctuations of the J-statistics (e.g., red trace in Fig. 6-B, panel 8) are due to the FDR correction procedure, which results in rejecting the null hypotheses only corresponding to links with strong enough coefficients at a given time step. Therefore, the stochastic fluctuations of the cross-history coefficients (e.g., red traces in Fig. S12-A) lead to the fluctuations of the deviance statistics around the statistical thresholds set by the FDR control procedure in our multiple hypothesis testing framework.

11. Assessing the Reliability of the AGC Inference Results from the Ferret A1-PFC Experiment via Surrogate Data Analysis

Given the lack of access to ground truth in the analysis of real data, it is crucial to assess the reliability of our results using carefully devised surrogate data. To this end, we generate two sets of data using random shuffling and network subsampling procedures, and thereby evaluate the consistency of our results.

Analysis of Surrogate Data from Random Shuffling. We first assess the reliability of the inferred AGC networks in the analysis of the ferret A1-PFC interaction through surrogate data obtained by random shuffling. To this end, we randomly shuffle the activity of single-units across different repetitions (14 repetitions in total), such that each repetition of a single unit would be randomly aligned with different repetitions of other single units recorded at different experimental periods. We then infer the AGC network patterns for each shuffled composition of the repetitions. Our goal is to investigate whether our AGC inference procedure detects any significant GC pattern from the shuffled data.

We repeat the random shuffling procedure for $R = 100$ trials, and compute the J-statistics for different links across the whole experiment. We test the reliability of the detected significant links from the original unshuffled data by comparing their J-statistics to those pooled from the randomly shuffled surrogate data. For brevity, we focus on two of the most notable GC links: the top-down $(8 \mapsto 3)$ link from PCF to A1 and the bottom-up $(4 \mapsto 8)$ link from A1 to PFC.

Fig. S13 shows the time course of the J-statistics for these two representative GC links inferred from both the original (red and green traces) and surrogate (gray traces) data. In each panel, the black solid trace represents the mean J-statistics across the $R = 100$ randomly shuffled repetitions, and the colored hulls indicate the corresponding 95% confidence regions. It can be observed that the mean J-statistics from the surrogate data do not surpass the small value of 0.06, while the originally detected J-statistics take large values in the range of $\in (0.7, 1)$. For instance, the value of $J_k^{(8\mapsto 3)}$ at $k = 300$ s is significantly higher than those from the surrogate data (One-tailed Z-test, $p < 0.0001$).

Moreover, the J-statistics of the surrogate data do not suggest any task-dependent behavior, as opposed to those from the original data. To illustrate this more precisely, suppose that the task-dependence behavior of the link $(8 \mapsto 3)$ were to be preserved in the surrogate data, i.e., this link would persist for blocks comparable in length to that of the original data. Given that this link is active with significant J-statistics for $\sim$ 120s, then it would be expected that the average J-statistics of this link for the surrogate data would be close to $120/420 \approx 0.28$. However, the p-value of this observation with respect to the distribution of the J-statistics over the entire duration of the surrogate data is given by $p = 0.0007$ (One-tailed Z-test).

This analysis verifies that the highly significant AGC links inferred from the data vanish under random shuffling of the repetitions, and are therefore highly specific to the correct temporal ordering of the repetitions in the experiment.

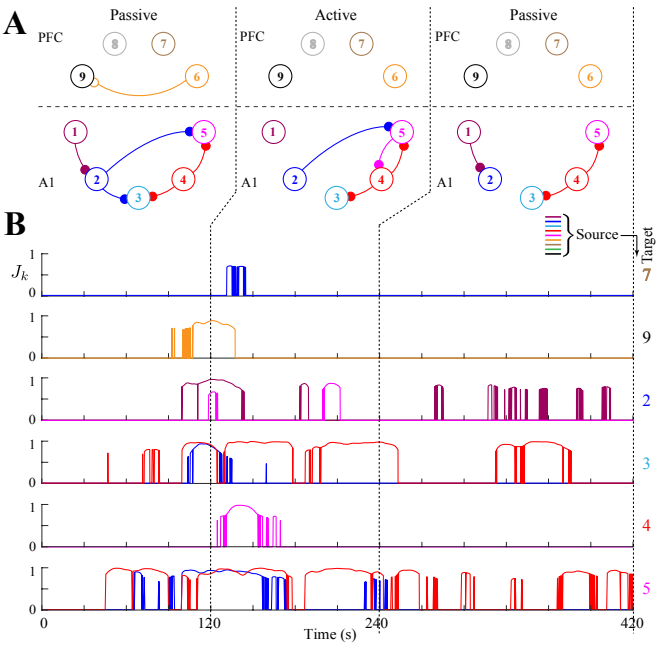

Fig. S14. Analysis of surrogate data from network subsampling in the ferret A1-PFC experiment, where unit 8 (shown in gray) is excluded from the analysis. As expected, the significant bottom-up and top-down links between A1 and PFC vanish.

