Information retention by synaptic learning rules in stochastic spiking neural networks

Rich Pang*
Neurotaxis, Madison, Wisconsin, USA

(Received 22 December 2025; accepted 7 July 2026; published xxxxxxxxxx)

Physical learning systems require a means to store information about the past. In the brain, information is thought to be stored in synaptic connections between neurons whose weights change in response to pre- and postsynaptic action potentials, or spikes, according to specific plasticity rules. To date, most models of synaptic plasticity are derived from *in vitro* studies in which spike times are under precise experimental control, but whether these models generalize to *in vivo* regimes of high spike time variability is not known. To shed light on this problem theoretically, here I estimate the mutual information between Poisson spike trains in a network of neurons and the synaptic weight updates they trigger via different plasticity rules. Although most synaptic plasticity models propose correlation- or coincidence-based rules based on the measurement of pre- and postsynaptic spike times, I show that such rules do not always maximize information storage. In the biologically important, energy-efficient operating regime of low firing rates and sparse plasticity, rules that simply count pre- and/or postsynaptic spikes in fact store more information per synapse than spike-timing-dependent rules that measure coincidence events. In such regimes, spikes are more common than coincidences hence better able to transmit information into synaptic weights. It may in turn be advantageous to encode information to be stored as spike counts, or time averages, rather than correlations. The theory suggests a normative account for recently discovered noncanonical hippocampal plasticity rules *in vivo* and predicts that spike-timing-dependent rules will be preferentially active in neural circuits operating at higher firing rates.

DOI: 10.1103/m8bw-9fzg

I. INTRODUCTION

Physical learning systems must store information using finite resolution components and under energetic constraints. In the brain, information is thought to be stored in synaptic connection weights between neurons, which update in response to pre- and postsynaptic action potentials, or spikes. Which update rules operate in the living brain and how these support efficient learning, however, are not well understood. Spike-timing-dependent plasticity (STDP) is a canonical proposed rule predicting that synaptic weights change in response to pre- and postsynaptic spike coincidences [1–5] [Fig. 1(a)]. Classic *in vitro* STDP experiments found that spikes must be paired within tens of milliseconds to change synaptic efficacies, with the order of the spikes determining the sign of the weight change [1,2]. Larger coincidence windows ($\sim$75 ms) were observed in brain slices from hippocampal area CA3—a region implicated in memory formation [6–8]—where timing but not order matters, which is thought to relate to recurrent connectivity in the area [4]. Rich theoretical work has also shown how STDP could serve important functions in biological network models [9–14].

Considerations of the brain's operating regimes *in vivo*, however, pose challenges for STDP's putative role in awake, behaving animals. In contrast to *in vitro* preparations in which spike timing is usually under precision experimental control, spike timing in the awake brain is extremely variable—often modeled as a Poisson process [15–21]—and the applicability of STDP in such a regime is not clear. Low firing rates *in vivo* [22–24] would also be expected to reduce coincidence events, limiting STDP's main input stream. Modern *in vivo* experiments in rodent hippocampus suggest an alternative rule—termed "behavioral timescale synaptic plasticity" (BTSP)—that is both qualitatively and quantitatively distinct from STDP [8,25–28]. In BTSP, (i) synaptic weights are thought to change in response to the coincidence of presynaptic spiking and a postsynaptic plateau potential—a long-lasting depolarization of the dendritic arbor—that does not require postsynaptic spiking; (ii) presynaptic spiking and the plateau must occur only within seconds of each other, unlike STDP's coincidence window of tens of milliseconds; and (iii) synaptic weights are thought to change saliently after a single plateau potential, whereas STDP typically requires many spike pairings to evoke significant weight changes. Plateau potentials are extremely sparse *in vivo*, at frequencies estimated at under 10^{-3} Hz within a single neuron, and with likely largely stochastic timing [27,29]. Functionally, plateau potentials are thought to serve as gating signals that determine when input-dependent synaptic plasticity can occur. Theoretical work suggests that BTSP exhibits advantages over traditional plasticity rules in storing memories and shaping neural dynamics [8,29–31]. However, the general ability of BTSP-like learning rules, in particular in comparison to STDP, to store information in energetically constrained systems with high spiking variability has received little formal investigation.

