**Title:** Criticality, stability, competition, and consolidation of new representations in brain networks

**Authors:** Quinton M. Skilling\(^1\), Daniel Maruyama\(^2\), Nicolette Ognjanovski\(^3\), Sara J. Aton\(^3\)*, Michal Zochowski\(^1,2,*\)

**Affiliations:**
\(^1\) Biophysics Program, University of Michigan, Ann Arbor, MI 48109
\(^2\) Department of Physics, University of Michigan, Ann Arbor, MI 48109
\(^3\) Department of Molecular, Cellular, and Developmental Biology, University of Michigan, Ann Arbor, MI 48109

* Correspondence to: michalz@umich.edu, saton@umich.edu

**Abstract**

The brain forms and stores distributed representations from sparse external input that compete for neuronal resources with already stored memory traces. It is unclear what dynamical properties of neural systems allow formation and subsequent consolidation of new, distributed memory representations under these conditions. Here we use analytical, computational, and experimental approaches to show that a dynamical regime near a phase-transition in neuronal network activity (i.e. criticality) may play an important role in this process. Our results reveal that near-critical dynamics are necessary to stabilize and store new sparsely driven representations when they compete with native network states.
Perhaps one of the most profound capabilities of the brain is the conversion of incoming input into singular representation that is subsequently consolidated into a long lasting memory. Memory representations, or engrams, are known to be distributed\cite{1,2}, robust, and long lasting. Two remarkable features of engrams are 1) that they seem to be driven initially by the activity of relatively small group of neurons (compared to the size of the brain as a whole), and 2) they are robust to competition with already stored (native) memory representations\cite{3} for neuronal resources. An important unanswered question is: what are the dynamical properties of neural systems that allow newly encoded information to successfully compete with existing representations towards the formation of new memory engrams?

There is a growing body of work suggesting that the brain self-adjusts to operate near criticality\cite{4-8}. This state is exemplified by a specific and universal scaling behavior of activity. While the Self-Organized Criticality (SOC) hypothesis is still being debated\cite{9}, work over the past decade has focused on: 1) providing experimental support for brain SOC and 2) demonstrating (using neural network models) the computational benefits of networks operating near criticality. Regarding the latter, it has been shown that critical dynamics supports optimized information transfer and memory storage capacity (among other network features)\cite{10-14}. More recent studies have shown that neuronal and synaptic responses to external input help neural networks transverse a phase transition toward criticality\cite{15,16}. Others have postulated how residing near criticality affects storage of new memories\cite{17}. Despite this, a unifying mechanistic link between criticality and memory formation is still lacking. Here, using various model systems, we first show that intrinsic properties of the critical state itself provide distinct dynamical advantages necessary for forming novel memory representations. We then show experimentally that our models’ predictions are born out in the hippocampal network in vivo, during de novo memory consolidation.

In this study, we first use a generic, auto-associative (attractor) neural network that embodies competition between native configurations and newly arriving external input. At the same time, the model can be driven towards criticality to show that a new representation driven by sparse input can be stabilized and stored only near that regime. Using mean field theory\cite{18}, we argue that this is due to different scaling properties of stability of natively stored representations vs. externally driven input near criticality, allowing the latter to be fractionally more stable within a critical regime. Furthermore, the subsequent storage of this new representation drives the network away from criticality, temporarily stabilizing the overall state of the network. Based on this prediction we introduce a new metric - Functional Network Stability (FuNS) - to measure persistence of network representations from spike trains. We then turn to spiking a neuron model displaying SOC, and show that increased stability of functional representations upon memory storage also preferentially occurs near a critical regime. Finally, using the same metric, we analyze in vivo recordings from mouse hippocampus after contextual fear
conditioning (CFC) and show that increased representational stability after conditioning predicts fear memory recall 24 h later.

**Criticality and stability of new memories in attractor neural networks**

The observations about the nature of memory representations led many to posit that memory behaves as a neural attractor, whereby representations are encoded as distributed network states. Attractor neural networks are capable of storing multiple representations as distributed states, with internal dynamics driving the system to adopt the state corresponding to the lowest free energy. We used a network known to exhibit critical behavior (manifesting itself through changes in stability of the stored representations) to investigate the network dynamics when a new input competes with already formed (native) representations. We modified the generic attractor network model to incorporate synaptic plasticity, and we show that the storage of a new, externally-driven representation in such a network preferentially occurs near criticality.

The studied system consists of ten thousand nodes, with binary states $S_i = \pm 1$, arranged in a small-world network ($P_{rewire} = 0.1$) with ~2% connectivity. On each step of integration, each node changes its state with probability $P(h_{ij}) = \frac{1}{1 + \exp(-2\beta|h_{ij}|)}$ to align with its observed synaptic input $h_{ij} = \frac{1}{k} \sum_j J_{ij} S_j$, where $k$ represents the number of presynaptic nodes and $J_{ij}$ is the weighted connection between the nodes. Here, $\beta = \frac{1}{T}$ is a control parameter which allows us to externally vary the system proximity to a phase transition by changing dynamics from low to high activity quantified by the average number of times a node switches its orientation; historically $T$ is referred to as temperature but, from a system brain dynamics point of view, it represents a collective of hidden variables that serve to adjust its dynamics to drive it to criticality.

We define two respectively uncorrelated states as $\{\xi_i^n \in [1, 10000]\}$ to represent a natively stored configuration ($\xi^n$) and a new configuration ($\xi^e$) to be learned (stored) by the system. We choose a set $M_{inp} = \{1, ..., M\}$ of $M$ random nodes from the network to represent persistent external drive of the new configuration by setting $S_i (i \in M_{inp}) = \xi_i^e$ to be constant throughout the simulation. Each connection stemming from one of these input nodes then has a weight of $J_{ij} = w^e \xi_i^e \xi_j^e$ ($i \in M_{inp}$ and $j \in [1, 10000]$), while a connection stemming from every other node has a weight of $J_{ij} = w^n \xi_i^n \xi_j^n$; here $w^{e/n}$ is the weight of a respective configuration. The effect is a competition between the local fields emanating from the newly driven state and the native state, stored in the rest of the network.