Analysis of Surrogate Data from Network Subsampling. Next, we assess the reliability of the inferred AGC interactions in the analysis of the ferret A1-PFC interaction through surrogate data obtained by network subsampling. To this end, we investigate the robustness of the inferred AGC patterns and their time course against excluding a single or a group of neurons from the observed ensemble. For brevity, we focus on the two bottom-up and top-down AGC links and assess their reliability under three different network subsampling scenarios:

Scenario 1. We first exclude the single-unit 8, the only unit in PFC with a significant GC link to A1, from the analysis. We explore the presence of any possible new inter-region GC interactions, and expect that the top-down and bottom-up GC links between PFC and A1 would vanish due to the exclusion of unit 8. Fig. S14 shows the resulting AGC network maps

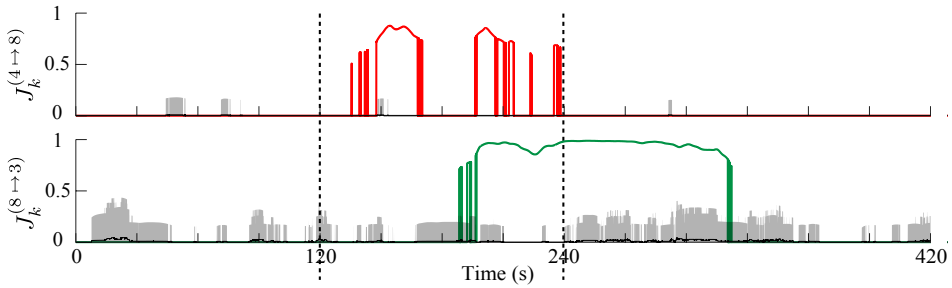

Fig. S13. Analysis of surrogate data from random shuffling of the repetitions in the ferret A1-PFC experiment. The J-statistics of the $(4 \mapsto 8)$ and $(8 \mapsto 3)$ links inferred from the original data are shown in red and green traces, respectively. The average J-statistics obtained from the randomly shuffled ensemble are shown by black traces, with 95% confidence regions shown by the gray hulls. The J-statistics inferred from the original data show a significant statistical separation from those obtain from the surrogate data.

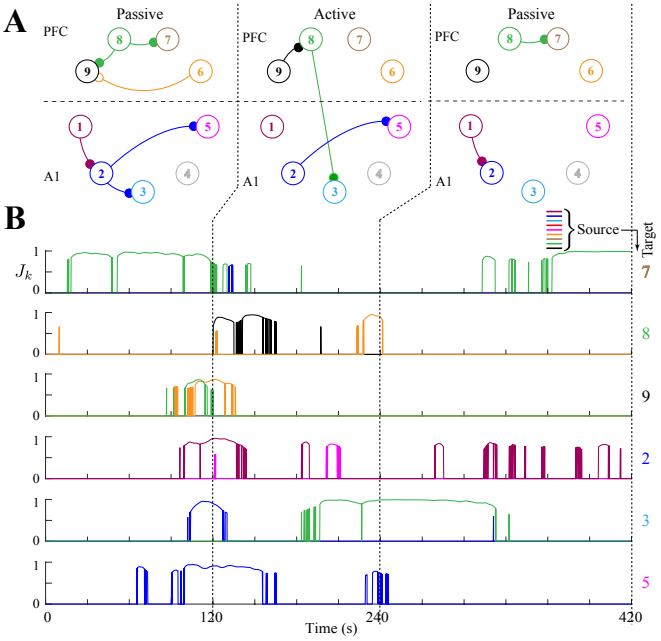

Fig. S15. Analysis of surrogate data from network subsampling in the ferret A1-PFC experiment, where unit 4 (shown in gray) is excluded from the analysis. As expected, the bottom-up link from A1 to PFC vanishes.

and the time courses of the corresponding J-statistics. Indeed, the significant bottom-up and top-down interactions between A1 and PFC vanish, while the rest of the networks within A1 and PFC remain unchanged. The only notable exception is a small transient link from 2 to 7 for $t \in [135, 140]$s.

