RESEARCH ARTICLE

Efficient "Shotgun" Inference of Neural Connectivity from Highly Sub-sampled Activity Data

Daniel Soudry[1]*, **Suraj Keshri**[2], **Patrick Stinson**[1], **Min-hwan Oh**[2], **Garud Iyengar**[2], **Liam Paninski**[1]

1 Department of Statistics, Department of Neuroscience, the Center for Theoretical Neuroscience, the Grossman Center for the Statistics of Mind, the Kavli Institute for Brain Science, and the NeuroTechnology Center, Columbia University, New York, New York, United States of America, **2** Department of Industrial Engineering and Operations Research, Columbia University, New York, New York, United States of America

* daniel.soudry@gmail.com

Abstract

Inferring connectivity in neuronal networks remains a key challenge in statistical neuroscience. The “common input” problem presents a major roadblock: it is difficult to reliably distinguish causal connections between pairs of observed neurons versus correlations induced by common input from unobserved neurons. Available techniques allow us to simultaneously record, with sufficient temporal resolution, only a small fraction of the network. Consequently, naive connectivity estimators that neglect these common input effects are highly biased. This work proposes a “shotgun” experimental design, in which we observe multiple sub-networks briefly, in a serial manner. Thus, while the full network cannot be observed simultaneously at any given time, we may be able to observe much larger subsets of the network over the course of the entire experiment, thus ameliorating the common input problem. Using a generalized linear model for a spiking recurrent neural network, we develop a scalable approximate expected loglikelihood-based Bayesian method to perform network inference given this type of data, in which only a small fraction of the network is observed in each time bin. We demonstrate in simulation that the shotgun experimental design can eliminate the biases induced by common input effects. Networks with thousands of neurons, in which only a small fraction of the neurons is observed in each time bin, can be quickly and accurately estimated, achieving orders of magnitude speed up over previous approaches.

Author Summary

Optical imaging of the activity in a neuronal network is limited by the scanning speed of the imaging device. Therefore, typically, only a small fixed part of the network is observed during the entire experiment. However, in such an experiment, it can be hard to infer from the observed activity patterns whether (1) a neuron A directly affects neuron B, or (2) another, unobserved neuron C affects both A and B. To deal with this issue, we propose a “shotgun” observation scheme, in which, at each time point, we observe a small changing

OPEN ACCESS

Citation: Soudry D, Keshri S, Stinson P, Oh M-h, Iyengar G, Paninski L (2015) Efficient "Shotgun" Inference of Neural Connectivity from Highly Sub-sampled Activity Data. PLoS Comput Biol 11(10): e1004464. doi:10.1371/journal.pcbi.1004464

Editor: Matthias Bethge, University of Tübingen and Max Planck Institute for Biologial Cybernetics, GERMANY

Received: December 8, 2014

Accepted: July 9, 2015

Published: October 14, 2015

Data Availability Statement: All code is available on https://github.com/danielso/Shotgun.

Funding: The work of DS was supported by the Gruss Lipper Charitable Foundation (http://www.eglcf.org). The work of LS was supported by 1) The army Research Office Multidisciplinary University Research Initiative W911NF-12-1-0594 (http://www.arl.army.mil/www/default.cfm?page=472), 2) The Defense Advanced Research Projects Agency W91NF-14-1-0269 and N66001-15-C-4032 (http://www.darpa.mil) 3) The National Science Foundation Faculty Early Career Development Program IOS-0641912 (http://www.nsf.gov) 4) The National Science Foundation IIS-1430239 (http://www.nsf.gov). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

subset of the neurons from the network. Consequently, many fewer neurons remain completely unobserved during the entire experiment, enabling us to eventually distinguish between cases (1) and (2) given sufficiently long experiments. Since previous inference algorithms cannot efficiently handle so many missing observations, we develop a scalable algorithm for data acquired using the shotgun observation scheme, in which only a small fraction of the neurons are observed in each time bin. Using this kind of simulated data, we show the algorithm is able to quickly infer connectivity in spiking recurrent networks with thousands of neurons.

This is a *PLOS Computational Biology Methods* paper

Introduction

It is now possible to image hundreds of neurons simultaneously at high spatiotemporal resolution [1] or tens of thousands of neurons at low spatiotemporal resolution [2]. The number of recorded neurons is expected to continue to grow exponentially [3]. This, in principle, provides the opportunity to infer the "functional" (or "effective") connectivity of neuronal networks, *i.e.* a statistical estimate of how neurons are affected by each other, and by a stimulus. The ability to accurately estimate large, possibly time-varying, neural connectivity diagrams would open up an exciting new range of fundamental research questions in systems and computational neuroscience [4]. Therefore, the task of estimating connectivity from neural activity can be considered one of the central problems in statistical neuroscience.

Naturally, such a central problem has attracted much attention in recent years (see section 8). Perhaps the biggest challenge here involves the proper accounting for the activity of unobserved neurons. Despite rapid progress in simultaneously recording activity in massive populations of neurons, it is still beyond the reach of current technology to simultaneously monitor a complete large network of spiking neurons at high temporal resolution. Since connectivity estimation relies on the analysis of the the activity of neurons in relation to their inputs, the inability to monitor all of these inputs can result in persistent errors in the connectivity estimation due to model miss-specification. More specifically, "common input" errors, in which correlations due to shared inputs from unobserved neurons are mistaken for direct, causal connections, plague most naive approaches to connectivity estimation. Developing a robust approach for incorporating the latent effects of such unobserved neurons remains an area of active research in connectivity analysis (see section 8).

In this paper we propose an experimental design which can greatly ameliorate these common-input problems. The idea is simple: if we cannot observe all neurons in a network simultaneously, perhaps we can instead observe many overlapping sub-networks in a serial manner over the course of a long experiment. Then we can use statistical techniques to patch the full estimated network back together, analogous to "shotgun" genetic sequencing [5]. Obviously, it is not feasible to purposefully sample from many distinct sub-networks at many different overlapping locations using multi-electrode recording arrays, since multiple re-insertions of the array would lead to tissue damage. However, fluorescence-based imaging of neuronal calcium [6, 7] (or, perhaps in the not-too-distant future, voltage [8]) makes this approach experimentally feasible.

For example, such a shotgun approach could be highly beneficial and relatively straightforward to implement using a 3D acousto-optical deflector microscope [1]. Using such a microscope, one can scan a volume of 400 × 400 × 500 μm, which contains approximately 8000 cells. In normal use, the microscope's 50kHz sampling rate allows for a frame rate of about 6Hz when scanning the entire volume. Unfortunately, this frame rate is too low for obtaining reliable connectivity estimates, which requires a frame rate of at least 30Hz [9]. However, we can increase the effective frame rate to 30Hz by using a shotgun approach. We simply divide the experimental duration into segments, where in each segment we scan only 20% of the network. As a side benefit of this shotgun approach, photobleaching and phototoxicity (two of the most important limitations on the duration of these experiments [10]) are reduced, since only a subset of the network is illuminated and imaged at any given time.

Connectivity estimation with missing observations is particularly challenging (section 9). Fortunately, as we show here, given the shotgun sampling scheme, we do not have to infer the unobserved spikes. We considerably simplify the network model loglikelihood using the expected loglikelihood approximation [11–13], and a generalized Central Limit Theorem (CLT) [14] argument to approximate the neuronal input as a Gaussian variable when the size of the network is large. This approximate loglikelihood and its gradients depend only on the empiric second order statistics of the spiking process (mean spike rate and spike correlations). Importantly, these approximate sufficient statistics can be calculated, even with partial observations, by simply "ignoring" any unobserved activity (section 3.6).

In order to obtain an accurate estimation of the connectivity, posterior distributions involving this simplified loglikelihood (along with various types of prior information about network connectivity) can be efficiently maximized. Using a sparsity inducing prior on the weights, we demonstrate numerically the effectiveness of our approach on simulated recurrent networks of spiking neurons. First, we demonstrate that the shotgun experimental design can largely eliminate the biases induced by common input effects (section 4). Then, we show that we can quickly infer connectivity for large networks, with a low fraction of neurons observed in each time bin (section 5). For example, our algorithm can be used to infer the connectivity of a sparse network with $O(10^3)$ neurons and $O(10^5)$ connections, given $O(10^6)$ time bins of spike data in which only 10% – 20% of the neurons are observed in each time bin. On a standard laptop, simulating such a network takes about half an hour, while inference takes a few minutes. This is faster than previous approaches by orders of magnitude, even when all spikes are observed (section 6.2). Our parameter scans suggest that our method is robust, and could be used for arbitrarily low observation ratios and an arbitrarily large number of neurons, given long enough experiments. We will discuss the outlook for experimental realizations of the proposed approach below, after presenting the basic methodology and simulated results. The supplementary material S1 Text contains the full details of the mathematical derivations and the numerical simulations.

Methods

1 Preliminaries

1.1 General Notation. A boldfaced letter $\mathbf{x}$ denotes a vector with components x_i, a boldfaced capital letter $\mathbf{X}$ denotes a matrix with components $X_{i,j}$, $\mathbf{X}^{(k)}$ denotes the k-th matrix in a list, and $\mathbf{X}_{\cdot,k}$ ($\mathbf{X}_{k,\cdot}$) the k-th column (row) vector of matrix $\mathbf{X}$. For $\mathbf{X} \in \mathbb{R}^{N\times T}$ we define the empiric average and variance

$$\langle X_{i,t}\rangle_T \quad \triangleq \quad \frac{1}{T}\sum_{t=1}^{T} X_{i,t}\ ;\ \mathrm{Var}_T(X_{i,t}) \triangleq \frac{1}{T}\sum_{t=1}^{T}(X_{i,t} - \langle X_{i,t}\rangle_T)^2$$

Note the above expressions do not depend on t, despite the t index, which is maintained for notational convenience. For any condition A, we make use of $\mathcal{I}\{A\}$, the indicator function ($\mathcal{I}\{A\} = 1$ if A holds, and zero otherwise). We define $\delta_{i,j} \triangleq \mathcal{I}\{i = j\}$, Kronecker's delta function. If $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu},\Sigma)$, then $\mathbf{x}$ is Gaussian random vector with mean $\boldsymbol{\mu}$ and covariance matrix Σ, and we denote its density by $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\Sigma)$.

1.2 Model. We use a discrete-time neural network. The neurons, indexed from $i = 1$ to N, produce spikes in time bins indexed from $t = 1$ to T. The spiking matrix $\mathbf{S}$ is composed of variables $S_{i,t}$ indicating the number of spikes neuron i produces at time bin t. We assume each neuron i generates spikes $S_{i,t} \in \{0,1\}$ according to a Generalized Linear neuron Model (GLM [15–17]), with a logistic probability function

$$P(S_{i,t} = 1|U_{i,t}) = f(U_{i,t}) \triangleq \frac{1}{1 + e^{-U_{i,t}}}\ , \tag{1}$$

depending on the the input $U_{i,t}$ it receives from other neurons, as well as from some external stimulus. Such a logistic function is adequate if any time bin rarely contains more than one spike (this is approximately true if the time bin is much smaller than the average inter-spike interval). The input to all the neurons in the network is therefore

$$\mathbf{U}_{\cdot,t} \triangleq \mathbf{W}\mathbf{S}_{\cdot,t-1} + \mathbf{b} + \mathbf{G}\mathbf{X}_{\cdot,t}\ , \tag{2}$$

where b_i is the (unknown) bias of neuron i; $\mathbf{X} \in \mathbb{R}^{D\times T}$ are the external inputs (with D being the number of inputs); $\mathbf{G} \in \mathbb{R}^{N\times D}$ is the input gain; and $\mathbf{W} \in \mathbb{R}^{N\times N}$ is the (unknown) network connectivity matrix. The diagonal elements $W_{i,i}$ of the connectivity matrix correspond to the post spike filter accounting for the cell's own post-spike effects (*e.g.*, refractory period), while the off-diagonal terms $W_{i,j}$ represent the connection weights from neuron j to neuron i. The bias b_i controls the mean spike probability (firing rate) of neuron i. The external input $\mathbf{X}$ can represent a direct (*e.g.*, light activated ion channels) or sensory stimulation of neurons in the network. The input gain $\mathbf{G}$ is a spatial filter that acts on the input $\mathbf{X}$. We assume that the initial spiking pattern is drawn from some fixed distribution $P(\mathbf{S}_{\cdot,0})$.

To simplify notation, we have assumed in Eq 2 that $\mathbf{U}_{\cdot,t}$ is only affected by spiking activity from the previous time bin ($\mathbf{W}\,\mathbf{S}_{\cdot,t-1}$). However, to include a longer history of the spiking activity, we can simply replace the vector $\mathbf{S}_{\cdot,t-1}$ in Eq 2 with the concatenation of the vectors $\mathbf{S}_{\cdot,t-1}$, . . .,$\mathbf{S}_{\cdot,t-k}$ and obtain similar results.

1.3 Task. Our goal is to infer the connectivity matrix $\mathbf{W}$, biases $\mathbf{b}$ and the stimulus gain $\mathbf{G}$. We assume that we have some prior information on the weights, and that we know N, and the external input $\mathbf{X}$. We noiselessly observe a subset of the generated spikes. For simplicity we initially ignore the problem of inferring spikes from the experimental data, which requires spike sorting or deconvolution of fluorescence traces. Later, we will address this issue of spike inference numerically (see also [9, 18] for a more systematic analysis of this issue). We use a binary matrix $\mathbf{O}$ to indicate which neurons were observed, so

$$O_{i,t} \triangleq \mathcal{I}[S_{i,t}\ \text{was observed}]\ . \tag{3}$$

Practically, if $O_{i,t} = 1$ neuron i was imaged for sufficiently long time and with a high enough frame rate around time bin t so that we can infer whether a spike occurred in time bin t with relative certainty.

2 Analytical results—Bayesian inference of the weights

We use a Bayesian approach to infer the unknown weights. Suppose initially, for simplicity, that all spikes are observed and that there is no external input ($\mathbf{G} = 0$). In this case, the log-posterior of the weights, given the spiking activity, is

$$\ln P(\mathbf{W}|\mathbf{S}, \mathbf{b}) = \ln P(\mathbf{S}|\mathbf{W}, \mathbf{b}) + \ln P_0(\mathbf{W}) + C, \tag{4}$$

where ln $P(\mathbf{S}|\mathbf{W},\mathbf{b})$ is the loglikelihood, $P_0(\mathbf{W})$ is some prior on the weights (we do not assume a prior on the biases **b**), and C is some unimportant constant which does not depend on **W** or **b**. Our aim is to find the Maximum A Posteriori (MAP) estimator for **W**, together with the Maximum Likelihood (ML) estimator for **b**, by solving

$$\max_{\mathbf{W},\mathbf{b}} \ln P(\mathbf{W}|\mathbf{S}, \mathbf{b}). \tag{5}$$

If **S** is fully observed, this problem can be straightforwardly optimized without requiring an approximation (though the optimization procedure can be slow). However, our goal is to provide an estimate when only a subset of **S** is observed. This cannot be easily done using standard method. To see this, we examine the likelihood of a GLM (recalling Eqs (1) and (2)),

$$\begin{aligned} & \ln P(\mathbf{S}|\mathbf{W}, \mathbf{b}) \\ = & \sum_{i=1}^{N}\sum_{t=1}^{T} \ln\left[\frac{e^{S_{i,t}U_{i,t}}}{1 + e^{U_{i,t}}}\right] \end{aligned} \tag{6}$$

$$= \sum_{i=1}^{N}\sum_{t=1}^{T}\left[S_{i,t}U_{i,t} - \ln\left(1 + e^{U_{i,t}}\right)\right]. \tag{7}$$

This likelihood (Eq 7), and its gradients, both contain a sum over weighted spikes in $U_{i,\,t}$ (the $\mathbf{WS}_{\cdot,\,t-1}$ term in Eq 2), that cannot be evaluated if some spikes are missing, unless the missing spikes are accurately inferred (section E in S1 Text). However, methods for inferring these missing spikes are typically slow, and do not scale well.