We quantify the similarity of the final state, in terms of the spin orientations, with both the new and native configurations using network overlap, defined as $m^{n/e}(t) = \left| \frac{1}{N} \sum_i S_i(t) \xi_i^{n/e}\right|$, where $N =$
10000 − M, and the corresponding mean-field approximation is derived for the steady state solutions as a function of the system's temperature (Fig. 1a). These solutions are

\[ m^n = \tanh(\beta w^n \frac{N}{N + M} m^n) \]  

for the native state and

\[ m^e = \tanh(\beta w^e \frac{M}{N + M}) \]  

for the new representation driven by the external input. Based on the overlap of these solutions and on the stability (to be described below; Fig. 1c), we define the critical point region to span an order of magnitude in T, starting at T = 10^{-1}. Below criticality and for N>>M the steady state of the network achieves a higher overlap with the native state rather than with the new representation. Near criticality, however, the overlap of the network state with the native state abruptly declines with increasing temperature, whereas the network’s overlap with the new memory attenuates significantly more slowly, leaving it dominating over the native state. For high values of temperature, the \( m^n/e \approx 0 \) are the only stable solutions. The energy of the network state associated with these solutions is \( E \sim \frac{1}{2} m^2 \), making states with higher overlap a preferred network state. These results indicate that for system dynamics residing in a subcritical regime, the native representation is a preferential solution for the network state and only near criticality is the fractional overlap with the new representation dominating.

Computer simulations (Fig. 1b top) coincide well with the mean-field approximation. In the subcritical (low temperature) regime the native state is systematically preferably retrieved over the new, externally driven state whereas in the supercritical regime (high temperature) neither representation is stable. However, within the critical region, the network overlap with the new representation is persistently fractionally greater than that with the native one.

We next investigated whether the network can stabilize this preferential state and store it as the new representation. To that effect network plasticity was introduced, allowing insight into whether, and over what dynamical regime (i.e. temperature range), the new configuration can be stored (i.e. the overlap with the new state will increase so that \( m^e \rightarrow 1 \) (Fig. 1b bottom)). The form of synaptic plasticity, \( \Delta j_{ij}(t) = \epsilon S_i(t) S_j(t) \) resembles state dependent plasticity similar to STDP in spiking neurons; here \( \Delta j_{ij}(t) \) denotes the change of connection strength between \( i^{th} \) and \( j^{th} \) elements at time \( t \) and \( S_i(t) \) denotes the state of the element at that time. The only temperature range at which the new representation is stabilized (i.e. \( m^e \approx 1 \)) is around the critical regime. This indicates that criticality can play a vital role in storing new representations that are driven by small (M<<N) external input. To see how the novel network state recruits neural resources near criticality we calculate the scaling of the overlap \( m^e \) as a
function of number of direct connections received from the input neurons for sub-critical, critical, and supercritical regimes. As expected from phenomenology of the phase transitions\textsuperscript{19} the new state has non-zero stability near criticality and its spatial correlation decays the slowest around the transition point (Fig. 1c). Thus the critical regime plays here a multifaceted role by: 1) rapidly declining overlap of native representations near the critical point, 2) maintaining fractional overlap of the externally driven representation (Fig. 1a, 1b top), and 3) allowing the new representation to stabilize the global network state (through emergence of long-distance correlations) so that through synaptic plasticity, the representation can be distributed more widely (Fig. 1b bottom). Importantly, we observed similar emergence of long-range correlations in a spiking network (Fig. 5c).

We further examined how overall network dynamics change with the storage of new representation by measuring the change in network representational stability from before to after learning. Here, representational stability is defined as $F = 1 - \overline{p}/p_{\text{max}}$, where $\overline{p}$ is the mean spin flip probability and $p_{\text{max}} = 0.5$ is the random spin flip probability (achieved naturally for high temperature). For stable representations, $F=1$ and for fast, noise driven changes, $F=0$ (Fig. 2a); these values correspond to the functional stability metric introduced below for spiking networks. As expected, before learning, the network transitions around the critical point from being inherently stable to highly unstable. During the storage of the novel configuration (driven by subsequent strengthening of the network’s connectivity), the critical point is shifted toward higher temperature values, resulting in a significant change of the network’s stability around the critical regime (Fig. 2b). We further investigated the relationship between changes in overlap of a new representation with changes in overall representational stability during learning. Namely, we measured these two quantities at near critical temperature during initial stages of memory consolidation (i.e. the first 20 iterations when synaptic plasticity is activated), (Fig 2c). We observe that memory overlap and network stability have a monotonic relationship.

Thus, successful storage of the new representation leads to a significant increase in network stability - this is an important prediction of the model, which we will test below against data from spiking neuron models (Fig. 5) and experimental recordings (Fig. 6).

Finally, to investigate robustness of the phenomenon, we measured the change in the maximal network overlap with the new configuration from pre-learning to post-learning for varying input properties. The maximum change in the overlap invariably occurs near criticality (Fig. 3a). We observed the consolidation of the new memory for as small set of input neurons as $M = 300$ (i.e. 3% of total number of neurons) and synaptic weight ratio of $W^e/W_n \sim 2.3$ (Fig. 3b); the former is consistent with experimental observations\textsuperscript{22, 23}.
Criticality promotes stability of representations in spiking neural networks.

