Inference, validation and predictions about statistics and propagation of cortical spiking in vivo

J. Wilting[1] & V. Priesemann[1,2,*]

[1]Max-Planck-Institute for Dynamics and Self-Organization, Am Faßberg 17, 37077 Göttingen, Germany; [2]Bernstein-Center for Computational Neuroscience, Göttingen, Germany
[*] viola.priesemann@ds.mpg.de

Electrophysiological recordings of spiking activity can only access a small fraction of all neurons simultaneously. This spatial subsampling has hindered characterizing even most basic properties of collective spiking in cortex. In particular, two contradictory hypotheses prevailed for over a decade: the first proposed an asynchronous irregular, the second a critical state. While distinguishing them is straightforward in models, we show that in experiments classical approaches fail to infer them correctly, because subsampling can bias measures as basic as the correlation strength. Deploying a novel, subsampling-invariant estimator, we find evidence that in vivo cortical dynamics clearly differs from asynchronous or critical dynamics, and instead occupies a narrow "reverberating" regime, consistently across multiple mammalian species and cortical areas. These results enabled us to predict cortical properties that are difficult or impossible to obtain experimentally, including responses to minimal perturbations, intrinsic network timescales, and the strength of external input compared to recurrent activation.

Introduction

When investigating spiking activity in neuronal networks, only a tiny fraction of all neurons can be recorded experimentally with millisecond precision. Such spatial subsampling fundamentally limits virtually any recording and hinders inferences about the collective dynamics of cortical networks.[1–4] In fact, even some of the most basic characteristics of cortical network dynamics are not known with certainty, such as the population Fano factor, or the fraction of spikes generated internally versus those triggered by input.

In particular, two contradicting hypotheses to describe network dynamics have competed for more than a decade, and are the subjects of ongoing scientific debate: One hypothesis suggests that collective dynamics are "asynchronous irregular"[5–7] (AI), i.e. neurons spike independently of each other and in a Poisson manner, which may reflect a balanced state.[8,9] The other hypothesis proposes that neuronal networks operate at criticality[10–17] and thus in a particularly sensitive state close to a phase transition. These hypotheses have distinct implications for the coding strategy of the brain. The typical balanced state minimizes redundancy,[18–22] supports fast network responses,[8] and shows vanishing autocorrelation time ($\tau \to 0$). In contrast, criticality in models optimizes performance in tasks that profit from extended reverberations of activity in the network,[23–29] because it is characterized by long-range correlations in space and time ($\tau \to \infty$). It has been proposed that τ reflects an integration window over past activity, thereby allowing brain networks to operate on specific timescales.[30–33] Timescales estimated from single neurons span hundreds of milliseconds,[34] but it is unclear how timescales of the full network can be inferred in the face of subsampling.

Surprisingly, there is experimental evidence for both AI and critical states in cortical networks, although both states are clearly distinct. Evidence for the AI state is based on characteristics of single neuron spiking resembling a Poisson process, i.e. exponential inter spike interval (ISI) distributionss and Fano factors F close to unity.[35] Moreover, spike count cross-correlations[36,37] are small. Evidence for criticality was typically obtained from a population perspective instead, and assessed neuronal avalanches, i.e. spatio-temporal clusters of activity,[1,10,38–41] whose sizes are expected to be power-law distributed if networks are critical.[42] Deviations from power-laws, typically observed for spiking activity in awake animals,[2,3,43,44] were attributed to subsampling effects.[1–4,45–47] Hence, different analysis approaches provided evidence for one or the other dynamical state's dominance.

We rely on a classic approach to probe the dynamical states of a system at steady state, namely applying minimal perturbations. Studying how perturbations cascade through a system enables the inference of numerous system properties. London and colleagues applied such a perturbation framework and estimated that one average $m = 28$ additional postsynaptic spikes are triggered by one extra spike in a presynaptic neuron from intracellular recordings.[48] From their complementary extracellular spike recordings, one can equally well estimate $m \approx 0.04\,\mathrm{Hz/neuron} \cdot 10\,\mathrm{ms} \cdot k = 0.6$: in the 10 ms subsequent to the perturbation, an increase of 0.04 Hz is observed for each neuron. Assuming that this 10 ms is an upper bound to directly activate any of the $k \approx 1500$ directly connected neurons, one obtains as as upper bound $m \approx 0.04\,\mathrm{Hz/neuron} \cdot 10\,\mathrm{ms} \cdot k = 0.6$. This vast range for estimates of m arises largely because such inferences are heavily influenced by subsampling. We here build on a subsampling-invariant approach presented in a companion study,[49] which allows us to resolve questions surrounding the contradictory results on cortical dynamics: (i) we establish an analytically tractable minimal model for *in vivo* like activity, which can interpolate from AI to critical dynamics; (ii) we estimate the dynamical state of cortical activity based on a novel, subsampling-invariant estimator;[49] (iii) we predict a number of network dynamical properties, which are experimentally accessible and allow to validate our approach; (iv) we predict a number of yet unknown network properties, including m, the expected number of spikes triggered by one additional spike, the emergent network timescale τ, the distribution of the total number of spikes triggered by a single extra action potential, and the fraction of activation that can be attributed to afferent external input to a cortical network.

Material and Methods

Minimal model of spike propagation

To gain an intuitive understanding of our mathematical approach, make a thought experiment in your favorite spiking network: apply one additional spike to an excitatory neuron, in analogy to the approach by London and colleagues[48]. How does the network respond to that perturbation? As a first order approximation, one quantifies the number of spikes that are triggered by this perturbation *additionally* in all postsynaptic neurons. This number may vary from trial to trial, depending on the membrane potential of the postsynaptic neurons; however, what interests us most is m, the *mean number of spikes triggered by the one extra spike*. Taking a mean-field approximation and assuming that the perturbation indeed is small, any of these triggered spikes in turn trigger spikes in their postsynaptic neurons in a similar manner, and thereby the perturbation may cascade through the system. Mathematically, such cascades can be described by a branching process.[50–52]

In the next step, assume that perturbations are started continuously at rate h, for example through afferent input from other brain areas or sensory modalities. As neurons presumably do not distinguish whether a postsynaptic potential was elicited from a neuron from within the network, or from afferent input, all spikes are assumed to have on average the same impact on the network dynamics. Together, this leads to the mathematical framework of a branching

network,[2,3,10,24,45] which can generate dynamics spanning AI and critical states depending on the input,[53] and hence is well suited to probe network dynamics *in vivo* (see Supp. 1 for details). Most importantly, this framework allows to infer m and other properties from the ongoing activity proper, because one treats any single spike as a minimal perturbation on the background activity of the network. Mathematical approaches to infer m are long known if the full network is sampled.[54,55] Under subsampling, however, it is the novel estimator described in[49] that for the first time allows an unbiased inference of m, even if only a tiny fraction of neurons is sampled. After inferring m, a number of quantities can be analytically derived, and others can be obtained by simulating a mean-field spiking model, which is constrained by the experimentally measured m and the spike rate.

The framework of branching networks can be interpreted as a stochastic description of spike propagation on networks, as outlined above. It can alternatively be taken as a *strictly phenomenological* approximation to network dynamics that enables us to infer details of network *statistics* despite subsampling. Independent of the perspective, the dynamics of the network is mainly governed by m (Fig. 1**a**). If an action potential only rarely brings any postsynaptic neuron above threshold ($m \gtrapprox 0$), external perturbations quickly die out, and neurons spike independently and irregularly, driven by external fluctuations h. In general, if one action potential causes less than one subsequent action potential on average ($m < 1$), perturbations die out and the network converges to a stable distribution, with increasing fluctuations and variance the closer m is to unity. If $m > 1$, perturbations may grow infinitely, potentially leading to instability. The critical state ($m = 1$) separates the stable (subcritical) from the unstable (supercritical) phase. When approaching this critical state from below, the expected size $\langle s \rangle$ and duration $\langle d \rangle$ of individual cascades or avalanches diverge: $\langle s \rangle \sim \frac{1}{m_c - m}$. Therefore, especially close to criticality, a correct estimate of m is vital to assess the risk that the network develops large, potentially devastating cascades, which have been linked to epileptic seizures,[56] either generically or via a minor increase in m.

Simulation. We simulated a branching network model by mapping a branching process[50] (Supp. 1) onto a fully connected network of $N = 10,000$ neurons.[24] An active neuron activated each of its k postsynaptic neurons with probability $p = m/k$. Here, the activated postsynaptic neurons were drawn randomly without replacement at each step, thereby avoiding that two different active neurons would both activate the same target neuron. Similar to the branching process, the branching network is critical for $m = 1$ in the infinite size limit, and subcritical (supercritical) for $m < 1$ ($m > 1$). We modeled input to the network at rate h by Poisson activation of each neuron at rate h/N. Subsampling[1] was applied to the model by sampling the activity of n neurons only, which were selected randomly before the simulation, and neglecting the activity of all other neurons.

If not stated otherwise, simulations were run for $L = 10^7$ time steps (corresponding to ~11 h). Confidence intervals were estimated according to[49] from $B = 100$ realizations of the network, both for simulation and experiments.

The reverberating branching networks were defined to match the respective experimental recording in the number of sampled neurons n, mean activity $\langle a_t \rangle$, and branching ratio m. Exemplarily for the cat recording, which happened to represent the median $\hat{m}$, this yielded $m = \hat{m} = 0.98$, $n = 50$, and $\langle a_t \rangle = 1.58$ per bin, from which $h = 0.032 \cdot N$ follows. The corresponding AI and near-critical networks were matched in n and $\langle a_t \rangle$, but set up with branching ratios of $m = 0$ or $m = 0.9999$ respectively. For all networks, we chose a full network size of $N = 10^4$.

In Figs. 2**c**, the reverberating branching network was also matched to the length of the cat recording of 295 s. To test for stationarity, the cat recordings and the reverberating branching network were split into 59 windows of 5 s each, before estimating m for each window. In Fig. 1**c**, subcritical and critical branching networks with $N = 10^4$ and $\langle A_t \rangle = 100$ were simulated,

and $n = 100$ units sampled.

Experiments

We evaluated spike population dynamics from recordings in rats, cats and monkeys. The rat experimental protocols were approved by the Institutional Animal Care and Use Committee of Rutgers University.[57,58] The cat experiments were performed in accordance with guidelines established by the Canadian Council for Animal Care.[59] The monkey experiments were performed according to the German Law for the Protection of Experimental Animals, and were approved by the Regierungspräsidium Darmstadt. The procedures also conformed to the regulations issued by the NIH and the Society for Neuroscience. The spike recordings from the rats and the cats were obtained from the NSF-founded CRCNS data sharing website.[57–60]

Rat experiments. In rats the spikes were recorded in CA1 of the right dorsal hippocampus during an open field task. We used the first two data sets of each recording group (ec013.527, ec013.528, ec014.277, ec014.333, ec015.041, ec015.047, ec016.397, ec016.430). The data-sets provided sorted spikes from 4 shanks (ec013) or 8 shanks (ec014, ec015, ec016), with 31 (ec013), 64 (ec014, ec015) or 55 (ec016) channels. We used both, spikes of single and multi units, because knowledge about the identity and the precise number of neurons is not required for the MR estimator. More details on the experimental procedure and the data-sets proper can be found in[57,58].

Cat experiments. Spikes in cat visual cortex were recorded by Tim Blanche in the laboratory of Nicholas Swindale, University of British Columbia.[59] We used the data set pvc3, i.e. recordings of 50 sorted single units in area 18.[60] We used that part of the experiment in which no stimuli were presented, i.e., the spikes reflected spontaneous activity in the visual cortex of the anesthetized cat. Because of potential non-stationarities at the beginning and end of the recording, we omitted data before 25 s and after 320 s of recording. Details on the experimental procedures and the data proper can be found in[59,60].

Monkey experiments. The monkey data are the same as in[3,61]. In these experiments, spikes were recorded simultaneously from up to 16 single-ended micro-electrodes ($\varnothing = 80\,\mu$m) or tetrodes ($\varnothing = 96\,\mu$m) in lateral prefrontal cortex of three trained macaque monkeys (M1: 6 kg ♀; M2: 12 kg ♂; M3: 8 kg ♀). The electrodes had impedances between 0.2 and $1.2\,\mathrm{M}\Omega$ at 1 kHz, and were arranged in a square grid with inter electrode distances of either 0.5 or 1.0 mm. The monkeys performed a visual short term memory task. The task and the experimental procedure is detailed in[61]. We analyzed spike data from 12 experimental sessions comprising almost 12.000 trials (M1: 5 sessions; M2: 4 sessions; M3: 3 sessions). 6 out of 12 sessions were recorded with tetrodes. Spike sorting on the tetrode data was performed using a Bayesian optimal template matching approach as described in[62] using the "Spyke Viewer" software.[63] On the single electrode data, spikes were sorted with a multi-dimensional PCA method (Smart Spike Sorter by Nan-Hui Chen).

Analysis

Temporal binning For each recording, we collapsed the spike times of all recorded neurons into one single train of population spike counts a_t, where a_t denotes how many neurons spiked in the t^{th} time bin Δt. If not indicated otherwise, we used $\Delta t = 4$ ms, reflecting the propagation time of spikes from one neuron to the next.

Multistep regression estimation of $\widehat{m}$ From these time series, we estimated $\widehat{m}$ using the MR estimator described in[49]. For $k = 1, \ldots, k_{\max}$, we calculated the linear regression slope $r_{k\,\Delta t}$ for the linear statistical dependence of a_{t+k} upon a_t. From these slopes, we estimated $\widehat{m}$ following the relation $r_{\delta t} = b \cdot \widehat{m}^{\delta t/\Delta t}$, where b is an (unknown) parameter that depends on the higher moments of the underlying process and the degree of subsampling. However, for an estimation of m no further knowledge about b is required.

Throughout this study we chose $k_{\text{max}} = 2500$ (corresponding to 10 s) for the rat recordings, $k_{\text{max}} = 150$ (600 ms) for the cat recording, and $k_{\text{max}} = 500$ (2000 ms) for the monkey recordings, assuring that k_{max} was always in the order of multiple autocorrelation times.

In order to test for the applicability of a MR estimation, we used a set of conservative tests[49] and included only those time series, where the approximation by a branching network was considered appropriate. For example, we excluded all recordings that showed an offset in the slopes r_k, because this offset is, strictly speaking, not explained by a branching network and might indicate non-stationarities. Details on these tests are found in[49]. Even with these conservative tests, we found the exponential relation $r_k = b\, m^{\delta t/\Delta t}$ expected for branching networks in the majority of experimental spike recordings (14 out of 21, Fig. S1).

Avalanche size distributions Avalanche sizes were determined similarly to the procedure described in[1,3]. Assuming that individual avalanches are separated in time, let $\{t_i\}$ indicate bins without activity, $a_{t_i} = 0$. The size s_i of one avalanche is defined by the integrated activity between two subsequent bins with zero activity:

$$s_i = \sum_{t=t_i}^{t_{i+1}} a_t. \tag{1}$$

From the sample $\{s_i\}$ of avalanche sizes, avalanche size distributions $p(s)$ were determined using frequency counts. For illustration, we applied logarithmic binning, i.e. exponentially increasing bin widths for s.