Scenario 2. Next, we exclude the single-unit 4 in A1, with a bottom-up link to PFC, and test the robustness of our method in terms of the new detected GC links, and expect that no new bottom-up links from A1 to PFC are discovered. Fig. S15 shows the resulting AGC network maps and the time courses of the corresponding J-statistics. As expected, the bottom-up link from unit 4 to 8 vanishes, while the rest of the AGC interactions, notably the top-down link from 8 to 3, remain unchanged.

Scenario 3. Finally, we consider a highly undersampled case where we restrict the observable set to the three units $\{3, 4, 8\}$

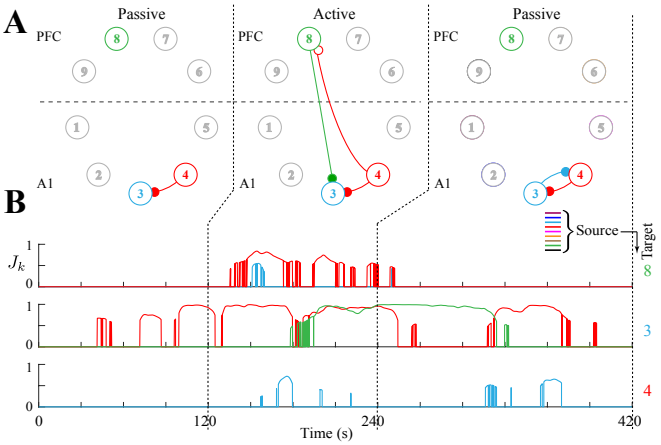

Fig. S16. Analysis of surrogate data from network subsampling in the ferret A1-PFC experiment, where only units 3, 4 and 8 are included in the analysis (all other units shown in gray). As expected, the significant bottom-up and top-down links and their respective time courses are preserved.

which are involved in the top-down and bottom-up interactions. We expect that the same bottom-up and top-down patterns between these units are discovered in the absence of all the other 6 neurons which did not exhibit any inter-region GC links. Fig. S16 shows the resulting AGC network maps and the time courses of the corresponding J-statistics. Indeed, the expected pattern of GC interaction between these three units is preserved, with the exception of a weak excitatory GC link from 3 to 4 with low statistical significance.

These results show that the inferred AGC maps and the times courses of the corresponding J-statistics, and notably those pertaining to the bottom-up and top-down network structure, are robust to network subsampling. Hence, they are specific to the interactions between the single-units under study in this experiment, and there is no evidence to believe that they are the byproduct of this particular observable subsampled network of 9 neurons.

12. Supporting Example: Ferret A1-PFC Interaction

In this section, we present the application of our proposed AGC inference on another instance of spike recordings from the same set of experiments on ferrets as described in the Applications section of the main manuscript, where the animal is performing a pure tone detection task (25).

Fig. S17 shows the results of our AGC inference for a selected experiment consisting of four main blocks: pre-active, active, and two post-active conditions. Each block is composed of 5 repetitions. Within each repetition, a complete set of 30 randomly permuted 1 sec-long TORCs was presented along with a randomized repetition of the target tone at $f = 8$ kHz. A total number of $C = 8$ single-units are detected through spike sorting (4 units in each region), whose spike trains are shown in Fig. S17-A. For graphical convenience, we only plotted the spike trains within the last repetition of each block. Fig. S17 shares the same structural format as Fig. 6. Fig. S17-B shows the time-course of the changes in the J-statistics associated with detected GC links, where each row represents the corresponding significant GC influences from all units to a target unit, which passed the BY FDR control procedure. Each single-unit along with its significant outgoing GC link is color-coded uniquely as shown on the right. Fig. S17-C depicts the detected changes in the pattern of G-causal links among the 8 single-units during three main blocks of the experiment. Three snapshots of the STRFs of all the four A1 units at the endpoints of the pre-active, active and post-active blocks are shown in Fig. S17-D, along with the target frequency $f = 8$ kHz indicated by a red arrow.