While the persistent dynamical properties of critical regions are well captured by the generic attractor neural network introduced above, we next examined how this dynamical transition region would affect a spiking network. We investigate changes in the functional stability in the presence and absence of a structural network heterogeneity for networks composed of spiking neurons. The structural heterogeneity is implemented here as a strengthened set of connections emanating from a predefined subset of neurons. This directly corresponds to the $M$ neurons being driven by external input in the attractor model. We find that increases in network stability and correlation length (analogous to those shown in Fig. 1c and Fig. 3, respectively) are again observed near criticality, when synaptic heterogeneities corresponding to memory traces are introduced. One important difference from the attractor model is the progression of the phase transition itself: here functional stability saturates for super critical states and decays to zero when subcritical. However, the pattern of behavior (e.g. long-range correlations) in the critical regime is consistent across systems.

To characterize the dynamics of functional connectivity patterns for neuronal spike trains, we developed a metric we refer to as functional network stability (FuNS; Fig. 4a-c). Briefly, spike trains for the network are divided into temporal bins of the shortest possible length necessary to provide a robust estimate of functional connectivity (Fig. 4a). Average minimum distance (AMD)$^{24}$ is then used to evaluate functional network connectivity between cell pairs within each temporal bin (Fig.). FuNS is quantified as the average cosine similarity given by $C_{AB} = \cos \theta_{AB} = \frac{<A,B>}{\sqrt{<A,A> \cdot <B,B>}}$, between functional network structures in adjacent temporal bins (denoted as $A$ and $B$ here), with $C_{AB} \rightarrow 1$ representing stable (unchanging) functional network structures (Fig. 4c; for more details see Methods).

We subsequently assessed how global dynamics near criticality affects FuNS. Here, we used a modified Bak-Tang-Wiesenfeld (BTW) model$^{25}$ as a proxy of integrate and fire neuronal network, to measure the distribution of spike-time avalanche lengths in the system as a function of coupling strength (Fig. 5a). Avalanches represent consecutive temporal activation sequences of neurons within the network. The BTW model is known to exhibit critical dynamics when connectivity strength is near unity. As expected, we found that the scaling of the avalanche lengths near criticality follows a power law distribution. We next compared FuNS in the presence and absence of a structural heterogeneity. Changes of FuNS as a function of the connectivity strength are shown in (Fig. 5b). The spiking model results are fully analogous to the ones presented for the attractor network (Fig. 1). The highest sensitivity of FuNS occurs just below the critical connectivity of 1.0. This identifies the optimal dynamical regime to be near, but just below criticality, coincident with a number of recent findings that in vivo brain dynamics operate in that regime$^{26}$. We also investigated how the representational stability between neurons changes as a
function of their distance from the heterogeneity (Fig. 5c). As expected from the properties of the systems near phase transitions and similarly to the attractor network described in previous section (see Fig. 1 & Fig. 2), the decay of the change in stability with the distance from heterogeneity is slowest near criticality.

Taken together, these results imply that, when near criticality, network heterogeneities (driven through external input and/or local changes to synaptic connectivity) change the scaling of the avalanche statistics preferentially towards avalanches of longer duration and encompassing activity of more neurons as exemplified by changing avalanche distributions and stability change as a function of distance from those heterogeneities. This change promotes increased representational stability in a distributed circuit and subsequently provides dynamical substrate for spike timing dependent plasticity and thus learning of the novel configuration.

Changes in network dynamics and functional stability predict subsequent memory formation after contextual fear conditioning in mice.

To provide experimental support for the above hypothesis, we investigated whether FuNS changes can be detected following learning-induced network reorganization in vivo. We hypothesized that changes in network dynamics due to synaptic plasticity in hippocampal area CA1 following single-trial contextual fear conditioning (CFC) is a plausible biological analogue of the modeled storage of a new representation. For this reason, we recorded the same population of CA1 neurons over a 24-h baseline and for 24 h following CFC to determine how functional network dynamics are affected by de novo memory formation.

C57BL/6J mice underwent either CFC (placement into a novel environmental context, followed 2.5 min later by a 0.75 mA foot shock; n = 4 mice), sham conditioning (placement in a novel context without foot shock; Sham; n = 4 mice), or CFC followed by 6 h of sleep deprivation (a manipulation known to disrupt fear memory consolidation; Sleep Dep (SD); n = 5 mice; see Methods for details). We first set out to characterize the dynamical changes in terms of event statistics for mice that form contextual fear memory (CFM) compared to those that do not; CFM is blocked by post-CFC sleep deprivation and is absent in sham conditioned mice. Avalanche statistics are calculated from LFP traces (Fig. 6a). Values of the branching parameter for varying specified event rates and bin sizes indicate that system dynamics are sub-sampled for all mice in all groups. Taking this into account, we calculate the fractional change in the branching parameter from pre- to post-learning; it shows clear separation between CFC mice and either SD or Sham mice (Fig. 6b). The branching parameter measures the rate at which new neurons are recruited into the avalanche and indicates the spread of correlated activity throughout the network. Mechanistically, its increase is an indicator of network-wide strengthening of connections (i.e.
formation of network structural heterogeneity). During CFM consolidation, CFC mice exhibit a persistent increase in branching parameter, while SD mice with experimentally disrupted CFM consolidation show a persistent decrease in branching parameter, both relative to Sham branching parameter statistics. By contrast, sham animals showed no statistical deviations. We speculate that the increase of branching in CFC mice is directly related to memory consolidation while the decrease in SD mice can be associated with departure from criticality and failure of consolidation. This hypothesis is supported by previous studies indicating that dynamics in the brain during sleep are closest to a critical regime. It is also generally consistent with studies that show deterioration of critical dynamics during long-term sleep deprivation in humans. We then measured FuNS changes after each manipulation by quantifying pairwise AMD on a minute-by-minute basis over the entire pre- and post-training intervals. In line with our theory that predicts that FuNS will increase significantly only near criticality, we observed an increase in FuNS over the 24h following CFC, which was most pronounced when comparing FuNS in slow wave sleep (SWS; Fig. 6c). We hypothesize that this may be due to two factors: 1) the aforementioned critical dynamics of the sleeping brain, and 2) the fact that SWS corresponds to internal network dynamics, rather than the changing external drive to the network experienced during wakefulness.