For each experiments, these empirical avalanche size distributions were compared to avalanche size distributions obtained in a similar fashion from three different matched models (see below for details). Model likelihoods $l(\{s_i\})\,|\,m)$ for all three models were calculated following[64], and we considered the likelihood ratio to determine the most likely model based on the observed data.

ISI distributions, Fano factors and spike count cross-correlations. For each experiment and corresponding reverberating branching network (subsampled to a single unit), ISI distributions were estimated by frequency counts of the differences between subsequent spike times for each channel.

We calculated the single unit Fano factor $F = \text{Var}[a_t]/\langle a_t\rangle$ for the binned activity a_t of each single unit, with the bin sizes indicated in the respective figures. Likewise, single unit Fano factors for the reverberating branching networks were calculated from the subsampled and binned time series.

From the binned single unit activities a_t^1 and a_t^2 of two units, we estimated the spike count cross correlation $r_{\text{sc}} = \text{Cov}(a_t^1, a_t^2)/\sigma_{a_t^1}\sigma_{a_t^2}$. The two samples a_t^1 and a_t^2 for the reverberating branching networks were obtained by sampling two randomly chosen neurons.

Results

Subsampling-invariant inference of the dynamical state

In a companion study[49] we showed that conventional estimators based on linear regression[54,55] significantly underestimate $\hat{m}$ when the system is subjected to subsampling (Fig. 1**c**), as it is always the case in electrophysiological recordings (Fig. 1**b**). The bias is considerable: For example, sampling 50 neurons or a single neuron in a branching network with $m = 0.99$ resulted in the wrong estimates $\hat{m}_{\text{Conv}} = 0.21$, or even $\hat{m}_{\text{Conv}} = 0.002$, respectively (Fig. 1**d**). Thus a process close to instability ($m = 0.99$) is mistaken as Poisson-like ($\hat{m}_{\text{Conv}} = 0.002 \approx 0$) just because the estimate from subsampled activity is taken as face value for the entire population. The same study presented a novel *multistep regression* estimator (MR estimator), which correctly characterizes the population dynamics via m even under strong subsampling, in principle even from single neurons. Importantly, one can estimate m even when sampling only a very small

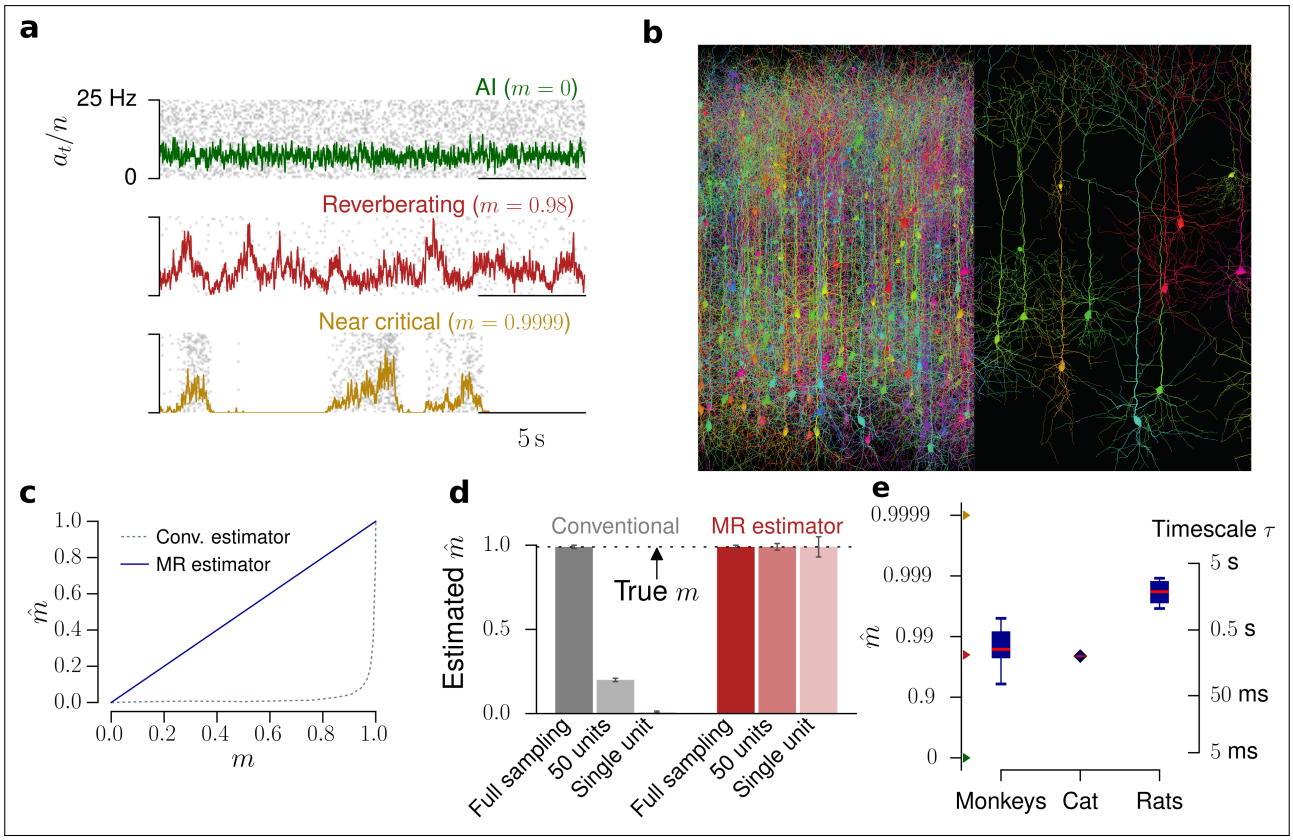

Figure 1: Spatial subsampling. **a**. Raster plot and population rate for networks with different spike propagation parameters. They exhibit vastly different dynamics, which readily manifest in the population activity. **b**. When assessing neuronal spiking activity, only a small subset of all neurons can be recorded. This spatial subsampling can hinder correct inference of collective properties of the whole network; figure created using TREES[65] and reproduced from[49]. **c**. Estimated branching ratio $\hat{m}$ as a function of the simulated branching ratio m, inferred from subsampled activity (100 out of 10,000 neurons). While the conventional estimator misclassified m from this subsampled observation (gray, dotted line), the novel multistep regression (MR) estimator returned the correct values **d**. For a reverberating branching network with $m = 0.98$, the conventional estimator inferred $\hat{m} = 0.21$ or $\hat{m} = 0.002$ when sampling 50 or 1 units respectively, in contrast to MR estimation, which returned the correct $\hat{m}$ even under strong subsampling. **e**. Using the novel MR estimator, cortical network dynamics in monkey prefrontal cortex, cat visual cortex, and rat hippocampus were consistently found to exhibit reverberating dynamics, with $0.94 < \hat{m} < 0.991$ (median $\hat{m} = 0.98$ over all experimental sessions, boxplots indicate median / 25% – 75% / 0% – 100% over experimental sessions per species). These correspond to network timescales between 80 ms and 2 s.

fraction of neurons and without knowing the network size N, the number of sampled neurons n, nor any moments of the underlying process.[49] This robustness makes the estimator an ideal tool for the analysis of neuronal network recordings.

Reverberating spiking activity *in vivo*

We analyzed *in vivo* spiking activity from Macaque monkey prefrontal cortex during a short term memory task,[61] anesthetized cat visual cortex with no stimulus,[59] and rat hippocampus during a foraging task.[57,58] We applied MR estimation to the binned population spike counts a_t of the recorded neurons of each session (see methods). In the continuous spectrum from AI ($m = 0$) to critical ($m = 1$), we identified a limited range of branching values *in vivo*: in the experiments $\hat{m}$ ranged from 0.963 to 0.998 (median $\hat{m} = 0.98$), corresponding to autocorrelation times between 100 ms and 2 s (median 247 ms, Figs. 1**e**, S1). This clearly suggests that spiking activity *in vivo* is neither AI-like, nor consistent with a critical state. Instead, it is poised in a regime

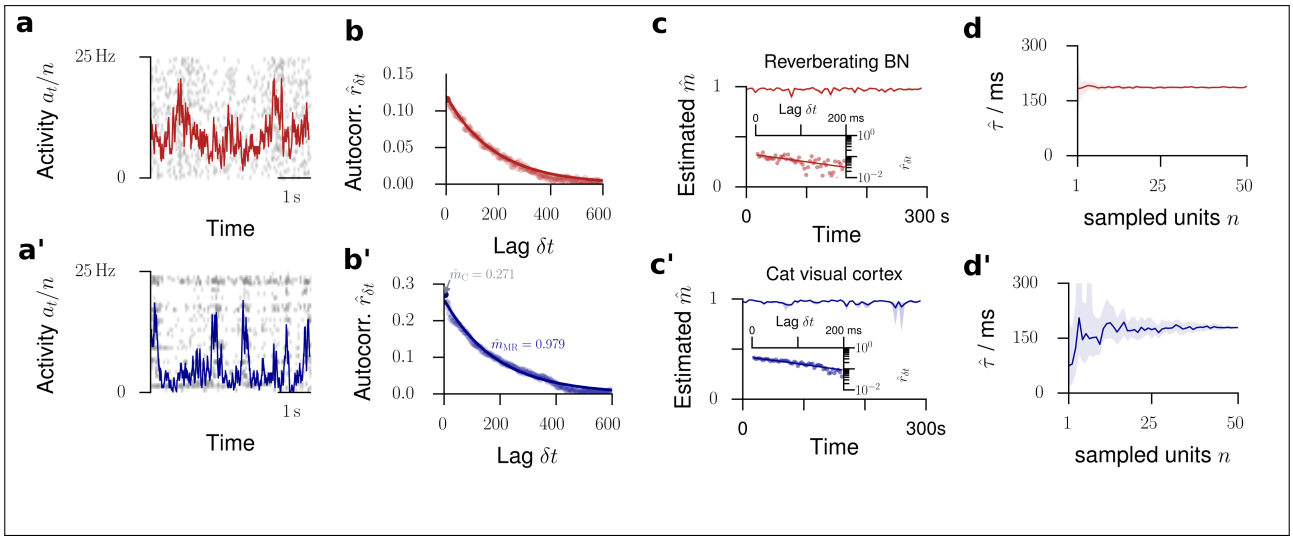

FIGURE 2: Validation of the model assumptions. The top row displays properties from a reverberating network model, the bottom row spike recordings from cat visual cortex. **a/a'**. Raster plot and population activity a_t within bins of $\Delta t = 4$ ms, sampled from $n = 50$ neurons. **b/b'**. Multistep regression (MR) estimation from the subsampled activity (5 min recording). The predicted exponential relation $r_{\delta t} \sim m^{\delta t/\Delta t} = \exp(-\delta t/\tau)$ provides a validation of the applicability of the model. The experimental data are fitted by this exponential with remarkable precision. **c/c'**. The estimated branching parameter $\hat{m}$ for 59 windows of 5 s length suggests stationarity of m over the entire recording (shaded area: 16% to 84% confidence intervals). The variability in $\hat{m}$ over consecutive windows was comparable for experimental recording and the matched network ($p = 0.09$, Levene test). Insets: MR estimation exemplified for one example window each. **d/d'**. When subsampling even further, MR estimation always returns the correct timescale $\hat{\tau}$ (or $\hat{m}$) in the model. In the experiment, this invariance to subsampling also holds, down to $n \approx 10$ neurons (shaded area: 16% to 84% confidence intervals estimated from 50 subsets of n neurons).

that, unlike critical or AI, does not maximize one particular property alone but may combine features of both (see discussion). Due to the lack of one prominent characterizing feature, we name it the *reverberating* regime, stressing that activity reverberates (different from the AI state) at timescales of hundreds of milliseconds (different from a critical state, where they can persist infinitely).

The reverberating state differs from criticality

On first sight, $\hat{m} = 0.98$ of the reverberating state may suggest that the collective spiking dynamics is very close to critical. Indeed, physiologically a $\Delta m \approx 1.6\%$ difference to criticality is small in terms of the effective synaptic strength. However, this apparently small difference in single unit properties has a large impact on the collective *dynamical* fingerprint and makes AI, reverberating, and critical states clearly distinct: (1) This distinction is readily manifest in the fluctuations of the population activity, where states with $m = 0.98$ and $m = 0.999$ are clearly different (Fig. 1**a**). (2) Consider the sensitivity to a small input, i.e. the susceptibility $\chi = \partial\langle A_t\rangle \,/\, \partial h = \frac{1}{1-m}$. The susceptibility diverges at criticality. A critical network is thus overly sensitive to input. In contrast, states with $m \approx 0.98$ assure sensitivity without instability. (3) Likewise, the $\Delta m \approx 1.6\%$ difference limits the intrinsic timescale of the dynamics to a few hundred milliseconds, while at criticality it approaches infinity. (4) Because of the divergences at criticality, network dynamics dramatically differ between $m = 0.9$, $m = 0.99$ or $m = 0.999$: for example, the differences in susceptibility (sensitivity) and variance are 100-fold. Because this has a strong impact on network dynamics and putative network function, finely distinguishing between dynamical states is both important and feasible even if the corresponding differences in effective synaptic strength (m) appear small.

Validity of the approach

There is a straight-forward verification of the validity of our phenomenological model: it predicts an exponential autocorrelation function $r_{\delta t}$ for the population activity a_t. We found that the activity in cat visual cortex (Figs. 2**a**,**a'**) is surprisingly well described by this exponential fit (Fig. 2**b**,**b'**). This validation holds to the majority of experiments investigated (14 out of 21, Fig. S1).

A second verification of our approach is based on its expected invariance under subsampling: We further subsampled the activity in cat visual cortex by only taking into account spikes recorded at a subset of n out of all available single units. As predicted (Fig. 2**d**), the estimates of $\hat{m}$, or equivalently of $\hat{\tau}$, coincided for any subset of single units. Only if the activity of less than 5 out of the available 50 single units was considered, the autocorrelation time was underestimated (Fig. 2**d'**), most likely because of the heterogeneity of cortical networks. These results demonstrate, however, that our approach gives consistent results from the activity of $n \geq 5$ neurons, which were available for all investigated experiments.

Origin of the activity fluctuations

The fluctuations found in cortical spiking activity, instead of being intrinsically generated, could in principle arise from non-stationarities, which could in turn lead to misestimation of m. This is unlikely for three reasons: First, we defined a set of conservative tests to reject recordings that show any signature of common non-stationarities. Even with these tests, we found the exponential relation $r_{\delta t} \sim m^{\delta t/\Delta t}$ expected for branching networks not only in cat visual cortex, but in the majority of experiments (14 out of 21, Fig. S1). Second, recordings in cat visual cortex were acquired in absence of any stimulation, excluding stimulus-related non-stationarities. Third, when splitting the spike recording into short windows, the window-to-window variation of $\hat{m}$ in the recording did not differ from that of stationary *in vivo*-like branching networks ($p = 0.3$, Figs. 2**c**,**c'**). The *in vivo*-like branching network by definition was set up with the same branching ratio m, spike rate $\langle a_t \rangle$, number of sampled neurons n, and duration as the experimental recording (e.g. for the cat $n = 50$, $m = 0.98$, $\bar{r} = 7.9$ Hz, recording of 295 s length). For these reasons the observed fluctuations likely reflect intrinsic timescales of the underlying collective network dynamics.