Here I address the problem of how synaptic plasticity unfolds in naturalistic operating regimes using information theory [32]. For three classes of plasticity rules, I estimate the total information stored in the final synaptic weights of a network of neurons about past spike trains produced by a

*Contact author: rkp.science@gmail.com

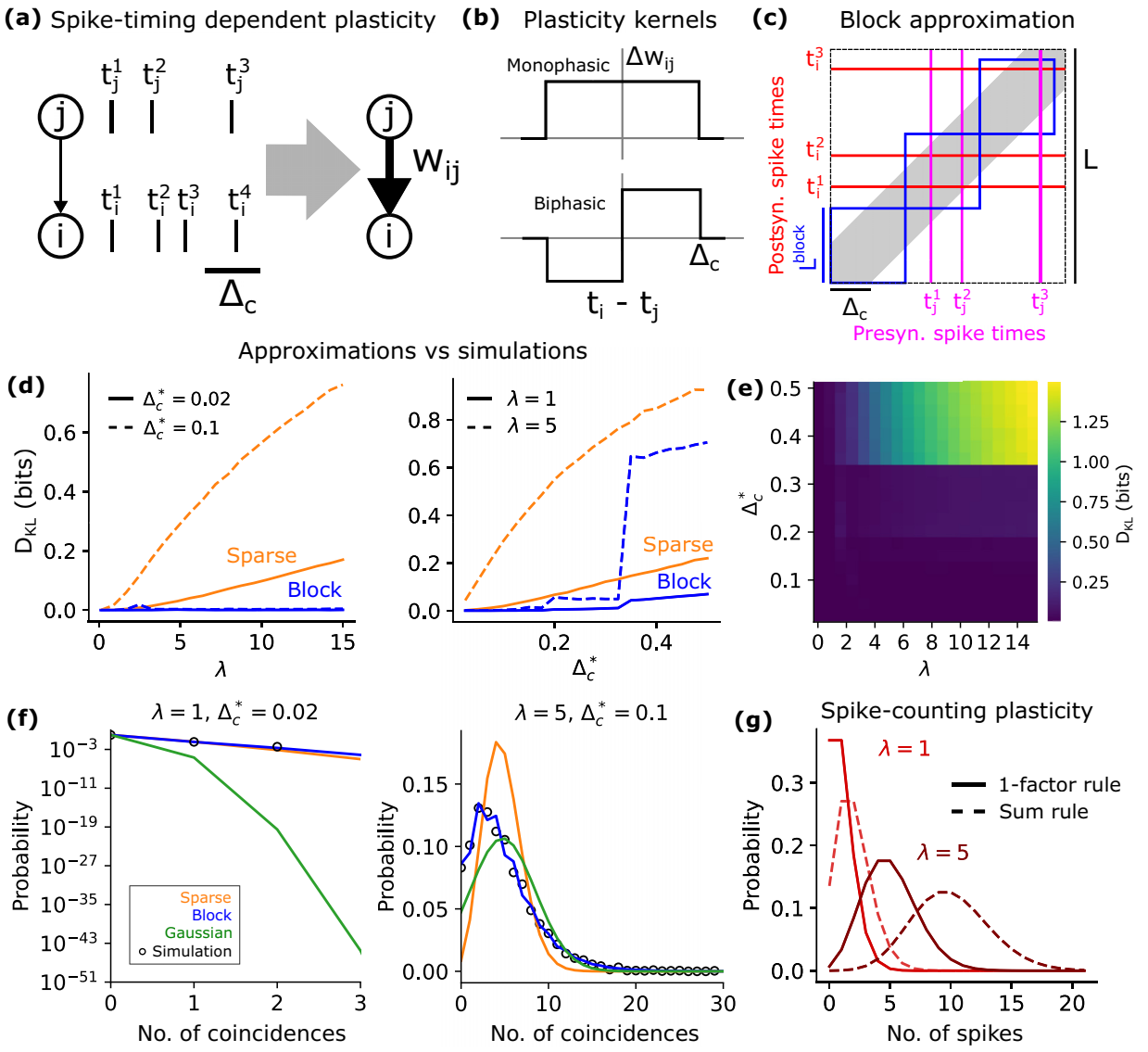

FIG. 1. Distributions of plasticity event counts at a single synapse. (a) Schematic of spike-timing-dependent plasticity (STDP) rule in which synaptic weights update in response to pre- and postsynaptic spike coincidences. (b) Schematics of rectangular STDP kernels used in the analysis. (c) Schematic of block approximation for counting STDP coincidence events. (d) Kullback-Leibler divergence (D_{KL}) between sparse and block approximations of the coincidence count distribution and directly simulated count distributions vs λ and Δ_c^*. Orange and blue lines indicate the sparse and block approximations, respectively. Solid and dashed lines indicate coincidence detection windows of $\Delta_c^* = 0.02$ and 0.1, respectively (left); or $\lambda = 1$ and 5, respectively (right). (e) Heat map of D_{KL} of block approximation vs simulated coincidence count distributions across a range of parameters. (f) Example sparse, block, and Gaussian approximations of coincidence count distributions vs simulated distributions. Orange, blue, and green lines indicate sparse, block, and Gaussian approximations, respectively, and black circles simulation results. (g) Distribution of counts of plasticity events corresponding to spikes, rather than coincidence events, which follow a Poisson distribution. Light and dark red lines indicate mean firing rates of $\lambda = 1$ and 5, respectively. Solid lines show spike count distribution for a one-factor rule (e.g., counting only presynaptic spikes) and dashed lines for the sum rule that counts the sum of pre- and postsynaptic spikes.

Poisson process. I then analyze plasticity under a simple model of energetic constraints on learning, which suggests salient advantages of a BTSP-like learning rule in energy-efficient information storage.

II. THEORETICAL ASSUMPTIONS

To make the analysis tractable while retaining key biological details I employ the following assumptions: (i) a synapse is plastic over a plasticity period of L seconds during which spiking activity in the pre- and/or postsynaptic neuron can modify its weight. This is supported by *in vivo* experimental work [8,25,26] suggesting that synapses are only significantly plastic for short, finite periods on the order of seconds. (ii) The pre- and postsynaptic neurons emit spikes sampled from independent homogeneous Poisson processes at rates r_j (pre) and r_i (post), as in standard models of *in vivo* spiking [15–20]. In particular, I assume that rate changes during the