In contrast, no change in SWS FuNS was seen in Sham mice or following the period of experimental sleep deprivation in in SD mice. Group differences in SWS FuNS were reflected in the behavior of the mice 24 h post-training, when CFM was assessed behaviorally. CFC mice showed an increase in context-specific freezing upon return to the conditioning environment, which was significantly greater than freezing in Sham and SD mice (Fig. 6d inset). Remarkably, SWS-specific FuNS increases for individual mice were predictive of their behavioral performance during subsequent memory assessment (Fig. 6d). Thus the formation of a behaviorally-accessible memory trace in vivo is accompanied by increased FuNS in the CA1 network.

These results indicate that during and just after CFC a set of synapses is strengthened resulting in correlated spiking avalanches encompassing more active cells. These avalanches, through increased functional stability of activated cells, provide a dynamical substrate for subsequent memory consolidation. When animals are sleep deprived the branching parameter is significantly lowered indicating smaller recruitment of cells into the avalanches, lower functional stability in the circuits, and subsequent failure of memory consolidation.

Discussion

The question we address in this paper is how relatively sparse input can dynamically compete with already stored representations to be stored and later consolidated into a distributed memory (engram). Through theory and computation we show that criticality may play a vital role in mediating...
stabilization and subsequent storage of the new memory as a distributed representation. Namely, we show in a reduced attractor network that only when the system is driven towards a critical point, the new representation (through the input sites) globally impinges its activity pattern on the network making it fractionally dominant as compared to the native representation (Fig. 1). This is primarily due to the fact that at criticality, the external input biases the state of the network towards the new representation, and the emergence of long-distance correlations allows it to spread throughout the system. Subsequently, state dependent synaptic plasticity allows for long-term storage (consolidation) of this new representation.

During the initial phase of consolidation we observe significant increase of network wide representational stability, and this increase can be directly correlated with the degree of memory consolidation (Fig. 2). Thus network dynamics near criticality, during presentation of novel stimuli, allows preferential stabilization of novel representation throughout the circuit. Moreover, increased representational stability becomes a predictor of successfully storing a novel configuration; the changes in stability are not observed in networks far away from criticality.

We further show that this property is consistent across systems and can be measured in the brain. Namely, we construct a metric (FuNS) that reports stability of functional representations for spiking, network-wide activity patterns and show that, similar to the results described above, FuNS increases occur only near the critical point for spiking network models, namely in a model with a known critical point (Fig. 5. Finally, through in vivo experimentation we show that: 1) FuNS is enhanced in CA1 during de novo memory encoding, 2) the degree of change of FuNS after CFC predicts subsequent memory performance, and 3) there is a systematic increase in branching parameter during normal memory consolidation, and a decrease during SD-induced memory disruption.

Taken together these results hint at a unifying mesoscopic mechanism underlying what is commonly referred to as “systems consolidation” - i.e., the formation of a widely distributed engram from a transient, discrete, and localized group of synaptic changes

Our results support the hypothesis that, if the system is near criticality, discrete external input representing features of a novel learning experience stabilizes the functional connectivity within the larger neural circuit in order to promote formation of a distributed engram. Dynamically driven, widespread network stabilization promotes coordinated consolidation of initially disjointed memory features (e.g. smell and color), and could later aid in the seemingly instantaneous, coordinated recall of these features.

Finally, this work underscores the importance of FuNS as a measurement of memory consolidation. Our in vivo recording data suggest that, contrary to reports of sequential replay or reactivation which typically occur only minutes following learning\textsuperscript{37, 38}, FuNS changes in the hippocampus are long-lasting (i.e., across the 24 h following training). The spatial extent and duration of FuNS changes are thus highly amenable to promoting systems memory consolidation. We believe that
FuNS could provide a mechanism to drive structural network changes (e.g., through spike timing-dependent plasticity) over widely-distributed networks. If (as recent data suggest) synaptic structures in CA1 are far more transient than synapses in the neocortex\textsuperscript{39}, the rapid dissemination of memory traces for longer-term storage outside the hippocampus is likely required for consolidation.
Methods

1. Simulations:

Dynamics and memory storage in an auto-associative memory model: A two-dimensional square network comprising 10000 nodes is simulated. Each node is binary with state given by $S_i = \pm 1$. A subset $M_{\text{inp}} = \{1, \ldots, M\}$ of these nodes are chosen to represent an external input, which is fully represented by the state $\{\xi_1^e, \xi_2^e, \ldots, \xi_{N+M}^e\}$, separate from the native state $\{\xi_1^n, \xi_2^n, \ldots, \xi_{N+M}^n\}$, and are not allowed to change throughout the duration of the simulation; the remaining $N$ neurons do not adhere to this restriction. Each node is connected to $\sim 2\%$ of the network, where connections are predominantly local with a probability of $p = 0.1$ of being rewired to a random target. The strength of each connection is defined by $J_{ij} = \frac{w^e}{n} \frac{\xi_i^e}{\xi_1^e} \frac{\xi_j^e}{\xi_1^e}$ with the superscripts given by $e$ (external input) only if $j \in M_{\text{inp}}$ and $n$ (native) otherwise; here $w$ is a weight value representing the relative strength of the external input to the native state. The total input to each node is then given by $h_i = \frac{1}{T} \sum_j J_{ij} S_j$, where the summation is over all $j$ connections. Finally, during each step of the simulation, a node aligns to the sign of its input with probability $P(h) = \frac{1}{1+\exp(-2\beta|h|)}$ where varying $\beta = \frac{1}{T}$ controls the dynamic regime (subcritical, critical, and supercritical for low, medium, and high $T$, respectively). All simulations are initialized to random conditions. Unless otherwise stated, all figures are results for an external input of $M = 700$ neurons and a weight ratio given by $\frac{w^e}{w^n} \sim 3.5$.