Timescales of the network and single units

The dynamical state described by m directly relates to an exponential autocorrelation function with an intrinsic network timescale of $\tau = -\Delta t/\ln m$. Exemplarily for the cat recording, $m = 0.98$ implies a network timescale of $\tau = 188$ ms, where we here chose $\Delta t = 4$ ms. While the autocorrelation function of the full network activity is expected to show an exponential decay, we showed that the autocorrelation of single neuron activity rapidly decreases at the timescale of a bin size (Fig. 3**a**). This rapid decrease is typically interpreted a lack of memory, overlooking that single neurons do not need to be equivalent to the network in terms of autocorrelation strength. Our theoretical results explain how this prominent dip comes about even in reverberating systems: because of the strong subsampling when considering single neuron activity, the strength of autocorrelation is decreased by a constant factor for any lag $\delta t \neq 0$. Ignoring the value at $\delta t = 0$, the floor of the autocorrelation function still unveils the exponential relation. Remarkably, the autocorrelogram of single units in cat visual cortex displayed precisely the shape of autocorrelation predicted for single neurons (compare Figs. 3**a** and **b**).

Although our results were largely invariant to further subsampling (provided $n \geq 5$, Fig. 2**d'**), the intrinsic timescales τ (or m) of single neurons differed from the network timescale, as one might expect in heterogeneous systems. We found that single neuron timescales were typically smaller than the network timescale (Fig. 3**c**, median $\tau = 85$ ms for single neurons in cat visual cortex versus $\tau = 180$ ms for the network, Figs. 2**d'**, S9**c**). Therefore, the network timescale inferred by our approach contributes further information about network dynamics compared to previous studies which only considered single neurons.[34]

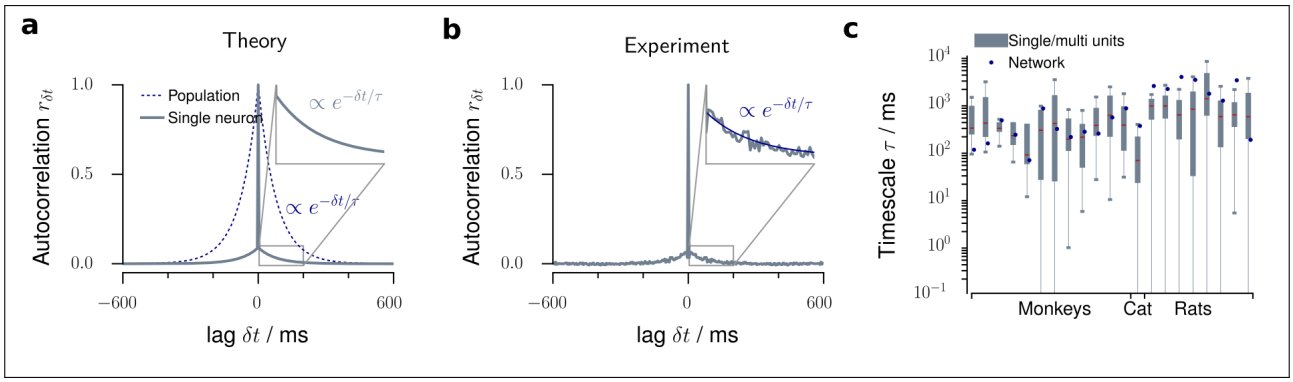

Figure 3: MR estimation and autocorrelation times. a. In a branching model, the autocorrelogram of the population activity is exponential with decay time τ (blue dotted line). In contrast, the autocorrelogram for single neurons shows a sharp drop from $r_0 = 1$ at lag $\delta t = 0$ s to the next lag $r_{\pm\Delta t}$ (gray solid line). We showed that this drop is a subsampling-induced bias. When ignoring the zero-lag value, the autocorrelation strength is decreased, but the exponential decay and even the value of the autocorrelation time τ of the network activity are preserved (inset). **b**. The autocorrelogram of single neuron activity recorded in cat visual cortex precisely resembles this theoretical prediction, namely a sharp drop and then an exponential decay. **c**. Single unit and population timescales for all experimental sessions. The boxplots indicate the distribution of timescales inferred from single unit activity (median in red / 25% – 75% / 2.5% – 97.5%), the blue dots the timescale inferred from the population activity of all sampled units.

Established methods are biased to identifying AI dynamics

On the population level, networks with different m are clearly distinguishable (Fig. 1**a**). Surprisingly, single neuron statistics, namely interspike interval (ISI) distributions, Fano factors, conventional estimation of m, and the autocorrelation $r_{\delta t}$, all returned signatures of AI activity regardless of the underlying network dynamics and cannot serve as a reliable indicator for the network's dynamical state.

First, exponential interspike interval (ISI) distributions are considered a strong indicator of Poisson-like firing. Surprisingly, the ISIs of single neurons in the *in vivo*-like branching network closely followed exponential distributions, which were determined mainly by the firing rate, and were almost indistinguishable from ISI distributions obtained from AI networks (Figs. 4**a**,**a'**, S2). This result was confirmed by coefficients of variation close to unity, as expected for exponential distributions (Fig. S2).

Second, the Fano factor F for the activity of single neurons was close to unity, a hallmark feature of irregular spiking,[35] in any network model (Fig. 5**g**, analytical result: Eq. (S8)) and for single unit activity across all units and experiments (Figs. 4**b**,**b'**, S3). Even when increasing the bin size to 4 s, the median Fano factor of single unit activity did not exceed $F = 10$ in any of the experiments, even in those with the longest reverberation. In contrast, for the full network the Fano factor rose to $F \approx 10^4$ for the *in vivo*-like branching network and diverged when approaching criticality (Fig. 5**g**, analytical result: Eq. (S4)).

Third, conventional regression estimators[54,55] are biased towards inferring irregular activity, as shown before. Here, conventional estimation yielded a median of $\hat{m} = 0.057$ for single neuron activity in cat visual cortex, in contrast to $\hat{m} = 0.954$ returned by MR estimation even from single unit recordings (Fig. S9).

Fourth, when examining the autocorrelation function of an experimental recording (Fig. 3**b**) the prominent decay of $r_{\delta t}$ prevails and hence single neuron activity appears uncorrelated in time.

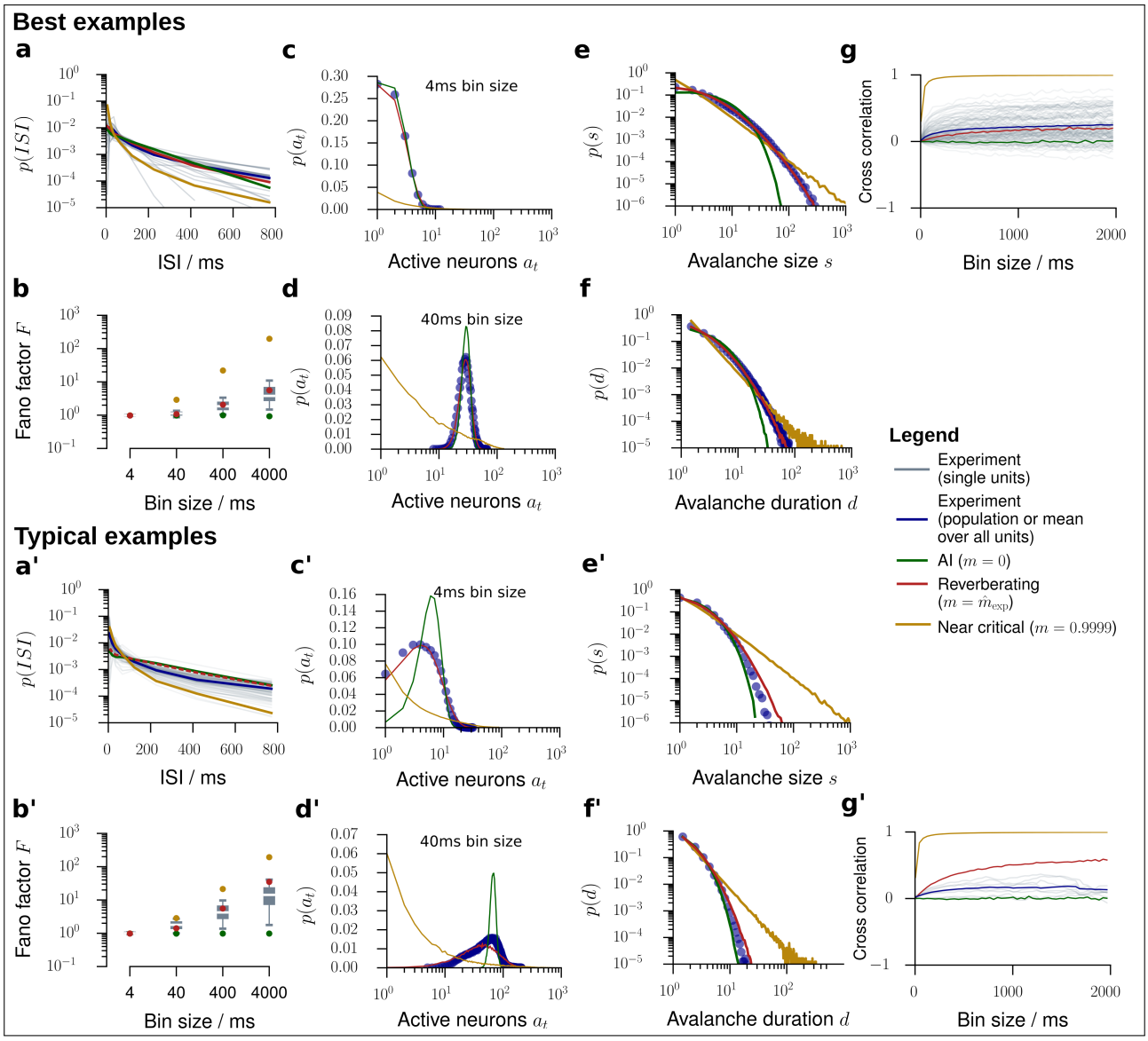

Figure 4: Model validation for *in vivo* spiking activity. We validated our model by comparing experimental results to predictions obtained from the *in vivo*-like, reverberating model, which was matched to the recording in the mean rate, inferred m, and number of recorded neurons. In general, the experimental results (gray or blue) were best matched by this reverberating model (red), compared to asynchronous irregular (AI, green) and critical (yellow) models. From all experimental sessions, best examples (top) and typical examples (bottom) are displayed. For results from all experimental sessions see Figs. S2 to S8. **a**/**a'**. Inter-spike-interval (ISI) distributions. **b**/**b'**. Fano factors of single neurons for bin sizes between 4 ms and 4 s. **c**/**c'**. Distribution of spikes per bin $p(a_t)$ at a bin size of 4 ms. **d**/**d'**. Same as **c**/**c'** with a bin size of 40 ms. **e**/**e'**. *In vivo* avalanche size distributions $p(s)$ for all sampled units. AI activity lacks large avalanches, near critical activity produces power-law distributed avalanches, even under subsampling. **f**/**f'**. *In vivo* avalanche duration distributions $p(d)$ for all sampled units. textbfg/**g'**. Spike count cross-correlations (r_{sc}) as a function of the bin size.

Cross-validation of model predictions

We compared the experimental results to an *in vivo*-like model, which was matched to the recording only in the average firing rate of single neurons, and in the inferred branching ratio m. Remarkably, this *in vivo*-like branching network could predict statistical properties not only of single neurons as shown before (ISI and Fano factor, see above), but also pairwise and

population properties. This prediction capability further underlines the usefulness of this simple model to approximate the ground state of cortical dynamics.

First, the model predicted the activity distributions, $p(a_t)$, better than AI or critical networks for the majority of experiments (15 out 21, Figs. 4**c**,**d**,**c'**,**d'**, S5, S6), both for the exemplary bin sizes of 4 ms and 40 ms. Hence, branching networks only matched in their respective first moments of the activity distributions (through the rate) and first moments of the spreading behavior (through m) in fact approximated all higher moments of the activity distributions $p(a_t)$.

Likewise, the model predicted the distributions of neural avalanches, i.e. spatio-temporal clusters of activity (Figs. 4**e**,**f**,**e'**, **f'**, S7, S8). Characterizing these distributions is a classic approach to assess criticality in neuroscience,[1,10] because avalanche size and duration distributions, $p(s)$ and $p(d)$ respectively, follow power laws in critical systems (yellow). In contrast, for AI activity, they are approximately exponential[66] (green). The matched branching networks predicted neither exponential nor power law distributions for the avalanches, but very well matched the distributions of the experiment (compare red and blue). Indeed, model likelihood[64] favored the *in vivo*-like branching network over Poisson and critical networks for the majority experiments (18 out of 21, Fig. S7). Our results here are consistent with those of spiking activity in awake animals, which typically do not display power laws.[2,3,43] In contrast, most evidence for criticality has been based on *coarse* measures of neural activity (LFP, EEG, BOLD; see[3] and references therein).

Last, the model predicted the pairwise spike count cross correlation r_{sc}. In experiments, r_{sc} is typically between 0.01 and 0.25, depending on brain area, task, and most importantly, the analysis timescale (bin size).[37] For the cat recording the model even correctly predicted the bin size dependence of r_{sc} from $\bar{r}_{\text{sc}} \approx 0.004$ at a bin size of 4 ms (analytical result: Eq. (S11)) to $\bar{r}_{\text{sc}} \approx 0.3$ at a bin size of 2 s (Fig. 4**g**). Comparable results were also obtained for some monkey experiments. In contrast, correlations in most monkey experiments and rat hippocampal neurons showed smaller correlation than predicted (Figs. 4**g'**, S4). It is very surprising that the model correctly predicted the cross-correlation even in some experiments, as m was inferred only from the *temporal* structure of the spiking activity alone, whereas r_{sc} characterizes spatial dependencies.

Overall, by only estimating the effective synaptic strength m from the *in vivo* recordings, higher-order properties like avalanche size distributions, activity distributions and in some cases spike count cross correlations could be closely matched using the generic branching network.

The dynamical state determines responses to small stimuli

After validating the model using a set of statistical properties that are well accessible experimentally, we now turn to making predictions for yet unknown properties, namely network responses to small stimuli. In the line of London and colleagues[48], assume that on a background of spiking activity one single extra spike is triggered. This spike may in turn trigger new spikes, leading to a cascade of additional spikes Δ_t propagating through the network. A dynamical state with branching ratio m implies that *on average*, this perturbation decays with time constant $\tau = -\Delta t / \log m$. Similar to the approach in[48], the evolution of the mean firing rate, averaged over a reasonable number of trials (here: 500) unveils the nature of the underlying spike propagation: depending on m, the rate excursions will last longer, the higher m (Figs. 5**a**,**b**,**c**, S10**a**). The perturbations are not deterministic, but show trial-to-trial variability which also depends on m (S10**b**).

