Quantitative relations among causality measures with applications to pulse-output nonlinear network reconstruction

Zhong-qi K. Tian[1,2,3,a], Kai Chen[1,2,3,a], Songting Li[1,2,3,b], David W. McLaughlin[5,6,7,8,b], Douglas Zhou[1,2,3,4,b]

[1]*School of Mathematical Sciences, Shanghai Jiao Tong University, Shanghai 200240, China*
[2]*Institute of Natural Sciences, Shanghai Jiao Tong University, Shanghai 200240, China*
[3]*Ministry of Education Key Laboratory of Scientific and Engineering Computing, Shanghai Jiao Tong University, Shanghai 200240, China*
[4]*Shanghai Frontier Science Center of Modern Analysis, Shanghai Jiao Tong University, Shanghai 200240, China*
[5]*Courant Institute of Mathematical Sciences, New York University, New York, NY 10012*
[6]*Center for Neural Science, New York University, New York, NY 10012*
[7]*Institute of Mathematical Sciences, New York University Shanghai, Shanghai 200122, China*
[8]*Neuroscience Institute of New York University Langone Health, New York University, New York, NY 10016*

[a]*Z.K.T. and K.C. contributed equally to this work.*
[b]*Correspondence: songting@sjtu.edu.cn (S.L.), david.mclaughlin@nyu.edu (D.W.M.) and zdz@sjtu.edu.cn (D.Z.)*

Abstract

The causal connectivity of a network is often inferred to understand the network function. It is arguably acknowledged that the inferred causal connectivity relies on the causality measure one applies, and it may differ from the network's underlying structural connectivity. However, the interpretation of causal connectivity remains to be fully clarified, in particular, how causal connectivity depends on causality measures and how causal connectivity relates to structural connectivity. Here, we focus on nonlinear networks with pulse signals as measured output, *e.g.*, neural networks with spike output, and address the above issues based on four intensively utilized causality measures, *i.e.*, time-delayed correlation coefficient, time-delayed mutual information, Granger causality, and transfer entropy. We theoretically show how these causality measures are related to one another when applied to pulse signals. Taking the simulated Hodgkin-Huxley neural network and the real mouse brain network as two illustrative examples, we further verify the quantitative relations among the four causality measures and demonstrate that the causal connectivity inferred by any of the four well coincides with the underlying network structural connectivity, therefore establishing a direct link between the causal and structural connectivity. We stress that the structural connectivity of networks can be reconstructed pairwise without conditioning on the global information of all other nodes in a network, thus circumventing the curse of dimensionality. Our framework provides a practical and effective approach for pulse-output network reconstruction.

Keywords

causality | correlation | mutual information | Granger causality | transfer entropy | neural networks

Significance Statement

Inferring network connectivity is a key challenge in many diverse scientific fields. We investigate networks with pulse signal as measured output and solve the above reverse-engineering issue by establishing a direct link between the network's causal connectivity and structural connectivity. Here, the causal connectivity can be inferred by any one of the four causality measures, *i.e.*, time-delayed correlation coefficient, time-delayed mutual information, Granger causality, and transfer entropy. We analytically reveal the relationship among these four measures and show that they are equally effective to fully reconstruct the network connectivity pairwise. Our work provides a practical framework to reconstruct the structural connectivity in general pulse-output nonlinear networks or subnetworks.

The structural connectivity of a network, such as a cortical network, is of great importance in understanding the cooperation and competition among nodes in the network Mišić and Sporns (2016); McLaughlin et al. (2000); Zhou et al. (2013a). However, it is often difficult to measure directly the structural connectivity of a network. On the other hand, with the development of experimental techniques in neuroscience, it has become feasible to record simultaneously, with high temporal resolution, the activities of many nodes in a network Stringer et al. (2019); Wheeler and Nam (2011); Stosiek et al. (2003). This provides a possibility to reveal the underlying connectivity of the network by analyzing the nodes' activity data and identifying the causal interactions (connectivity) among them Seth (2008); Honey et al. (2009; 2010); Suárez et al. (2020); Chen et al. (2022).

There are difficulties in measuring these causal interactions: One of the most widely used statistical indicator for interaction identification is the correlation coefficient Benesty et al. (2009); Eisen et al. (1998); Ito et al. (2011) that characterizes linear dependence between two nodes. The correlation coefficient is symmetric, and thus cannot distinguish the driver-recipient relation to recover the causal connectivity Stevenson et al. (2008). To solve this, the time-delayed correlation coefficient (TDCC) Bedenbaugh and Gerstein (1997); Ito et al. (2011) was introduced to detect the direction of causal connectivity. However, TDCC, as a linear measure, may fail to capture causal interactions in nonlinear networks. As a nonlinear model-free generalization of TDCC, time-delayed mutual information (TDMI) Vastano and Swinney (1988); Schreiber (2000); Frenzel and Pompe (2007) was proposed to measure the flow of information in nonlinear networks. Despite their mathematical simplicity and computational efficiency, TDCC and TDMI cannot exclude the historical effect of signals and may encounter the issue of over estimation Frenzel and Pompe (2007); Schreiber (2000); McLaughlin et al. (2004). Granger causality (GC) Granger (1969); Bressler and Seth (2011); Barnett et al. (2018) and transfer entropy (TE) Schreiber (2000); Bossomaier et al. (2016); Borge-Holthoefer et al. (2016) were two other measures that were introduced to detect causal connectivity with the exclusion of the signal's own historical effects. GC is based on linear regression models that assumes the causal relation can be revealed by analyzing low-order statistics of signals (up to the variance). Consequently, the validity of GC for nonlinear networks is in general questionable Li et al. (2018). In contrast, TE is a nonparametric information-theoretic measure that quantifies the causal interactions with no assumption of interaction models. However, it requires the estimation of the probability distribution of dynamical variables conditioned on the historical information in networks, which makes TE suffer from the curse of dimensionality in practical applications to network systems with many nodes Runge et al. (2012); Newell and Cheng (2016); Bach (2017). It is known that the causal connectivity inferred by different causality measures can be inconsistent with each other Marbach et al. (2010); De Smet and Marchal (2010); Zou et al. (2009). With one exception [TE has been proven to be equivalent to GC for Gaussian variables Barnett et al. (2009)], there is little understanding of the relationships between these measurement techniques, including when they accurately (or inaccurately) predict causal interaction (connectivity).

The determination of structural connectivity from causal interactions faces even more fundamental difficulties: First, causal interactions may not be the result of direct structural connections between the nodes of the network, but rather the result of correlations arising from common external stimuli, or correlations arising from indirect connections ($X \to Y$, $Y \to Z$; resulting in correlations between X and Z, yet without direct structural connection between X and Z). Second, causal connectivity, as inferred from any of the above four measures, is statistical rather than structural Koch et al. (2002); Seth (2005); Schiele et al. (2013); Zhou et al. (2013b; 2014), i.e., causal connectivity quantifies direct statistical correlation among network nodes, whereas structural connectivity corresponds to physical connections among network nodes. In general, it is not clear if structural connectivity can be reconstructed from causal connectivity.

In this work we resolve the above issues for an important class of nonlinear networks, which we term *pulse-output networks* and which include *spiking neural networks*. The activity of each node in a pulse-output network can be described as a stochastic binary time series for the presence/absence of a pulse (spike) in each time window, e.g., spike train. Under this description, we illustrate mathematically that the four causal connectivity measures (TDCC, TDMI, GC, and TE) can be represented by one another, to their leading order in a perturbation expansion. Next, both by simulations of a Hodgkin-Huxley (HH) neural network and with experimentally measured data from mouse cortical network Allen Institute for Brain Science (2016), we verify that the mathematical relations among the four causality measures remain valid for representative samples of pulse-output networks. More importantly, for pulse-output networks, we demonstrate that the underlying structural connectivity can be recovered from the causal connectivity, itself inferred from any one of four causality measures. For pulse-output networks of both simulated Hodgkin-Huxley neural networks and a real mouse cortical network Allen Institute for Brain Science (2016), we show that the potential problems with the recovery of structural connectivity from causal connectivity, e.g., confounders and hidden nodes, can be resolved.

We emphasize that the *pulse-output* nature of a spiking neuronal network allows one to represent the neuronal signal as a binary time series, with random spike times. With the utility of this stochastic binary representation, our analytical framework shows, using only pairwise information between neurons, (i) the establishment of mathematical relationships between four common measures of causal connectivity, and (ii) the accurate pairwise predictions of structural connectivity from causal connectivity, without conditioning on the global information from all other nodes. Thus, the reconstruction circumvents the curse of dimensionality and can be applied to the reconstruction of structural connectivity in large-scale pulse-output nonlinear systems or subsystems.

Results

Concepts of generalized pairwise TDCC, TDMI, GC, and TE

Consider a nonlinear network of N nodes with dynamics given by

$$\frac{d\mathbf{Z}}{dt} = \mathbf{F}(\mathbf{Z}, t), \tag{1}$$

where $\mathbf{Z} = (Z_1, Z_2, ..., Z_N)$. We focus on the application of TDCC, TDMI, GC, and TE to each pair of nodes without conditioning on the rest of nodes in the network, accounting for the practical constraint that conditional causality measures in general require the information of the whole network that is often difficult to observe. For the ease of illustration, we denote a pair of nodes as $X = Z_i$ and $Y = Z_j$, and their measured time series as $\{x_n\}$ and $\{y_n\}$, respectively.

TDCC Bedenbaugh and Gerstein (1997); Ito et al. (2011), as a function of time delay m, is defined by

$$C(X, Y; m) = \frac{\mathrm{cov}(x_n, y_{n-m})}{\sigma_x \sigma_y},$$

where "cov" represents the covariance, σ_x and σ_y are the standard deviations of $\{x_n\}$ and $\{y_n\}$, respectively. A positive (negative) value of m indicates the calculation of causal value from Y to X (from X to Y), and nonzero $C(X,Y;m)$ indicates the existence of causal interaction between X and Y. Without loss of generality, we consider the case of positive m in the following discussions, that is, the causality measure from Y to X.