Synaptic plasticity: A learning rule was implemented in order to facilitate the consolidation of a new representation to memory. The form of plasticity is generic and state dependent, resembling spike timing dependent plasticity (STDP) in spiking systems. During each time step in the learning phase of the simulation, the synaptic connectivity between all nodes changes by some small amount following $J_{ij} = J_{ij} + \varepsilon S_i S_j$ where $\varepsilon = 0.1$.

Modified Bak-Tang-Wiesenfeld model: A modified version of the Bak-Tang-Wiesenfeld (BTW) model was used as a simplified IAF model corresponding to a 3D network with $N = 12 \times 12 \times 12 = 1728$ cells (Fig. 5). Each cell had a connectivity radius of $r = 1$ with non-periodic
boundary conditions; connections were rewired to a small-world network with rewiring probability \( p = 0.1 \). The energy of each cell \((i)\), is given by \( E_i(t + 1) = E_i(t) + \sum_{j=1}^{n} A \ast E_j^*(t) + \partial(x - i) \) where the terms in the RHS correspond to the summation of the present value of the cell’s energy and all of its inputs. The summation term represents the total inputs from connected, spiking cells \((E_j^*)\) multiplied by the corresponding connectivity strength \(A\). The connectivity strength was uniform over the entire network except for a 5x5x5 cell sub-network, centrally placed, where \(A = 2.0\), representing a region of heterogeneity.

Following BTW, a cell \( x \) is randomly chosen with probability \( 1/N \) to receive a unit of external input of energy using the delta function. Here, \( A \) is strictly positive thus causing \( E_i \) to increase in time until a threshold is met whereby it “topples,” sending information to its connected cells before resetting in the following time step, i.e. \( E_i(t + 1) = 0, \text{ if } E_i(t) \geq 6 \). Once a cell is reset by this mechanism, it enters a refractory time of 20 time steps during which it cannot send or receive information. Similarly, if at least one cell fired in the previous time step, then no external input is provided to the system, i.e. \( x = 0 \) and thus \( \partial(x - i) = 0 \).

This model resembles integrate and fire neuronal model without leakage current – \( E(t) \) corresponds to discretely integrated neuronal voltage, while toppling of the element mimics emission of the spike; similarly to the neuronal model the toppling occurs when the threshold is reached, persists for short period of time under which the cell goes into refractory period.

2. **Data analysis:**

**Functional connectivity metrics using Average Minimum Distance (AMD):** Given a list of spike trains \( \{S_1, S_2, ..., S_n\} \) for \( n \) neurons, where \( S_i \) contains all the firing times for neuron \( i \), the functional relationship, \( FC_{ij} \), of the \( i^{th} \) and \( j^{th} \) neurons is evaluated by comparing the average temporal closeness of spike trains \( S_i \) and \( S_j \) to the expected sampling distance of train \( S_j \), \( \mu_j \). Values are normalized by dividing this difference by the expected variation in sampling distance of train \( S_j \), \( \sigma_j \).

The average distance from spike train \( S_i \) to \( S_j \) is given by the AMD, where we define \( AMD_{ij} = \frac{1}{N_i} \sum_k \Delta t_k^i \), here \( N_i \) is the number of events in \( S_i \) and \( \Delta t_k^i \) is the temporal distance between event \( k \) in \( S_i \) to the nearest event in \( S_j \). The expected values, \( \mu_j \) and \( \sigma_j \), are calculated by
integrating over the sampling minimum distance distribution of $S_j$. The expected mean sampling distance and standard deviation can be simply expressed by noting that over an individual interspike interval (ISI) of length $L$, the first two moments of minimal distance are given by $\mu = \langle MD^L \rangle = (1/4)L$ and $\langle (MD^L)^2 \rangle = (1/12)L^2$. Taking into account the probability of sampling an interval of length $L$ for a spike train of length $T$ is $(L/T)$, we can combine the values from intervals to give $\mu_j = \langle MD_j \rangle = \sum_{(L)} p_L \mu^L = \frac{1}{T} \sum_{(L)} \frac{L^2}{4}$, and $\langle (MD_j)^2 \rangle = \frac{1}{T} \sum_{(L)} \frac{L^3}{12}$. The expected standard deviation is given by $\sigma_j^2 = \langle (MD_j)^2 \rangle - \langle MD_j \rangle^2$. Knowing now how to evaluate all the components, we calculate the functional connectivity between neurons $i$ and $j$ as:

$$FC_{ij} = \sqrt{N_i} \frac{\mu_i - AMD_{ij}}{\sigma_j}.$$  

The functional connectivity values represent the significance of the average temporal closeness of $S_i$ to $S_j$ after taking into account the spiking distributions of $S_j$. The calculations of $\mu_j$ and $\sigma_j$ allow for rapid estimation of functional connectivity and are accurate approximations so long as the spike timings are well determined. Furthermore, the methods used here can be readily generalized to more complex measures, for instance a quadratic or exponential distance.

**Functional stability as a measure of network dynamics:** Functional connectivity networks can be calculated for multiple time segments and then compared using cosine similarity to get a measure of how functional relationships change over time. Cosine similarity, $C_{AB} = \cos \theta_{AB} = \frac{<A,B>}{\sqrt{<A,A>*<B,B>}}$, is a measure of the overlap between two values, with 1 denoting no change in the network and 0 indicating the two networks are not related. (See Fig. 4) When the similarity is found for a series of adjacent time windows, the average of these values denotes the average value of change which we refer to as functional stability. The length of the time bins is constrained at the low end by the spike counts of the analyzed neurons. We used the shortest time bins possible where each neuron contains at least 10 spikes, so that the functional connectivity values are based upon the full set of neurons. Longer time-windows will average out functional dynamics in favor of long term permanent structure.