Unless $m > 1$, the theory of branching networks ensures that perturbations will die out eventually after a duration d, having accumulated a total of $\Delta = \sum_{t=1}^{d} \Delta_t$ extra spikes in total. This perturbation size Δ and duration d follow specific distributions,[50] which are determined by m: they are power law distributed in the critical state, with a cutoff for any $m < 1$ (Fig. 5**f**, Supplementary Figs. S10**c**,**d**). These distributions imply a characteristic perturbation size $\langle \Delta \rangle$ (Fig. 5**d**), which diverges at the critical point. The variability of the perturbation sizes is also

determined by m and also diverges at the critical point (inset of Fig. 5**d** and Supplementary Fig. S10**e**).

Taken together, these results imply that the closer a neuronal network is to criticality, the more sensitive it is to external perturbations and can better amplify small stimuli. At the same time, these networks also show larger trial-to-trial variability. For typical cortical networks, we found that the response to one single extra spike will on average comprise between 20 and 1000 additional spikes in total (Figs. 5**e**).

The dynamical state determines network susceptibility and variability

Moving beyond single spike perturbations, our model gives precise predictions for the network response to continuous stimuli. If extra action potentials are triggered at rate h in the network, the network will again amplify these external activations, depending on m. Provided an appropriate stimulation protocol, this rate response could be measured and our prediction tested in experiments (Fig. S10**g**). The susceptibility $\mathrm{d}r/\mathrm{d}h$ diverges at the critical transition and is unique to a specific branching ratio m. We predict that typical cortical networks will amplify a small, but continuous increase of the input rate about 50-fold (Fig. S10**h**, red).

While the mean activity is determined by the network input and its susceptibility, the network activity fluctuates around this mean value. The magnitude of these fluctuations in relation to the mean can be described by the network Fano factor $F = \mathrm{Var}[A_t]\,/\,\langle A_t\rangle$ (Fig. 5**g**). This quantity cannot be directly inferred from experimental recordings, because the Fano factor of subsampled populations severely underestimates the network Fano factor, as shown before. We here used our *in vivo*-like model to obtain estimates of the network Fano factor: for a bin size of 4 ms it is about $F \approx 40$ and rises to $F \approx 4000$ for bin sizes of several seconds.

Distinguishing afferent and recurrent activation

Last, our model gives an easily accessible approach to solving the following question: given a spiking neuronal network, which fraction of the activity is generated by recurrent activation from within the network, and which fraction can be attributed to external, afferent excitation? The branching model readily provides an answer: the fraction of externally generated activity is $h/\langle A\rangle = 1-m$ (Fig. 5**h**). In this framework, AI-like networks are completely driven by external input currents or noise, while reverberating networks generate a substantial fraction of their activity intrinsically. For the experiments investigated in this study, we inferred that between 0.1% and 7% of the activity are externally generated (median 2%, Fig. 5**i**). While this view may be simplistic given the complexity of neuronal network activity, keep in mind that "all models are wrong, but some are useful".[67] Here, the model has proven to provide a good first order approximation and therefore promises to make reasonable predictions on properties of spiking networks.

Discussion

Our results resolve contradictions between AI and critical states

Our results for spiking activity *in vivo* suggest that network dynamics shows AI-like statistics, because under subsampling the observed correlations are underestimated. In contrast, typical experiments assessing criticality potentially overestimated correlations by sampling from overlapping populations (LFP, EEG) and thereby hampered a fine distinction between critical and subcritical states.[68] By employing for the first time a consistent, quantitative estimation, we provided evidence that *in vivo* spiking population dynamics reflects a reverberating state, i.e. it lives in a narrow regime around $m = 0.98$. This result is supported by the findings by Dahmen and colleagues:[69] based on distributions of covariances, they inferred that cortical networks should operate in a regime below criticality. Given the generality of our results across different species, brain areas, and cognitive states, our results suggest self-organization to this regime as a general organization principle for neural network dynamics.

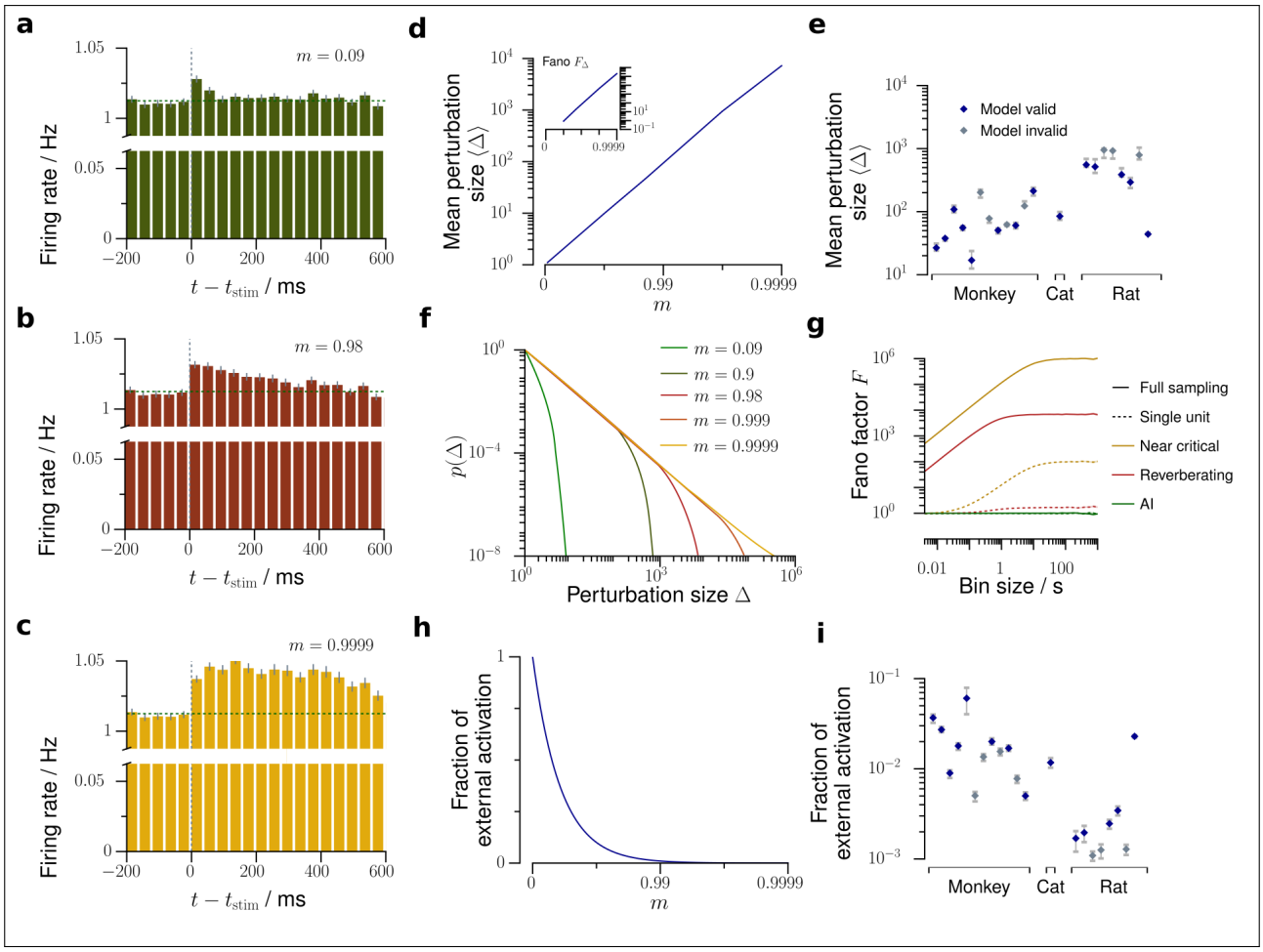

Figure 5: Predictions about network activity. Using our *in-vivo*-like, reverberating model, we can predict several network properties, which are yet very hard or impossible to obtain experimentally. **a – c.** In response to one single extra spike, a perturbation propagates in the network depending on the branching ratio m and can be observed as a small increase of the average firing rate of the sampled neurons, here simulated for 500 trials (see also London et al. [48]). This increase of firing rate decays exponentially, with the decay time τ being determined by m. The perturbation **a** is rapidly quenched in the asynchronous irregular network, **b** decays slowly over hundreds of milliseconds in the reverberating state, or **c** persists almost infinitely in the critical state. **d.** The average perturbation size $\langle \Delta \rangle$ and Fano factor F_Δ (inset) increase strongly with m. **e.** Average total perturbation sizes predicted for each spike recording of mammalian cortex (errorbars: 5% – 95% confidence intervals). **f.** Distribution $p(\Delta)$ of total perturbation sizes Δ. The asynchronous irregular networks show approximately Poisson distributed, near critical networks power-law distributed perturbation sizes. textbfg. Bin size dependent Fano factors of the activity, here exemplarily shown for the asynchronous irregular ($m = 0$, green), representative reverberating ($m = 0.98$, red), and near critical ($m = 0.9999$, yellow) networks. While the directly measurable Fano factor of single neurons (dotted lines) underestimates the Fano factor of the whole network, the model allows to predict the Fano factor of the whole network (solid lines). **h.** The fraction of the externally generated spikes compared to all spikes in the network strongly decreases with larger m. **i.** Fraction of the externally generated spikes predicted for each spike recording of mammalian cortex (errorbars as in **e**).

The reverberating state combines features of AI and critical regimes

Operating in a reverberating state, which is between AI and critical, may combine the computational advantages of the two dynamical states: (1) AI networks react to external input rapidly, and show very little reverberation of the input. In contrast, criticality is associated with "critical slowing down", i.e. performing any computation might take overly long. The $m = 0.98$ state

shows intermediate timescales of a few hundred milliseconds. These reverberations may carry short term memory and allow to integrate information over limited timescales.[27,70] (2) Criticality has been associated with maximal processing capacity. However, a number of everyday tasks, e.g. memory recall, require only sufficient capacity for survival and reproduction rather than maximum capacity.[71] Thus maximizing this one property alone is most likely not necessary from an evolutionary point of view. One particular example manifest from our results is the trade-off between sensitivity and reliability: while the critical state maximizes sensitivity by amplifying small stimuli (Fig. 5**h**), this sensitivity comes at the cost of increased trial-to-trial variability (Fig. 5**i**) and therefore may hinder reliable responses.[72] (3) Criticality in a branching process marks the transition to unstable dynamics. These instabilities have been associated with epilepsy.[56] The prevalence of epilepsy in humans[73,74] supports our results that the brain indeed operates biophysically still close to instability, but keeps a sufficient safety-margin to make seizures sufficiently unlikely.[3] This is in line with our results that the effective synaptic strength is close to, but not at $m = 1$.

More complex network models

Cortical dynamics is clearly more complicated than a simple branching network. For example, heterogeneity of neuronal morphology and function, non-trivial network topology, and the complexity of neurons themselves are likely to have a profound impact on the population dynamics. However, we showed that *statistics* of cortical networks are well approximated by a branching network. Therefore, we interpret branching networks as a *statistical* approximation of spike propagation, which can capture dynamics as complex as cortical activity. By using branching networks, we draw on the powerful advantage of analytical tractability, which allowed for basic insight into dynamics and stability of cortical networks.

It is a logical next step to refine the model by including additional relevant parameters, guided by the results obtained from the well-understood estimator. For example, our results show that networks with balanced excitation and inhibition,[8,75,76] which became a standard model of neuronal networks,[77] should be extended to incorporate the network reverberations observed *in vivo*. Possible candidate mechanisms are increased coupling strength or inhomogeneous connectivity. Both have already been shown to exhibit rate fluctuations with timescales of several hundred milliseconds.[78–80]

Likewise, neuron models of spike responses typically model normally distributed network synaptic currents, which originate from the assumption of uncorrelated Poisson inputs. Our results suggest that this input should rather exhibit reverberating properties with timescales of a few hundred milliseconds to reflect input from cortical neurons *in vivo*.

Deducing network properties from the tractable model

Using the tractable model, we could predict and validate network properties, such as distributions of avalanche sizes and durations, interspike intervals, or activities. Given the experimental agreement with these predictions, we deduced further properties, which are impossible or difficult to assess experimentally and gave insight into more complex questions about network responses: how do perturbations propagate within the network and how susceptible is the network to external stimulation?

One particular question we could address is the following: which fraction of network activity is attributed to external or recurrent, internal activation? We inferred that about 98% of the activity are generated by recurrent excitation. However, note that this result likely depends on the brain area and cognitive state investigated: For layer 4 of primary visual cortex in awake mice, Reinhold and colleagues[81] concluded that the fraction of recurrent cortical excitation rises to only about 72% and cortical activity dies out with a timescale of about 12 ms after thalamic silencing. Their numbers agree perfectly well with our phenomenological model: a timescale of 12 ms implies that the fraction of recurrent cortical activation is $m \approx 0.71$, just as found experimentally. Under anesthesia, in contrast, they report timescales of several hundred mil-

liseconds, in agreement with our results. These differences show that the fraction of external activation may strongly depend on cortical area, layer, and cognitive state. The novel estimator can in future contribute to a deeper insight into these differences, because it allows for a straight-forward assessment of afferent versus recurrent activation without the requirement of thalamic or cortical silencing.

Acknowledgments

JW received support from the Gertrud-Reemstma-Stiftung. VP received financial support from the German Ministry for Education and Research (BMBF) via the Bernstein Center for Computational Neuroscience (BCCN) Göttingen under Grant No. 01GQ1005B, and by the German-Israel-Foundation (GIF) under grant number G-2391-421.13. JW and VP received financial support from the Max Planck Society.

Competing interests

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

[1] V. Priesemann et al. "Subsampling effects in neuronal avalanche distributions recorded in vivo." *BMC Neurosci.* 10 (2009), 40.

[2] T. L. Ribeiro et al. "Spike Avalanches Exhibit Universal Dynamics across the Sleep-Wake Cycle". *PLoS One* 5.11 (2010). Ed. by O. Sporns, e14129.

[3] V. Priesemann et al. "Spike avalanches in vivo suggest a driven, slightly subcritical brain state." *Front. Syst. Neurosci.* 8.June (2014), 108.

[4] T. L. Ribeiro et al. "Undersampled critical branching processes on small-world and random networks fail to reproduce the statistics of spike avalanches." *PLoS One* 9.4 (2014), e94992.

[5] B. D. Burns and A. C. Webb. "The Spontaneous Activity of Neurones in the Cat's Cerebral Cortex". *Proc. R. Soc. B Biol. Sci.* 194.1115 (1976), 211–223.

[6] W. R. Softky and C Koch. "The highly irregular firing of cortical cells is inconsistent with temporal integration of random EPSPs." *J. Neurosci.* 13.1 (1993), 334–350.