To circumvent these issues, we will show the loglikelihood can be approximated with a simple form, under a few reasonable assumptions. Importantly, this simple form can be easily calculated even if there are missing observations (the full derivation is in section 2.1). Using an extension of the techniques in [11–13], we develop an approximation to the likelihood based on the law of large numbers (the "expected loglikelihood" approximation) together with a generalized Central Limit Theorem (CLT) argument [14], in which we approximate the neuronal input to be Gaussian near the limit $N \to \infty$; then we calculate the "profile likelihood" $\max_{\mathbf{b}}$ln $P(\mathbf{S}|\mathbf{W},\mathbf{b})$, in which the bias term has been substituted for its maximizing value. The end result is

$$\max_{\mathbf{b}} \ln P(\mathbf{S}|\mathbf{W}, \mathbf{b}) \approx T\sum_{i=1}^{N}\left[\sum_{j=1}^{N}\left[W_{i,j}\Sigma_{i,j}^{(1)}\right] - h(m_i)\sqrt{1 + \frac{\pi}{8}\sum\nolimits_{k,j} W_{i,j}\Sigma_{k,j}^{(0)}W_{i,k}}\right], \tag{8}$$

where we defined the mean spike probability, spike covariance, and the entropy function, respectively:

$$m_i \triangleq \langle S_{i,t} \rangle_T \tag{9}$$

$$\Sigma_{i,j}^{(k)} \triangleq \langle S_{i,t} S_{j,t-k} \rangle_T - m_i m_j \tag{10}$$

$$h(m_i) \triangleq -m_i \ln m_i - (1-m_i) \ln (1-m_i)\,. \tag{11}$$

A few comments:

1. Importantly, the profile loglikelihood (Eq 8) depends only on the first and second order moments of the spikes $\mathbf{m}$ and $\Sigma^{(k)}$ for $k \in \{0,1\}$. When all of the neurons in the network are observed, these moments can be computed directly, and therefore the empirical moments are approximate sufficient statistics, whose value contains all the information needed to compute any estimate of $\mathbf{W}$. As we explain in section 3, these empirical moments can be estimated even if only a subset of the spikes is observed.

2. As we show in section A in S1 Text, the profile loglikelihood (Eq 8) is concave, so it is easy to maximize the log-posterior and obtain the MAP estimate of $\mathbf{W}$. This can be done orders of magnitude faster than in the standard MAP estimate (section 6.2), since Eq 8 does not contain a sum over time, as the original loglikelihood (Eq 7). Moreover, the optimization problem of finding the MAP estimate can be parallelized over the rows of $\mathbf{W}$.

$$\max_{\mathbf{b}} \ln P(\mathbf{W}|\mathbf{S},\mathbf{b}) = \sum_i \max_{\mathbf{b}} \ln P(\mathbf{W}_{i,\cdot}|\mathbf{S},\mathbf{b}), \tag{12}$$

   because the profile loglikelihood (Eq 8) decomposes over the rows of $\mathbf{W}$, as does the L1 prior we will use here (Eq 46).

3. As we show in section A in S1 Text, we can straightforwardly differentiate Eq 8 to analytically obtain the gradient, Hessian, and even the maximizer of this profile loglikelihood, which is the maximum likelihood estimate of $\mathbf{W}$. However, due to the nature of the integral approximation we make in Eq 14, more accurate results are obtained if we first differentiate the original loglikelihood (Eq 7), and then use the expectation approximation (together with the generalized CLT argument). This results in an adjustment of the loglikelihood gradient (section D in S1 Text).

4. A novel aspect of this work is that we apply the Expected LogLikelihood (ELL) approximation to a GLM with a bounded logistic rate function (Eq 1), which allows us to infer connectivity in *recurrent* neural networks. In contrast, previous works that used the ELL approximation [11–13] focused on single neuron responses, with an emphasis on either a Poisson neuron model with an exponential rate function, or simpler linear Gaussian models. Such models are less suitable for recurrent neural networks. Exponential rate functions cause instability, as the activity tends to to diverge, unless both the weights and the time bins are small. Linear networks are not a very realistic model for a neural network, and do not perform well in inferring synaptic connectivity [19].

5. Though we assumed a logistic neuron model (Eq 1), similar results can be derived for any spiking neuron model for which $1-f(x) = f(-x)$. This is explained in section A.3 S1 Text.

6. Though we assumed the network does not have a stimulus ($\mathbf{G} = 0$), one can be incorporated into the inference procedure. To do so, we treat the stimulus $\mathbf{X}_{\cdot,t}$ simply as the activity of additional, fully observed, neurons (albeit $X_{i,t} \in \mathbb{R}$ while $S_{i,t} \in \{0,1\}$). Specifically, we define

a new "spikes" matrix $\mathbf{S}^{\text{new}} \triangleq (\mathbf{S}^\top, \mathbf{X}^\top)^\top$, a new connectivity matrix

$$\mathbf{W}^{\text{new}} \triangleq \begin{pmatrix} \mathbf{W} & \mathbf{G} \\ 0^{D\times N} & 0^{D\times D} \end{pmatrix},$$

and a new observation matrix $\mathbf{O}^{\text{new}} \triangleq (\mathbf{O}^\top, 1^{T\times D})^\top$. Repeating the derivations for $\mathbf{S}^{\text{new}}$,$\mathbf{W}^{\text{new}}$ and $\mathbf{O}^{\text{new}}$, we obtain the same profile loglikelihood. Once it is used to infer $\mathbf{W}^{\text{new}}$, we extract the estimates of $\mathbf{W}$ and $\mathbf{G}$ from their corresponding blocks in $\mathbf{W}^{\text{new}}$.

7. For simplicity and efficiency, we chose to focus on MAP estimates. However, other types of estimators and Bayesian approaches (*e.g.*, MCMC, variational Bayes) might be used with this approximate loglikelihood, and should be explored in future work.

2.1 Derivation of the simplified loglikelihood (Eq 8). Recall Eqs (1) and (2) with $\mathbf{G} = 0$. Combining both for times $t = 1, \cdots, T$ and neurons $i = 1, \ldots, N$, we obtain

$$\begin{aligned}
&\ln P(\mathbf{S}|\mathbf{W},\mathbf{b}) \\
&= \sum_{i=1}^{N}\sum_{t=1}^{T}[S_{i,t}U_{i,t} - \ln(1+e^{U_{i,t}})], \\
&= T\sum_{i=1}^{N}[\langle S_{i,t}U_{i,t}\rangle_T - \langle \ln(1+e^{U_{i,t}})\rangle_T] \\
&\overset{(1)}{\approx}\dot{T}\sum_{i=1}^{N}[\langle S_{i,t}U_{i,t}\rangle_T - \int \ln(1+e^{x})\mathcal{N}(x|\langle U_{i,t}\rangle_T, \mathrm{Var}_T(U_{i,t}))dx] \\
&\overset{(2)}{\approx} T\sum_{i=1}^{N}\left[\langle S_{i,t}U_{i,t}\rangle_T - \sqrt{1+\pi\mathrm{Var}_T(U_{i,t})/8}\,\ln\left(1+\exp\left(\frac{\langle U_{i,t}\rangle_T}{\sqrt{1+\pi\mathrm{Var}_T(U_{i,t})/8}}\right)\right)\right] \\
&\overset{(3)}{=} T\sum_{i=1}^{N}\sum_{j=1}^{N}W_{i,j}\Sigma^{(1)}_{i,j} + m_i b_i \\
&- \sqrt{1+\pi\textstyle\sum_{k,j}W_{i,j}\Sigma^{(0)}_{k,j}W_{i,k}/8}\,\ln\left(1+\exp\left(\frac{\sum_{k=1}^{N}W_{i,k}m_k + b_i}{\sqrt{1+\pi\sum_{k,j}W_{i,j}\Sigma^{(0)}_{k,j}W_{i,k}/8}}\right)\right),
\end{aligned} \tag{13}$$

where we used the following:

1. The neuronal input, as a sum of N variables, converges to a Gaussian distribution, in the limit of large N, under rather general conditions [14]. Formally, we need to make sure these are fulfilled for our approximate method to work, which can become even more challenging with the addition of (arbitrary) external inputs. However, such a generalized CLT-based approximation tends to work quite well even when the neuronal input is not strictly Gaussian [20–22]. This robustness is demonstrated numerically in our simulations.

2. The integral approximation

$$\int_{-\infty}^{\infty}\log(1+e^{x})\mathcal{N}(x|\mu,\sigma^2)dx \approx \sqrt{1+\pi\sigma^2/8}\,\log\left(1+\exp\left(\frac{\mu}{\sqrt{1+\pi\sigma^2/8}}\right)\right), \tag{14}$$

from Eq 8 from [23]. This approximation is valid on a limited range, and is inaccurate for low μ. However, this can be corrected by adjusting by the gradient, as we explain in section D S1 Text.

3. Eq 2 for the neuronal input and Eqs (9)–(10) for the spike statistics, which yields

$$\begin{aligned}\langle U_{i,t}\rangle_T &= \sum_{k=1}^{N} W_{i,k}m_k + b_i\\ \mathrm{Var}_T(U_{i,t}) &= \sum_{k=1}^{N}\sum_{j=1}^{N} W_{i,j}\Sigma^{(0)}_{k,j}W_{i,k}\,.\end{aligned}$$

Though the loglikelihood in Eq 13 has already become tractable (and depends only on the sufficient statistics from Eqs (9)–(10), we can simplify it further by maximizing it over **b**. To do so, we equate the derivative of the simplified loglikelihood (Eq 13) to zero

$$\frac{d}{db_i}\ln P(\mathbf{S}|\mathbf{W},\mathbf{b}) \quad = \quad 0\,.$$

Solving this equation, we obtain

$$b_i = \sqrt{1+\pi\textstyle\sum_{k,j}W_{i,j}\Sigma^{(0)}_{k,j}W_{i,k}/8}\,\ln\left(\frac{m_i}{1-m_i}\right) - \sum_{k=1}^{N}W_{i,k}m_k\,. \tag{15}$$

Substituting this maximizer into Eq 13, we obtain Eq 8.

3 Observation schemes

As we showed in section 2, in order to infer network connectivity, we just need to estimate the first and second empiric spike statistics, defined in Eqs (9)–(10). These statistics cannot be calculated exactly if some observations are missing; in this case they must be estimated, as we discuss in section 3.6 below. First, though, it is useful to discuss a few concrete examples of the partial network observation schemes we are considering (Fig 1). We discuss the pros and cons of each scheme in terms of both inferential and experimental constraints.

3.1 Fixed subset observations. The simplest and most commonly used observation scheme is simply to image a fixed subset of the network. To increase the size of the subset, we must image with a lower frame rate. However, since we observe only a subset of the neurons and neuronal pairs (Fig 1A, 1B), we can estimate only a subset of the mean spike probabilities **m** and spike covariances $\Sigma^{(k)}$. If we attempt to infer connectivity using only these incomplete empirical moments (*i.e.*, by "pretending" the unobserved neurons do not exist), a persistent bias may be present in our estimate. This is due to the common input problem, as discussed in more depth in section 4 below.

3.2 Serial subset observations. An alternative natural approach is to continuously shift the observed subset (Fig 1C, 1D): observe a given volume of tissue, then move the microscope (or specimen) to the left a bit, then repeat this procedure in a scanning fashion. However, under this approach some neuron pairs are still never observed. Specifically, as can be seen in Fig 1D, if the size of the scanning block is k, we do not observe neuron pairs for which $|i–j| > k$ (*i.e.*, $\langle O_{i,\,t}\, O_{j,\,t-1}\rangle_T = 0$ for these pairs). Since we do not observe all pairs, it is not always possible to infer the spike covariances ($\Sigma^{(k)}$, Eq 10), which may be required for inferring connectivity.

3.3 Fully randomized subsets. In order to accurately infer spike covariances, we examine a different observation method. If we randomly generate our observations ($O_{i,\,t} = 1$ with probability p_{obs}, and otherwise $O_{i,\,t} = 0$), then all neuron pairs are uniformly observed (*i.e.*, $\langle O_{i,t}O_{j,t-1}\rangle_T = p^2_{\mathrm{obs}}, \forall i, j$); Fig 1E, 1F. In this case, it is easy to estimate both **m** and $\Sigma^{(k)}$, as