The total duration of $\mathcal{T} = 600$ s is binned by $\Delta = 1$ ms, and segmented by windows of length $W = 25$ bins. We applied the ℓ_1-PPF$_1$ adaptive filter to the spiking data of all single-units, where we selected a forgetting factor of $\beta = 0.9995$, a step size $\varsigma = \frac{1-\beta}{5W}$, $L = 20$ number of iterations per step, and regularization parameters tuned for each unit separately via two-fold even-odd cross validation. We consider the same dynamic GLM model to capture the spiking statistics as in the previous analysis, with the modulation coefficients accounting for both the ensemble spike history and stimuli. For the stimulus modulation, we consider a vectorized STRF array of size $I \times J$, with $I = 50$ time lags and $J = 50$ frequency bins in logarithmic scale represented by a Gaussian time-frequency dictionary (1), capturing the effect of the reference

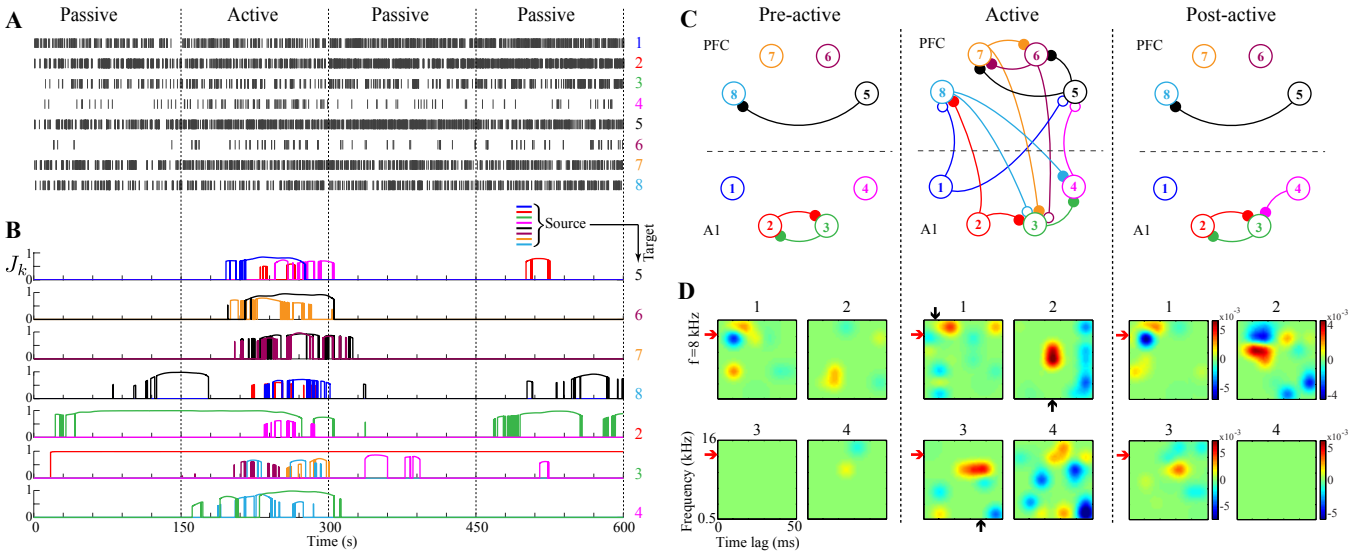

Fig. S17. Dynamic inference of GC links among single-units in the ferret PFC and A1 during a series of auditory tasks. (A) Sample of spike train data from eight cortical neurons (first 4 from A1 and second 4 from PFC) in passive listening and active task conditions, (B) inferred time-course of changes for the significant GC links through J-statistics, (C) inferred AGC maps during pre-task passive, active task, and post-task passive conditions. The excitatory and inhibitory nature of each GC link is represented by solid and hollow circles, respectively, D) snapshots of the STRF of A1 units at the endpoints of the three blocks of the experiment. Note the selective reduction in inhibition at 8 kHz (target tone frequency) in A1 cell 1 during behavior (downward arrow, middle panel). Adapted with permission from ref. 25.

acoustic stimuli spectrogram. As for the ensemble history dependence, we select $M_H^{\text{cross}} = 3$ cross-history and $M_H^{\text{self}} = 21$ self-history components associated with respective non-overlapping spike counting windows of $W_H^{\text{cross}} = [20, 30, 50]$ and $W_H^{\text{self}} = [20, 30, 50, \ldots, 50]$ bins. The FDR is controlled at the rate $\alpha = 0.1$, testing for $|\mathcal{C}| = 56$ possible GC links among the 8 single-units.