[7] R. B. Stein et al. "Neuronal variability: noise or part of the signal?" *Nat. Rev. Neurosci.* 6.5 (2005), 389–397.

[8] C. v. Vreeswijk and H Sompolinsky. "Chaos in Neuronal Networks with Balanced Excitatory and Inhibitory Activity". *Science.* 274.5293 (1996), 1724–1726.

[9] N. Brunel. "Dynamics of networks of randomly connected excitatory and inhibitory spiking neurons". *J. Physiol. Paris* 94.5-6 (2000), 445–463.

[10] J. M. Beggs and D. Plenz. "Neuronal Avalanches in Neocortical Circuits". *J. Neurosci.* 23.35 (2003), 11167–11177.

[11] a. Levina et al. "Dynamical synapses causing self-organized criticality in neural networks". *Nat. Phys.* 3.12 (2007), 857–860.

[12] A. Levina et al. "Phase Transitions towards Criticality in a Neural System with Adaptive Interactions". *Phys. Rev. Lett.* 102.11 (2009), 118110.

[13] D. R. Chialvo. "Emergent complex neural dynamics". *Nat. Phys.* 6.10 (2010), 744–750.

[14] J. M. Beggs and N. Timme. "Being critical of criticality in the brain". *Front. Physiol.* 3 JUN.June (2012), 1–14.

[15] D. Plenz and E. Niebur, eds. *Criticality in Neural Systems.* Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA, 2014.

[16] G. Tkačik et al. "Thermodynamics and signatures of criticality in a network of neurons". *Proc. Natl. Acad. Sci.* 112.37 (2015), 11508–11513.

[17] J. Humplik and G. Tkačik. "Probabilistic models for neural populations that naturally capture global coupling and criticality". *PLoS Comput. Biol.* 13.9 (2017), 1–26.

[18] H. B. Barlow. "Possible Principles Underlying the Transformations of Sensory Messages". In: *Sens. Commun.* Ed. by W. A. Rosenblith. The MIT Press, 2012, 217–234.

[19] J. J. Atick. "Could information theory provide an ecological theory of sensory processing?" *Netw. Comput. neural Syst.* 3.2 (1992), 213–251.

[20] A. J. Bell and T. J. Sejnowski. "The 'independent components' of natural scenes are edge filters". *Vision Res.* 37.23 (1997), 3327–3338.

[21] J. H. van Hateren and A. van der Schaaf. "Independent component filters of natural images compared with simple cells in primary visual cortex". *Proc. R. Soc. London B Biol. Sci.* 265.1394 (1998), 359–366.

[22] A. Hyvärinen and E. Oja. "Independent component analysis: Algorithms and applications". *Neural Networks* 13 (2000), 411–430.

[23] N. Bertschinger and T. Natschläger. "Real-time computation at the edge of chaos in recurrent neural networks". *Neural Comput.* 16.7 (2004), 1413–1436.

[24] C. Haldeman and J. Beggs. "Critical Branching Captures Activity in Living Neural Networks and Maximizes the Number of Metastable States". *Phys. Rev. Lett.* 94.5 (2005), 058101.

[25] O. Kinouchi and M. Copelli. "Optimal dynamical range of excitable networks at criticality". *Nat. Phys.* 2.5 (2006), 348–351.

[26] X. R. Wang et al. "Fisher information at the edge of chaos in random Boolean networks". *Artif. Life* 17.4 (2011), 315–329.

[27] J. Boedecker et al. "Information processing in echo state networks at the edge of chaos". *Theory Biosci.* 131.3 (2012), 205–213.

[28] W. L. Shew and D. Plenz. "The functional benefits of criticality in the cortex." *Neuroscientist* 19.1 (2013), 88–100.

[29] B. Del Papa et al. "Criticality meets learning: Criticality signatures in a self-organizing recurrent neural network". *PLoS One* 12.5 (2017), 1–22.

[30] H. Jaeger. "Harnessing Nonlinearity: Predicting Chaotic Systems and Saving Energy in Wireless Communication". *Science.* 304.5667 (2004), 78–80.

[31] C. Huang and B. Doiron. "Once upon a (slow) time in the land of recurrent neuronal networks...". *Curr. Opin. Neurobiol.* 46 (2017), 31–38.

[32] R. Chaudhuri et al. "A Large-Scale Circuit Mechanism for Hierarchical Dynamical Processing in the Primate Cortex". *Neuron* 88.2 (2015), 419–431.

[33] C. J. Honey et al. "Slow Cortical Dynamics and the Accumulation of Information over Long Timescales". *Neuron* 76.2 (2012), 423–434.

[34] J. D. Murray et al. "A hierarchy of intrinsic timescales across primate cortex." *Nat. Neurosci.* 17.12 (2014), 1661–3.

[35] R. R. de Ruyter van Steveninck et al. "Reproducibility and Variability in Neural Spike Trains". *Science.* 275.5307 (1997), 1805–1808.

[36] A. S. Ecker et al. "Supporting Online Material for Decorrelated Neuronal Firing in Cortical Microcircuits". *Science.* 327.584 (2010), 26–31.

[37] M. R. Cohen and A. Kohn. "Measuring and interpreting neuronal correlations." *Nat. Neurosci.* 14.7 (2011), 811–819.

[38] V Pasquale et al. "Self-organization and neuronal avalanches in networks of dissociated cortical neurons." *Neuroscience* 153.4 (2008), 1354–69.

[39] N. Friedman et al. "Universal Critical Dynamics in High Resolution Neuronal Avalanche Data". *Phys. Rev. Lett.* 108.20 (2012), 208102.

[40] E. Tagliazucchi et al. "Criticality in large-scale brain fmri dynamics unveiled by a novel point process analysis". *Front. Physiol.* 3.February (2012), 1–12.

[41] O. Shriki et al. "Neuronal avalanches in the resting MEG of the human brain." *J. Neurosci.* 33.16 (2013), 7079–90.

[42] P. Bak et al. "Self-organized criticality: An explanation of the 1/f noise". *Phys. Rev. Lett.* 59.4 (1987), 381–384.

[43] C. Bédard et al. "Does the 1/f frequency scaling of brain signals reflect self-organized critical states?" *Phys. Rev. Lett.* 97.11 (2006), 1–4.

[44] G. Hahn et al. "Neuronal avalanches in spontaneous activity in vivo." *J. Neurophysiol.* 104.6 (2010), 3312–22.

[45] V. Priesemann et al. "Neuronal avalanches differ from wakefulness to deep sleep–evidence from intracranial depth recordings in humans." *PLoS Comput. Biol.* 9.3 (2013), e1002985.

[46] M. Girardi-Schappo et al. "Critical avalanches and subsampling in map-based neural networks coupled with noisy synapses". *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.* 88.2 (2013), 1–5.

[47] A. Levina and V. Priesemann. "Subsampling scaling". *Nat. Commun.* 8.May (2017), 1–9.

[48] M. London et al. "Sensitivity to perturbations in vivo implies high noise and suggests rate coding in cortex." *Nature* 466.7302 (2010), 123–7.

[49] J. Wilting and V. Priesemann. "Inferring collective dynamical states from widely unobserved systems". *Nat. Commun.* 9.1 (2018), 2325.

[50] T. E. Harris. *The Theory of Branching Processes*. Springer Berlin, 1963.

[51] C. R. Heathcote. "A Branching Process Allowing Immigration". *J. R. Stat. Soc. Ser. B* 27.1 (1965), 138–143.

[52] A. G. Pakes. "Branching Processes with Immigration". *J. Appl. Probab.* 8.1 (1971), 32.

[53] J. Zierenberg et al. "Homeostatic plasticity and external input shape neural network dynamics". *Phys. Rev. X* In Press (2018).

[54] C. C. Heyde and E. Seneta. "Estimation Theory for Growth and Immigration Rates in a Multiplicative Process". *J. Appl. Probab.* 9.2 (1972), 235.

[55] C. Z. Wei. "Convergence rates for the critical branching process with immigration". *Stat. Sin.* 1 (1991), 175–184.

[56] C. Meisel et al. "Failure of adaptive self-organized criticality during epileptic seizure attacks." *PLoS Comput. Biol.* 8.1 (2012), e1002312.

[57] K Mizuseki et al. *Multi-unit recordings from the rat hippocampus made during open field foraging*. 2009.

[58] K Mizuseki et al. "Theta Oscillations Provide Temporal Windows for Local Circuit Computation in the Entorhinal-Hippocampal Loop". *Neuron* 64.2 (2009), 267–280.

[59] T. Blanche. *Multi-neuron recordings in primary visual cortex*. 2009.

[60] T. J. Blanche and N. V. Swindale. "Nyquist interpolation improves neuron yield in multiunit recordings." *J. Neurosci. Methods* 155.1 (2006), 81–91.

[61] G. Pipa et al. "Performance- and stimulus-dependent oscillations in monkey prefrontal cortex during short-term memory." *Front. Integr. Neurosci.* 3.October (2009), 25.

[62] F. Franke et al. "An online spike detection and spike classification algorithm capable of instantaneous resolution of overlapping spikes." *J. Comput. Neurosci.* 29.1-2 (2010), 127–48.

[63] R. Pröpper and K. Obermayer. "Spyke Viewer: a flexible and extensible platform for electrophysiological data analysis." *Front. Neuroinform.* 7.November (2013), 26.

[64] A. Clauset et al. "Power-Law Distributions in Empirical Data". *SIAM Rev.* 51.4 (2009), 661–703.

[65] H. Cuntz et al. "One Rule to Grow Them All: A General Theory of Neuronal Branching and Its Practical Application". *PLoS Comput. Biol.* 6.8 (2010), e1000877.

[66] V. Priesemann and O. Shriki. "Can a time varying external drive give rise to apparent criticality in neural systems?" *In preparation.* ().

[67] G. E. P. Box. "Robustness in the strategy of scientific model building". In: *Robustness Stat.* Ed. by R. L. Launer and G. N. Wilkinson. Vol. 1. Academic Press, 1979, 201–236.

[68] J. Pinheiro Neto and V. Priesemann. "Coarse sampling bias inference of criticality in neural system". *In review.* ().

[69] D. Dahmen et al. "Distributions of covariances as a window into the operational regime of neuronal networks". *Prepr. http//arxiv.org/abs/1605.04153* (2016).

[70] X. J. Wang. "Probabilistic decision making by slow reverberation in cortical circuits". *Neuron* 36.5 (2002), 955–968.

[71] A. Lazar. "SORN: a Self-organizing Recurrent Neural Network". *Front. Comput. Neurosci.* 3.October (2009), 1–9.

[72] L. L. Gollo. "Coexistence of critical sensitivity and subcritical specificity can yield optimal population coding". *J. R. Soc. Interface* 14.134 (2017), 20170207.

[73] J. W. Sander. "The epidemiology of epilepsy revisited". *Curr. Opin. Neurol.* 16.2 (2003), 165–170.

[74] A. Neligan et al. "The epidemiology of the epilepsies". In: *Handb. Clin. Neurol.* Vol. 107. 2012, 113–133.

[75] C. van Vreeswijk and H Sompolinsky. "Irregular firing in cortical circuits with inhibition/excitation balance". *Comput. Neurosci. Trends Reserach, 1997* (1997), 209–213.

[76] N. Brunel. "Dynamics of networks of randomly connected excitatory and inhibitory spiking neurons". *J. Physiol. Paris* 94.5-6 (2000), 445–463.

[77] D. Hansel and C. van Vreeswijk. "The Mechanism of Orientation Selectivity in Primary Visual Cortex without a Functional Map". *J. Neurosci.* 32.12 (2012), 4049–4064.

627 [78] A. Litwin-kumar and B. Doiron. "Slow dynamics and high variability in balanced cortical networks with clus-
628 tered connections". *Nat. Neurosci.* 15.11 (2012), 1498–1505.

629 [79] S. Ostojic. "Two types of asynchronous activity in networks of excitatory and inhibitory spiking neurons." *Nat*
630 *Neurosci* 17.4 (2014), 594–600.

631 [80] J. Kadmon and H. Sompolinsky. "Transition to chaos in random neuronal networks". *Phys. Rev. X* 5.4 (2015),
632 1–28.

633 [81] K. Reinhold et al. "Distinct recurrent versus afferent dynamics in cortical visual processing." *Nat. Neurosci.*
634 18.12 (2015), 1789–1797.

Supplementary material for "Inference, validation and predictions about statistics and propagation of cortical spiking in vivo" by J. Wilting and V. Priesemann

Supp. 1 Branching processes

In a branching process (BP) with immigration[50–52] each unit i produces a random number $y_{t,i}$ of units in the subsequent time step. Additionally, in each time step a random number h_t of units immigrates into the system (drive). Mathematically, BPs are defined as follows:[50,51] Let $y_{t,i}$ be independently and identically distributed non-negative integer-valued random variables following a law Y with mean $m = \langle Y \rangle$ and variance $\sigma^2 = \mathrm{Var}[Y]$. Further, Y shall be non-trivial, meaning it satisfies $\mathrm{P}[Y = 0] > 0$ and $\mathrm{P}[Y = 0] + \mathrm{P}[Y = 1] < 1$. Likewise, let h_t be independently and identically distributed non-negative integer-valued random variables following a law H with mean rate $h = \langle H \rangle$ and variance $\xi^2 = \mathrm{Var}[H]$. Then the evolution of the BP A_t is given recursively by

$$A_{t+1} = \sum_{i=1}^{A_t} y_{t,i} + h_t, \tag{S1}$$

i.e. the number of units in the next generation is given by the offspring of all present units and those that were introduced to the system from outside.

The stability of BPs is solely governed by the mean offspring m. In the subcritical state, $m < 1$, the population converges to a stationary distribution A_∞ with mean $\langle A_\infty \rangle = h/(1-m)$. At criticality ($m = 1$), A_t asymptotically exhibits linear growth, while in the supercritical state ($m > 1$) it grows exponentially.

We will now derive results for the mean, variance, and Fano factor of subcritical branching processes. Following previous results, taking expectation values of both sides of Eq. (S1) yields $\langle A_{t+1} \rangle = m \langle A_t \rangle + h$. Because of stationarity $\langle A_{t+1} \rangle = \langle A_t \rangle = \langle A_\infty \rangle$ and the mean activity is given by

$$\langle A_\infty \rangle = \frac{h}{1-m}. \tag{S2}$$

In order to derive an expression for the variance of the stationary distribution, observe that by the theorem of total variance, $\mathrm{Var}[A_{t+1}] = \langle \mathrm{Var}[A_{t+1} \,|\, A_t] \rangle + \mathrm{Var}[\langle A_{t+1} \,|\, A_t \rangle]$, where $\langle \cdot \rangle$ denotes the expected value, and $A_{t+1} \,|\, A_t$ conditioning the random variable A_{t+1} on A_t. Because A_{t+1} is the sum of independent random variables, the variances also sum: $\mathrm{Var}[A_{t+1} \,|\, A_t] = \sigma^2 A_t + \xi^2$. Using the previous result for $\langle A_\infty \rangle$ one then obtains

$$\mathrm{Var}[A_{t+1}] = \xi^2 + \sigma^2 \frac{h}{1-m} + \mathrm{Var}[mA_t + h] = \xi^2 + \sigma^2 \frac{h}{1-m} + m^2 \mathrm{Var}[A_t].$$

Again, in the stationary distribution $\mathrm{Var}[A_{t+1}] = \mathrm{Var}[A_t] = \mathrm{Var}[A_\infty]$ which yields

$$\mathrm{Var}[A_\infty] = \frac{1}{1-m^2} \left(\xi^2 + \sigma^2 \frac{h}{1-m} \right), \tag{S3}$$

The Fano factor $F_{A_t} = \mathrm{Var}[A_t] / \langle A_t \rangle$ is easily computed from (S2) and (S3):

$$F_{A_t} = \frac{\xi^2}{h(1+m)} + \frac{\sigma^2}{1-m^2}. \tag{S4}$$

Interestingly, the mean rate, variance, and Fano factor all diverge when approaching criticality (given a constant input rate h): $\langle A_\infty \rangle \to \infty$, $\mathrm{Var}[A_\infty] \to \infty$, and $F_{A_t} \to \infty$ as $m \to 1$.