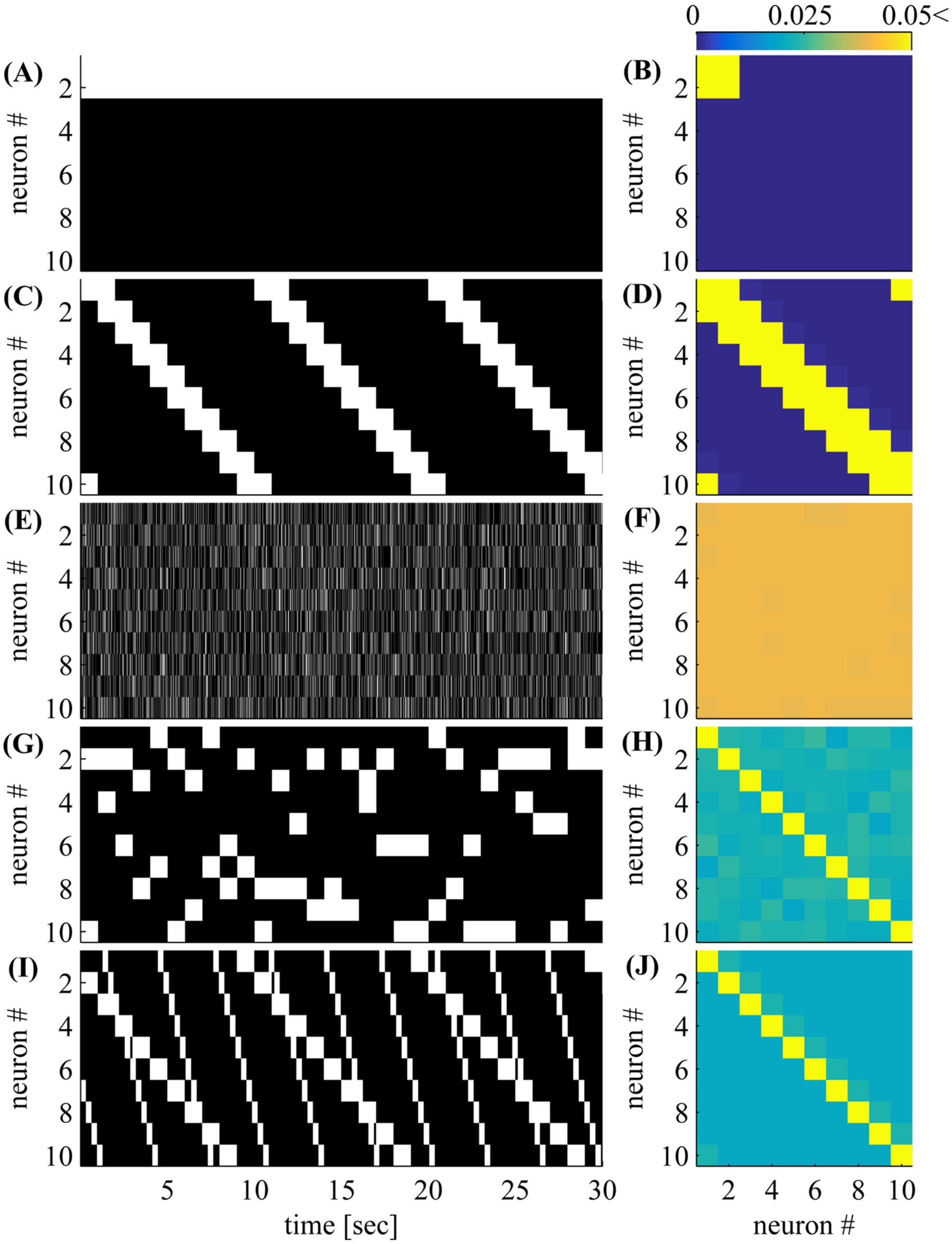

Fig 1. Observation scheme examples. In each scheme (the different rows) we observe two out of ten neurons in each time bin: **(A,B)** Fixed subset **(C,D)** Serial **(E,F)** Fully Random, **(G,H)** Random Blocks, and **(I, J)** double serial. *Left* (A,C,E,G,I): A sample of the observations **O** demonstrating the scanning method (a zero-one matrix, Eq 3). *Right* (B,D,F,H,J): empirical frequency of observed neuron pairs $\langle O_{i,t}O_{j,t-1}\rangle_T = \frac{1}{T}\sum_{t=1}^{T} O_{i,t}O_{j,t-1}$, with saturated colors to accentuate differences between methods (all values above 0.05 are shown in yellow). In the "fixed" scheme, some neurons are never observed. In the "serial" scheme some neuronal pairs are never observed. In all other schemes, all neuronal pairs are observed, so we can estimate the empirical moments using Eqs (16)–(17) and infer connectivity. In the two bottom schemes, observations are collected in persistent blocks, so neuron pairs which are close to the diagonal are observed more often.

doi:10.1371/journal.pcbi.1004464.g001

explained in section 3.6 below. However, this observation scheme is experimentally infeasible, as we cannot infer spikes from fluorescence traces if neurons are observed for too short a time.

3.4 Persistent block observations. To facilitate spike inference from fluorescence traces, we can randomly select a block of $p_{\text{obs}}N$ neurons from the network, and observe this block for a sufficiently long time and with a sufficiently high frame rate, so that all spikes within the block can be inferred accurately (Fig 1G, 1H). Again, we can easily estimate both $\mathbf{m}$ and $\Sigma^{(k)}$ (section 3.6).

Randomly selecting blocks can be technically challenging. Within the field of view, this can be done using an acousto-optical deflector [24] or a spatial light modulator [25]. Enlarging the field of view of these methods remains an open experimental challenge, however.

Of course, non-random block scanning approaches are also possible. An alternative approach could be to employ light sheet methods [2, 26] and then slowly rotate the angle the light sheet forms as it passes through the specimen; as this angle changes, we will collect statistics involving groups of neurons in different planes. The estimation of $\mathbf{m}$ and $\Sigma^{(k)}$ would remain straightforward in this case.

3.5 Double serial scanning. Finally, we note that combinations of the above schemes are possible. One such approach uses two simultaneous serial scans. For example, we can use a lexicographic scheme with two scanners–we first observe a fixed subset with one scanner, and, at the same time, serially observe the remaining blocks of the network with another scanner. Then, we move the first scanner to a different area, and perform another full scan with the second. We continue this way until we have completed a full scan of the network with the first scanner. Alternatively, we do not have to wait until the second scanner has finished a complete scan of the network. Instead, we can continuously scan with both scanners. If the scanning periods are incommensurate (*i.e.*, their ratio is an irrational number) then eventually, we will observe all neuron pairs, as illustrated in Fig 1I, 1J. Such a dual-scanning scheme would of course pose some significant engineering challenges, but recent progress in imaging technology provides some hope that these challenges will be surmountable [27].

3.6 Moment estimation. Next, we explain how the empirical moments can be estimated when there are missing observations. Perhaps the simplest estimate of these mean spike probabilities ignores any missing observations and just re-normalizes the empirical sums accordingly:

$$\tilde{m}_i \triangleq \frac{\langle O_{i,t}S_{i,t}\rangle_T}{\langle O_{i,t}\rangle_T}\,. \tag{16}$$

This estimate is consistent (*i.e.*, $\tilde{m}_i \to m_i$ when $T \to \infty$), since

$$\frac{\langle O_{i,t}S_{i,t}\rangle_T}{\langle O_{i,t}\rangle_T} \overset{(1)}{\to} \frac{\langle O_{i,t}\rangle_T\langle S_{i,t}\rangle_T}{\langle O_{i,t}\rangle_T} \overset{(2)}{\to} \langle S_{i,t}\rangle_T = m_i$$

where we have assumed that

1. The observation process is uncorrelated with the spikes.

2. $\langle O_{i,\,t}\rangle_T$ converges to a strictly positive limit $\forall i$.

The first condition is typically the case in most experiments. The second condition implies we observe each neuron for a large number of time bins; importantly, this condition does *not*

Table 1. Basic notation.

N	Total number of neurons
T	Total number of time bins
p_{obs}	Empiric observation probability—the mean fraction of neurons observed at eachtime bin
p_{conn}	Network sparsity—the average probability that two neurons are directlyconnected
$\mathbf{S}$	$N \times T$ matrix of spike activity
$\mathbf{W}$	$N \times N$ matrix of synaptic connection weights
$\mathbf{U}$	$N \times T$ matrix of neuronal inputs
$\mathbf{b}$	N×1 vector of neuronal biases
$\mathbf{O}$	$N \times T$ binary matrix denoting when neurons are observed
$\mathbf{m}$	N×1 vector of mean spike probability (firing rates)
$\mathbf{\Sigma}^{(k)}$	$N \times N$ matrix of mean spike covariances with lag k

doi:10.1371/journal.pcbi.1004464.t001

hold in the fixed subset observation scheme. Similarly,

$$\tilde{\Sigma}^{(k)}_{i,j} \triangleq \frac{\langle O_{i,t} O_{j,t-k} S_{i,t} S_{j,t-k} \rangle_T}{\langle O_{i,t} O_{j,t-k} \rangle_T} - \tilde{m}_i \tilde{m}_j \rightarrow \Sigma^{(k)} , \tag{17}$$

if we additionally assume that

3. $\langle O_{i,\,t}\, O_{j,\,t-k} \rangle_T$ converges to a strictly positive limit $\forall i, j$ and $\forall k \in \{0,1\}$.

This third condition implies that we observe each neuron pair (either at the same time, or delayed) for a large number of time bins (recall that this condition does not hold in the serial observation scheme).

This direct approach is simple and computationally quite cheap. However, it seems to ignore potentially useful information: if we could "fill in" the activity in the unobserved bins of the $S_{i,\,t}$ matrix, we would increase our effective sample size and therefore estimate the required moments more accurately. We have experimented with a couple approximate Bayesian methods for filling in this missing information (as detailed in more depth in the section E S1 Text), and somewhat surprisingly have found that they do not greatly improve the estimation accuracy, while imposing significant computational cost. See section 6.1 below for further details.

Results

Our goal in this section is to demonstrate numerically that connectivity can be inferred, efficiently and accurately, from highly sub-sampled spike data. Readers can find the basic notation used in this section in Table 1. In section 4, we give a qualitative demonstration that the shotgun approach can be used to significantly decrease the usual persistent bias resulting from common inputs. In section 5, we perform quantitative tests to show that our estimation method is effective and robust; we perform parameter scans for various network sizes, observation probabilities, firing rates and connection sparsities. In section 6, we show that our Expected LogLikelihood (ELL) based estimation method is efficient, both statistically and computationally. Finally, in section 7, we infer connectivity from fluorescence measurements.

4 The common input problem

In this section we use a toy network with $N = 50$ neurons to visualize the common input problem, and its suggested solution—the "shotgun" approach.

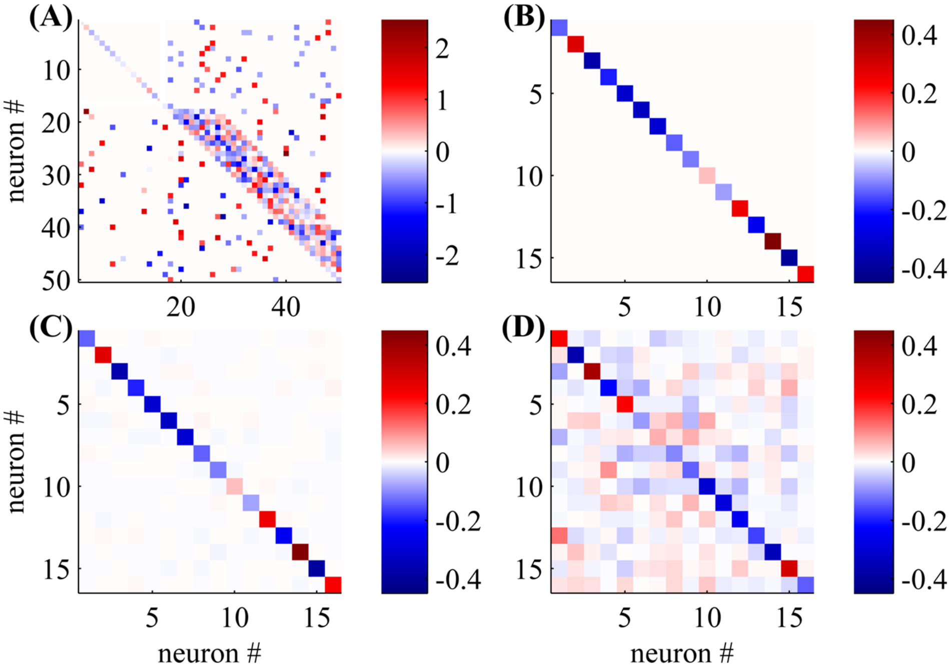

Fig 2. Visualization of the persistence of the common input problem, despite a large amount of spiking data, and its suggested solution—the shotgun approach. (A) The true connectivity—the weight matrix **W** of a network with $N = 50$ neurons. **(B)** A zoomed-in view of the top 16 neurons in A (upper left white rectangle in A). **(C)** The same zoomed-in view of the top 16 neurons in the ML estimate of the weight matrix **W** (Eq 25), where we used the shotgun (random blocks) observation scheme on the whole network, with a random observation probability of $p_{\text{obs}} = 16/50$. **(D)** The ML estimator of the weight matrix **W** of the top 16 neurons if we observe only these neurons. Note the unobserved neurons cause false positives in connectivity estimation. These "spurious connections" do not vanish even when we have a large amount of spike data. In contrast, the shotgun approach (C), does not have these persistent errors, since it spreads the same number of observations evenly over the network. $T = 5{\cdot}10^8$, $b_i \sim \mathcal{N}(-0.5, 0.1)$.

doi:10.1371/journal.pcbi.1004464.g002

Errors caused by common inputs are particularly troublesome for connectivity estimation, since they can persist even as $T \to \infty$. Therefore, for simplicity, we work in a regime where the experiment is long and data is abundant ($T = 5{\cdot}10^8$ timebins). In this regime, any prior information we have on the connectivity becomes unimportant so we simply use the Maximum Likelihood (ML) estimator. We chose the weight matrix **W** to illustrate a "worst-case" common input condition (Fig 2A). Note that the upper-left third of **W** is diagonal (Fig 2B): *i.e.*, neurons $i = 1, \ldots, 16$ share no connections to each other, other than the self-connection terms $W_{i,\,i}$. However, we have seeded this **W** with many common-input motifs, in which neurons i and j (with $i, j \leq 16$) both receive common input from neurons k with $k \geq 17$.

If we use a "shotgun" approach and observe the whole network with $p_{\text{obs}} = 16/50$ with a fully random observation scheme, we obtain a good ML estimate of the network connectivity, including the 16×16 upper-left submatrix (Fig 2C). Now, suppose instead we concentrate all our observations on these 16 neurons, so that $p_{\text{obs}} = 1$ within that sub-network, but the other neurons are unobserved. If common input was not a problem, our estimation quality should improve on that submatrix (since we have more measurements per neuron). However, if common noise is problematic, then we will "hallucinate" many nonexistent connections (i.e., off-diagonal terms) in this submatrix. Fig 2D illustrates this phenomenon. In contrast to the shotgun case, the resulting estimates are significantly corrupted by the common input effects.