**Neuronal Avalanche statistics:** Neuronal avalanches are calculated from model spiking data and from experimental LFP recordings following previous studies. Briefly, the time series in question are divided into discrete, regular time windows and probed for neural activity. An
avalanche is defined as a series of events in consecutive time windows across channels. The avalanche is bounded by quiescent windows and the lifetime of an avalanche is the total number of time windows constituting an avalanche. For model spiking data, an event is the occurrence of a spike and the time window is based on the time step of integration. Avalanche calculations from continuous LFP data is more involved. The trace is subjected to area-under-the-curve integration only for negative potentials, bounded by the zero potential line. The deflection threshold is chosen to produce an invariable event rate R across all channels ensuring that each channel equally contributes to avalanche statistics. The time window, or bin width, is chosen to be \( \text{bw} = \alpha \overline{IEI} \) where \( \alpha \) is a scalar value and \( \overline{IEI} = \frac{1}{RN} \) is the inter-event interval with \( N \) representing the number of channels. Here, \( R = 2 \text{ Hz} \) and \( \alpha = \{ \frac{1}{32}, \frac{1}{16}, \frac{1}{8}, \frac{1}{4}, 1, 2, 4 \} \).

**Branching Parameter Calculation for Varying Event Rate-based Bin Sizes**

The avalanche branching parameter \( \sigma \) phenomenologically represents the average cascade level of an avalanche and it provides means for characterization of systems dynamical regime, attaining values \( \sigma < 1, \sigma = 1, \) and \( \sigma > 1 \) for subcritical, critical, and supercritical dynamics respectively.

The branching parameter calculation follows work by others\(^4^\) and is given by:

\[
\sigma = \frac{1}{\sum_i L_i} \sum_i^{n} \sum_{j=1}^{t_i} \frac{a_j^i}{a_{j-1}^i}
\]

Where \( a \) is the number of events in a given time bin, \( L_i \) is the total lifetime of \( i^{th} \) avalanche, \( t \) represents number of activity bins \( a \) in \( i^{th} \) avalanche and \( n \) is the total number of avalanches analyzed.

3. **Chronic in vivo recording and contextual fear conditioning:**

Male C57BL6/J mice (Jackson, aged 2-5 months) were implanted with driveable headstages containing two bundles of 7 stereotrodes each (spaced 1 mm apart) for single-unit and local field potential (LFP), and silver-plated wires for nuchal electromyographic (EMG) recording. LFP and EMG signals were used to assign behavioral states (wake, SWS, REM sleep) in 5-s epochs throughout the recording period. Mice were individually housed (in standard caging with
beneficial environmental enrichment including nesting material, manipulanda, and treats) during post-operative recovery and subsequent behavioral experiments. Lights were maintained on a 12hr: 12hr light:dark cycle, and food and water were available ad lib, throughout all procedures. All housing and experimental procedures were approved by the University Committee on Use and Care of Animals at the University of Michigan.

Following a 1-week recovery period, mice were habituated to daily handling (5-10 min/day) for 3 days. During this habituation period, stereotrodes were gradually lowered into CA1 until stable neuronal recordings (with characteristic spike waveforms continuously present on individual recording channels for more than 24 hr) were obtained. After this, no changes to electrode position were made throughout subsequent experimental procedures. All mice underwent a 24-hr baseline recording starting at lights on. At lights on the following day, mice underwent single-trial contextual fear conditioning (CFC) or sham conditioning (Sham, n = 4)\textsuperscript{32}. Mice were placed into a standard conditioning chamber (Med Associates) with patterned Plexiglass walls and a metal grid floor. All mice were allowed to freely explore the novel chamber over the 3-min training session; CFC mice (but not sham mice) received a 2-s footshock (0.75 mA) after the first 2.5 min. At the end of 3 min in the conditioning chamber, mice were returned to their home cage for a 24 hr post-conditioning recording period. CFC mice were subdivided into two groups - one which was allowed ad lib sleep (CFC, n = 5), and a second which was sleep deprived by gentle handling for the first six hours following training (a manipulation which is sufficient to disrupt contextual fear memory consolidation\textsuperscript{31-34}; Sleep Dep, n = 5). 24 hr following training, at lights on, mice were returned to the conditioning chamber for a 5-min assessment of contextual fear memory. This was calculated as the change in context-specific freezing between testing and training trials (i.e., % time spent freezing at test — % time spent freezing at baseline [pre-shock]).