In contrast to the linear measure TDCC, TDMI is a model-free method being able to characterize nonlinear causal interactions Vastano and Swinney (1988); Schreiber (2000); Frenzel and Pompe (2007). TDMI from Y to X is defined by

$$I(X,Y;m) = \sum_{x_n, y_{n-m}} p(x_n, y_{n-m}) \log \frac{p(x_n, y_{n-m})}{p(x_n)p(y_{n-m})}, \tag{2}$$

where $p(x_n, y_{n-m})$ is the joint probability distribution of x_n and y_{n-m}, $p(x_n)$ and $p(y_{n-m})$ are the corresponding marginal probability distributions. $I(X,Y;m)$ is non-negative and vanishes if and only if x_n and y_{n-m} are independent Frenzel and Pompe (2007). Nonzero $I(X,Y;m)$ implies the existence of causal interaction from Y to X for a positive m.

It has been noted that TDCC and TDMI could overestimate the causal interactions when a signal has a long memory Frenzel and Pompe (2007); Schreiber (2000); McLaughlin et al. (2004). As an alternative, GC was proposed to overcome the issue of overestimation based on linear regression Granger (1969); Guo et al. (2008); Bressler and Seth (2011). The auto-regression for X is represented by $x_{n+1} = a_0 + \sum_{i=1}^{k} a_i x_{n+1-i} + \varepsilon_{n+1}$, where $\{a_i\}$ are the *estimated* auto-regression coefficients and ε_{n+1} is the residual. By including the historical information of Y with a message length l and a time-delay m, the joint regression for X is represented by $x_{n+1} = \tilde{a}_0 + \sum_{i=1}^{k} \tilde{a}_i x_{n+1-i} + \sum_{j=1}^{l} \tilde{b}_j y_{n+2-m-j} + \eta_{n+1}$, where $\{\tilde{a}_i\}$ and $\{\tilde{b}_j\}$ are the *estimated* joint regression coefficients, and η_{n+1} is the corresponding residual. If there exists a causal interaction from Y to X, then the prediction of X using the linear regression models shall be improved by additionally incorporating the historical information of Y. Accordingly, the variance of residual η_{n+1} is smaller than that of ε_{n+1}. Based on this concept, the GC value from Y to X is defined by

$$G_{Y\to X}(k,l;m) = \log \frac{\mathrm{Var}(\varepsilon_{n+1})}{\mathrm{Var}(\eta_{n+1})}.$$

The GC value is also non-negative and vanishes if and only if $\{\tilde{b}_j\} = 0$, *i.e.*, the variance of residual ε_{n+1} for X cannot be reduced by including the historical information of Y. Note that, by introducing the time-delay parameter m, the GC analysis defined above generalizes the conventional GC analysis, as the latter corresponds to the special case of $m = 1$.

GC assumes that the causal interaction can be fully captured by the variance reduction in the linear regression models, which is valid for Gaussian signals but not for more general signals. As a nonlinear extension of GC, TE was proposed to describe the causal interaction from the information theoretic perspective Schreiber (2000). The TE value from Y to X is defined by

$$\begin{aligned} T_{Y\to X}(k,l;m) = \sum_{x_{n+1}, x_n^{(k)}, y_{n+1-m}^{(l)}} & p(x_{n+1}, x_n^{(k)}, y_{n+1-m}^{(l)}) \\ & \cdot \log \frac{p(x_{n+1} | x_n^{(k)}, y_{n+1-m}^{(l)})}{p(x_{n+1} | x_n^{(k)})}, \end{aligned} \tag{3}$$

where the shorthand notation $x_n^{(k)} = (x_n, x_{n-1}, \ldots, x_{n-k+1})$ and $y_{n+1-m}^{(l)} = (y_{n+1-m}, y_{n-m}, \ldots, y_{n+2-m-l})$, k, l indicate

the length (order) of historical information of X and Y, respectively. Similar to GC, the time-delay parameter m is introduced that generalizes the conventional TE, the latter of which corresponds to the case of $m = 1$. TE is non-negative and vanishes if and only if $p(x_{n+1}|x_n^{(k)}, y_{n+1-m}^{(l)}) = p(x_{n+1}|x_n^{(k)})$, *i.e.*, the uncertainty of x_{n+1} is not affected regardless of whether the historical information of Y is taken into account.

In this work, we investigate the mathematical relations among TDCC, TDMI, GC, and TE by focusing on nonlinear networks described by Eq. 1 with pulse signals as measured output, *e.g.*, the spike trains measured in neural networks. Consider a pair of nodes X and Y in the network of N nodes, and denote their pulse-output signals by

$$w_x(t) = \sum_l \delta(t - \tau_{xl}) \quad \text{and} \quad w_y(t) = \sum_l \delta(t - \tau_{yl}), \tag{4}$$

respectively, where $\delta(\cdot)$ is the Dirac delta function, and $\{\tau_{xl}\}$ and $\{\tau_{yl}\}$ are the output time sequences of nodes X and Y determined by Eq. 1, respectively. With the sampling resolution of Δt, the pulse-output signals are measured as binary time series $\{x_n\}$ and $\{y_n\}$, where $x_n = 1$ ($y_n = 1$) if there is a pulse signal, *e.g.*, a spike generated by a neuron, of X (Y) occurred in the time window $[t_n, t_n + \Delta t)$, and $x_n = 0$ ($y_n = 0$) otherwise, *i.e.*,

$$x_n = \int_{t_n}^{t_n+\Delta t} w_x(t)dt \quad \text{and} \quad y_n = \int_{t_n}^{t_n+\Delta t} w_y(t)dt, \tag{5}$$

and $t_n = n\Delta t$. Note that the value of Δt is often chosen to make sure that there is at most one pulse signal in a single small enough time window. In the stationary state, the responses x_n and y_n can be viewed as stochastic processes when the network is driven by stochastic external inputs. In such a case, for the sake of simplicity, we denote $p_x = p(x_n = 1)$, $p_y = p(y_n = 1)$, and define $\Delta p_m = \frac{p(x_n=1, y_{n-m}=1)}{p(x_n=1)p(y_{n-m}=1)} - 1$, measuring the dependence between x_n and y_{n-m}.

Mathematical relation between TDMI and TDCC

For the relation between TDCC and TDMI when applied to nonlinear networks with pulse-output signals, we prove the following theorem:

Theorem 1. *For nodes X and Y with pulse-output signals given in Eqs. 4 and 5, we have*

$$I(X,Y;m) = \frac{C^2(X,Y;m)}{2} + O(\Delta t^2 \Delta p_m^3), \tag{6}$$

where the symbol "O" stands for the order.

Proof. The basic idea is to Taylor expand TDMI in Eq. 2 with respect to the term $\frac{p(x_n, y_{n-m})}{p(x_n)p(y_{n-m})} - 1$ (the detailed derivation can be found in *SI Appendix, Supporting Information Text 1B*), then we arrive at the following expression:

$$\begin{aligned} I(X,Y;m) &= \sum_{x_n, y_{n-m}} p(x_n, y_{n-m}) \log\left(1 + \frac{p(x_n, y_{n-m})}{p(x_n)p(y_{n-m})} - 1\right) \\ &= \frac{[p(x_n = 1, y_{n-m} = 1) - p_x p_y]^2}{2(p_x - p_x^2)(p_y - p_y^2)} + O(\Delta t^2 \Delta p_m^3). \end{aligned}$$

Since TDCC can be written as

$$C(X,Y;m) = \frac{p(x_n = 1, y_{n-m} = 1) - p_x p_y}{\sqrt{(p_x - p_x^2)(p_y - p_y^2)}}, \tag{7}$$

we have

$$I(X,Y;m) = \frac{C^2(X,Y;m)}{2} + O(\Delta t^2 \Delta p_m^3).$$

□

Mathematical relation between GC and TDCC

We next derive the relation between GC and TDCC as follows:

Theorem 2. *For nodes X and Y with pulse-output signals given in Eqs. 4 and 5, we have*

$$G_{Y\to X}(k,l;m) = \sum_{i=m}^{m+l-1} C^2(X,Y;i) + O(\Delta t^3 \Delta p_m^2). \tag{8}$$

Proof. From the definition, GC can be represented by the covariances of the signals Barnett et al. (2009) as

$$G_{Y\to X}(k,l;m) = \log \frac{\Gamma(x_{n+1}|x_n^{(k)})}{\Gamma(x_{n+1}|x_n^{(k)} \oplus y_{n+1-m}^{(l)})}, \tag{9}$$

where $\Gamma(\mathbf{x}|\mathbf{y}) = \mathrm{cov}(\mathbf{x}) - \mathrm{cov}(\mathbf{x},\mathbf{y})\mathrm{cov}(\mathbf{y})^{-1}\mathrm{cov}(\mathbf{x},\mathbf{y})^T$ for random vectors $\mathbf{x}$ and $\mathbf{y}$, $\mathrm{cov}(\mathbf{x})$ and $\mathrm{cov}(\mathbf{y})$ denote the covariance matrix of $\mathbf{x}$ and $\mathbf{y}$, respectively, and $\mathrm{cov}(\mathbf{x},\mathbf{y})$ denote the cross-covariance matrix between $\mathbf{x}$ and $\mathbf{y}$. The symbol T is the transpose operator and $\oplus$ denotes the concatenation of vectors.