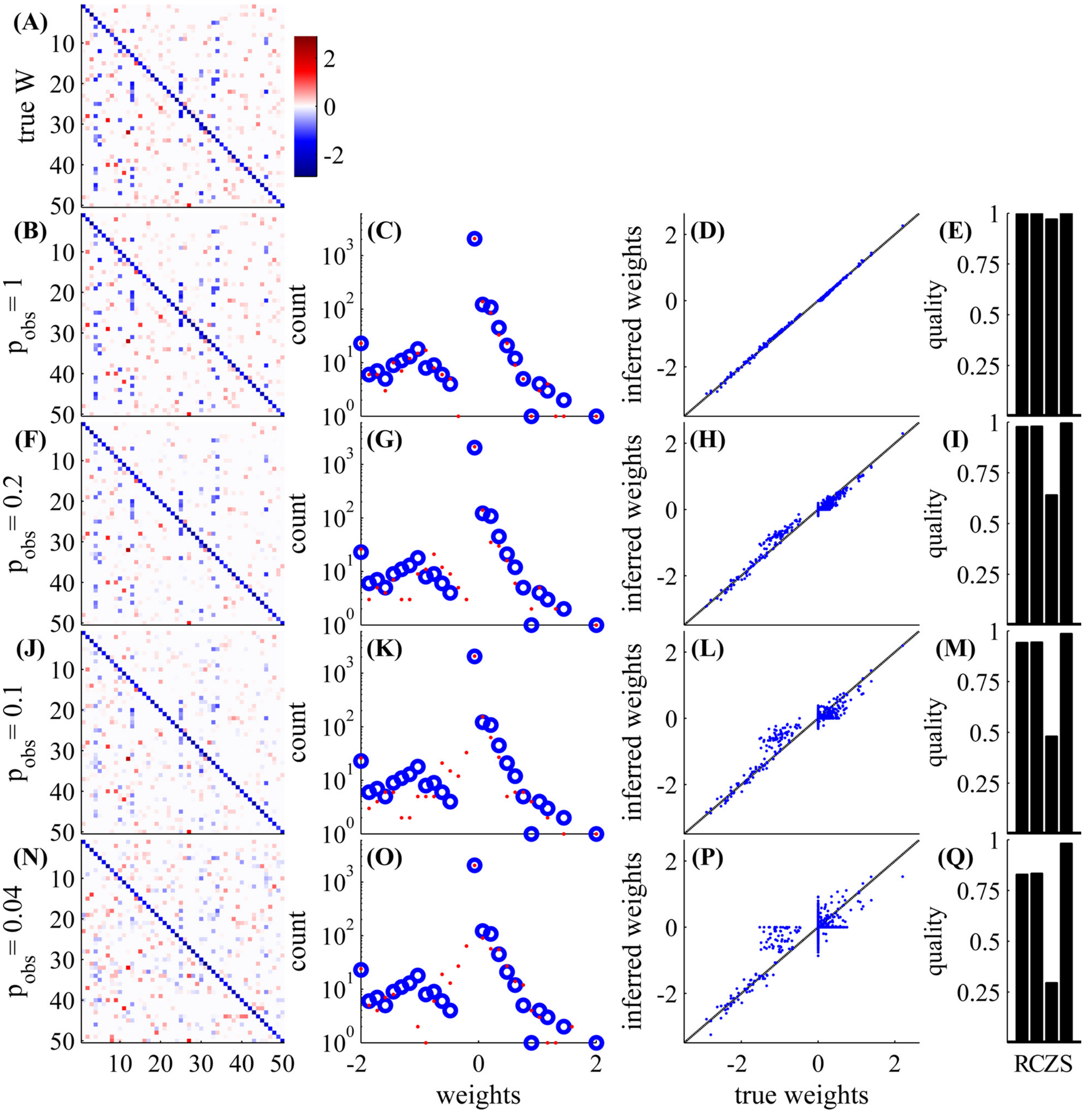

Fig 3. Network connectivity can be well estimated even with low observation ratios. With $N = 50$ neurons, and an experiment length of 5.5 hours, we examine various observation probabilities: $p_{obs} = 1,0.2,0.1,0.04$. *Left* **(A,B,F,J,N)**: weight matrix (either true or estimated). *Middle left* **(C,G,K,O)**: non-zero weights histogram (blue—true, red—estimated). *Middle right* **(D,H,L,P)**: inferred weight vs. true weight. *Right* **(E,I,M,Q)**: quality of estimation—S = sign detection, Z = zero detection, C = correlation, $\mathrm{R} = \sqrt{\mathrm{R}^2}$ (for exact definitions see Eqs 40–43 in S1 Text); higher values correspond to better estimates. In the first row, we have the true weight matrix **W**. In the other rows we have the inferred **W**—the MAP estimate of the weight matrix with the L1 prior (section C in S1 Text), with λ chosen so that the sparsity of the inferred **W** matches that of **W**. Estimation is possible even with very low observation ratios; in the lowest row we observe only 2 neurons out of 50 in each time bin. The weights on the diagonal are estimated better because we observe them more often in double serial scanning scheme (Fig 1I, 1J).

doi:10.1371/journal.pcbi.1004464.g003

5 Connectivity estimation—quantitative analysis

Next, we quantitatively test the performance of the Maximum A Posteriori (MAP) estimate of the network connectivity matrix **W** using a detailed network model with biologically plausible parameters from the mouse visual cortex. Details on the network parameters, simulation details and definitions of the quality measures are given in section B in S1 Text. We use the inference method described in section 2, with a sparsity inducing prior (section C in S1 Text) on a simulated network with GLM neurons (Eqs (1)–(2)).

First, in Fig 3, we examine a small GLM network with $N = 50$ observed neurons, with an experiment length of 5.5 hours. As can be seen, the weight matrix can be very accurately estimated for high values of observation probability p_{obs}, and reasonably well even for low value of p_{obs}. For example, even if $p_{\text{obs}} = 0.04$, and *only two neurons* are observed in each timestep, we get a correlation of $C \approx 0.84$ between inferred weights and the true weights, and the signs of the non-zero weights are only wrong only for 4 weights (out of 448 non-zero weights). When p_{obs} is decreased, the variance of the estimation increases, more weak weights are inferred as zero weights (and vice versa), and we also see more "shrinkage" of the non-diagonal weights (a decreased magnitude of the non-zero weights) due to the $L1$ penalty imposed on them (Eq 47 in S1 Text).

In Fig 4 we demonstrate that our method works well even if the neuron model is not a GLM, as we assume, but a Leaky Integrate and Fire (LIF) neuron model (Fig 4). The model mismatch results in a weight mismatch by a global multiplicative constant, and in a worse estimate of the diagonal weights, due to the hard reset in the LIF model. Besides these issues, inference results are both qualitatively and quantitatively similar to results in the GLM network

In Fig 5, we examine another GLM network with $N = 1000$ observed neurons, which is closer to the scale of the number of recorded neurons in current calcium imaging experiments (see activity simulation in Fig S1 S1 Text). The experiment duration is again 5.5 hours. Results are qualitatively the same as the case of $N = 50$, except performance is somewhat decreased (as we have more parameters to estimate). Additional information is available in Fig 6. On the left (A,D,G), we see that the algorithm converges properly to a single solution. In the middle panels (B,E,H), we see that for $p_{\text{obs}} = 1$ we have very good performance (in terms of area under the ROC), but this performance declines for the excitatory weights as p_{obs} decreases. The inhibitory weights are correctly detected much better than the excitatory weights. This is because most excitatory weights are much weaker, as can be seen on the right column (C,F,I). In that column, we observe that strong weights are more easily detected than weak weights. Specifically, around the median of the excitatory weight distribution (0.178), we detected 99.9%, 34.9% and 16.1% of all the weights, when $p_{\text{obs}} = 1{,}0.2$ and 0.1, respectively.

Next, in Fig 7 we quantify how inference performance changes with parameters. We vary the number of neurons, N, observation probability p_{obs}, mean firing rate m and connection sparsity p_{conn}. For the given parameters N, p_{obs} and m, performance monotonically improves when T increases. These scans suggest we can maintain a good quality of connectivity estimation for arbitrarily large or small values of N or p_{obs}, respectively—as long as we sufficiently increase T. Note there is a lower bound on T, below which estimation does not work. Looking at Fig 7, we find that approximately, this lower bound scales as

$$T \propto \frac{N}{p_{\text{obs}}^2} \,. \tag{18}$$

Above this lower bound, estimation quality gradually improves with T. Moreover, in order to maintain good estimation quality (up to some saturation level) above this bound, T should be

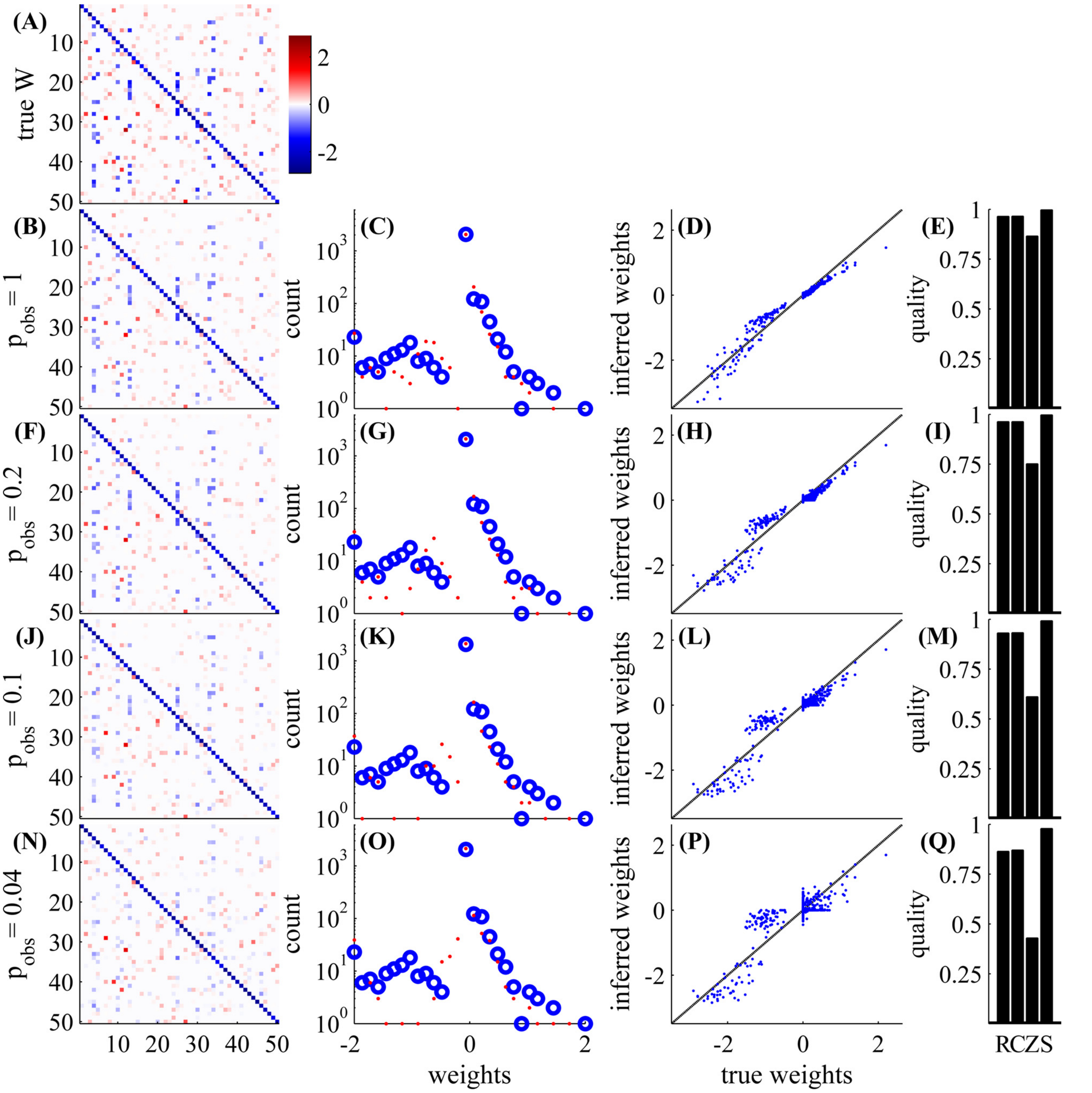

Fig 4. Network connectivity can be reasonably estimated, even with model mismatch. The panels (**A-Q**) are the same as Fig 3, where instead of a logistic GLM (Eq 1), we used a stochastic leaky integrate and fire neuron model (in discrete time). In this model, $V_{i,t} = (\gamma V_{i,t-1}+(1-\gamma)U_{i,t}+\varepsilon_{i,t})\mathcal{I}[S_{i,t-1} = 0]$ (**U** defined in Eq 2), $S_{i,t+1} = \mathcal{I}[V_{i,t} > 0.5]$. We used $\varepsilon_{i,t} \sim \mathcal{N}(0,1)$ as a white noise source. Also, we set $\gamma = 20\text{ms}^{-1}$, similar to the inverse of the membrane's voltage average integration timescale [60]. The weights were estimated up to a global multiplicative constant (resulting from the model mismatch), which was adjusted for in the figure. We conclude that our estimation method is robust to modeling errors, except perhaps the diagonal weights—their magnitudes were somewhat over-estimated due to the reset mechanism (which effectively increases self inhibition).

doi:10.1371/journal.pcbi.1004464.g004

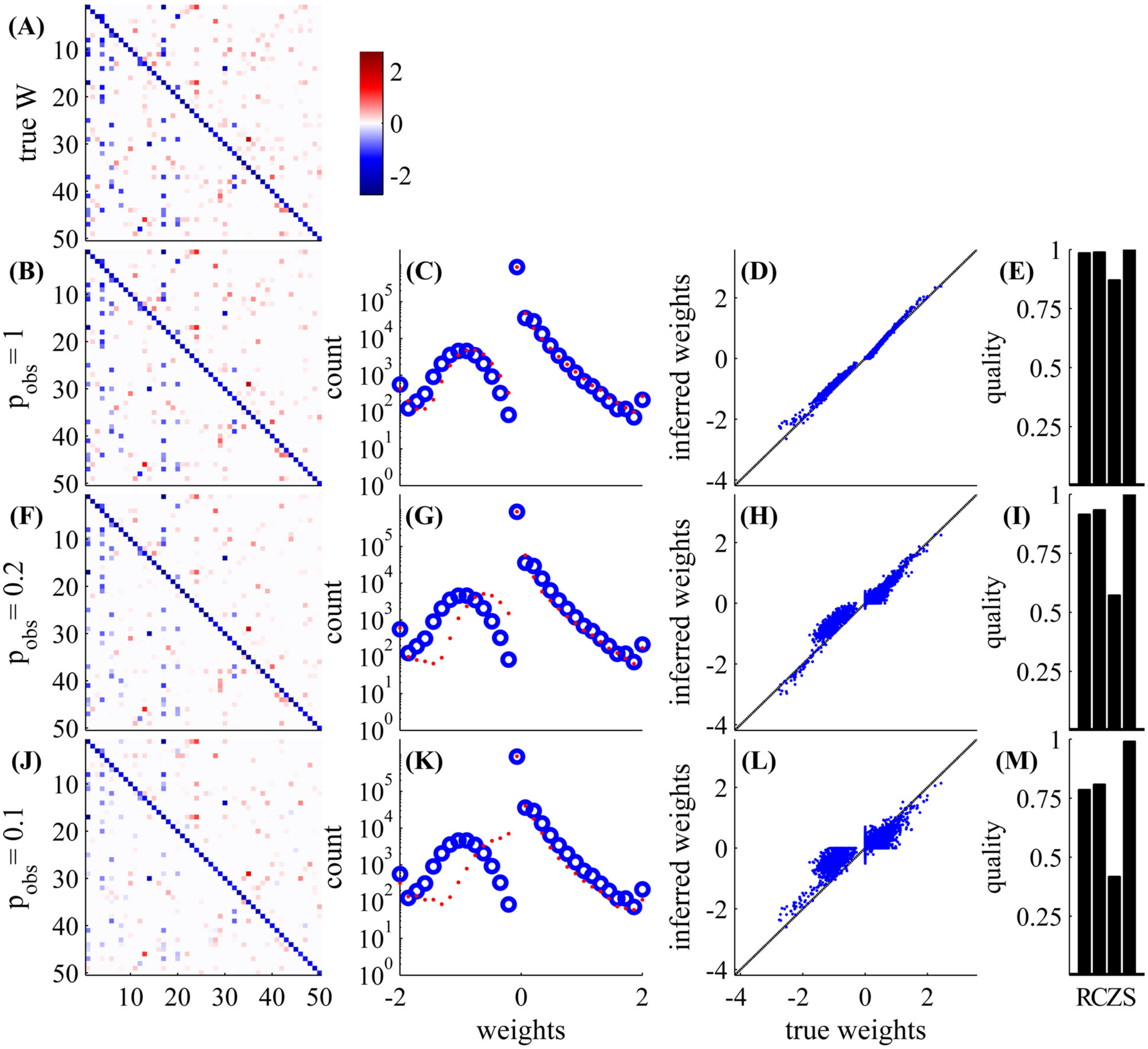

Fig 5. Network connectivity can be well estimated even with a large network. The panels (**A-M**) are arranged in columns as in Fig 3, except now we have $N = 1000$ observed neurons and additional 200 unobserved (this is the same simulation as in Fig S1), and $p_{\text{obs}} = 1$, 0.2 and 0.1. In the left (A,B,F,J) and middle right columns (D,H,L) we show a random subset of 50 neurons out of 1000, to improve visibility. The other columns show statistics for all observed neurons. On a standard laptop, the network simulation takes about half an hour, and the connectivity estimate can be produced in minutes. Therefore, our algorithm is scalable, and much faster than standard GLM based approaches, as we explain in section 6.2.