These results were derived without assuming any particular law for Y or H. Although the limiting behavior of BPs does not depend on it,[50–52] fixing particular laws allows to simplify these expressions further.

We here chose Poisson distributions with means m and h for Y and H respectively: $y_{t,i} \sim \mathrm{Poi}(m)$ and $h_t \sim \mathrm{Poi}(h)$. We chose these laws for two reasons: (1) Poisson distributions allow for non-trivial offspring distributions with easy control of the branching ratio m by only one parameter. (2) For the brain, one might assume that each neuron is connected to k postsynaptic neurons, each of which is excited with probability p, motivating a binomial offspring distribution with mean $m = k\,p$. As in cortex k is typically large and p is typically small, the Poisson limit is a reasonable approximation. Choosing these distributions, the variance and Fano factor become

$$\begin{aligned} \mathrm{Var}[A_t] &= h \,/\, ((1-m)^2(1+m)), \\ F_{A_t} &= 1 \,/\, (1-m^2). \end{aligned} \tag{S5}$$

Both diverge when approaching criticality ($m = 1$).

Supp. 2 Subsampling

A general notion of subsampling was introduced in Wilting and Priesemann [49]. The subsampled time series a_t is constructed from the full process A_t based on the three assumptions: (i) The sampling process does not interfere with itself, and does not change over time. Hence the realization of a subsample at one time does not influence the realization of a subsample at another time, and the conditional *distribution* of $(a_t|A_t)$ is the same as $(a_{t'}|A_{t'})$ if $A_t = A_{t'}$. However, even if $A_t = A_{t'}$, the subsampled a_t and $a_{t'}$ do not necessarily take the same value. (ii) The subsampling does not interfere with the evolution of A_t, i.e. the process evolves independent of the sampling. (iii) *On average* a_t is proportional to A_t up to a constant term, $\langle a_t \,|\, A_t \rangle = \alpha A_t + \beta$.

In the spike recordings analyzed in this study, the states of a subset of neurons are observed by placing electrodes that record the activity of the same set of neurons over the entire recording. This implementation of subsampling translates to the general definition in the following manner: If n out of all N neurons are sampled, the probability to sample a_t active neurons out of the actual A_t active neurons follows a hypergeometric distribution, $a_t \sim \mathrm{Hyp}(N, n, A_t)$. As $\langle a_t \,|\, A_t = j \rangle = j\,n \,/\, N$, this representation satisfies the mathematical definition of subsampling with $\alpha = n \,/\, N$. Choosing this special implementation of subsampling allows to derive predictions for the Fano factor under subsampling and the spike count cross correlation. First, evaluate $\mathrm{Var}[a_t]$ further in terms of A_t:

$$\begin{aligned} \mathrm{Var}[a_t] &= \langle \mathrm{Var}[a_t \,|\, A_t] \rangle + \mathrm{Var}[\langle a_t \,|\, A_t \rangle] \\ &= n \langle \frac{A_t}{N} \frac{N - A_t}{N} \frac{N-n}{N-1} \rangle + \mathrm{Var}[\frac{n}{N} A_t] \\ &= \frac{1}{N} \frac{n}{N} \frac{N-n}{N-1} \left(N \langle A_t \rangle - \langle A_t^2 \rangle \right) + \frac{n^2}{N^2} \mathrm{Var}[A_t] \\ &= \frac{n}{N^2} \frac{N-n}{N-1} \left(N \langle A_t \rangle - \langle A_t \rangle^2 \right) + \left(\frac{n^2}{N^2} - \frac{n}{N^2} \frac{N-n}{N-1} \right) \mathrm{Var}[A_t]. \end{aligned} \tag{S6}$$

This expression precisely determines the variance $\mathrm{Var}[a_t]$ under subsampling from the properties $\langle A_t \rangle$ and $\mathrm{Var}[A_t]$ of the full process, and from the parameters of subsampling n and N. We now show that the Fano factor approaches and even falls below unity under strong subsampling, regardless of the underlying dynamical state m. In the limit of strong subsampling ($n \ll N$) Eq. (S6) yields:

$$\mathrm{Var}[a_t] \approx \frac{n}{N^2}\left(N\langle A_t\rangle - \langle A_t\rangle^2\right) + \frac{n^2-n}{N^2}\mathrm{Var}[A_t]. \tag{S7}$$

Hence the subsampled Fano factor is given by

$$F_{a_t} = \frac{\mathrm{Var}[a_t]}{\langle a_t\rangle} \approx 1 - \frac{\langle A_t\rangle}{N} + \frac{n-1}{N}\frac{\mathrm{Var}[A_t]}{\langle A_t\rangle} = 1 - \frac{\langle A_t\rangle - (n-1)F_{A_t}}{N}. \tag{S8}$$

Interestingly, when sampling a single unit ($n = 1$) the Fano factor of that unit becomes completely independent of the Fano factor of the full process:

$$F_{a_t} = 1 - \langle A_t\rangle/N = 1 - \langle a_t\rangle/n = 1 - R, \tag{S9}$$

where $R = \langle a_t\rangle/n$ is the mean rate of a single unit.

Based on this implementation of subsampling, we derived analytical results for the cross-correlation between the activity of two units on the time scale of one time step. The pair of units is here represented by two independent samplings a_t and $\tilde{a}(t)$ of a BP A_t with $n = 1$, i.e. each represents one single unit. Because both samplings are drawn from identical distributions, their variances are identical and hence the correlation coefficient is given by $r_{\mathrm{sc}} = \mathrm{Cov}(a_t, \tilde{a}(t))\,/\mathrm{Var}[a_t]$. Employing again the law of total expectation and using the independence of the two samplings, this can be evaluated:

$$\mathrm{Cov}(a_t, \tilde{a}(t)) = \langle\langle a_t\,\tilde{a}(t)\,|\,A_t\rangle\rangle_{A_t} - \langle\langle a_t\,|\,A_t\rangle\rangle^2_{A_t} = \frac{1}{N^2}\mathrm{Var}[A_t], \tag{S10}$$

with the first inner expectation being taken over the joint distribution of a_t and $\tilde{a}(t)$. Using Eq. (S7), one easily obtains

$$r_{\mathrm{sc}} = \frac{\mathrm{Var}[A_t]}{N\langle A_t\rangle - \langle A_t\rangle^2} = \frac{F_{A_t}}{N - \langle A_t\rangle} = \frac{F_{A_t}}{N\,(1-R)} \tag{S11}$$

with the mean single unit rate $R = \langle A_t\rangle/N$. For subcritical systems, the Fano factor F_{A_t} is much smaller than N, and the rate is typically much smaller than 1. Therefore, the cross-correlation between single units is typically very small.

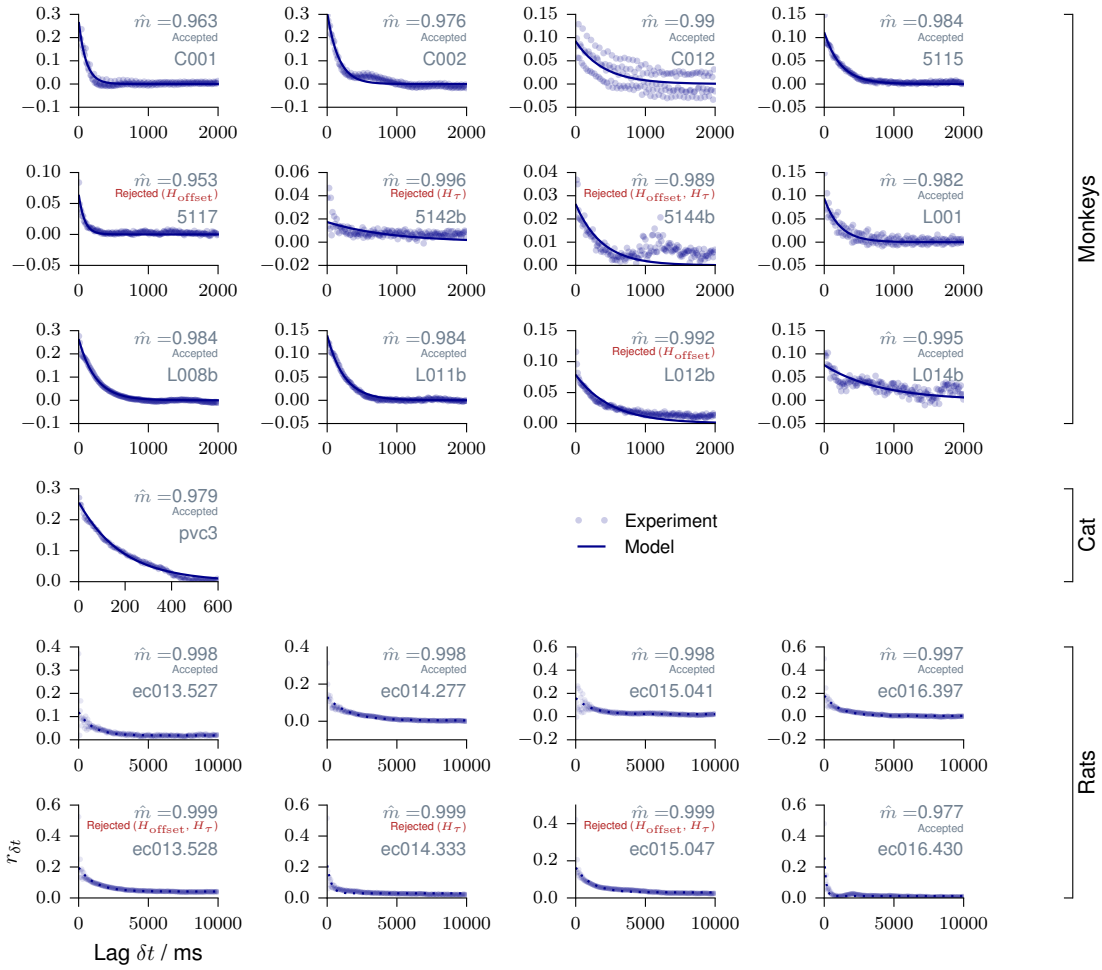

Figure S1: MR estimation for individual recording sessions. Reproduced from[49]. MR estimation is shown for every individual animal. The consistency checks are detailed in[49]. Data from monkey were recorded in prefrontal cortex during an working memory task. The third panel shows a oscillation of r_k with a frequency of 50 Hz, corresponding to measurement corruption due to power supply frequency. Data from anesthetized cat were recorded in primary visual cortex. Data from rat were recorded in hippocampus during a foraging task. In addition to a slow exponential decay, the slopes r_k show the ϑ-oscillations of 6 – 10 Hz present in hippocampus.

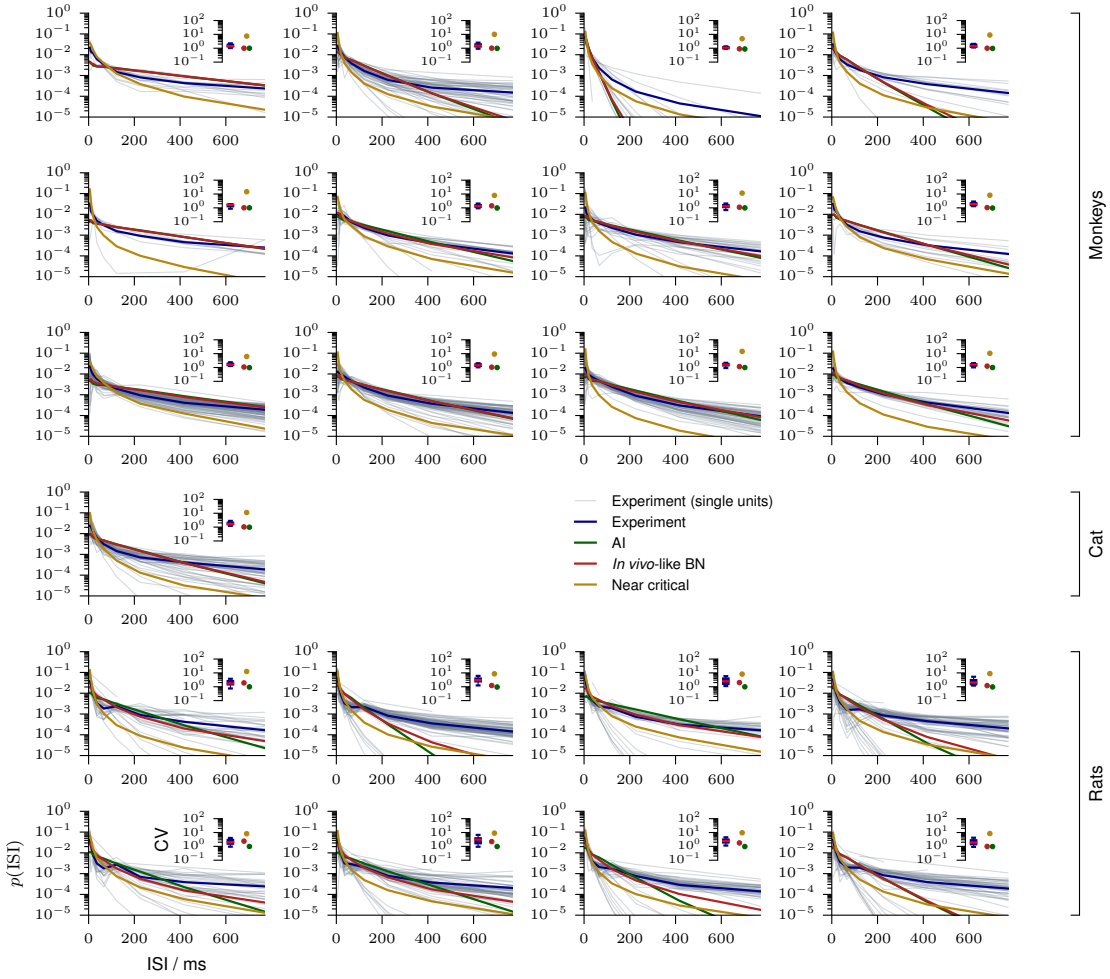

Figure S2: Interspike interval distribution for individual recording sessions. Interspike interval (ISI) distributions are shown for individual units of each recording (gray), for the average over units of each recording (blue), as well as for the matched models, either AI (green), *in vivo*-like (red), or near critical (yellow). The insets show the corresponding coefficients of variation (CV). For every experiment AI and *in vivo*-like models are virtually indistinguishable by the ISI distributions.