We first prove that the auto-correlation function (ACF) of binary time series $\{x_n\}$ as a function of time delay takes the order of Δt (see *SI Appendix, Supporting Information Text 1C* for the proof, Fig. S1). Accordingly, we have

$$\mathrm{cov}(x_n^{(k)}) = \sigma_x^2(\mathbf{I} + \hat{\mathbf{A}}),$$

where $\hat{\mathbf{A}} = (\hat{a}_{ij}), \hat{a}_{ij} = O(\Delta t)$, and $\mathbf{I}$ is the identity matrix. Hence,

$$\begin{aligned}\Gamma(x_{n+1}|x_n^{(k)}) =& \sigma_x^2 - \frac{1}{\sigma_x^2}\mathrm{cov}(x_{n+1}, x_n^{(k)})(\mathbf{I} - \hat{\mathbf{A}}) \\ &\cdot \mathrm{cov}(x_{n+1}, x_n^{(k)})^T + O(\Delta t^5).\end{aligned} \tag{10}$$

In the same way, we have

$$\begin{aligned}&\Gamma(x_{n+1}|x_n^{(k)} \oplus y_{n+1-m}^{(l)}) \\ =& \sigma_x^2 - \frac{1}{\sigma_x^2}\mathrm{cov}(x_{n+1}, x_n^{(k)})(\mathbf{I} - \hat{\mathbf{A}})\mathrm{cov}(x_{n+1}, x_n^{(k)})^T \\ &- \frac{1}{\sigma_y^2}\mathrm{cov}(x_{n+1}, y_{n+1-m}^{(l)})(\mathbf{I} - \hat{\mathbf{B}})\mathrm{cov}(x_{n+1}, y_{n+1-m}^{(l)})^T \\ &+ O(\Delta t^4 \Delta p_m^2).\end{aligned} \tag{11}$$

where $\hat{\mathbf{B}} = (\hat{b}_{ij}), \hat{b}_{ij} = O(\Delta t)$. Substituting Eqs. 10 and 11 into Eq. 9 and Taylor expanding Eq. 9 with respect to Δt, we can obtain

$$G_{Y\to X}(k,l;m) = \sum_{i=m}^{m+l-1} C^2(X,Y;i) + O(\Delta t^3 \Delta p_m^2). \tag{12}$$

The detailed derivation of Eqs. 10, 11, and 12 can be found in *SI Appendix, Supporting Information Text 1C.* □

Mathematical relation between TE and TDMI

From the definitions of TE and TDMI, TE can be regarded as a generalization of TDMI conditioning on the signals' historical information additionally. To rigorously establish their relationship, we require that $\left\|x_{n+1}^{(k+1)}\right\|_0 \leq 1$ and $\left\|y_{n+1-m}^{(l)}\right\|_0 \leq 1$ in the definition of TE given in Eq. 3, where $\left\|\cdot\right\|_0$ denotes the l_0 norm of a vector, *i.e.*, the number of nonzero elements in a vector. This assumption indicates that the length of historical information used in the TE framework is shorter than the minimal time interval between two consecutive pulse-output signals. With this condition, we mathematically establish the following theorem:

Theorem 3. *For nodes X and Y with pulse-output signals given in Eqs. 4 and 5, under the assumption that* $\left\|x_{n+1}^{(k+1)}\right\|_0 \leq 1$ *and* $\left\|y_{n+1-m}^{(l)}\right\|_0 \leq 1$*, we have*

$$T_{Y\to X}(k,l;m) = \sum_{i=m}^{m+l-1} I(X,Y;i) + O(\Delta t^3 \Delta p_m^2), \tag{13}$$

where $T_{Y\to X}$ *is defined in Eq. 3.*

Proof. To simplify the notation, we denote $x_n^{(k)} = (x_n, x_{n-1}, \ldots, x_{n-k+1})$ and $y_{n+1-m}^{(l)} = (y_{n+1-m}, y_{n-m}, \ldots, y_{n+2-m-l})$ as x^- and y^-, respectively. From Eq. 3, we have

$$\begin{aligned} T_{Y\to X}(k,l;m) &= \sum_{x_{n+1},x^-,y^-} p(x_{n+1},x^-,y^-) \log \frac{p(x_{n+1}|x^-,y^-)}{p(x_{n+1}|x^-)} \\ &= \sum_{i=m}^{m+l-1} I(X,Y;i) + \mathscr{A} + \mathscr{B}, \end{aligned} \tag{14}$$

where

$$\mathscr{A} = \sum_{x_{n+1},y^-} p(x_{n+1},y^-) \log \frac{p(y^-|x_{n+1})}{\prod_j p(y_j|x_{n+1})} \frac{\prod_j p(y_j)}{p(y^-)}$$

and

$$\mathscr{B} = \sum_{x_{n+1},x^-,y^-} p(x_{n+1},x^-,y^-) \log \frac{p(x_{n+1}|x^-,y^-)}{p(x_{n+1}|y^-)} \frac{p(x_{n+1})}{p(x_{n+1}|x^-)}$$

and $\prod_j$ in $\mathscr{A}$ represents $\prod_{j=n+2-m-l}^{n+1-m}$. The detailed derivation of Eq. 14 can be found in *SI Appendix, Supporting Information Text 1D.* $\mathscr{A}$ and $\mathscr{B}$ in Eq. 14 consist of multiple terms and the leading order of each term can be analytically calculated. For the sake of illustration, we derive the leading order of one of these terms and the leading order of the rest terms can be estimated in a similar way. Under the assumption that $\left\|x_{n+1}^{(k+1)}\right\|_0 \leq 1$ and $\left\|y_{n+1-m}^{(l)}\right\|_0 \leq 1$, the number of nonzero components is at most one in $x_{n+1}^{(k+1)}$ and $y_{n+1-m}^{(l)}$. Without loss of generality, we assumed $x_{n+1} = 1$ and $y_{n+1-m} = 1$. In such a case, we can obtain the following expression in $\mathscr{A}$,

$$p(x_{n+1},y^-)\log\frac{p(y^-|x_{n+1})}{\prod_j p(y_j|x_{n+1})}\frac{\prod_j p(y_j)}{p(y^-)}\Bigg|_{x_{n+1}=1,y_{n+1-m}=1}$$
$$= p_x p_y^2 \sum_{j\neq n+1-m}\Delta p_{n+1-j} + O(\Delta t^3\Delta p_m^2).$$

We can further show that the leading order of all the terms in $\mathscr{A}$ cancel each other out (*SI Appendix, Supporting Information Text 1D*), thus we have $\mathscr{A} = O(\Delta t^3\Delta p_m^2)$. Similarly, we can also show $\mathscr{B} = O(\Delta t^3\Delta p_m^2)$ (*SI Appendix, Supporting Information Text 1D*), and thus

$$T_{Y\to X}(k,l;m) = \sum_{i=m}^{m+l-1} I(X,Y;i) + O(\Delta t^3\Delta p_m^2).$$

□

Mathematical relation between GC and TE

From Theorems 1-3, we can prove the following theorem (see details in *SI Appendix, Supporting Information Text 1E*):

Theorem 4. *For nodes X and Y with pulse-output signals given in Eqs. 4 and 5, under the assumption that* $\left\|x_{n+1}^{(k+1)}\right\|_0 \leq 1$ *and* $\left\|y_{n+1-m}^{(l)}\right\|_0 \leq 1$*, we have*

$$G_{Y\to X}(k,l;m) = 2T_{Y\to X}(k,l;m) + O(\Delta t^2\Delta p_m^3) \tag{15}$$

where $T_{Y\to X}$ *is defined in Eq. 3.*

Note that in Theorems 3 and 4 we require the condition $\left\|x_{n+1}^{(k+1)}\right\|_0 \leq 1$ and $\left\|y_{n+1-m}^{(l)}\right\|_0 \leq 1$ to establish the relations of causality measures. However, the relations of the four causality measure remain to hold approximately in the absence of this condition, as will be discussed below. We summarize the relations among the four causality measures in Fig. 1.

Mathematical relations of causality measures verified in HH neural networks

To verify the relations among the causality measures derived above, as an illustrative example, we apply generalized pairwise TDCC, TDMI, GC, and TE to the HH neural network described in *Materials and Methods*. We first consider a pair of neurons denoted by X and Y with unidirectional connection from Y to X in an HH network containing 10 excitatory neurons driven by homogeneous Poisson inputs. Let $\{\tau_{xl}\}$ and $\{\tau_{yl}\}$ be the ordered spike times of neuron X and Y in the HH network respectively and denote their spike trains as $w_x(t) = \sum_l \delta(t-\tau_{xl})$ and $w_y(t) = \sum_l \delta(t-\tau_{yl})$, respectively. With a sampling resolution of Δt, the spike train is measured as a binary time series as described above. To numerically verify the above theorems, we then check the order of the remainders in Eqs. 6, 8, 13, and 15 in terms of Δt and Δp_m. Note that Δp_m, the measure of the dependence between X and Y, is insensitive to sampling resolution Δt (*SI Appendix*, Fig. S2). Therefore, by varying sampling interval Δt and coupling strength S (linearly related to Δp_m) respectively, the orders of the remainders are consistent with those derived in Eqs. 6, 8, 13, and 15 as shown

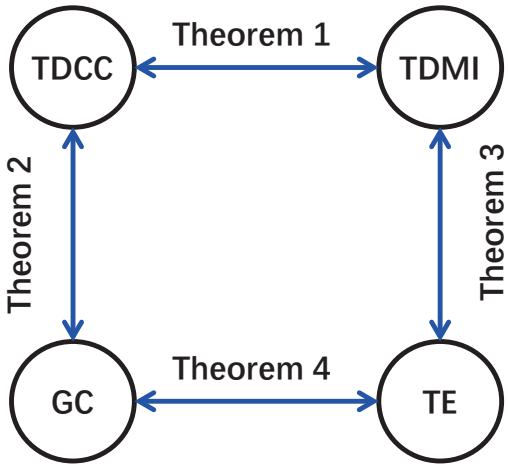

Figure 1: Mathematical relations among TDCC, TDMI, GC, and TE estiblished by four theorems.

in Fig. 2. In addition, Fig. 3 verifies the relations among the causality measures by changing other parameters. For example, in Eqs. 8, 13, and 15, the four causality measures are proved to be independent of the historical length k, which is numerically verified in Fig. 3A. And although the values of GC and TE rely on the historical length l in Fig. 3B, the mathematical relations among the four causality measures revealed by Theorems 1-4 hold for a wide range of l. Recall that, in order to establish rigorously the relations of the causality measures of Theorems 3 and 4, our proofs required the assumption $\left\|x_{n+1}^{(k+1)}\right\|_0 \leq 1$ and $\left\|y_{n+1-m}^{(l)}\right\|_0 \leq 1$. This assumption can often be satisfied: For example, in the above simulations, the memory time of the neuron is about 20 ms, while the inter-spike interval is around 100 ms. However, even when the assumption breaks down in the regime of high firing rate, the mathematical relations among the four causality measures remain to hold approximately, as examined numerically in *SI Appendix*, Fig. S3.