doi:10.1371/journal.pcbi.1004464.g005

scaled as

$$T \propto \frac{N}{p_{\text{obs}}^2 m^2} . \tag{19}$$

This scaling can be explained intuitively. Our main sufficient statistic is the partially observed spike covariance $\tilde{\mathbf{\Sigma}}^{(k)}$ (Eq 17). Each component (i, j) of $\tilde{\mathbf{\Sigma}}^{(k)}$ contains a sum of all the observed spike pairs ($T\langle O_{i,t} O_{j,t-k} S_{i,t} S_{j,t-k}\rangle_T$) divided by the number of observed neurons ($T\langle O_{i,t} O_{j,t-k}\rangle_T$). The total number of observed neuron pairs is approximately NTp_{obs}^2 (ignoring

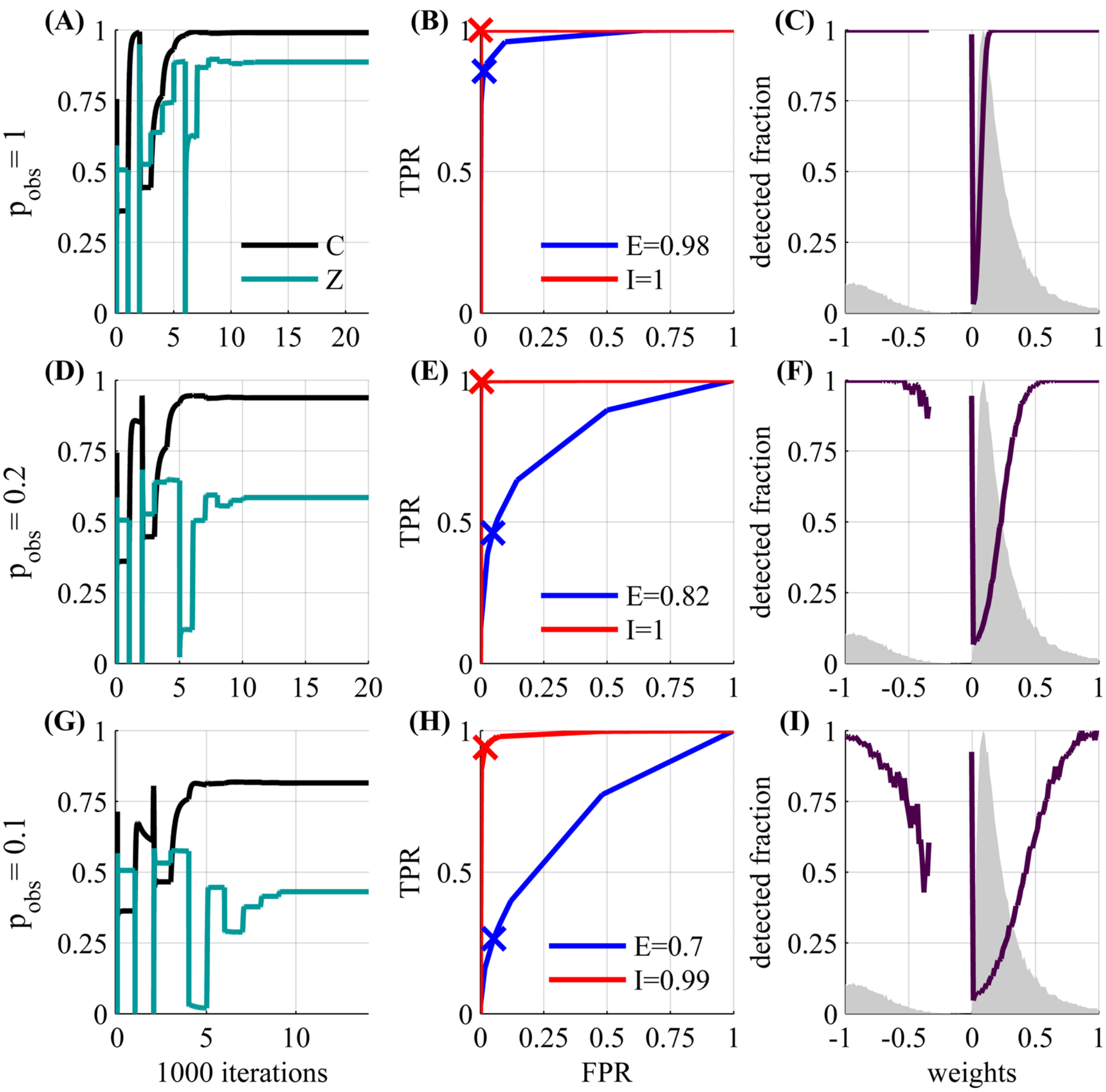

Fig 6. Statistical Analysis. We use the same network as in Fig 5 with $N = 1000$ and $p_{obs} = 1$, 0.2 and 0.1. *Left* **(A,D,G)**: convergence of performance. Recall that we use the FISTA algorithm (section C.1 in S1 Text) inside an outer loop that sets the regularization parameter according to sparsity (section C.2 in S1 Text). Therefore convergence is non-monotonic, and "jumps" each time the parameter is changed. Each time this happens, it takes about a thousand iterations until convergence. *Middle* **(B,E,H)**: Receiver Operating Characteristic (ROC) curve, showing the trade-off between the false positive rate and the true positive rate (FPR and TPR, Eqs 44 and 45 in S1 Text, receptively) in detecting excitatory (blue) or inhibitory weights (red)—*i.e.*, inferring a non zero weight, with the right sign. The curve illustrates the classification performance as the discrimination threshold is varied by changing λ, the L1 regularization parameter (Eq 46 in S1 Text). The 'x' marks the performance for λ chosen by our algorithm. For each case (excitatory/inhibitory), the measures E and I are the area under the curve—values close to 1 (0.5) indicate good (bad) performance. Performance is significantly better for the inhibitory weights, since they are typically stronger, and we can more easily distinguish non-zero weights. We see this explicitly on the *Right* **(C,F,I)**: Magenta line—fraction of weights detected with the correct sign (-1,0, or 1) as a function of weight value. Line stops if less than 30 weight values exist in that range. For clarity, we added the non-zero weight distribution (shaded gray area, scaled to fit panel) and zoomed on the range [−1, 1]. Weights with magnitude larger then 1 were perfectly detected. Small weights are harder to detect at low p_{obs}.

doi:10.1371/journal.pcbi.1004464.g006

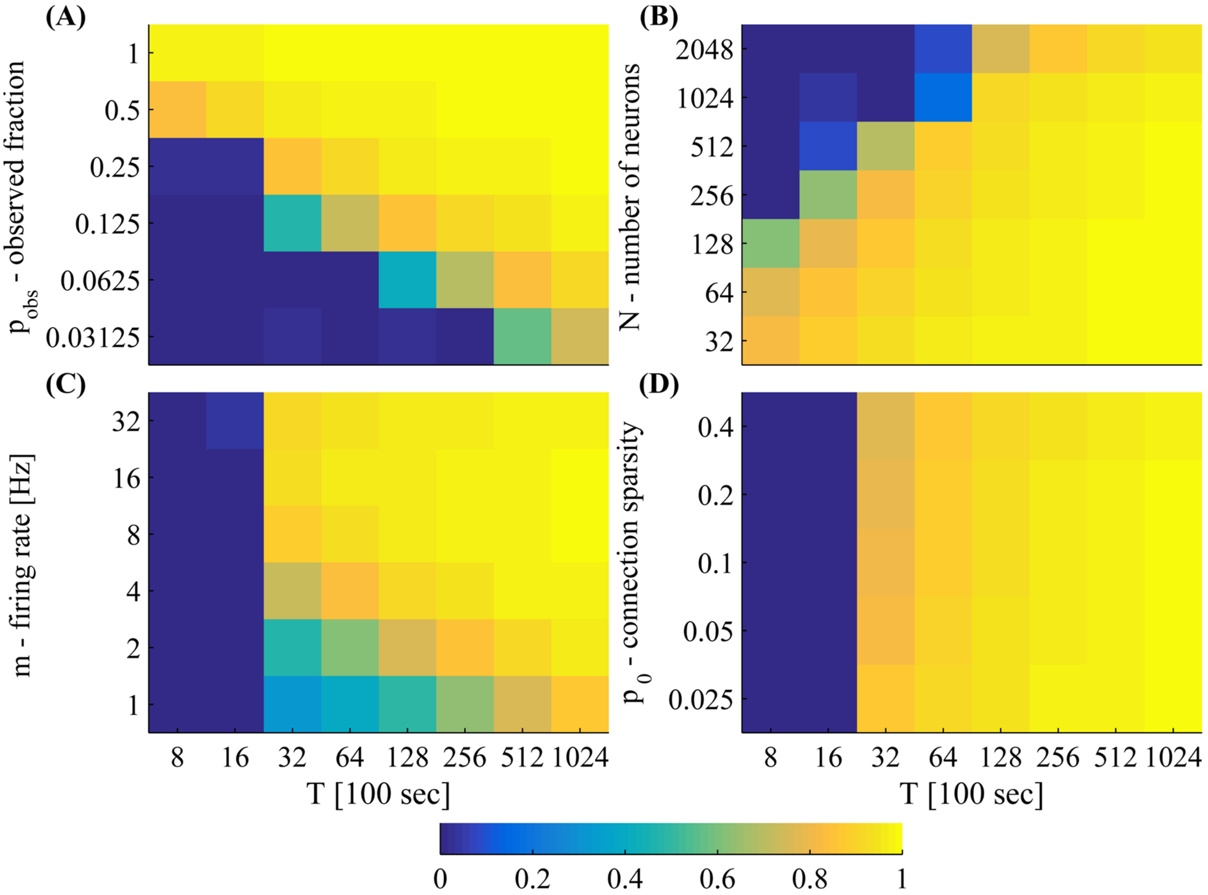

Fig 7. Parameter scans show that *C*, the correlation between true and estimated connectivity, monotonically increases with *T* in various parameter regimes. We scan over **(A)** observation probability p_{obs}, **(B)** network size N, **(C)** mean firing rate m (similar to the firing rate of the excitatory neurons—inhibitory neurons fire approximately twice as fast), and **(D)** connection sparsity parameter p_0 (which is proportional to actual connection sparsity p_{conn}—see Eq 37 in S1 Text) and experiment duration T. Other parameters (when these are not scanned): $p_{\text{obs}} = 0.2$, $N = 500$.

doi:10.1371/journal.pcbi.1004464.g007

observation correlations), and the total number of observed spike pairs is approximately $NTp_{\text{obs}}^2 m^2$ (ignoring spike correlations, and assuming the firing rate is not very high), where T is measured in time bins. The total number of components in $\tilde{\Sigma}^{(k)}$ is N^2. Therefore, in each component of $\tilde{\Sigma}^{(k)}$, the average number of observed neuron pairs is Tp_{obs}^2/N, while the average number of observed spike pairs is approximately $Tp_{\text{obs}}^2 m^2/N$ (except on the diagonal of $\Sigma^{(0)}$, where we have Tp_{obs}/N neuron pairs and $Tp_{\text{obs}}\, m/N$ spikes). We conclude that the number of both observed neuron pairs and spike pairs must be above a certain threshold so that inference will be able to work properly. Above these thresholds, performance improves further when p_{conn} is decreased (Fig 7), as this reduces the effective number of parameters we are required to estimate. For analytic results on this issue see [28].