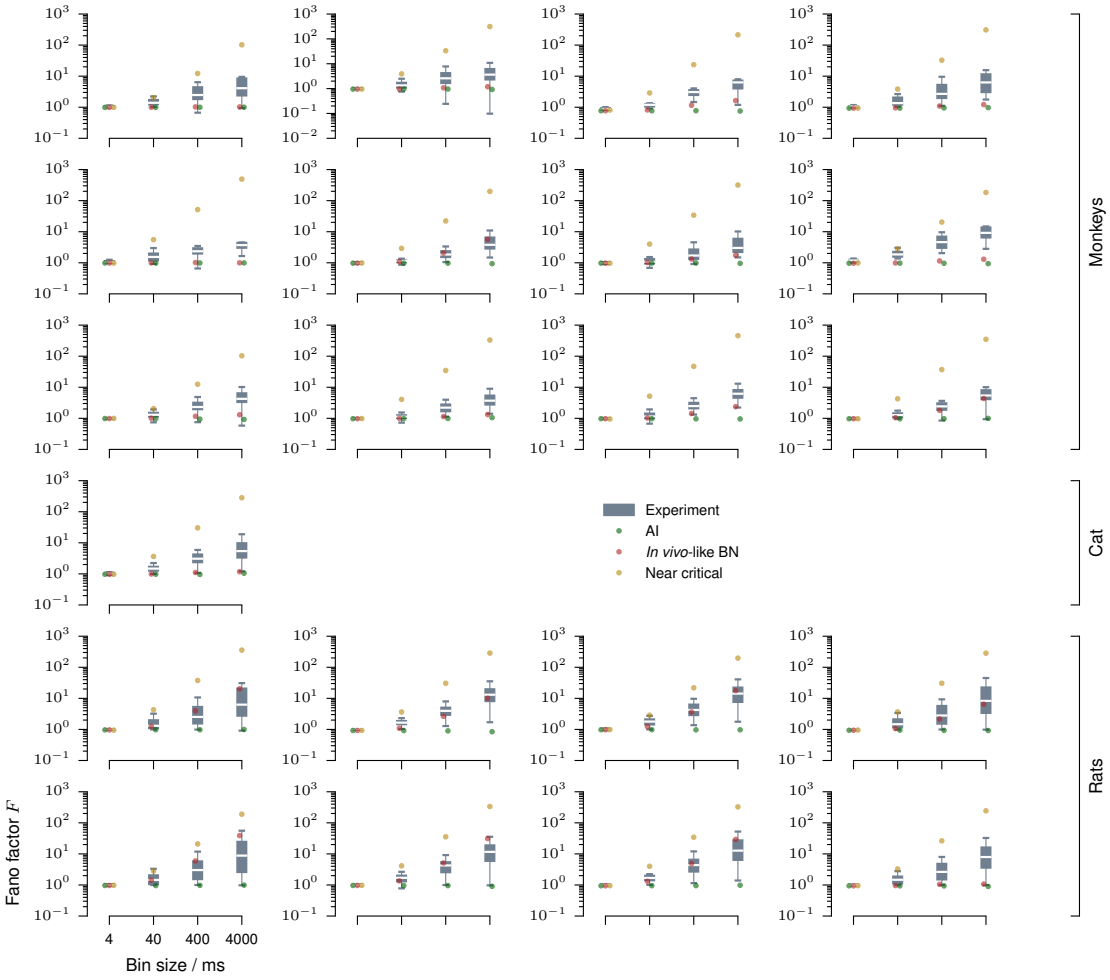

Figure S3: Fano factors for individual recording sessions. Fano factors are shown for individual single or multi units of every recording (gray boxplots, median / 25% – 75%, 2.5% – 97.5%), as well as for the matched models, either AI (green), *in vivo*-like (red), or near critical (yellow).

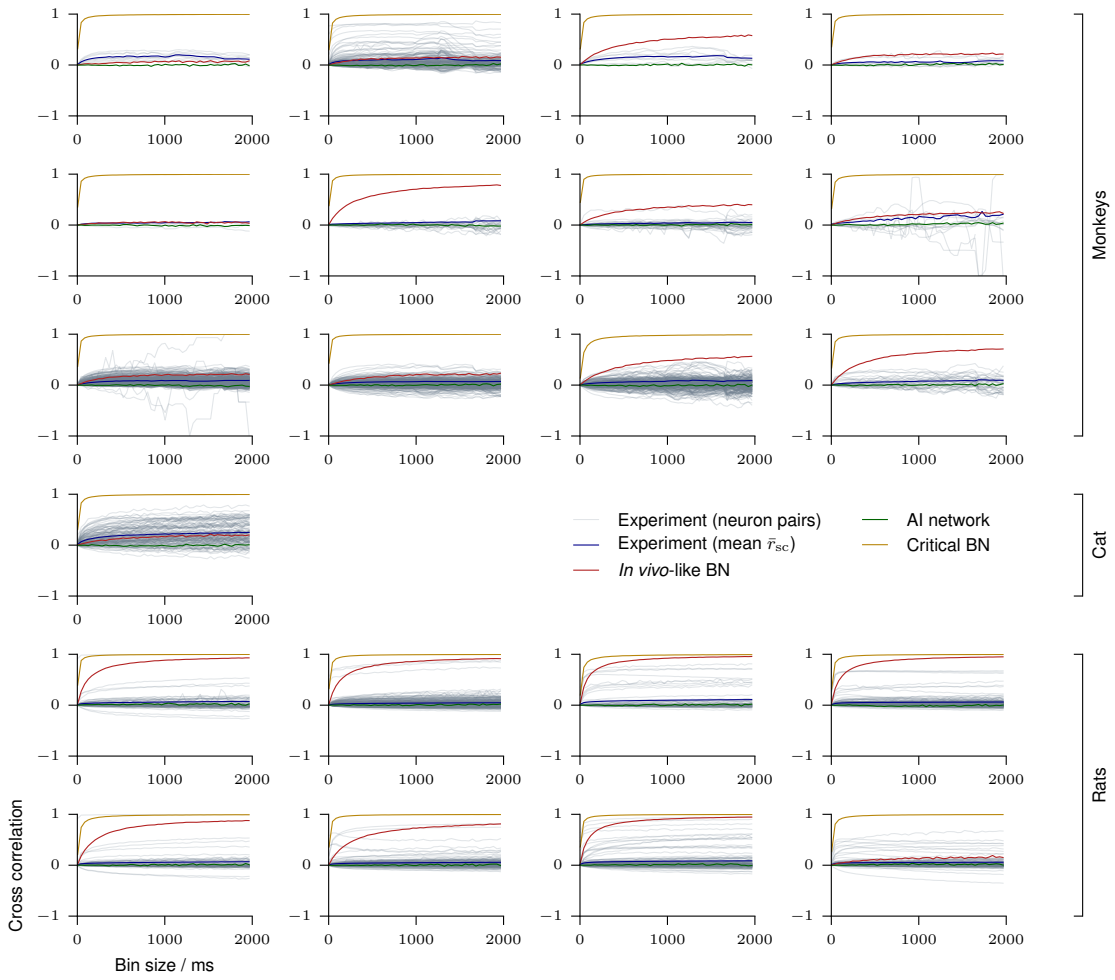

Figure S4: Cross correlations for individual recording sessions. Spike count cross correlations (r_{sc}) are shown for every neuron pair (gray) and the ensemble average (blue) of each recording, for bin sizes from 1 ms to 2s. Cross correlations are also shown for the matched models, either AI (green), *in vivo*-like (red), or near critical (yellow).

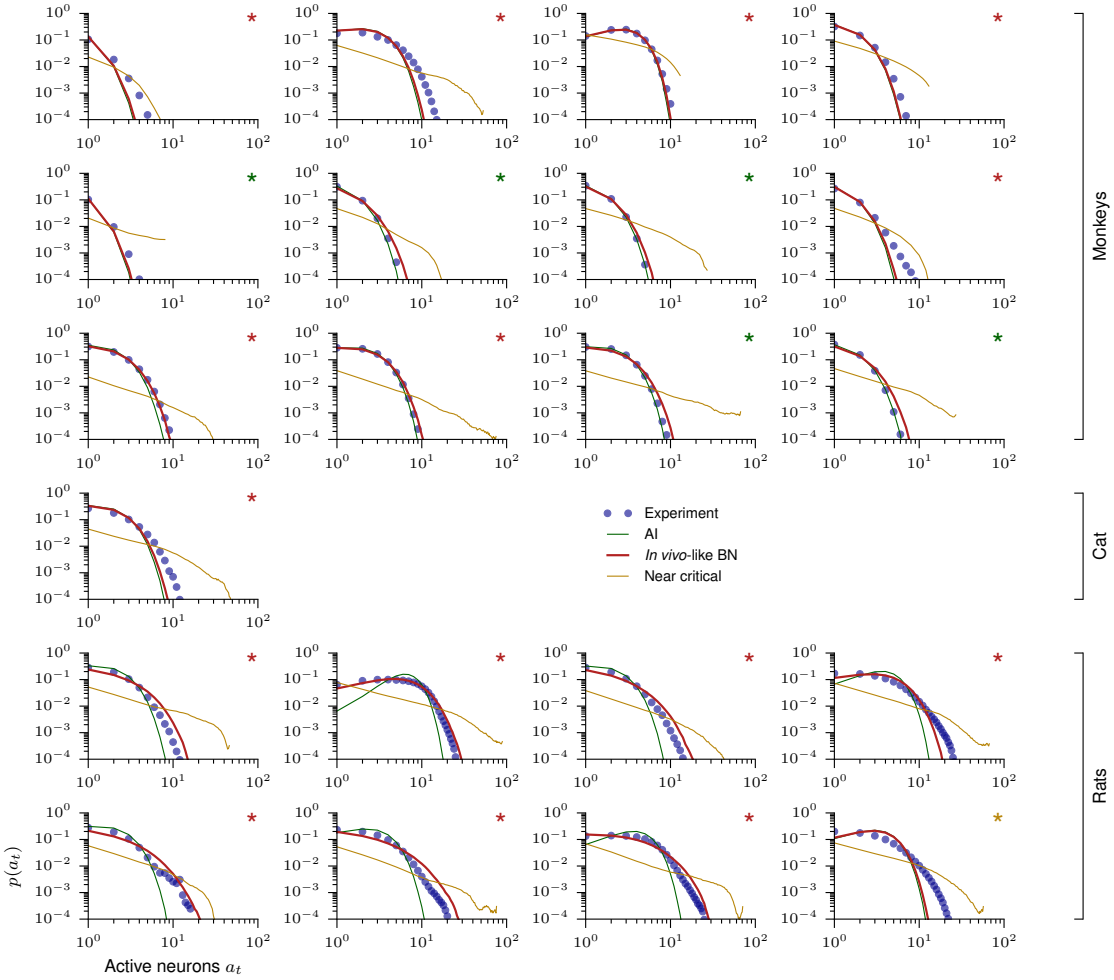

Figure S5: Activity distributions (4 ms bin size). Activity distributions are shown for every recording for a bin size of 4 ms (blue). Activity distributions for the matched models, either AI (green), *in vivo*-like (red), or near critical (yellow) are also shown. The color of the asterisk indicates which of the three models yielded the highest likelihood for the data following[64].

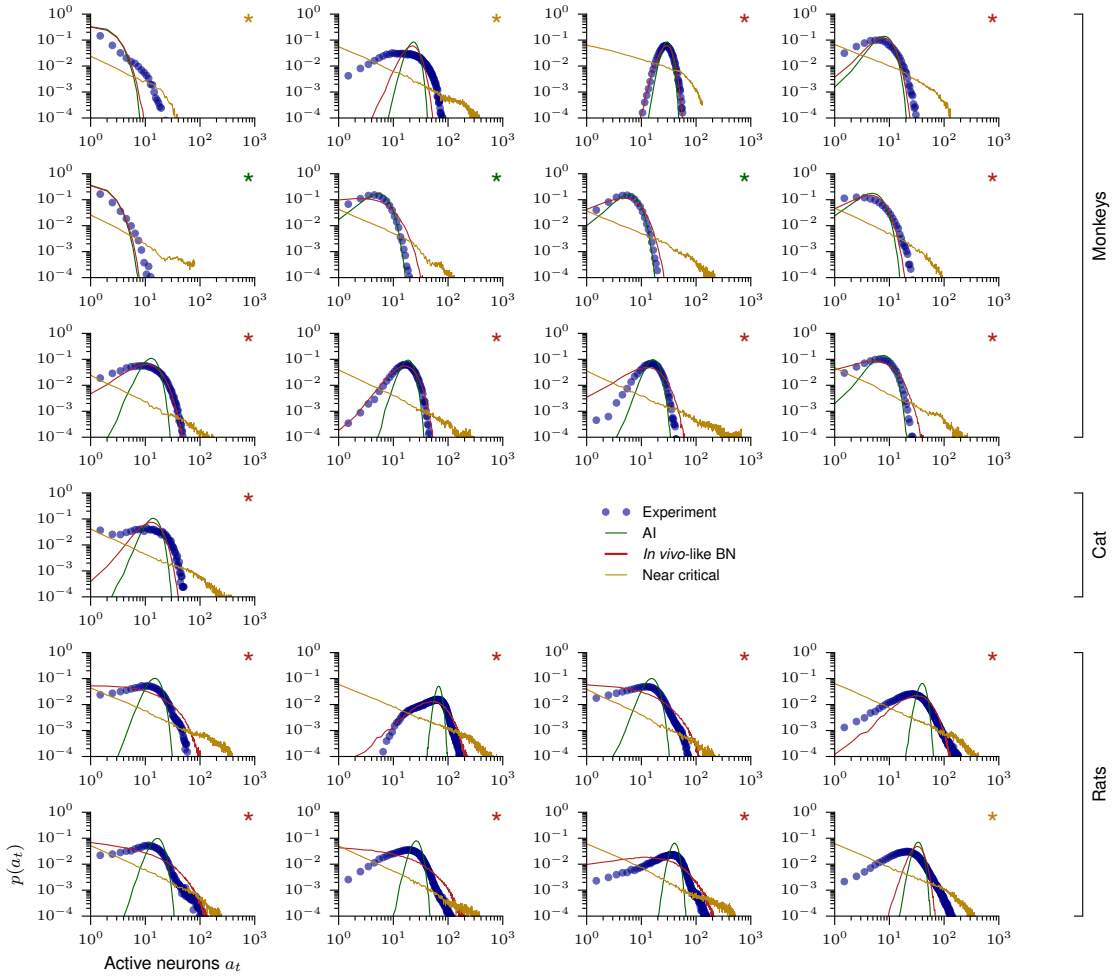

Figure S6: Activity distributions (40 ms bin size). Activity distributions are shown for every recording, for a bin size of 40 ms (blue). Activity distributions for the matched models, either AI (green), *in vivo*-like (red), or near critical (yellow) are also shown. The color of the asterisk indicates which of the three models yielded the highest likelihood for the data following[64].