We next verify the mathematical relations among the causality measures for the parameter of time delay m by fixing parameters k and l. In principle, the value of k and l in GC and TE shall be determined by the historical memory of the system. To reduce the computational cost Gourévitch and Eggermont (2007); Vicente et al. (2011); Ito et al. (2011), we take $k = l = 1$ for all the results below. It turns out that this parameter choice works well for pulse-output networks because of the short memory effect in general, as will be further discussed later. Fig. 3C shows the mathematical relations hold for a wide range of the time-delay parameter used in computing the four causality measures (see more examples in *SI Appendix*, Fig. S4).

We further examine the robustness of the mathematical relations among TDCC, TDMI, GC, and TE by scanning the parameters of the coupling strength S between the HH neurons and external Poisson input strength and rates. As shown in Fig. 3D, the values of the four causality measures with different coupling strength are very close to one another. Their relations also hold for a wide range of external Poisson input parameters (*SI Appendix*, Fig. S5). From the above, the mathematical relations among TDCC, TDMI, GC, and TE described in Theorems 1-4 are verified in the HH network.

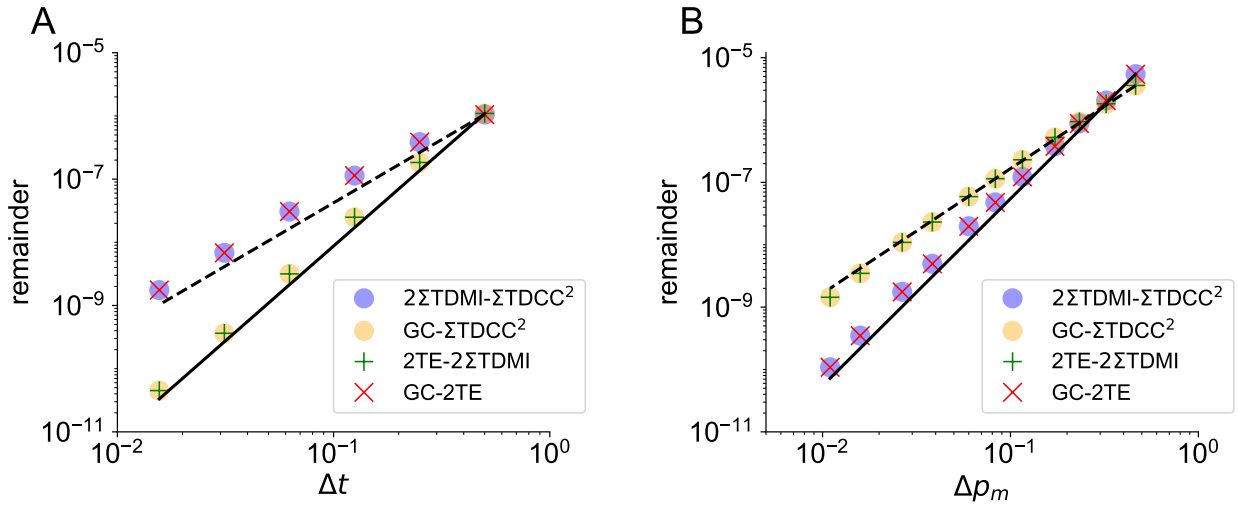

Figure 2: Numerical verification of the convergence order of the remainders in terms of (*A*) Δt and (*B*) Δp_m. The convergence orders for Δt in (*A*) and Δp_m in (*B*) agree well with Theorems 1-4 ($R^2 > 0.998$). The gray dashed and solid lines indicate the 2nd-order and 3rd-order convergence, respectively. The four causal measures are calculated from a pair of unidirectionally connected neurons in an HH network of 10 excitatory neurons randomly connected with probability 0.25. The parameters are set as $k = l = 5$ and $m = 6$ (time delay is 3 ms), $S = 0.02$ mS·cm^{-2} in (*A*), and $\Delta t = 0.5$ ms in (*B*).

Relation between structural connectivity and causal connectivity in HH neural networks

We next discuss the relation between the inferred causal connectivity and the structural connectivity. Note that the causal connectivity inferred by these measures is statistical rather than structural Koch et al. (2002); Seth (2005); Schiele et al. (2013), *i.e.*, the causal connectivity quantifies the direct statistical correlation or dependence among network nodes, whereas the structural connectivity corresponds to physical connections among network nodes. Therefore, a precise relationship between causal connectivity and structural connectivity has been unclear. In Fig. 3*C*, the peak causal value from Y to X (at time delay around 3 ms, $m = 6$) is greater than the significance level (the gray dashed line in Fig. 3*C*), while the causal value from X to Y is not. Based on this, the inferred direct causal connections between X and Y are consistent with the underlying structural connections. From now on, we adopt peak causal values, $m = 6$, to represent the causal connectivity unless noted explicitly.

To investigate the validity of this consistency in larger networks, we further investigate a larger HH network (100 excitatory neurons) with random connectivity structure and homogeneous coupling strength (*SI Appendix*, Fig. S6A). As shown in Fig. 4*A*, the distributions of all four causal values across all pairs of neurons virtually overlapped, which again verify their mathematical relations given by Theorems 1-4. In addition, as the network size increases from 10 to 100 neurons, the distributions of the causality measures in Fig. 4*A* exhibits a bimodal structure with a clear separation of orders. By mapping the causal values with the structural connectivity, we find that the right bump of the distributions with larger causal values corresponded to connected pairs of neurons, while the left bump with smaller causal values corresponded to unconnected pairs. The well separation of the two modals indicates that the underlying structural connectivity in the HH network can be accurately estimated from the causal connectivity.

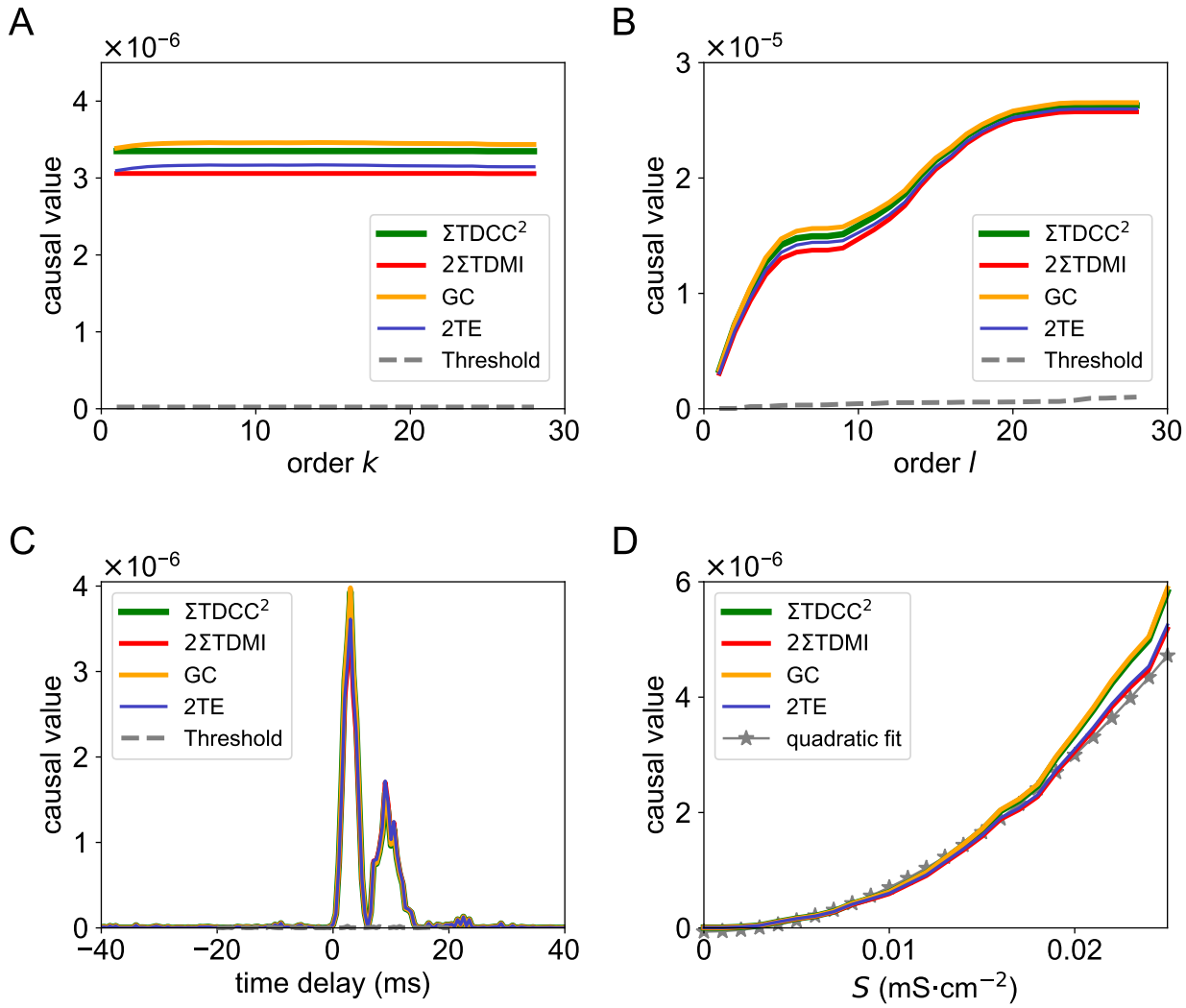

Figure 3: Dependence of causal values on parameters of (*A*) order k, (*B*) order l, (*C*) time delay, and (*D*) coupling strength S obtained from the same pair of neurons in the HH network in Fig. 2. In (*C*), a positive (negative) time delay indicates the calculation of causal values from neuron Y to neuron X (from X to Y). The green curve represents the summation of squared TDCC $C(X,Y;m)$, the red curve represents twice of the summation of TDMI $I(X,Y;m)$, the orange curve stands for GC $G_{Y\to X}(k,l;m)$, and the blue curve stands for twice of TE $T_{Y\to X}(k,l;m)$. The curves virtually overlap in (A)-(D) (all significantly greater than those of randomly surrogate time series, $p<0.05$). The gray dashed curve in (A)-(C) is the significance level of causality for two unconnected neurons in the same HH network. The gray star curve in (D) is the quadratic fit of causal values with respect to different S. The parameters are set as (*A*): $l=1$, $S=0.02$ mS·cm^{-2}, $\Delta t=0.5$ ms, $m=6$; (*B*): $k=1$, $S=0.02$ mS·cm^{-2}, $\Delta t=0.5$ ms, $m=6$; (*C*): $k=l=1$, $S=0.02$ mS·cm^{-2}, $\Delta t=0.5$ ms; (*D*): $k=l=1$, $\Delta t=0.5$ ms, $m=6$.