If our goal is to infer all the input connections of a single neuron, then performance can be significantly improved if we always observe the output of that neuron. This is demonstrated in Fig 8. In this figure, we examine a single neuron with $O(10^4)$ observed inputs, $O(10^3)$ of which are non-zero (implementation details in S1 Text, section B.2). The inputs are partially observed (with $p_{\text{obs}} = 1{,}0.1{,}0.01$), but we always observe the output neuron. Therefore, the average

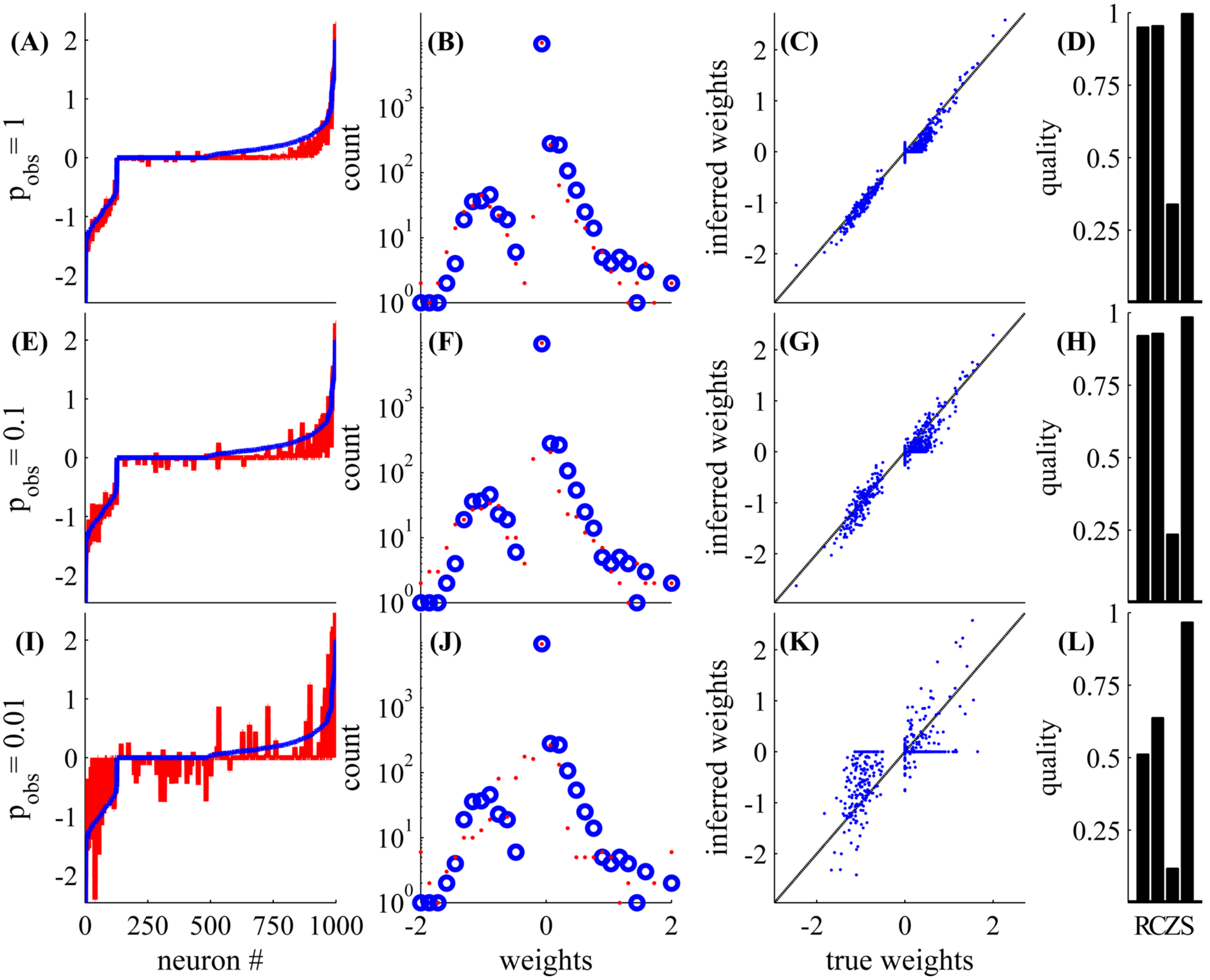

Fig 8. Inferring input connectivity to a single neuron with many inputs and low observation ratios. The panels **B-D**,**F-H**, and **J-L** are arranged in columns as in Fig 3. In the left column **(A,E,I)** we show a sample of 1000 input weight values from the true (blue) and inferred weights (red), sorted according to the value of the true weights. In the other panels, we show all the weights. We have $N = 10626$ observed inputs, 968 of which have non-zero weights. The output neuron is always observed, while 83% of the input neurons are only partially observed with $p_{obs} = 1,0.1,0.01$. The rest are never observed. Other network parameters are the same as before (*e.g.*, $T = 5.5$ hours), and the firing rate of the output neuron was 2.8Hz. More implementation details in S1 Text, section B.2.

doi:10.1371/journal.pcbi.1004464.g008

number of observed neuron pairs and spike pairs in $\Sigma^{(1)}$ is increased to NTp_{obs} and $NmTp_{obs}$, respectively. This can improve the scaling relations in Eqs (18) and (19) to T/p_{obs} and $T/(p_{obs}\, m^2)$, respectively, if the off-diagonal terms of $\Sigma^{(0)}$ are not too strong (since the number of observations for these components still scales with $\propto p_{obs}^2$). Thus, in Fig 8, we see that even when $p_{obs} = 0.01$ it is still possible to estimate strong weights with some accuracy, despite the large number of connections.

6 Expected LogLikelihood-based estimation: accuracy and speed

6.1 Statistical efficiency. The results of the previous section (mostly, Fig 3, for $p_{obs} = 1$, as well as our parameter scans) demonstrate numerically that our Expected LogLikelihood-based (ELL) estimation method is effective given sufficiently large observation times T. However, it is

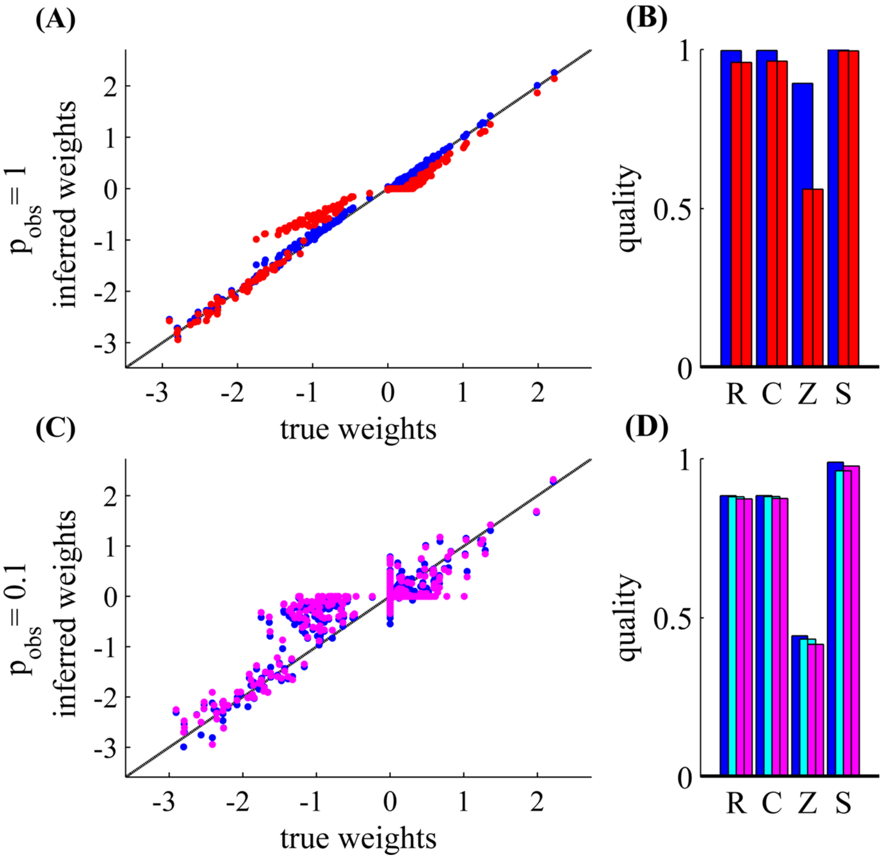

Fig 9. Expected LogLikelihood (ELL) based estimation is statistically efficient. *Top* **(A,B)**: The ELL-based method (blue) compares favorably to the standard MAP estimate (red) when spikes are fully observed (using the same L1 prior). *Bottom* **(C,D)**: We compare the ELL based method (blue) to the Expectation Maximization (EM) approach, when only 10% of the spikes are observed. We show the results after one (cyan) and two (magenta) EM steps. The EM steps do not improve over the ELL-based method. Parameters: $N = 50$, $T = 1.4$ hours. For the Gibbs sampling we used a single sample after a burn-in period of 30 samples, as we used in our EM simulations without the ELL-based initialization (section E S1 Text).

doi:10.1371/journal.pcbi.1004464.g009

still not clear if our approximations hurt the statistical efficiency of our estimation. Specifically, can we get a significantly smaller error with the same T, if we did not use any approximations? We give numerical evidence in Fig 9 that this is not the case. All the parameters are as described in S1 Text, section B.1, except we used a random blocks observation scheme (see Fig 1G, 1H) and did not have any unobserved neurons.

First, we examine the case in which all the spikes are observed ($p_{obs} = 1$). We compare our ELL-based estimate with standard MAP optimization of the full likelihood (with the same L1 prior). We implement the latter by plugging the weight gradients of the loglikelihood (Eq 7) in the same optimization algorithm we use for the ELL-based estimate (section C.1 in S1 Text), together with the bias gradients. As can be seen in Fig 9A, 9B, the approximate ELL-based MAP estimate (blue) actually slightly outperforms the accurate MAP estimate (red, which exhibits more shrinkage). These results support the validity of our approximations (See [13] for further discussion of how the ELL approximation can in some cases improve the MAP estimate).

Next, we demonstrate numerically that we can safely ignore missing spikes without increasing estimation error, when $p_{\text{obs}} < 1$. Specifically, we want to verify the efficiency of using the "partial" empirical moments ($\tilde{\mathbf{m}}$ and $\tilde{\boldsymbol{\Sigma}}$) instead of the full sufficient statistics ($\mathbf{m}$ and Σ), as detailed in section 3.6. These full sufficient statistics can be potentially inferred "correctly" using Bayesian inference techniques such as Markov chain Monte Carlo (MCMC) (section E in S1 Text). In Fig 9C, 9D, we demonstrate that inferring these missing spikes does not improve our estimation. We compare the ELL-based estimate (blue), to the estimate produced by initializing with the ELL-based estimate and then performing one (red) and two (magenta) Expectation-Maximization (EM) steps [29]. In more detail, to perform the first EM step we first estimate $\mathbf{W}$ using the ELL-based method, then Gibbs sample the missing spikes (section E.1.1 in S1 Text), assuming this $\mathbf{W}$ is the true connectivity, and then infer $\mathbf{W}$ again using the ELL-based method. For the second step, we initialize $\mathbf{W}$ with the first step, Gibbs sample the missing spikes, and infer $\mathbf{W}$, again using the ELL-based method. This Monte Carlo EM procedure should converge to a local optimizer of the full log-posterior, assuming a sufficient number of MCMC samples are obtained in each iteration [29]. Nonetheless, empirically, we see that these EM steps do not help to improve estimation quality here. In addition, using the standard MAP estimate of $\mathbf{W}$ (instead of the ELL-based estimate) in the EM steps does not qualitatively change these results (data not shown).

6.2 Computational efficiency. An important advantage of the ELL-based method is that it enables extremely fast MAP estimation of the weights (Eq 5). In the standard MAP estimate, we need to calculate all the N^2 components of the gradient of the original loglikelihood (Eq 7). In total, this requires $O(N^3\,T)$ operations in each iteration of the optimization procedure. We also find $O(N^3\,T)$ operations-per-step in other standard estimation methods we tested for $\boldsymbol{W}$ (MCMC and variational Bayes, see section E.1.2 in S1 Text). In contrast, in the ELL-based method, the first step is to calculate $\tilde{\mathbf{m}}$ and $\tilde{\boldsymbol{\Sigma}}$, which takes $O(N^2\,T)$ operations, but we only need to do this once (this usually takes much less time than the simulation of the network activity, which also takes $O(N^2\,T)$ operations). Once these are calculated, we need only $O(N^3)$ operations to calculate the loglikelihood (or its gradient) in each iteration of the optimization algorithm. This results in orders of magnitude improvements in estimation speed over the standard MAP estimate from the original loglikelihood.

For example, in the simulation we show in Fig 9A, 9B, (where $N = 50$ and $T = 5{\cdot}10^5$ time bins) it takes about 11 seconds to run the optimization algorithm using the ELL-based method: approximately one second to calculate the empirical moments, and 10 seconds for the optimization algorithm to converge. A single step of calculating the gradient and updating the weights (in the internal FISTA loop) took 0.002 sec. A similar step took 0.7 sec in the standard MAP estimation. In total, the algorithm took about 4 hours to converge (taking more iterations than the ELL-based method).

As another example, in [30], it takes $O(10^5)$ CPU hours using a computer cluster to estimate the connectivity of a network (where $N = 1000$ and $T = 3.6 \cdot 10^7$ time bins). In our case (where $N = 1000$ and $T = 2 \cdot 10^6$ time bins), a similar simulation on a standard laptop (Fig 5) takes about half an hour to generate the spikes, together with the sufficient statistics, and a few more minutes to perform the estimation for a given prior distribution. While our model is slightly simpler than that of [30], most of this massive improvement in speed is due to the differences in the inference methods used.

Lastly, we note that Gibbs sampling the spikes also requires $O(N^3\,T)$ operations in each time step (Eq 78 in S1 Text). For example, in the simulation behind Fig 9C, 9D, each Gibb steps for all the spikes took about *2.5* minutes. All the steps took in total about 80 minutes. Therefore, ignoring the missing spikes, instead of sampling them, greatly improves computational speed.

7 Fluorescence-based inference

In a real imaging experiment, we would not have direct access to spikes, as we have assumed for simplicity so far. Next, we test the estimation quality when we only have direct access to the fluorescence traces of activity (Fig 10A). The fluorescence traces were generated using a model of GCaMP6f calcium fluorescence indicator. Implementation details are described in section B.3 in S1 Text. As can be seen in Fig 10A, 10B, our spike inference algorithm works reasonably well, both in high and low noise regimes. We then infer network connectivity both from the inferred spikes and the true spikes. As can be seen in Fig 10C-10H, using the inferred spikes usually somewhat reduces estimation performance. This is due to the temporal inaccuracy in the spike estimation. For example, in the inhibitory neurons, the higher firing rates result in more missing spikes in the inference. This causes shrinkage in the magnitude of the inferred weights, since the cross-correlation is weakened by these missing spikes. Combining this information into the inference algorithm (as in [9]), it may be possible to correct for this; we have not pursued this question further here. However, even at low observation probabilities (p_{obs} = 0.1), strong weights are inferred reasonably well, and the sign of synapse is usually inferred correctly for almost all nonzero weights. Therefore, weight inference is still possible at low firing rates, using current generation fluorescence imaging methods.

Discussion

8 Previous works

Neural connectivity inference has attracted much attention in recent years. One approach to this problem is direct anatomical tracing [31, 32]. However, this method is computationally challenging [33]; moreover, the magnitudes of the synaptic connections (which also vary over time [34]) currently cannot be inferred this way. Another approach, on which we focus here, aims to infer the synaptic connectivity from neural activity. This activity can be either action potentials (“spikes”) [16, 35–37] or calcium fluorescence traces [9, 18, 38–40] which are approximately a noisy and filtered version of the spikes.

Various inference procedures have been suggested for this purpose. Some works use model-free empirical scores [38, 40, 41]. Others assume an explicit generative model for the network activity [16, 30, 35–37], and then infer connectivity by estimating model parameters. So far, only few works have validated the connectivity estimate with some form of “ground truth”. Gerhard et al. [19] inferred small scale anatomical connectivity, comparing different methods. A Generalized Linear Model (GLM) approach was successful, while linear models and model-free approaches failed. Volgushev et al. [42] estimated the weights of fictitious synapses (injected current). Again, a GLM-based approach outperformed simple correlation-based approaches. Lastly, Latimer et al. [43] was able to infer the magnitude of intracellular synaptic conductances, using a modified GLM. These results indicate that a GLM-based approach should be the method of choice for estimating synaptic connectivity.