**Spike data analysis:** Electrophysiological signals were digitized and differentially filtered as spike and LFP data as described previously\textsuperscript{32} using Omniplex hardware and software; single-unit spike data was discriminated using Offline Sorter software (Plexon). The firing of individual neurons was tracked throughout each experiment on the basis of spike waveform, relative spike amplitude on the two stereotrode recording channels, positioning of spike wave-form clusters in
three-dimensional principal component space, and neuronal subclass (e.g., FS interneurons vs. principal neurons). Only those neurons that were reliably discriminated and continuously recorded across 24-hr baseline and 24-hr post-conditioning recording periods were included in analyses of network stability.
References
1. Hubener, M. & Bonhoeffer, T. Searching for engrams. *Neuron* **67**, 363-371 (2010).
2. Ryan, T., Roy, D., Pignatelli, M., Arons, A. & Tonegawa, S. Engram cells retain memory under retrograde amnesia. *Science* **348**, 1007-1013 (2015).
3. Mahajan, G. & Mehta, A. Competing synapses with two timescales as a basis for learning and forgetting. *Europhysics Letters* **95** (2011).
4. Beggs, J. M. & Plenz, D. Neuronal Avalanches in Neocortical Circuits. *J Neurosci* **23**, 11167-11177 (2003).
5. Chialvo, D. Emergent complex neural dynamics. *Nature Physics* **6**, 744-750 (2010).
6. Dehghani, N. et al. Avalanche analysis from multielectrode ensemble recordings in cat, monkey, and human cerebral cortex during wakefulness and sleep. *Front. Physiol* **3**, 302 (2012).
7. Hesse, J. & Gross, T. Self-organized criticality as a fundamental property of neural systems. *Front. Syst. Neurosci* **8**, 166 (2014).
8. Markovic, D. & Gros, C. Power laws and self-organized criticality in theory and nature. *Phys. Rep.* **536**, 33 (2014).
9. Touboul, J. & Destexhe, A. Can power-law scaling and neuronal avalanches arise from stochastic dynamics? *PLoS One* **5**, doi:10.1371/journal.pone.0008982 (2010).
10. Kinouchi, O. & Copelli, M. Optimal dynamical range of excitable networks at criticality. *Nature Physics* **2**, 348–351 (2006).
11. Haldemann, C. & Beggs, J. M. Critical Branching Captures Activity In Living Neural Networks and Maximizes the Number of Metastable States. *Phys Rev Lett.* **94**, 058101 (2005).
12. Shew, W. L., Yang, H., Petermann, T., Roy, R. & Plenz, D. Neuronal Avalanches Imply Maximum Dynamic Range in Cortical Networks at Criticality. *J. Neurosci.* **29**, 15595-15600 (2009).
13. Shew, W. L., Yang, H., Yu, S., Roy, R. & Plenz, D. Information Capacity and Transmission are Maximized in Balanced Cortical Networks with Neuronal Avalanches. *J Neurosci.* **31**, 55-63 (2011).
14. Shew, W. L. & Plenz, D. The Functional Benefits of Criticality in the Cortex. *Neuroscientist* **19**, 88-100 (2013).
15. Shew, W. L. et al. Adaptation to sensory input tunes visual cortex to criticality. *Nature Physics* **11**, 659-664 (2015).
16. Levina, A., Herrmann, J. & Geisel, T. Dynamical synapses causing self-organized criticality in neural networks. *Nature Physics* **3**, 857-860 (2007).
17. de Arcangelis, L. Herrmann, H.J. Learning as a phenomenon occurring in a critical state. *PNAS* 107(9): 3977-3981 (2010).
18. Hopfield, J. Neural networks and physical systems with emergent collective computational abilities. *PNAS* **79**, 2554-2558 (1982).
19. Amit, D. J. *Modeling Brain Function: The world of attractor neural networks*. (Cambridge University Press, Cambridge, England, 1989).
20. Feldman, D. The spike-timing dependence of plasticity. *Neuron* **75**, 556-571 (2012).
21. Landau, L. D. & Lifshitz, E. M. *Statistical Physics*. 3 edn, Vol. 5 (Elsevier Science and Technology, New York, USA, 2011).
22. Liu, X. et al. Optogenetic stimulation of a hippocampal engram activates fear memory recall. *Nature* **484**, 381-385 (2012)
23. Redondo, R. et al. Bidirectional switch of the valence associated with a hippocampal contextual memory engram. *Nature* **513**, 426-430 (2014).
24. Feldt, S., Waddell, J., Hetrick, V., Berke, J. & Zochowski, M. Functional clustering algorithm for the analysis of dynamic network data. *Phys Rev E* **75**, 056104 (2009).
25. Bak, P., Tang, C. & Wiesenfeld, K. Self-Organized Criticality: An Explanation of 1/f Noise. *Phys. Rev. Lett.* **59**, 381-384 (1987).
26. Priesemann, V. et al. Spike Avalanches in vivo suggest a driven, slightly subcritical brain state. *Front. Syst. Neurosci.* **8**, 108 (2014).
27. Izhikevich, EM. *Dynamical systems in Neuroscience*. MIT Press, 2007.
28. Shin, C. & Kim, S. Self-organized criticality and scale-free properties in emergent functional neural networks. *Phys. Rev. E.* **74**, 4 (2006).
29. Tronson, N. et al. Segregated Populations of Hippocampal Principal CA1 Neurons Mediating Conditioning and Extinction of Contextual Fear. *J Neurosci.* **29**, 3387-3394 (2009).
30. Daumas, S., Halley, H., Frances, B., & Lassalie JM. Encoding, consolidation and retrieval of contextual memory: different involvement of dorsal CA3 and CA1 hippocampal subregions. *Learn Mem.* **12**, 375-382 (2005).
31. Vecsey, C. et al. Sleep deprivation impairs cAMP signalling in the hippocampus. *Nature* **461**, 1122-1125 (2009).
32. Ognjanovski, N., Maruyama, D., Lashner, N., Zochowski, M. & Aton, S. CA1 hippocampal network activity changes during sleep-dependent memory consolidation. *Front. Syst. Neurosci.* **8**, 61 (2014).
33. Prince, T. et al. Sleep deprivation during a specific 3-hour time window post-training impairs hippocampal synaptic plasticity and memory. *Neurobiol. Learn. Mem.* **109**, 122-130 (2014).
34. Graves, L., Heller, E., Pack AI & Abel, T. Sleep deprivation selectively impairs memory consolidation for contextual fear conditioning. *Learn. Mem.* **10**, 168-176 (2003).
35. Priesemann, V., Valderrama, M., Wibral, M. & Le Van Quyen, M. Neuronal Avalanches Differ From Wakefulness to Deep Sleep - Evidence from Intracranial Depth Recordings in Humans. *PLoS Comput. Biol.* **9**, e1002985 (2013).
36. Meisel, C., Olbrich, E., Shriki, O. & Achermann, P. Fading signatures of critical brain dynamics during sustained wakefulness in humans. *J Neurosci.* **33**, 9 (2013).
37. Louie, K. & Wilson, M. Temporally Structured Replay of Awake Hippocampal Ensemble Activity during Rapid Eye Movement Sleep. *Neuron* **29**, 145-156 (2001).
38. Wilson, M. & McNaughton, B. Reactivation of Hippocampal Ensemble Memories During Sleep. *Science* **265**, 676-679 (1994).
39. Attardo, A., Fitzgerald, J. & Schnitzer, M. Impermanence of dendritic spines in live adult CA1 hippocampus. *Nature* **523**, 592-596 (2015).
40. Priesemann, V., Munk, M.H.J., and Wibral, M. Subsampling effects in neuronal avalanche distributions recorded *in vivo*. *BMC Neurosci.* **10**: 40 (2009).
Acknowledgments:
The work described above was supported by NIH NIBIB grant R01EB018297.