The performance of this reconstruction approach can be quantitatively characterized by the receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC) Fawcett (2006); Marbach et al. (2010); Carter et al. (2016) (see *Materials and Methods*). It is found that the AUC value becomes 1 when applying any of these four causality measures

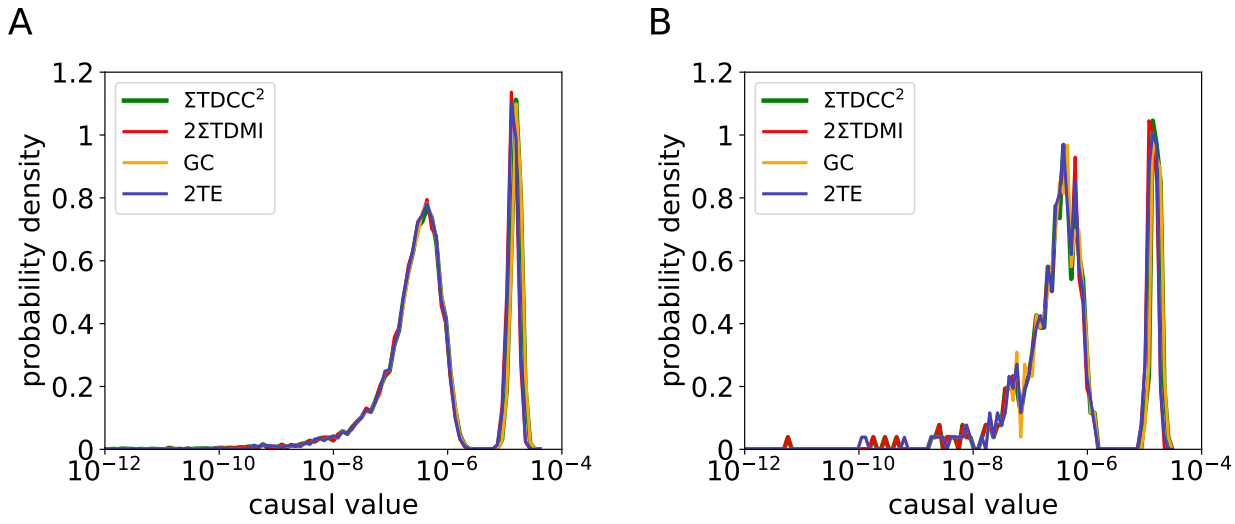

Figure 4: Distributions of causal values in an HH network of 100 excitatory neurons randomly connected with probability 0.25. (*A*) The distribution of causal values of each pair of neurons in the whole network. (*B*) The distribution of causal values of each pair of neurons in an HH subnetwork of 20 excitatory neurons. The parameters are set as $k = l = 1$, $S = 0.02$ mS·cm^{-2}, $\Delta t = 0.5$ ms, and $m = 6$. The colors are the same as those in Fig. 3 and the curves nearly overlap.

(*SI Appendix*, Fig. S6B), which indicates that the structural connectivity of the HH network could be reconstructed with 100% accuracy. We point out that the reconstruction of network connectivity based on causality measures is achieved by calculating the causal values between each pair of neurons that requires no access to the activity data of the rest of neurons. Therefore, this inference approach can be applied to a subnetwork when the activity of neurons outside the subnetwork is not observable. For example, when a subnetwork of 20 excitatory HH neurons is observed, the structural connectivity of the subnetwork can still be accurately reconstructed without knowing the information of the rest 80 neurons in the full network as shown in Fig. 4*B*. In such a case, the AUC values corresponding to the four causality measures are 1 (*SI Appendix*, Fig. S6C). In addition, we have also shown that, for HH networks with *heterogeneous* coupling strength (for example, the coupling strength follows a log-normal distribution), accurate reconstructions of network structural connectivity can still be achieved (*SI Appendix*, Fig. S7).

Mechanism underlying network connectivity reconstruction by causality measures

We next demonstrate the mechanism underlying the validity of pairwise inference of pulse-output signals in the reconstruction of network structural connectivity. It has been frequently noticed that pairwise causal inference may potentially fail to distinguish direct interactions from indirect ones Stevenson et al. (2008). For example, in the case that $Y \to W \to X$ where "$\to$" denotes a direct connection, the indirect interaction from Y to X may possibly be mis-inferred as a direct interaction via pairwise causality measures. However, pulse-output signals circumvent such spurious inferences as explained below. Here we take TDCC as an example to elucidate the underlying reason of successful reconstruction. If we denote $\delta p_{Y\to X} = p(x_n = 1|y_{n-m} = 1) - p(x_n = 1|y_{n-m} = 0)$ as the increment of

probability of generating a pulse output for neuron X induced by a pulse-output signal of neuron Y at m time step earlier, we have TDCC $C(X,Y;m) = \delta p_{Y\to X}\sqrt{\frac{p_y-p_y^2}{p_x-p_x^2}}$ through Eq. 7. For the case of $Y \to W \to X$, we can derive $\delta p_{Y\to X} = O(\delta p_{Y\to W}\cdot\delta p_{W\to X})$, and further $C(X,Y;m) = O\left(C(W,Y;m)\cdot C(X,W;m)\right)$ (see derivation details about the relations among δp, Δp_m, and causal measures in *SI Appendix, Supporting Information Text 2*, Figs. S8A-B). Because the influence of a single pulse-output signal is often small (*e.g.*, in the HH neural network with physiologically realistic coupling strengths, we obtain $|\delta p| < 0.01$ from simulation data), the causal value $C(X,Y;m)$ due to the indirect interaction is significantly smaller than $C(W,Y;m)$ or $C(X,W;m)$ by direct interactions as depicted in Fig. 4*A*.

We also note that the increment δp is linearly dependent on the coupling strength S (*SI Appendix*, Fig. S8C), thus we establish a mapping between the causal and structural connectivity, in which the causal value of TDCC is proportional to the coupling strength S between two neurons. The mapping between causal and structural connectivity for TDMI, GC, and TE are also established in a similar way, in which the corresponding causal values are proportional to S^2 as shown in Fig. 3*D*. Therefore, the application of pairwise causality measures to pulse-output signals is able to successfully reveal the underlying structural connectivity of a network. In summary, for pulse-output networks, structural connectivity can be accurately inferred from causal connectivity in a pair-wise manner that overcomes computational issues of high dimensionality. Thus, the method is potentially applicable to experimentally measured data from large-scale biological networks or subnetworks as discussed below.

Network connectivity reconstruction with physiological experimental data

Next, we apply all four causality measures to experimental data to address the issue of validity of their mathematical relations and reconstruction of the network structural connectivity. Here, we analyze the *in vivo* spike data recorded in the mouse cortex from Allen Brain Observatory Allen Institute for Brain Science (2016) (see *Materials and Methods*). By applying the four causality measures, we infer the causal connectivity of those cortical networks. As the underlying structural connectivity of the recorded neurons in experiments is unknown, we first detect putative connected links from the distribution of causality measures, and then follow the same procedures as previously described in the HH model case using ROC and AUC in signal detection theory Fawcett (2006); Marbach et al. (2010); Carter et al. (2016) (see *Materials and Methods*) to quantify the reconstruction performance.

Because we have demonstrated the equivalence of the four causality measures, we will use TE as a representative causality measure to demonstrate a way to detect putative connections. As we have shown above, the TE values are proportional to S^2. In addition, previous experimental works observed that the coupling strength S follows the log-normal distribution both for intra-areal and inter-areal cortical networks in mouse and monkey brains Buzsáki and Mizuseki (2014). Thus, the distribution of TE for connected pairs should follow the log-normal distribution. In addition, we assume that the distribution of TE for unconnected pairs also obeys the log-normal distribution (see *Materials and Methods*).

To validate the above assumption of two log-normal distributions of TE values, we consider an HH network where neurons are randomly connected and the network coupling strength follows the log-normal distribution as observed in experiment Buzsáki and Mizuseki (2014). In this case, although the distribution of log TE values $\log_{10} T_{Y\to X}$ in Fig. 5*A* for synaptically connected and unconnected pairs of neurons overlap with each other, it can be well fitted by the summation of two log-normal distributions of TE values. Importantly, the fitted distributions of TE of connected and