The task of inferring synaptic connectivity is severely hindered by technical limitations on the number of neurons that can be simultaneously observed with sufficient quality. Typically, the scanning speed of the imaging device is limited, so we cannot cover the entire network with a high enough frame rate and signal-to-noise ratio to infer spikes from the observed fluorescence traces. Previous studies indicate that at low frame rates (below 30Hz [9]), synaptic connectivity cannot be inferred. In such low frame rate regimes, one may use spike correlations or simple dynamical systems as a coarse measure of effective connectivity (*e.g.*, [44]), but such

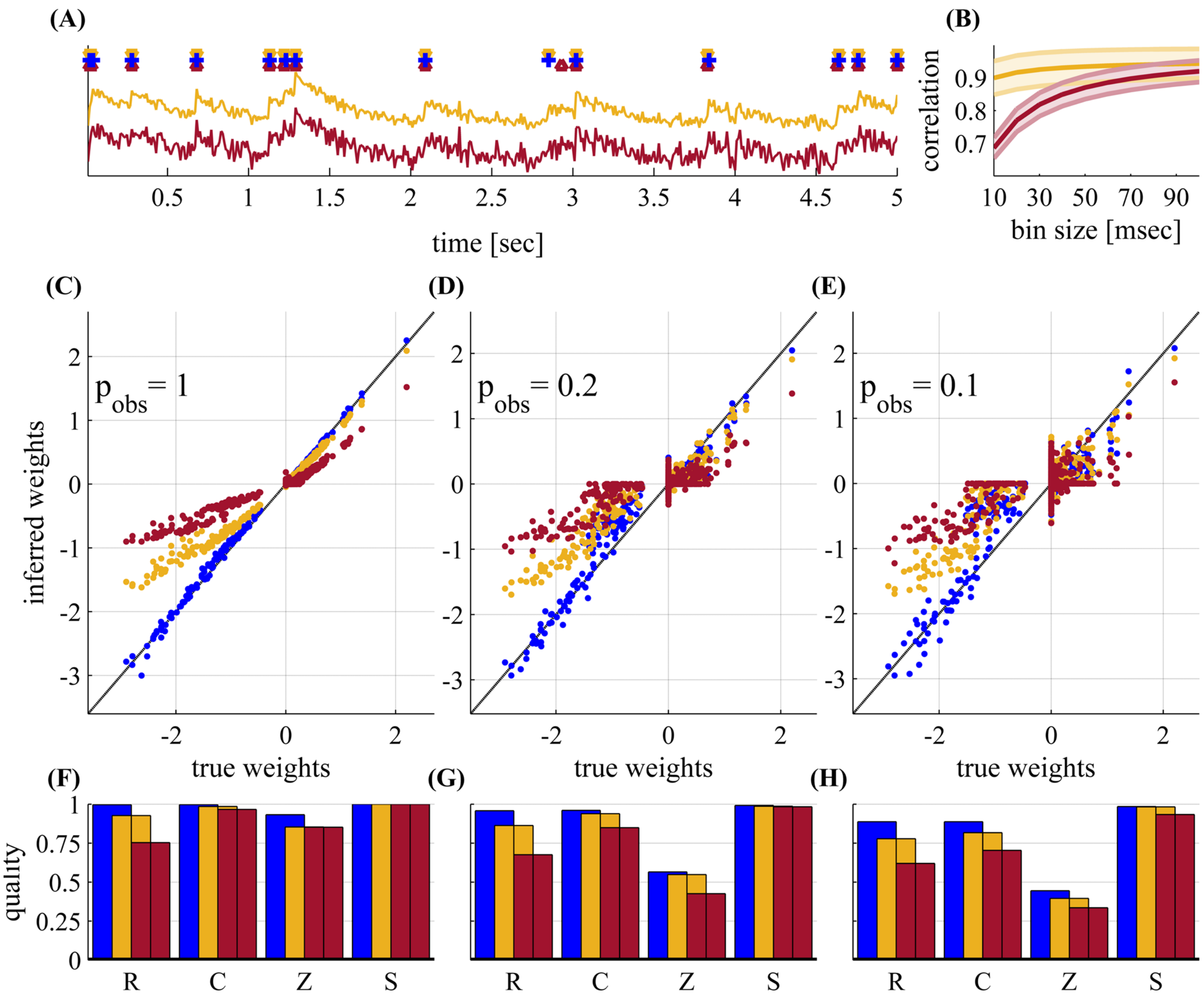

Fig 10. Inferring connectivity from fluorescence measurements from a network with *N* = 50 observed neurons—for different noise regimes: none (blue), low (yellow, snr = 0.2) and high (brown, snr = 0.4). (A) A short sample showing the fluorescence traces, for both noise regimes. On top we show the actual spikes (blue cross), and inferred spikes for low / high noise (yellow/brown triangles). **(B)** Correlation (population mean ± std) between actual and inferred spikes,for both low and high noise regimes. We bin the spikes (both actual and inferred) at various time bin sizes (x axis) and calculate the correlation using the definition of *C* (Eq 41, only for spikes instead of weights). Spikes are reasonably well estimated, given the noisy fluorescence traces. **(C-E)** Estimated weights vs. true weights for $p_{\mathrm{obs}} = 1$, 0.2 and 0.1. **(F-H)** Quality of inference for the C-E, respectively. Blue—spikes are directly measured, Yellow / brown—spikes are inferred from the respective fluorescence traces. Inhibitory weights exhibit more “shrinkage” due to their higher firing rate, which makes it harder to infer spikes from fluorescence. The mean firing rate is 3Hz, and *T* = 5.5 hours.

doi:10.1371/journal.pcbi.1004464.g010

measures are not claimed to predict synaptic connectivity, only provide a statistical description of the network dynamics.

Therefore, common approaches to infer connectivity of a neural network focus all the observations in one experiment on a small part of the network, in which all neurons are fully observed at a high frame rate. However, unobserved input into this sub-network can generate significant error in the estimation, and this error does not vanish with longer experiments. Various works aimed to deal with this persistent error: [17] inferred connectivity in a simulated two-neuron network in which one neuron was never observed; [45] inferred connectivity in a simulated network with two observed neurons and an unobserved common input; [46] inferred unobserved common input in an experimentally recorded network of 250 neurons using a GLM network with latent variables; [47] inferred connectivity in a simulated network with 100 neurons where 20–50 were never observed, with a varying degree of success.

9 The shotgun approach

To help deal with the "common input" problem, we propose a "shotgun" approach, in which we reconstruct network connectivity by serially observing small parts of the network—where each part is observed at a high frame rate for a limited duration. Thus, despite the limited scanning speed of the imaging device, by using this method, we can extend the number of the neurons covered by the scanning device and effectively decrease the number (and therefore the effect) of unobserved common inputs. Additionally, as only a small part of the network is illuminated together, this method can potentially reduce phototoxicity and photobleaching, and allow long, possibly chronic [48], imaging experiments.

9.1 Inferring correlations. Though our goal is to infer synaptic connections, we first discuss the closely related goal of inferring correlations between neurons. It is straightforward to infer these correlations from sub-sampled shotgun data when all neuron pairs can be observed together for long enough durations. We simply have to "ignore" any unobserved activity (section 3.6). We therefore suggest several observation schemes that might be used to eventually observe a much larger fraction of neuron pairs in the network (section 3). For example, we show that this can be done using two serial scanners with incommensurate periods (Fig 1I, 1J). If two scanning systems are combined on the same microscope, it can increase the effective frame rate above the critical 30Hz level [9] and allow successful weight reconstruction given long enough experiments. Alternatively, if we can use two moving microscopes to implement this scheme [27], the "effective field of view", could be expanded to any region that is not visually obstructed (such as deep regions in the tissue). This expansion can be arbitrarily large, again, as long as the experimental duration is long enough to compensate.

It may be also possible to infer correlations even if not all neuron pairs are observed (e.g., in a serial scanning scheme). For example, the methods discussed in [44, 49] might be helpful, if the fraction of observed neurons is not too low. In [44], in which experimentally recorded spikes are divided into two minimally overlapping blocks, the covariance matrix could only be accurately completed if more than 60% of the neurons were observed in each block (so, in total, 68% of all neuron pairs). Another covariance matrix completion method loosely requires that the size of the overlapping regions between the blocks must be larger than the rank of the full matrix [49]. It is not yet clear when these conditions apply in a physiologically relevant regime. And so, it remains to be seen if such methods could be used when only small fraction of all neurons is observed in each block.

9.2 Inferring connections. As we discussed in the previous section, it relatively straightforward to infer neuronal correlations, given enough observed neural pairs. It is also relatively easy to infer a linear-Gaussian model of the network activity with missing observations [44, 50], since we can analytically integrate out any unknown observations. However, as mentioned earlier (section 8) when inferring actual synaptic connectivity from real data, correlation and linear-based methods are inferior to a GLM-based approach.

Connectivity estimation with missing observations in a GLM is particularly challenging. Standard inference methods (maximum likelihood or maximum a posteriori) cannot be used, since the GLM likelihood cannot be evaluated without first inferring the missing spikes. However, exact Bayesian inference of the unobserved spikes is generally intractable. Therefore, previous works approximated the unobserved spikes through sampling [9, 17, 51], using Markov Chain Monte-Carlo (MCMC) methods on a GLM. However, such methods typically do not scale well for large networks. In fact, even if all the spikes are observed, inferring network connectivity using GLMs is very slow—taking about 10^5 CPU hours for a network with a thousand neurons in the recent work of [30].

In order to infer connectivity from this type of sub-sampled data we developed an Expected LogLiklihood (ELL) based method, which approximates the loglikelihood and its gradients so they depend on the spikes only through easily estimated second order statistics. By ignoring missing spikes in these statistics, we can infer neural network connectivity even when the spike data is (heavily) sub-sampled. This way we avoid the task of inferring the unobserved spikes, which requires computationally expensive latent variable approaches (section E) in S1 Text, as in [9, 17, 39, 51]. Even when all neurons are observed, the computational complexity drastically improves (section 6.2)—from $O(N^3\ TK)$ in standard algorithms, to $O(N^3\ K+N^2\ T)$ in the ELL-based method, where K is the number of iterations in the algorithm (K did not increase in the ELL-based estimation).

9.3 Numerical results. We demonstrate numerically (section 3.6) that such a double serial scanning method can be used to estimate the synaptic connectivity of a spiking neural network with connectivity roughly similar to that of the mouse visual cortex. We show that the inference is possible even if the spike data is sub-sampled at arbitrarily low observation ratios (*e.g.*, 10% in a network model with $N = 1000$ neurons, Figs 5 and 6, or 1% in single neuron model with $O(10^4)$ inputs, Fig 8); if the actual neuron model is not a GLM (a LIF model, Fig 4); and if fluorescence traces are observed instead of spikes (Fig 10). We perform parameter scans to examine the robustness of our method, and find the amount of data required for accurate shotgun reconstruction (Fig 7). Additionally, we confirm the accuracy and efficiency of our ELL-based method, in comparison to existing methods (Fig 9).

These results indicate that by using the shotgun observation scheme, we can remove the persistent bias resulting from the common input problem (Fig 2). Therefore, the limited scanning speed of imaging devices is not a fundamental obstacle hindering connectivity estimation. A complete removal of the bias is possible only if all the neurons in the network are observed together with all inputs to the network for a sufficient length of experimental time. However, in most experimental setups, some neurons will never be observed. Therefore, some persistent bias may remain. We modelled such a small bias in all simulations by adding a small number of neurons which are never observed. As we demonstrate numerically, this did not have a strong effect on our results. Stronger common inputs may require the incorporation of latent variables in the model, as in [46]; this is conceptually straightforward, and is an important direction for future research.

Clearly, the most important test for a connectivity inference method is on experimental data. Typically, on real data, performance is almost never as good as in simulations. Moreover, our numerical results suggest that, though our method is clearly much faster and more scalable than previous approaches, it still requires a substantial amount of data (hours). For low amounts of data (e.g., due to low observation ratios) it is likely to capture only the strongest connections accurately (Fig 6). These limitations are not properties of the method, but rather properties of the problem at hand and the type and amount of data typically available. For example, it was previously demonstrated in [42] that a significant amount of data is required to infer weak weights. However, there are a few potential extensions to the inference method that may significantly improve performance, as we explain next.

10 Challenges and future directions

We showed here that the proposed method is capable of incorporating prior information about the sparsity of synaptic connections. More specific information could be included. An abundance of such prior information is available for both connection probabilities and synaptic weight distributions as a function of cell location and identity [52]. Cutting edge labeling and tissue preparation methods such as Brainbow [53] and CLARITY [54] are beginning to provide rich anatomical data about "potential connectivity" (*e.g.*, the degree of coarse spatial overlap between a given set of dendrites and axons) that can be incorporated into these priors. Exploiting such prior information can significantly improve inference quality, as demonstrated in previous network inference papers [9, 17, 55]. For example, by adjusting the L1 regularization parameters, we can reflect such additional priors: that the probability of having a connection between two neurons typically decreases with the distance between two neurons, and that it is affected by the neuronal type.

Another way to improve connectivity estimates is to use stimulus information. For example, increasing the firing rate can improve quality (Eq 19 and Fig 7), up to a limit. If the firing rate is too high, it becomes harder to infer spikes from fluorescence. A more sophisticated spatio-temporal stimulus scheme can potentially lead to significant improvements in estimation quality [56]. The type of stimulus used can also affect performance. Sensory stimulation usually affects the measured network indirectly, potentially through many layers of neuronal processing. This may result in undesirable common input ("noise correlations"). Optogenetic stimulation does not have this problem, since it stimulates neurons directly by using light sensitive ion channels. However, this type of optical stimulation can potentially interfere with optical recording. Such cross-talk can be minimized by using persistent ion channels [57] (which require only a brief optical stimulus to be activated), or more sophisticated types of stimulation schemes [58, 59]. Such optogenetic approaches, coupled with the inference and experimental design methods described here, have the potential to lead to significantly improved connectivity estimates.

Even if all the neuronal inputs are eventually observed, if the observation probability p_{obs} is low then the variance due to the unobserved inputs may still be high, since, at any given time, most of the inputs to each neuron will be unobserved (see also [28]). As a result, the duration of the experiment required for accurate inference increases quadratically with the inverse of the observation probability (Eqs (18)–(19) and Fig 7), and weak weights become much harder to infer (Fig 6). Note this variance may be significantly reduced if we only aim to infer the input connections to only a few neurons (Fig 8). However, in many cases we wish to infer the entire network. In those cases the variance issue will persist, for any fixed observation strategy that does not take into account any prior information on the network connectivity.

However, there might be a significant improvement in performance if we can focus the observations on synaptic connections which are more probable. This way, we can effectively reduce input noise from unobserved neurons, and improve the signal to noise ratio. As a simple example, suppose we know the network is divided into several disconnected components. In this case, we should scan each sub-network separately, *i.e.*, there is no point in interleaving spike observations from two disconnected sub-networks. How should one focus observations in the more general case, making use of past observations in an online manner? Again, we leave this "active learning" problem as an important direction for future research.

11 Conclusions

In this work we suggest a “shotgun” experimental design, in which we infer the connectivity of a neural network from highly sub-sampled spike data. This is done in order to overcome experimental limitations stemming from the bounded scanning speed of any imaging device.

To do this, we develop a statistical expected loglikelihood-based Bayesian method. This method formally captures the intuitive notion that empiric spike correlations and mean spike rates are approximately the sufficient statistics for connectivity inference. Exploiting these sufficient statistics, our method has two major advantages over previous related approaches: (1) it is orders of magnitude faster (2) it can be used even when the spike data is massively sub-sampled.