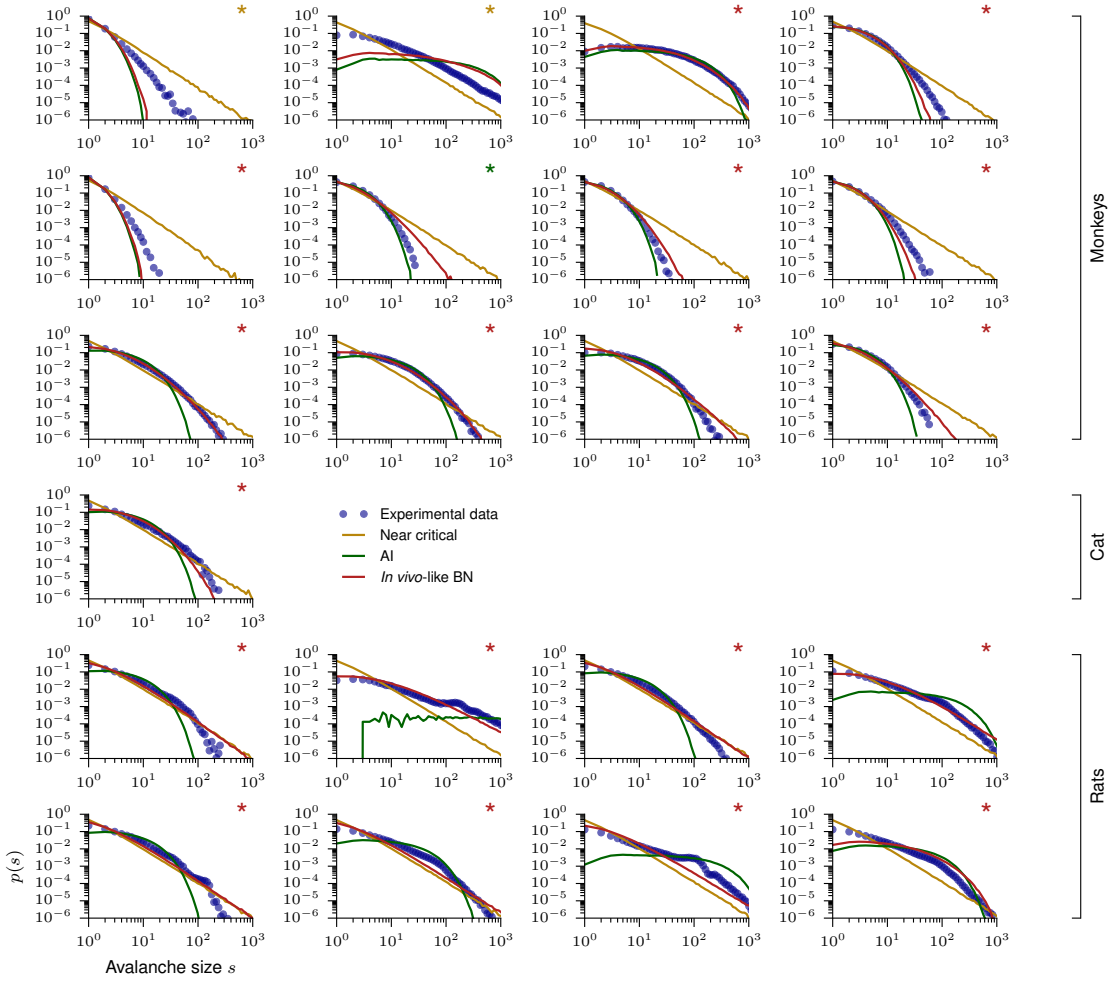

Figure S7: Avalanche size distribution for individual recording sessions. Avalanche size distributions are shown for every recording (blue) and for matched models, either AI (green), *in vivo*-like (red), or near critical (yellow). The color of the asterisk indicates which of the three models yielded the highest likelihood for the data following[64].

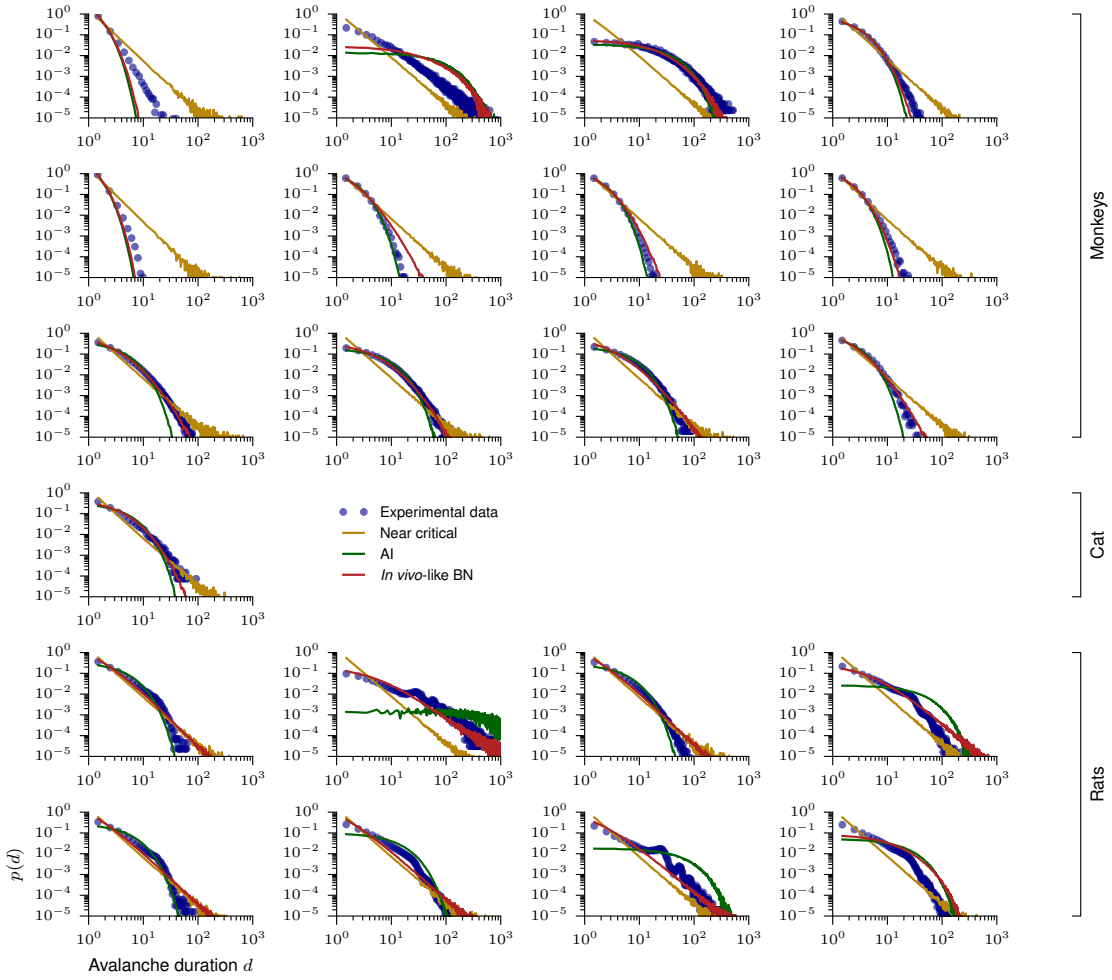

Figure S8: Avalanche duration distribution for individual recording sessions. Avalanche duration distributions are shown for every recording (blue) and for matched models, either AI (green), *in vivo*-like (red), or near critical (yellow).

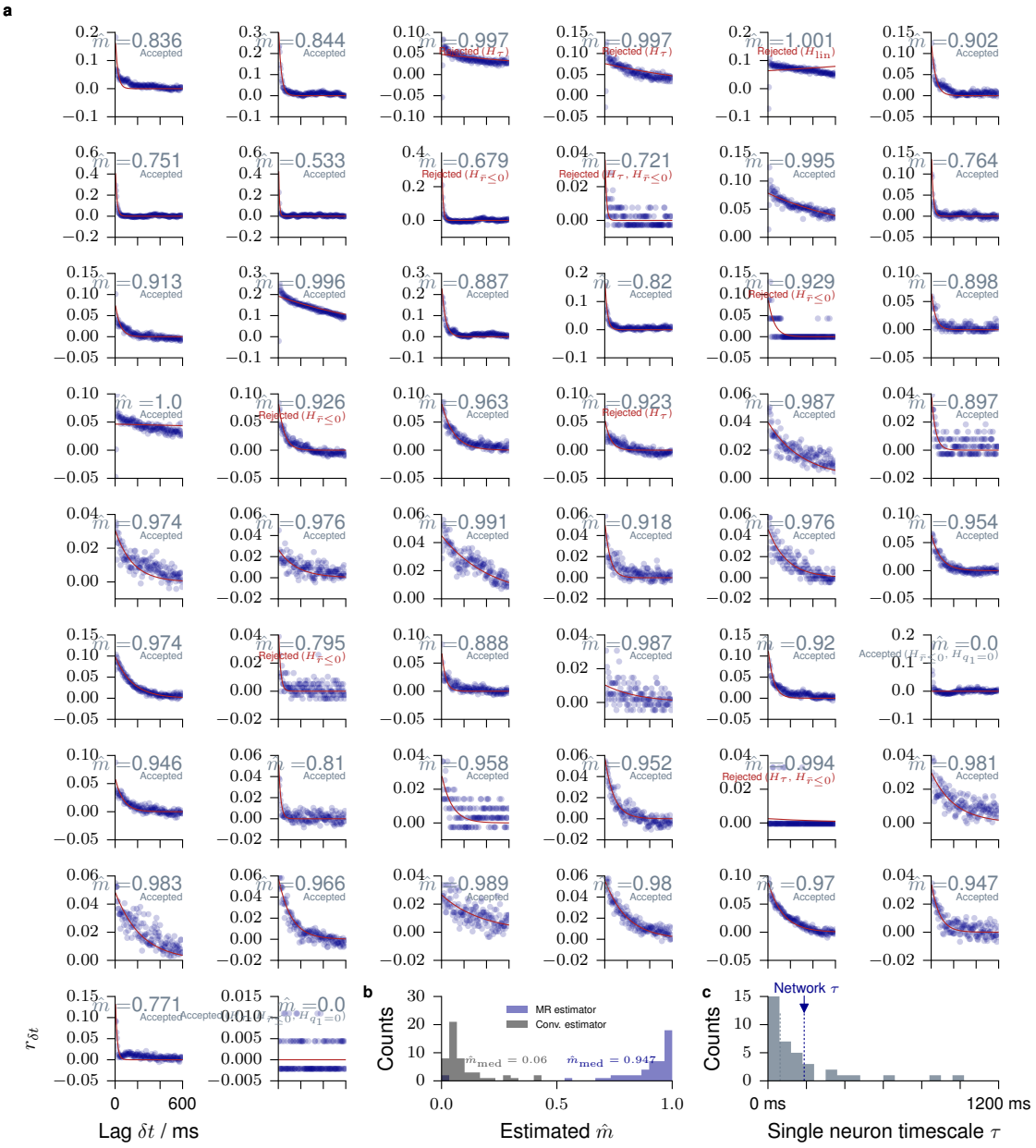

Figure S9: MR estimation from single neuron activity (cat). Modified from[49]. MR estimation is used to estimate $\hat{m}$ from the activity a_t of a single units in cat visual cortex. **a.** Each panel shows MR estimation for one of the 50 recorded units. Autocorrelations decay rapidly in some units, but long-term correlations are present in the activity of most units. The consistency checks are detailed in[49]. **b.** Histogram of the single unit branching ratios $\hat{m}$, inferred with the conventional estimator and using MR estimation. The difference between these estimates demonstrates the subsampling bias of the conventional estimator, and how it is overcome by MR estimation. **c.** Histogram of single unit timescales with their median (gray dotted line) and the timescale of the dynamics of the whole network (blue dotted line).

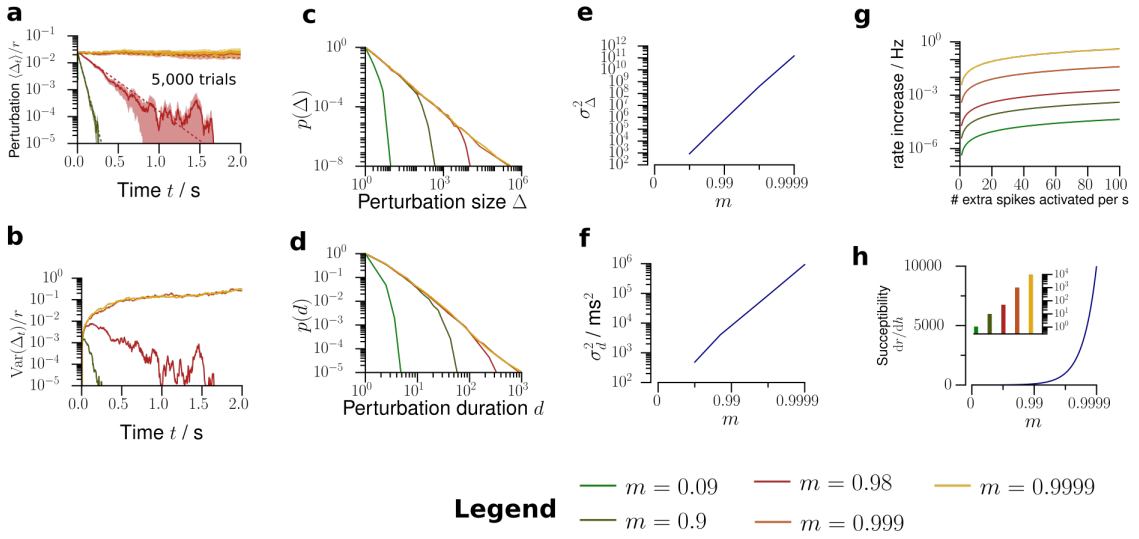

Figure S10: Further predictions about network activity. **a**. The model predicts that the perturbation decays exponentially with decay time $\tau = -\Delta t / \log m$. **b** The variance across trials of the perturbed firing rate has a maximum, whose position depends on m. **c**. Depending on m, the model predicts the distributions for the total number of extra spikes Δ generated by the network following a single extra spike. **d**. Likewise, the model predicts distributions of the duration d of these perturbations. **e**. Variance of the total perturbation size as a function of m. **f**. Variance of the total perturbation duration as a function of m. **g**. Increase of the network firing rate as a function of the rate of extra neuron activations for different m. **h**. Amplification (susceptibility) $\mathrm{d}r/\mathrm{d}h$ of the network as a function of the branching ratio m.