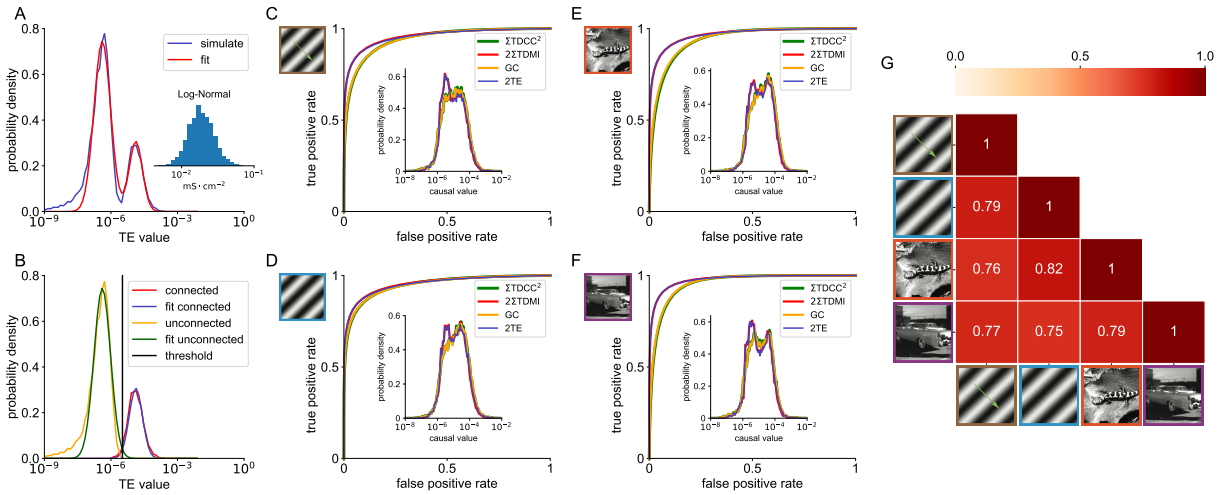

Figure 5: Reconstruction of structural connectivity by the assumption of mixed log-normal distribution of causal values. (A) Distribution of TE values in an HH network of 100 excitatory neurons. The entry A_{ij} in the adjacency matrix follows a Bernoulli distribution with probability of 0.25 being 1. For the connected pairs, $e.g.$, $A_{ij} = 1$, the corresponding coupling strength from neuron j to neuron i is sampled from a log-normal distribution. The blue and red curves are the simulated and fitted distributions, respectively. Inset: histogram of structural coupling strength, S. (B) Distributions of TE values from connected and unconnected pairs of HH neurons in (A). The red and orange curves are the simulated TE values from connected and unconnected pairs respectively, while the blue and green curves are the fitted TE values from connected and unconnected pairs respectively which are obtained from the fitting in (A). The parameters are the same as those in Fig. 4. (C-F): ROC curves of the network composed by the recorded neurons in experiment under visual stimuli of (C) drifting gratings, (D) static gratings, (E) natural scenes, and (F) natural movie, with AUC values equal to 0.950, 0.951, 0.953, 0.967. Insets: the distribution of causal values of each pair of neurons in the whole network. The parameters are set as $k = 1, l = 5$, $\Delta t = 1$ ms, and $m = 1$. The colors are the same as those in Fig. 3. (G) The similarity matrix of TE values over all neuron pairs across different stimuli conditions. For each pair of stimuli conditions, the similarity between the inferred TE values is measured by their correlation coefficient. The similarity matrix is symmetric with respect to its diagonal of ones, and the lower triangular part of the matrix is shown in the heatmap.

unconnected pairs agree well with those of the true connectivity set up in simulation as shown in Fig. 5*B*. Thus, we take the log-normal fitted causality measure distributions as the ground truth of structural connectivity to evaluate the performance of network connectivity reconstruction, *e.g.*, the AUC value is 0.997 in Fig. 5*B*. Practically, an optimal inference threshold determined as the intersection of the two fitted distribution curves Dayan and Abbott (2001) can be used for network connectivity reconstruction, which is indicated by the vertical solid line in Fig. 5*B*. In addition, the above results are robust for a variety of coupling strength distributions, including uniformly distributed and Gaussian distributed coupling strength with AUC values higher than 0.95 (*SI Appendix*, Fig. S7).

We then apply the same ROC analysis to all the four causality measures when analyzing experimental data under different visual stimuli conditions. As shown in the insets of Figs. 5*C*-*F*, the distribution of TDCC, TDMI, GC, and TE values are close to each other, which again verifies their mathematical relations given by Theorems 1-4. Under the assumption of mixed log-normal distribution (*SI Appendix*, Figs. S9A-D), we first infer the ground truth of the network structural connectivity, and then evaluate the reconstruction performance of TDCC, TDMI, GC, and TE by the ROC curves as shown in Figs. 5*C*-*F* with AUC values greater than 0.95. To investigate the consistency of reconstruction across different stimuli conditions, we compute the correlation coefficient of causal values for each pair of stimuli. As shown in Fig. 5*G*, our reconstruction achieves relatively high consistency across conditions with correlation coefficient higher than 0.75. With causal values, we also infer the binary adjacency matrix under each of four different stimuli conditions using the optimal inference threshold (see *Materials and Methods*) from the ROC analysis. Note that almost all of the causal values of inconsistently inferred connections fall into the overlapping region of the connected and unconnected distributions, *i.e.*, the causal values from connected (unconnected) pairs but less (greater) than the threshold (green curve in *SI Appendix*, Figs. S9E-H) which are generally error prone. We point out that such error prone generally exists as we can see similar phenomena for HH networks (*SI Appendix*, Fig. S10).

Discussion

In this work, inspired by the *pulse-output* nature of spiking neuronal networks, we have built an analytical framework upon the stochastic binary representation of neuronal interactions. It allows us, using only pairwise information between neurons, to (i) establish the mathematical relationships between four widely used measures (TDCC, TDMI, GC, and TE) of causal connectivity, and (ii) accurately reconstruct physical (structural) connectivity from causal connectivity in simulated HH neuronal networks. In addition, our framework provides the clarification and reduction of the major possible obstacles induced by confounders and unmeasured hidden nodes in conventional network reconstructions. Finally, we have used this analysis of pulse-output signals to reconstruct the structural connectivity of the real neuronal network in mouse brain from experimentally measured spike-train data and have achieved promising performances.

We emphasize two key features of pulse-output signals in eliminating of the curse of dimensionality that lead to effective network reconstruction via our framework: (i) the short timescale of auto-correlation, and (ii) the weakness of indirect causalities. On the one hand, with small time step Δt commonly used in experiment, the short auto-correlation timescale protects the inferred causality from the corruption of the self-memory of time series. Therefore, the short auto-correlation length overcomes the curse of dimensionality in the estimation of probability density function by reducing the order of conditioned signal history (e.g., $k = l = 1$). On the other hand, because of the spike nature

of the pulse-output signal, the causality value of direct connection is several orders of magnitude larger than that of indirect connection, which distinguishes direct connections from indirect ones. These two features make our framework a practical approach for pulse-coupled network reconstruction. In contrast, if these causality measures are directly applied to continuous-valued signals, *e.g.*, voltage time series of a neuronal network, the mathematical relations derived in our theorems do not hold (*SI Appendix*, Fig. S11A) and the network reconstruction may also fail. For instance, TDCC and TDMI gives incorrect reconstruction of the structural connectivity due to the strong self-correlation of continuous-valued time series (*SI Appendix*, Fig. S11B).

In the main text, we have illustrated the effectiveness of the four causality measures by taking the examples of an excitatory HH neural network receiving uncorrelated external Poisson drive (see *Materials and Methods*). In fact, as discussed below, these methods apply much more broadly for pulse-output networks, including networks in different dynamical states, networks receiving correlated inputs, networks with both excitatory and inhibitory neurons, networks with different neuronal models.

First, oscillations and synchronizations are commonly observed in the biological brain network, as shown in *SI Appendix*, Fig. S12A. Due to the fake causality between neurons introduced by the strong synchronous state, conventional reconstruction frameworks fail to capture the true structural connectivity. However, with our framework, high inference accuracy (AUC > 0.88) can still be achieved (*SI Appendix*, Figs. S12B-C). Furthermore, by applying a desynchronized sampling method that only samples the pulse-output signals in asynchronous time intervals (*SI Appendix*, Fig. S12D), we can again perfectly reconstruct the network (AUC > 0.99, *SI Appendix*, Figs. S12E-F). The relation between percentages of desynchronized downsampling and AUC values is shown in *SI Appendix*, Fig. S13.

Second, external inputs to the network in the brain can often be correlated. In such a case, the synchronized states may be observed similarly as in previous cases (*SI Appendix*, Fig. S14A). Nevertheless, our framework can still achieve high inference accuracy (AUC > 0.99 with desynchronized downsampling methods, or AUC > 0.88 without downsampling, *SI Appendix*, Figs. S14 B-F) if the external inputs are moderately correlated, *e.g.*, correlation coefficient less than 0.35 in our simulation case.

Third, in the case of HH networks with both excitatory and inhibitory neurons, we obtain similarly accurate reconstructions (AUC > 0.99 with desynchronized downsampling methods, or AUC > 0.71 without downsampling, *SI Appendix*, Fig. S15).

Last but not least, we also apply our framework of reconstruction to other types of neuronal networks, including leaky integrate-and-fire (LIF) networkAbbott (1999), Izhikevich networkIzhikevich (2003), FitzHugh-Nagumo networkFitzHugh (1961), and Morris-Lecar networkMorris and Lecar (1981) (see *Materials and Methods*). The results of all these networks can be seen in *SI Appendix*, Fig. S16. Our framework works well, with clear two-modal distributions of causality values (*SI Appendix*, Fig. S16B, S16E, S16H, S16K) and high reconstruction performance (AUC > 0.98, *SI Appendix*, Fig. S16C, S16F, S16I, S16L).

Materials and Mechods

Reconstruction performance evaluation

For binary inference of structural connectivity, analysis based on receiver operating characteristic (ROC) curves is adopted to evaluate the reconstruction performance in this work. The following two scenarios are considered.

With known structural connectivity

The conventional procedures for ROC curves analysis can be naturally applied towards data with true labels, i.e. structural connectivity in our binary reconstruction case. The area under the ROC curve (AUC) quantifies how well causality measures can distinguish structurally connected node pairs from unconnected node pairs. If AUC is close to 1, the distribution of the causality values of directly connected node pairs and that of unconnected ones are well distinguishable, i.e., the performance of binary reconstruction is good. If the AUC is close to 0.5, the distribution of those two kinds of causality values are virtually indistinguishable, meaning the performance of reconstruction is close to a random guess.

Without known structural connectivity

Conceptually, the causality measures for synaptically connected and unconnected pairs of neurons can be fitted by the log-normal distribution with different parameters, *i.e.*, $\log_{10} T^{(\mathrm{con})}_{Y\to X} \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $\log_{10} T^{(\mathrm{uncon})}_{Y\to X} \sim \mathcal{N}(\mu_2, \sigma_2^2)$ where the superscripts "con" and "uncon" indicates a direct synaptic connection and no synaptic connection from Y to X, respectively. So motivated, we fit the overall distribution with the summation of two log-normal distributions

$$\frac{p_c}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right) + \frac{1-p_c}{\sqrt{2\pi\sigma_2^2}} \exp\left(-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right),$$

where p_c is the proportional coefficient between two Gaussian distributions $\log_{10} T^{(\mathrm{con})}_{Y\to X}$ and $\log_{10} T^{(\mathrm{uncon})}_{Y\to X}$. After that, the fitted distribution of two types of pairs are regarded as the true labels of structural connectivity. The similar procedures, as the previous case with known structural connectivity, are applied to obtain the ROC curve and the AUC value to evaluate the reconstruction performance.