We show that by using a double serial scanning scheme, all spike rates and correlations can be eventually inferred (and therefore neural connectivity). We demonstrate numerically that our method works efficiently in a simulated model with highly sub-sampled data and thousands of neurons. We conclude that the limited scanning speed of an imaging device recording neuronal activity is not a fundamental barrier which prevents consistent estimation of network connectivity.

Supporting Information

S1 Text. Technical appendix with full mathematical derivations and algorithmic details. (PDF)

Acknowledgments

The authors are thankful to Eftychios Pnevmatikakis, Ari Pakman and Ben Shababo for their help and support, and to Ran Rubin, Christophe Dupre, Lars Buesing and Yuriy Mishchenko for their helpful comments.

Author Contributions

Conceived and designed the experiments: DS LP. Performed the experiments: DS SK PS MO. Analyzed the data: DS. Contributed reagents/materials/analysis tools: GI. Wrote the paper: DS SK PS MO GI LP. Derived mathematical results: DS LP.

References

1. Katona G, Szalay G, Maák P, Kaszás A, Veress M, Hillier D, et al. Fast two-photon in vivo imaging with three-dimensional random-access scanning in large tissue volumes. Nat Methods. 2012; 9(2):201–208. doi: 10.1038/nmeth.1851 PMID: 22231641
2. Ahrens MB, Orger MB, Robson DN, Li JM, Keller PJ. Whole-brain functional imaging at cellular resolution using light-sheet microscopy. Nat Methods. 2013 May; 10(5):413–420. doi: 10.1038/nmeth.2434 PMID: 23524393
3. Stevenson IH, Kording KP. How advances in neural recording affect data analysis. Nat Neurosci. 2011 Feb; 14(2):139–142. doi: 10.1038/nn.2731 PMID: 21270781
4. Alivisatos AP, Chun M, Church GM, Greenspan RJ, Roukes ML, Yuste R. The Brain Activity Map Project and the Challenge of Functional Connectomics. Neuron. 2012; 74(6):970–974. doi: 10.1016/j.neuron.2012.06.006 PMID: 22726828
5. Venter JC, Adams MD, Sutton GG, Kerlavage AR, Smith HO, Hunkapiller M. Shotgun sequencing of the human genome. Science. 1998; 280:1540–1542. doi: 10.1126/science.280.5369.1540 PMID: 9644018
6. Reddy GD, Kelleher K, Fink R, Saggau P. Three-dimensional random access multiphoton microscopy for functional imaging of neuronal activity. Nat Neurosci. 2008; 11(6):713–720. doi: 10.1038/nn.2116

7. Grewe BF, Langer D, Kasper H, Kampa BM, Helmchen F. High-speed in vivo calcium imaging reveals neuronal network activity with near-millisecond precision. Nat Methods. 2010 May; 7(5):399–405. doi: 10.1038/nmeth.1453 PMID: 20400966

8. Hochbaum DR, Zhao Y, Farhi SL, Klapoetke N, Werley Ca, Kapoor V, et al. All-optical electrophysiology in mammalian neurons using engineered microbial rhodopsins. Nat Methods. 2014 Jun; 11 (8):825–833. doi: 10.1038/nmeth.3000 PMID: 24952910

9. Mishchenko Y, Vogelstein JT, Paninski L. A Bayesian approach for inferring neuronal connectivity from calcium fluorescent imaging data. Ann Appl Stat. 2011; 5(2B):1229–1261. doi: 10.1214/09-AOAS303

10. Hoebe RA, Van der Voort HTM, Stap J, Van Noorden CJF, Manders EMM. Quantitative determination of the reduction of phototoxicity and photobleaching by controlled light exposure microscopy. J Microsc. 2008 Jul; 231(Pt 1):9–20. doi: 10.1111/j.1365-2818.2008.02009.x PMID: 18638185

11. Park IM, Pillow JW. Bayesian Spike-Triggered Covariance Analysis. In: Neural Inf Process Syst; 2011. p. 1–9.

12. Sadeghi K, Gauthier JL, Field GD, Greschner M, Agne M, Chichilnisky EJ, et al. Monte Carlo methods for localization of cones given multielectrode retinal ganglion cell recordings. Network. 2013; 24(1):27–51. doi: 10.3109/0954898X.2012.740140 PMID: 23194406

13. Ramirez AD, Paninski L. Fast inference in generalized linear models via expected log-likelihoods. J Comput Neurosci. 2014 Apr; 36(2):215–234. doi: 10.1007/s10827-013-0466-4 PMID: 23832289

14. Diaconis P, Freedman D. Asymptotics of graphical projection pursuit. Ann Stat. 1984; 12(3):793–815. doi: 10.1214/aos/1176346703

15. Brillinger D. Maximum likelihood analysis of spike trains of interacting nerve cells. Biol Cyberkinetics. 1988; 59:189–200. doi: 10.1007/BF00318010

16. Rigat F, de Gunst M, van Pelt J. Bayesian modelling and analysis of spatio-temporal neuronal networks. Bayesian Anal. 2006; 1:733–764. doi: 10.1214/06-BA124

17. Pillow JW, Latham P. Neural characterization in partially observed populations of spiking neurons. In: Neural Inf Process Syst; 2007. p. 1–9.

18. Lütcke H, Gerhard F, Zenke F, Gerstner W, Helmchen F. Inference of neuronal network spike dynamics and topology from calcium imaging data. Front Neural Circuits. 2013; 7(Dec):201. doi: 10.3389/fncir.2013.00201 PMID: 24399936

19. Gerhard F, Kispersky T, Gutierrez GJ, Marder E, Kramer M, Eden U. Successful reconstruction of a physiological circuit with known connectivity from spiking activity alone. PLoS Comput Biol. 2013 Jul; 9 (7):e1003138. doi: 10.1371/journal.pcbi.1003138 PMID: 23874181

20. Teh YW, Newman D, Welling M. A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. In: Neural Inf Process Syst; 2006. p. 1–9.

21. Ribeiro F, Opper M. Expectation propagation with factorizing distributions: a Gaussian approximation and performance results for simple models. Neural Comput. 2011 Apr; 23(4):1047–1069. doi: 10.1162/NECO_a_00104 PMID: 21222527

22. Soudry D, Hubara I, Meir R. Expectation backpropagation: parameter-free training of multilayer neural networks with continuous or discrete weights. In: Neural Inf Process Syst. Montreal; 2014. p. 963–971.

23. Wang SI, Manning CD. Fast dropout training. Int Conf Mach Learn. 2013; 28:118–126.

24. Cotton RJ, Froudarakis E, Storer P, Saggau P, Tolias AS. Three-dimensional mapping of microcircuit correlation structure. Front Neural Circuits. 2013; 7:151. doi: 10.3389/fncir.2013.00151 PMID: 24133414

25. Nikolenko V, Watson B, Araya R, Woodruff A, Peterka D, Yuste R. SLM microscopy: scanless two-photon imaging and photostimulation using spatial light modulators. Front Neural Circuits. 2008; 2:5. doi: 10.3389/neuro.04.005.2008 PMID: 19129923

26. Bouchard MB, Voleti V, Mendes CS, Lacefield C, Grueber WB, Mann RS, et al. Swept confocally-aligned planar excitation (SCAPE) microscopy for high-speed volumetric imaging of behaving organisms. Nat Photonics. 2015 Jan; 9(2):113–119. doi: 10.1038/nphoton.2014.323 PMID: 25663846

27. Lecoq J, Savall J, Vučinić D, Grewe BF, Kim H, Li JZ, et al. Visualizing mammalian brain area interactions by dual-axis two-photon calcium imaging. Nat Neurosci. 2014 Nov; 17(12):1825–1829. doi: 10.1038/nn.3867 PMID: 25402858

28. Mishchenko Y. Consistency of the complete neuronal population connectivity reconstructions using shotgun imaging; 2015. In prep.

29. McLachlan G, Krishnan T. The EM algorithm and extensions. Wiley-Interscience; 2007.

30. Zaytsev YV, Morrison A, Deger M. Reconstruction of recurrent synaptic connectivity of thousands of neurons from simulated spiking activity; 2015. http://arxiv.org/abs/1502.04993

31. Denk W, Briggman KL, Helmstaedter M. Structural neurobiology: missing link to a mechanistic understanding of neural computation. Nat Rev Neurosci. 2012 May; 13(5):351–358. PMID: 22353782

32. Takemura Sy, Bharioke A, Lu Z, Nern A, Vitaladevuni S, Rivlin PK, et al. A visual motion detection circuit suggested by Drosophila connectomics. Nature. 2013 Aug; 500(7461):175–181. doi: 10.1038/nature12450 PMID: 23925240

33. Helmstaedter M. Cellular-resolution connectomics: challenges of dense neural circuit reconstruction. Nat Methods. 2013 Jun; 10(6):501–507. doi: 10.1038/nmeth.2476 PMID: 23722209

34. Minerbi A, Kahana R, Goldfeld L, Kaufman M, Marom S, Ziv NE. Long-term relationships between synaptic tenacity, synaptic remodeling, and network activity. PLoS Biol. 2009; 7(6):e1000136. doi: 10.1371/journal.pbio.1000136 PMID: 19554080

35. Nykamp DQ. Reconstructing stimulus-driven neural networks from spike times. In: Neural Inf Process Syst. vol. 15; 2003. p. 309–316.

36. Paninski L. Maximum likelihood estimation of cascade point-process neural encoding models. Netw Comput Neural Syst. 2004 Nov; 15(4):243–262. doi: 10.1088/0954-898X/15/4/002

37. Memmesheimer RM, Rubin R, Olveczky BP, Sompolinsky H. Learning precisely timed spikes. Neuron. 2014 May; 82(4):925–938. doi: 10.1016/j.neuron.2014.03.026 PMID: 24768299

38. Stetter O, Battaglia D, Soriano J, Geisel T. Model-free reconstruction of excitatory neuronal connectivity from calcium imaging signals. PLoS Comput Biol. 2012 Jan; 8(8):e1002653. doi: 10.1371/journal.pcbi.1002653 PMID: 22927808

39. Fletcher AK, Rangan S. Scalable inference for neuronal connectivity from calcium imaging. In: Neural Inf Process Syst; 2014. p. 1–9.

40. Mohler G. Learning convolution filters for inverse covariance estimation of neural network connectivity. In: Neural Inf Process Syst; 2014. p. 1–9.

41. Kispersky T, Gutierrez G, Marder E. Functional connectivity in a rhythmic inhibitory circuit using Granger causality. Neural Syst Circuits. 2011; 1:9. doi: 10.1186/2042-1001-1-9 PMID: 22330428

42. Volgushev M, Ilin V, Stevenson IH. Identifying and tracking simulated synaptic inputs from neuronal firing: insights from in vitro experiments. PLOS Comput Biol. 2015; 11(3):e1004167. doi: 10.1371/journal.pcbi.1004167 PMID: 25823000

43. Latimer KW, Chichilnisky EJ, Rieke F, Pillow JW. Inferring synaptic conductances from spike trains with a biophysically inspired point process model. In: Neural Inf Process Syst; 2014. p. 954–962.

44. Turaga S, Buesing L, Packer AM, Dalgleish H, Pettit N, Hausser M, et al. Inferring neural population dynamics from multiple partial recordings of the same neural circuit. In: Neural Inf Process Syst; 2013. p. 1–9.

45. Nykamp DQ. Pinpointing connectivity despite hidden nodes within stimulus-driven networks. Phys Rev E. 2008 Aug; 78(2):021902. doi: 10.1103/PhysRevE.78.021902

46. Vidne M, Ahmadian Y, Shlens J, Pillow JW, Kulkarni J, Litke AM, et al. Modeling the impact of common noise inputs on the network activity of retinal ganglion cells. J Comput Neurosci. 2012; 33(1):97–121. doi: 10.1007/s10827-011-0376-2 PMID: 22203465

47. Tyrcha J, Hertz J. Network inference with hidden units. Math Biosci Eng. 2014 Feb; 11(1):149–165. doi: 10.3934/mbe.2014.11.149 PMID: 24245678

48. Aramuni G, Griesbeck O. Chronic calcium imaging in neuronal development and disease. Exp Neurol. 2013 Apr; 242:50–56. doi: 10.1016/j.expneurol.2012.02.008 PMID: 22374357

49. Bishop W, Byron M. Deterministic Symmetric Positive Semidefinite Matrix Completion. In: Neural Inf Process Syst; 2014. p. 1–9.

50. Pakman A, Huggins J, Smith C, Paninski L. Fast penalized state-space methods for inferring dendritic synaptic connectivity. J Comput Neurosci. 2014 Sep; 36(3):415–443. doi: 10.1007/s10827-013-0478-0

51. Mishchenko Y, Paninski L. Efficient methods for sampling spike trains in networks of coupled neurons. Ann Appl Stat. 2011; 5(3):1893–1919. doi: 10.1214/11-AOAS467

52. Song S, Sjöström PJ, Reigl M, Nelson S, Chklovskii DB. Highly nonrandom features of synaptic connectivity in local cortical circuits. PLoS Biol. 2005 Mar; 3(3):507–519. doi: 10.1371/journal.pbio.0030068

53. Lichtman JW, Livet J, Sanes JR. A technicolour approach to the connectome. Nat Rev Neurosci. 2008 Jun; 9(6):417–422. doi: 10.1038/nrn2391 PMID: 18446160

54. Chung K, Wallace J, Kim S, Kalyanasundaram S, Andalman AS, Davidson TJ, et al. Structural and molecular interrogation of intact biological systems. Nature. 2013 Apr; 497:332–337. doi: 10.1038/nature12107 PMID: 23575631

55. Jonas E, Kording K. Automatic discovery of cell types and microcircuitry from neural connectomics; 2014. http://arxiv.org/abs/1407.4137

56. Shababo B, Brooks P, Pakman A, Paninski L. Bayesian inference and online experimental design for mapping neural microcircuits. In: Neural Inf Process Syst; 2013. p. 1–9.

57. Berndt A, Yizhar O, Gunaydin LA, Hegemann P, Deisseroth K. Bi-stable neural state switches. Nat Neurosci. 2009 Feb; 12(2):229–234. doi: 10.1038/nn.2247 PMID: 19079251

58. Rickgauer JP, Deisseroth K, Tank DW. Simultaneous cellular-resolution optical perturbation and imaging of place cell firing fields. Nat Neurosci. 2014 Nov; 17(12):1816–1824. doi: 10.1038/nn.3866 PMID: 25402854

59. Packer AM, Russell LE, Dalgleish HWP, Häusser M. Simultaneous all-optical manipulation and recording of neural circuit activity with cellular resolution in vivo. Nat Methods. 2014; 12(2):140–146. doi: 10.1038/nmeth.3217 PMID: 25532138

60. Tripathy SJ, Savitskaya J, Burton SD, Urban NN, Gerkin RC. NeuroElectro: a window to the world's neuron electrophysiology data. Front Neuroinform. 2014 Jan; 8:40. doi: 10.3389/fninf.2014.00040 PMID: 24808858