Author Contributions:
Q.M.S. Performed modeling and part of analysis, participated in writing the manuscript
D.M. Developed part of analysis tools, performed part of simulations and participated in writing the manuscript.
N.J. Performed experiments and part of analysis.
S.J.A. Conceived and designed the experiments, performed part of analysis and participated in writing the manuscript.
M.Z. Conceived modeling and analysis, wrote the manuscript.

Competing Interests
Authors declare no competing interests.

Materials and Correspondence
Request for materials and correspondence should be addressed to Michal Zochowski (michalz@umich.edu) and Sara Aton (saton@umich.edu)
Figure 1. External input dominates network dynamics at criticality.
a) Mean-field solution of the overlap of the network steady state with native (black) and new (red) representations. Inset shows an example numerical solution used to obtain the traces.

b) (Top) Network simulations when connection plasticity is absent. The new representation (red) fractionally overlaps with the network state at criticality but not in subcritical (low temperature) or supercritical (high temperature) regimes; (Bottom) imposing a state dependent plasticity rule increases the overlap of the system with that of the new representation only at criticality.

c) Overlap of states of individual neurons with novel representation before learning as a function of number of direct connections to the input cells at criticality (red), subcritical regime (blue) and supercritical regime (black). Long distance correlations emerge at criticality. Error bars represent standard error over 30 simulations.
Figure 2. Overall changes in stability of network states as a function of temperature.

a) Comparison of stability of the network pre-learning (black) to post-learning (red).
b) Change of network stability as a function of temperature. Significant changes in network stability are only observed near criticality.

c) Changes of novel memory overlap as a function of overall changes in network stability. Error bars represent standard error over 30 simulations.
Figure 3. Change in overlap from pre- to post-learning as a function of temperature, input size and pattern weight.

a) The change in overlap for the new representation peaks in the critical regime (gray bar).
b) The maximum overlap change (color) as a function of external-to-native weight ratio and increasing the input size. Even very small external input can be consolidated if the weight ratio is high enough. Asterisk indicates parameter values used to generate all other related figures.
Figure 4. Functional network stability (FuNS) as a metric of functional connectivity dynamics.

a) An example of firing raster consisting of four neurons partitioned into seven time bins.
b) Functional connectivity matrices, based on spike trains in each time segment, estimated using the AMD metric. Color denotes connection strength; blue and red represent weak and strong connections, respectively.

c) Cosine similarity of adjacent temporal representations.
Figure 5. Criticality promotes stability change in a spiking network.

a) Scaling of avalanche sizes as a function of connectivity strength in the network. Critical exponent is established for connectivity strength of one.
b) (top) Average change in stability due to the presence of a structural heterogeneity, as a function of coupling strength, for network without heterogeneity (blue) and with the heterogeneity (green); (bottom) difference between the two curves. The maximal sensitivity of functional stability is just below the critical point. For critical and supercritical values the change in stability rapidly vanishes. Simulations are averaged over 30 independent network realizations.

c) The attenuation of stability difference as a function of mean connectivity distance between the heterogeneity and other network sites. The colors correspond to coupling strength outside the heterogeneity. We observe that the slowest attenuation of stability change happens near but below critical dynamics (A=1). The curves for A≥1 overlap at zero for all distances.
Figure 6. FuNS and branching parameter changes in the hippocampal network predict fear memory consolidation.
a) Comparison of avalanche scaling during 6 hours pre- and post-CFC for conditioned (light blue), and conditioned and subsequently sleep deprived (SD) animals (green).

b) Average percent difference of branching parameter from 24h baseline ($\Delta \sigma = \frac{2(\sigma_{post} - \sigma_{pre})}{\sigma_{post} + \sigma_{pre}}$) following Sham conditioning (novel environment context, no foot shock, dark blue), CFC (contextual fear conditioning; novel environmental context + foot shock, light blue), and SD (CFC followed by a 6-hr period of sleep deprivation, green). The change of branching parameter is calculated as a function of bin size used for avalanche statistics to show robustness of the result against this parameter.

c) Average change in CA1 hippocampus FuNS (from 24-hr baseline) for the same conditions as described above. Values are averaged over the entire 24-hr post-training period (regardless of behavioral state), over post-training SWS alone, and over post-training wake alone. * indicates p < 0.05, Holm-Sidak post hoc test vs. Sham and Sleep Dep mice.

d) Post-training changes in SWS FuNS vs. observed changes in freezing behavior for individual mice; $R^2 = 0.58$, $p < 0.005$ Pearson correlation. Inset: Behavioral measurement of fear memory consolidation (% change in context-specific freezing) for mice in the three experimental groups. * indicates $p < 0.05$, Holm-Sidak post hoc test vs. Sham and Sleep Dep mice.

e) A native state (left) represents stored representations (previous mouse experiences), while external input (center) corresponds to new memory, which is to be stored (mouse CFC in analyzed experiments). Only if network operates near-criticality can the new memory be consolidated (right) – successful consolidation is subsequently measured by mice freezing pattern when placed in the training cage.