HH model

The dynamics of the ith neuron of an HH network is governed by Sun et al. (2009)

$$\begin{aligned} C\frac{dV_i}{dt} &= -G_L(V_i - V_L) + I_i^{\mathrm{Na}} + I_i^{\mathrm{K}} + I_i^{\mathrm{input}} \\ I_i^{\mathrm{Na}} &= -G_{\mathrm{Na}} m_i^3 h_i (V_i - V_{\mathrm{Na}}) \\ I_i^{\mathrm{K}} &= -G_{\mathrm{K}} n_i^4 (V_i - V_{\mathrm{K}}) \end{aligned}$$

$$\frac{dz_i}{dt} = (1-z_i)\alpha_z(V_i) - z_i\beta_z(V_i), \quad \text{for } z = m, h, n,$$

where C and V_i are the neuron's membrane capacitance and membrane potential (voltage), respectively; m_i, h_i, and n_i are gating variables; $V_{\mathrm{Na}}, V_{\mathrm{K}}$, and V_{L} are the reversal potentials for the sodium, potassium, and leak currents, respectively; $G_{\mathrm{Na}}, G_{\mathrm{K}}$, and G_{L} are the corresponding maximum conductances; and α_z and β_z are the rate variables. The detailed dynamics of the gating variables m, h, n and the choice of parameters can be found in Ref. Dayan and Abbott (2001) and *SI Appendix, Supporting Information Text 3*. The input current $I_i^{\mathrm{input}} = -G_i(t)(V_i - V_E)$ where $G_i(t)$ is the input conductance defined by $G_i(t) = f\sum_l H(t-s_{il}) + \sum_j A_{ij} S \sum_l H(t-\tau_{jl})$ and V_E is the reversal potential of excitation. Here, s_{il} is the lth spike time of the external Poisson input with strength f and rate ν, $\mathbf{A} = (A_{ij})$ is the adjacency matrix with $A_{ij} = 1$ indicating a direct connection from neuron j to neuron i and $A_{ij} = 0$ indicating no connection there. S is the coupling strength, and τ_{jl} is the lth spike time of the jth neuron. The spike-induced conductance change $H(t)$ is defined by $H(t) = \frac{\sigma_d \sigma_r}{\sigma_d - \sigma_r}\left[\exp\left(-\frac{t}{\sigma_d}\right) - \exp\left(-\frac{t}{\sigma_r}\right)\right]\Theta(t)$, where σ_d and σ_r are the decay and rise time scale, respectively, and $\Theta(\cdot)$ is the Heaviside function. When the voltage V_j reaches the firing threshold, V_{th}, the jth neuron generates a spike at this time, say τ_{jl}, and it will induce the ith neuron's conductance change if $A_{ij} = 1$.

LIF model

The dynamics of the ith neuron in a leaky integrate-and-fire (LIF) network is governed by Newhall et al. (2010); Gu et al. (2018)

$$C\frac{dV_i}{dt} = -G_L(V_i - V_L) + I_i^{\mathrm{input}}$$
$$V_i(t) = V_{\mathrm{reset}} \quad \text{if} \quad V_i(t) \geq V_{\mathrm{th}},$$

where C and V_i are the membrane capacitance and membrane potential. V_L and G_L are the reversal potential and conductance for leak currents. Compared with HH model, LIF model drops terms of nonlinear sodium and potassium current, and the input current is given by $I_i^{\mathrm{input}} = f\sum_l \delta(t-s_{il}) + \sum_j A_{ij} S \sum_l \delta\left(t-\tau_{jl}\right)$, where $\delta(\cdot)$ is the Dirac delta function, s_{il} is the lth spike time of the external Poisson input with strength f, and rate ν, τ_{jl} is the lth spike time of jth neuron with strength S. And $\mathbf{A} = (A_{ij})$ is the adjacency matrix defined the same as that in the HH model. When the voltage reaches the threshold V_{th}, the ith neuron will emit a spike to all its connected post-synaptic neurons, and then reset to V_{reset} immediately. In numerical simulation, we use quantities: $C = 1$, $V_{\mathrm{reset}} = -65$ mV, $V_{\mathrm{th}} = -40$ mV, and the leakage conductance is set to be $G_L = 0.05\,\mathrm{ms}^{-1}$ corresponding to the membrane time constant of 20 ms.

Izhikevich model

The dynamics of the ith neuron in an Izhikevich network is governed by Izhikevich (2003)

$$\tau \frac{dV_i}{dt} = 0.04V_i^2 + 5V_i + 140 - u_i + I_i^{\text{input}}$$
$$\frac{du_i}{dt} = a(bV_i - u_i),$$

where V_i is the membrane potential and u_i is the recovery variable describing the force that drives V_i towards resting state. τ is the time constant of V_i. a and b describe the time scale and sensitivity (with respect to V_i) of u_i. I_i^{input} is the driving current containing the external Poisson input and the synaptic input from other neurons, defined by $I_i^{\text{input}} = f\sum_l \exp(t - s_{il})\Theta(t - s_{il}) + \sum_j A_{ij} S \sum_l \exp\left(t - \tau_{jl}\right)\Theta\left(t - \tau_{jl}\right)$, where $\Theta(\cdot)$ is the Heaviside function and parameters are defined similarly as those in the LIF model. When the voltage reaches the threshold V_{th}, the neuron emits a spike to all its post-synaptic neurons, and then reset V_i to c, and u_i to $u_i + d$. In numerical simulation, we use quantities: $\tau = 1$ ms, $a = 0.02\,\text{ms}^{-1}$, $b = 0.2\,\text{mV}^{-1}$, $c = -65$ mV, $d = 8$, $V_{\text{th}} = 30$ mV.

FitzHugh-Nagumo model

The dynamics of the ith neuron in a FitzHugh-Nagumo network is governed by FitzHugh (1961)

$$\tau_V \frac{dV_i}{dt} = V_i - \frac{V_i^3}{3} - W_i + I_i^{\text{input}}$$
$$\tau_W \frac{dW_i}{dt} = V_i + a - bW_i,$$

where V_i and W_i describe the membrane potential and recovery variable, respectively. τ_V and τ_W are their corresponding time constants. I_i^{input} is the driving current containing the external Poisson input and the synaptic input from other neurons, defined by $I_i^{\text{input}} = f\sum_l \delta(t - s_{il}) + \sum_j A_{ij} S \sum_l \delta\left(t - \tau_{jl}\right)$, where parameters are defined similarly as those in the LIF model. When the voltage reaches the threshold V_{th}, the neuron emits a spike to all its post-synaptic neurons. In numerical simulation, we set quantities: $a = 0.7$, $b = 0.8$, $V_{\text{th}} = 0$, $\tau_V = 1\,\text{ms}$, and $\tau_W = 12.5\,\text{ms}$.

Morris-Lecar model

The dynamics of the ith neuron in a Morris-Lecar network is governed by Morris and Lecar (1981)

$$\begin{aligned} C\frac{dV_i}{dt} =& -G_{\text{Ca}} M_i^{\infty}\left(V_i - V_{\text{Ca}}\right) - G_K W_i\left(V_i - V_{\text{K}}\right) \\ & - G_{\text{L}}\left(V_i - V_{\text{L}}\right) + I_i^{\text{input}} \\ \frac{dW_i}{dt} =& \frac{1}{\tau_i^W}\left(W_i^{\infty} - W_i\right), \end{aligned}$$

where C and V_i are the neuron's membrane capacitance and membrane potential, respectively; $V_{\text{Ca}}, V_{\text{K}}$, and V_{L} are the reversal potentials for the calcium, potassium, and leak currents, respectively; $G_{\text{Ca}}, G_{\text{K}}$ and G_{L} are the corresponding maximum conductances; and W_i is the neuron's recovery variable (normalized K^+ conductance). M_i^{∞} and W_i^{∞} are the voltage-dependent equilibrium value of the normalized conductance of calcium and potassium, respectively, defined by

$$M_i^\infty = 0.5\,(1+\tanh[(V_i - V_1)/V_2])$$
$$W_i^\infty = 0.5\,(1+\tanh[(V_i - V_3)/V_4])$$

where V_1, V_2, V_3, and V_4 are the constant parameters. τ_i^W is a voltage-dependent time constant of W_i, defined by

$$\tau_i^W = \tau_0 \left(\cosh \frac{V_i - V_3}{2V_4}\right)^{-1},$$

where τ_0 is a temperature-dependent parameter, fixed as a constant in simulation. I_i^{input} is the driving current containing the external Poisson input and the synaptic input from other neurons, defined by $I_i^{\text{input}} = f\sum_l \delta(t - s_{il}) + \sum_j A_{ij} S \sum_l \delta\left(t - \tau_{jl}\right)$, where parameters are defined similarly as those in LIF model. When the voltage reaches threshold V_{th}, the neuron emits a spike to all its post-synaptic neurons. In numerical simulation, we set parameters as: $G_{\text{Ca}} = 4\,\text{mS}\cdot\text{cm}^{-2}$, $V_{\text{Ca}} = 120\,\text{mV}$, $G_{\text{K}} = 8\,\text{mS}\cdot\text{cm}^{-2}$, $V_{\text{K}} = -80\,\text{mV}$, $G_{\text{L}} = 2\,\text{mS}\cdot\text{cm}^{-2}$, $V_{\text{L}} = -60\,\text{mV}$, $C = 20\,\text{mF}\cdot\text{cm}^{-2}$, $V_1 = -1.2$ mV, $V_2 = 18$ mV, $V_3 = 12$ mV, $V_4 = 17.4$ mV, $\tau_0 = 15$ ms, and $V_{\text{th}} = 0$ mV.

Neurophysiological data

The public spike train data is from Allen brain observatory Allen Institute for Brain Science (2016), accessed via the Allen Software Development Kit (AllenSDK) Allen Institute for Brain Science (2018). Specifically, the data labeled with sessions-ID 715093703 was analyzed in this work. The 118-day-old male mouse passively received multiple visual stimuli from one of four categories, including drift gratings, static gratings, natural scenes and natural movies. The single neuronal activities, i.e. spike trains, were recorded from multiple brain areas, including APN, CA1, CA3, DG, LGd, LP, PO, VISam, VISl, VISp, VISpm, and VISrl, using 6 Neuropixel probes. For each category of stimulus, the recording lasts for more than 20 minutes. 884 sorted spike trains were recorded and 156 of those were used for causality analysis with signal-to-noise ratio greater than 4 and mean firing rate greater than 0.08 Hz.