Fig. S17 reveals significant task-relevant changes in the pattern of G-causal interactions among the units within or across the PFC and A1 regions. The most striking observation is the identification of 4 bottom-up and 4 top-down GC links during active attentive behavior, which verifies the functional interaction (in the sense of Granger) between the higher-level PFC and the lower-level cortical region involved in active listening. The most significant and persistent bottom-up GC links, e.g. $(1 \mapsto 5)$, belong to the A1 unit 1, whose STRF characteristics show a frequency-selective suppression around the target frequency. As can be observed in Fig. S17-D, this A1 unit exhibits significant task-related plasticity (26), as its suppressive response to the target frequency vanishes entirely during the active attentive behavior (downward arrow, mid. panel, Fig. S17-D) while it G-causally influences the higher level PFC units in an inhibitory fashion. Interestingly, unit 1 retrieves its original pre-active STRF after the active task is over. In addition to the detected inter-region GC links, several instances of task-relevant changes in GC links within A1 (e.g., $3 \mapsto 2$; see upward arrows, mid. panel) or within PFC (e.g., $5 \mapsto 6$) occur during active behavior. In addition, the pattern of GC links within PFC changes dramatically during active attentive behavior as compared to the passive conditions.

13. Experimental Procedures

Surgery. Two hours before the mouse surgery, 0.1 cc dexamethasone (2 mg/ml, VetOne) was injected subcutaneously to reduce brain swelling during craniotomy. Anesthesia is induced with 4% isoflurane (Fluriso, VetOne) with a calibrated vaporizer (Matrx VIP 3000). During surgery, isoflurane level was reduced to and maintained at a level of 1.5%–2%. Body temperature of the animal is maintained at 36.0 degrees Celsius during surgery. Hair on top of head of the animal was removed using Hair Remover Face Cream (Nair), after which Betadine (Purdue Products) and 70% ethanol was applied sequentially 3 times to the surface of the skin before removing the skin. Soft tissues and muscles were removed to expose the skull. Then a custom designed 3D printed stainless headplate was mounted over left auditory cortex and secured with C&B-Metabond (Parkell). A craniotomy with a diameter of around 3.5 mm was then performed over left auditory cortex. A three layered cover slip was used as cranial window, which is made by gluing (NOA71, Norland Products) 2 pieces of 3 mm coverslips (64-0720 (CS-3R), Warner Instruments) with a 5 mm coverslip (64–0700 (CS-5R), Warner Instruments). Cranial window was quickly dabbed in kwik-sil (World Precision Instruments) before mounted 3 mm coverslips facing down onto the brain. After kwik-sil cured, Metabond was applied to secure the position of the cranial window. Synthetic Black Iron Oxide (Alpha Chemicals) was then applied to the hardened Metabond surface. 0.05 cc Cefazolin (1 gram/vial, West Ward Pharmaceuticals) was injected subcutaneously when entire procedure was finished. After the surgery the animal was kept warm under heat light for 30 minutes for recovery before returning to home cage. Medicated water (Sulfamethoxazole and Trimethoprim Oral Suspension, USP 200 mg/40 mg per 5 ml, Aurobindo Pharms USA; 6 ml solution diluted in 100 ml water) substitute normal drinking water for 7 days before any imaging was performed.

Awake two-photon imaging. Spontaneous activity data of population of layer 2/3 auditory cortex (A1) neurons is collected from adult (3-month old) Thy1-GCaMP6s female mouse implanted with chronic window following the above procedure, using two-photon imaging. Acquisition is performed using a two-photon microscope (Thorlabs Bscope 2) equipped with a Vision 2 Ti:Sapphire laser (Coherent), equipped with a

GaAsP photo detector module (Hamamatsu) and resonant scanners enabling faster high-resoluation scanning at 30–60 Hz per frame. The excitation wavelength was 920 nm. Regions ($\sim 300\ \mu m^2$) within A1 were scanned at 30 Hz through a 20x, 0.95 NA water-immersion objective (Olympus). During imaging the animal was head-fixed and awake. The microscope was rotated 45 degrees and placed over the left A1 where window was placed. An average image of field of view was generated by choosing a time window where minimum movement of the brain was observed and used as reference image for motion correction using TurboReg plugin in ImageJ. GCaMP6s positive cells are selected manually by placing a ring like ROI over each identified cell. Neuropil masks were generated by placing a 20 μm radius circular region over each cell yet excluding all cell soma regions. Traces of soma and neuropil were generated by averaging image intensity within respective masks at each time point. A ratio of 0.7 was used to correct for neuropil contamination. All procedures were approved by the University of Maryland Institutional Animal Care and Use Committee.

1. Sheikhattar A, Fritz JB, Shamma SA, Babadi B (2016) Recursive sparse point process regression with application to spectrotemporal receptive field plasticity analysis. *IEEE Trans. on Signal Processing* 64(8):2026–2039.
2. Van de Geer S, Bühlmann P, Ritov Y, Dezeure R (2014) On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Stat.* 42(3):1166–1202.
3. Haykin SS (1996) *Adaptive Filter Theory*. (Upper Saddle River, NJ, USA: Prentice Hall).
4. Kazemipour A, Wu M, Babadi B (2017) Robust estimation of self-exciting generalized linear models with application to neuronal modeling. *IEEE Trans. on Signal Processing* 65(14):3733–3748.
5. Davidson RR, Lever WE (1970) The limiting distribution of the likelihood ratio statistic under a class of local alternatives. *Sankhyā: The Indian Journal of Statistics, Series A* pp. 209–224.
6. Crowder MJ (1976) Maximum likelihood estimation for dependent observations. *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 45–53.
7. Billingsley P (2008) *Probability and measure*. (John Wiley & Sons).
8. Bickel PJ, Ritov Y, Ryden T (1998) Asymptotic normality of the maximum-likelihood estimator for general hidden markov models. *The Annals of Statistics* 26(4):1614–1635.
9. Douc R, Moulines E, Rydén T (2004) Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime. *The Annals of statistics* 32(5):2254–2304.
10. Wilks SS (1938) The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics* 9(1):60–62.
11. Wald A (1943) Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Trans. Amer. Math. Soc.* 54(3):426–482.
12. Tan W (1977) On the distribution of quadratic forms in normal random variables. *Canadian Journal of Statistics* 5(2):241–250.
13. Chipman JS, Rao M (1964) Projections, generalized inverses, and quadratic forms. *Journal of Mathematical Analysis and Applications* 9(1):1–11.
14. Smith A, Brown EN (2003) Estimating a state-space model from point process observations. *Neural Computation* 15(5):965–991.
15. Amos D (1974) Computation of modified Bessel functions and their ratios. *Mathematics of Computation* 28(125):239–251.
16. Baricz Á (2015) Bounds for Turánians of modified Bessel functions. *Expositiones Mathematicae* 33(2):223–251.
17. Shumway RH, Stoffer DS (1982) An approach to time series smoothing and forecasting using the EM algorithm. *Journal of time series analysis* 3(4):253–264.
18. Sheikhattar A, Fritz JB, Shamma SA, Babadi B (2015) Adaptive sparse logistic regression with application to neuronal plasticity analysis. *2015 49th Asilomar Conference on Signals, Systems and Computers* (IEEE, New York), pp. 1551–1555.
19. Babadi B, Kalouptsidis N, Tarokh V (2010) SPARLS: The sparse RLS algorithm. *IEEE Trans. on Signal Processing* 58(8):4013–4025.
20. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)* pp. 1–38.
21. Kim S, Putrino D, Ghosh S, Brown EN (2011) A Granger causality measure for point process models of ensemble neural spiking activity. *PLoS Comput Biol* 7(3):e1001110.
22. Okatan M, Wilson MA, Brown EN (2005) Analyzing functional connectivity using a network likelihood model of ensemble neural spiking activity. *Neural computation* 17(9):1927–1961.
23. Glantz SA (2012) *Primer of Biostatistics*. (McGraw-Hill Medical, New York).
24. Guo S, Seth AK, Kendrick KM, Zhou C, Feng J (2008) Partial Granger causality—eliminating exogenous inputs and latent variables. *J. Neurosci. Methods* 172(1):79–93.
25. Sheikhattar A, Miran S, Fritz JB, Shamma SA, Babadi B (2016) Probing the functional circuitry underlying auditory attention via dynamic Granger causality analysis. *2016 50th Asilomar Conference on Signals, Systems and Computers* (IEEE, New York), pp. 593–597.
26. Fritz J, Shamma S, Elhilali M, Klein D (2003) Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. *Nature neuroscience* 6(11):1216–1223.