Data Availability

Implementation of causality estimation, network simulation, and processing scripts for Allen Brain Observatory data are all available through GitHub (`https://github.com/neoneuron/causal4`).

Acknowledgements

This work was supported by Science and Technology Innovation 2030 - Brain Science and Brain-Inspired Intelligence Project with Grant No. 2021ZD0200204 and the Lingang Laboratory Grant No. LG-QS-202202-01 (S.L., D.Z.,); National Natural Science Foundation of China Grant 12271361 (S.L.); National Natural Science Foundation of China with Grant No. 12071287, 12225109 (D.Z.), Shanghai Municipal Science and Technology Major Project 2021SHZDZX0102 and the Student Innovation Center at Shanghai Jiao Tong University (Z.K.T, K.C., S.L., D.Z.)

References

Abbott LF. (1999). Lapicque's introduction of the integrate-and-fire model neuron (1907). Brain research bulletin *50*, 303—304.

Allen Institute for Brain Science (2016). Allen brain observatory.

Allen Institute for Brain Science (2018). Allen SDK documentation.

Bach F. (2017). Breaking the curse of dimensionality with convex neural networks. Journal of Machine Learning Research *18*, 1–53.

Barnett L., Barrett A.B., and Seth A.K. (2009). Granger causality and transfer entropy Are equivalent for gaussian variables. Physical Review Letters *103*, 238701.

Barnett L., Barrett A.B., and Seth A.K. (2018). Solved problems for Granger causality in neuroscience: A response to Stokes and Purdon. Neuroimage *178*, 744–748.

Bedenbaugh P., and Gerstein G.L. (1997). Multiunit normalized cross correlation differs from the average single-unit normalized correlation. Neural Computation *9*, 1265–1275.

Benesty J., Chen J., Huang Y., and Cohen I. (2009). Pearson correlation coefficient. In Springer Top. Signal Process. (Springer), vol. 2, pp. 1–4.

Borge-Holthoefer J., Perra N., Gonçalves B., González-Bailón S., Arenas A., Moreno Y., and Vespignani A. (2016). The dynamics of information-driven coordination phenomena: A transfer entropy analysis. Science Advances *2*, e1501158.

Bossomaier T., Barnett L., Harré M., and Lizier J.T. (2016). An Introduction to Transfer Entropy: Information Flow in Complex Systems (Springer).

Bressler S.L., and Seth A.K. (2011). Wiener-granger causality: A well established methodology. Neuroimage *58*, 323–329.

Buzsáki G., and Mizuseki K. (2014). The log-dynamic brain: How skewed distributions affect network operations. Nature Reviews Neuroscience *15*, 264–278.

Carter J.V., Pan J., Rai S.N., and Galandiuk S. (2016). ROC-ing along: Evaluation and interpretation of receiver operating characteristic curves. Surg. (United States) *159*, 1638–1645.

Chen Y., Rosen B.Q., and Terrence J. Sejnowski (2022). Dynamical differential covariance recovers directional network structure in multiscale neural systems. Proceedings of the National Academy of Sciences *119*, e2117234119.

Dayan P., and Abbott L.F. (2001). Theoretical Neuroscience, vol. 806 (Cambridge, MA: MIT Press).

De Smet R., and Marchal K. (2010). Advantages and limitations of current network inference methods. Nature Reviews Microbiology *8*, 717–729.

Eisen M.B., Spellman P.T., Brown P.O., and Botstein D. (1998). Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences of the United States of America *95*, 14863–14868.

Fawcett T. (2006). An introduction to ROC analysis. Pattern Recognition Letters *27*, 861–874.

FitzHugh R. (1961). Impulses and physiological states in theoretical models of nerve membrane. Biophysical Journal *1*, 445–466.

Frenzel S., and Pompe B. (2007). Partial mutual information for coupling analysis of multivariate time series. Physical Review Letters *99*, 204101.

Gourévitch B., and Eggermont J.J. (2007). Evaluating information transfer between auditory cortical neurons. Journal of Neurophysiology *97*, 2533–2543.

Granger C.W.J. (1969). Investigating causal relations by econometric models and cross-spectral methods. Econometrica : journal of the Econometric Society *37*, 424.

Gu Q., Tian Z., Kovacic G., and Cai D. (2018). The dynamics of balanced spiking neuronal networks under poisson drive is not chaotic. Frontiers in Computational Neuroscience *12*, 47.

Guo S., Wu J., Ding M., and Feng J. (2008). Uncovering interactions in the frequency domain. Plos Computational Biology *4*, e1000087.

Honey C.J., Sporns O., Cammoun L., Gigandet X., Thiran J.P., Meuli R., and Hagmann P. (2009). Predicting human resting-state functional connectivity from structural connectivity. Proceedings of the National Academy of Sciences *106*, 2035–2040.

Honey C.J., Thivierge J.P., and Sporns O. (2010). Can structure predict function in the human brain? Neuroimage *52*, 766–776.

Ito S., Hansen M.E., Heiland R., Lumsdaine A., Litke A.M., and Beggs J.M. (2011). Extending transfer entropy improves identification of effective connectivity in a spiking cortical network model. PLoS One *6*, e27431.

Izhikevich E.M. (2003). Simple model of spiking neurons. IEEE Transactions on Neural Networks *14*, 1569–1572.

Koch M.A., Norris D.G., and Hund-Georgiadis M. (2002). An investigation of functional and anatomical connectivity using magnetic resonance imaging. Neuroimage *16*, 241–250.

Li S., Xiao Y., Zhou D., and Cai D. (2018). Causal inference in nonlinear systems: Granger causality versus time-delayed mutual information. Physical Review E: Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics *97*, 52216.

Marbach D., Prill R.J., Schaffter T., Mattiussi C., Floreano D., and Stolovitzky G. (2010). Revealing strengths and weaknesses of methods for gene network inference. Proceedings of the National Academy of Sciences of the United States of America *107*, 6286–6291.

McLaughlin D.W., Shapley R., Shelley M., and Wielaard D.J. (2000). A neuronal network model of macaque primary visual cortex (V1): Orientation selectivity and dynamics in the input layer 4C alpha. Proceedings of National Academy of Sciences of the United States of America *97*, 8087–8092.

McLaughlin P.W., Narayana V., Kessler M., McShan D., Troyer S., Marsh L., Hixson G., and Roberson P.L. (2004). The use of mutual information in registration of CT and MRI datasets post permanent implant. Brachytherapy *3*, 61–70.

Mišić B., and Sporns O. (2016). From regions to connections and networks: New bridges between brain and behavior. Current Opinion in Neurobiology *40*, 1–7.

Morris C., and Lecar H. (1981). Voltage oscillations in the barnacle giant muscle fiber. Biophysical Journal *35*, 193–213.

Newell E.W., and Cheng Y. (2016). Mass cytometry: Blessed with the curse of dimensionality. Nature Immunology *17*, 890–895.

Newhall K.A., Kovačič G., Kramer P.R., and Cai D. (2010). Cascade-induced synchrony in stochastically driven neuronal networks. Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys. *82*, 41903.

Runge J., Heitzig J., Petoukhov V., and Kurths J. (2012). Escaping the curse of dimensionality in estimating multivariate transfer entropy. Physical Review Letters *108*, 258701.

Schiele F., Van Ryn J., Canada K., Newsome C., Sepulveda E., Park J., Nar H., and Litzenburger T. (2013). A specific antidote for dabigatran: Functional and structural characterization. Blood *121*, 3554–3562.

Schreiber T. (2000). Measuring information transfer. Physical Review Letters *85*, 461–464.

Seth A. (2005). Causal connectivity of evolved neural networks during behavior. Netw. Comput. Neural Syst. *16*, 35–54.

Seth A.K. (2008). Causal networks in simulated neural systems. Cognitive neurodynamics *2*, 49–64.

Stevenson I.H., Rebesco J.M., Miller L.E., and Körding K.P. (2008). Inferring functional connections between neurons. Current Opinion in Neurobiology *18*, 582–588.

Stosiek C., Garaschuk O., Holthoff K., and Konnerth A. (2003). In vivo two-photon calcium imaging of neuronal networks. Proceedings of the National Academy of Sciences of the United States of America *100*, 7319–7324.

Stringer C., Pachitariu M., Steinmetz N., Reddy C.B., Carandini M., and Harris K.D. (2019). Spontaneous behaviors drive multidimensional, brainwide activity. Science (80-.). *364*.

Suárez L.E., Markello R.D., Betzel R.F., and Misic B. (2020). Linking structure and function in macroscale brain networks. Trends in Cognitive Sciences *24*, 302–315.

Sun Y., Zhou D., Rangan A.V., and Cai D. (2009). Library-based numerical reduction of the Hodgkin–Huxley neuron for network simulation. Journal of Computational Neuroscience *27*, 369–390.

Vastano J.A., and Swinney H.L. (1988). Information transport in spatiotemporal systems. Physical Review Letters *60*, 1773–1776.

Vicente R., Wibral M., Lindner M., and Pipa G. (2011). Transfer entropy-a model-free measure of effective connectivity for the neurosciences. Journal of Computational Neuroscience *30*, 45–67.

Wheeler B.C., and Nam Y. (2011). In vitro microelectrode array technology and neural recordings. Critical Reviews in Biomedical Engineering *39*, 45–61.

Zhou D., Rangan A.V., Mclaughlin D.W., and Cai D. (2013a). Spatiotemporal dynamics of neuronal population response in the primary visual cortex. Proceedings of National Academy of Sciences of the United States of America *110*, 9517–9522.

Zhou D., Xiao Y., Zhang Y., Xu Z., and Cai D. (2013b). Causal and structural connectivity of pulse-coupled nonlinear networks. Physical Review Letters *111*, 54102.

Zhou D., Xiao Y., Zhang Y., Xu Z., and Cai D. (2014). Granger causality network reconstruction of conductance-based integrate-and-fire neuronal systems. PloS one *9*, e87636.

Zou C., Denby K.J., and Feng J. (2009). Granger causality vs. dynamic Bayesian network inference: A comparative study. BMC Bioinformatics *10*, 122.