plasticity period are negligible, in part because although weight changes may be induced during this period, they may not be significantly expressed until later [4,33]. (iii) Weight changes correspond to counts of discrete plasticity events, such that the distribution of final weight changes has support over the integers. I further assume that each potentiation event triggers a weight update of size $+1$ and each depression event an update of size -1. (iv) Within a synapse's plasticity period, there is no noise in the weight updates. Given a learning rule, all variation in the final weights is determined by variation in the spike trains (however, note that in the final section I will consider stochasticity in when the plasticity periods occur). The finite resolution of the synapses comes from the discretization of their support. Biological deviations from these assumptions I treat as higher-order effects, which I revisit in Sec. VIII.

Our goal is to compute the mutual information $\mathrm{MI}[\{w_{ij}\};\{t_k^l\}]$ between the final weight changes $\{w_{ij}\}$ and the spike trains $\{t_k^l\}$ during the plasticity period, where i, j, k index neurons and l spike times. Most generally I allow w_{ij}, the change in a synaptic weight from its original value following its plasticity period, to be any integer; although in most of the analyses w_{ij} will be non-negative. To proceed, I leverage assumption (iv). Since the weights are a deterministic function of the spikes in the plasticity period, we have

$$\mathrm{MI}\left[\{w_{ij}\};\left\{t_k^l\right\}\right] \equiv H[\{w_{ij}\}] - H\left[\{w_{ij}\}|\left\{t_k^l\right\}\right] = H[\{w_{ij}\}], \tag{1}$$

where $H[X] \equiv -\sum_x p(x)\log p(x)$ is the Shannon entropy of the distribution $p(x)$, and $H[X|Y] \equiv \mathrm{E}_y[-\sum_x p(x|y)\log p(x|y)]$ is the conditional entropy [32]. The final equality in Eq. (1) holds because when $\{w_{ij}\}$ is a deterministic function of $\{t_k^l\}$, the conditional entropy $H[\{w_{ij}\}|\{t_k^l\}]$ vanishes since there is no additional variability in the weights given the spike times. Thus, for a given learning rule it suffices to compute the entropy of the final weight distribution produced in response to samples of spike trains. I now proceed to estimate this quantity for three classes of learning rules: an STDP-like learning rule; a one-factor rule that counts only presynaptic (or equivalently only postsynaptic) spikes; as well as a straightforward but little explored sum rule that counts the sum of pre- and postsynaptic spikes. I treat the case of stochastic plasticity periods, in which the equality in Eq. (1) no longer holds, in Sec. VII.

III. PLASTICITY EVENTS AT A SINGLE SYNAPSE

The first technical challenge in the analysis is computing h_1^{STDP}, the entropy of the distribution of final synaptic weights produced by pre- and postsynaptic spike trains for a single synapse w_{ij} via STDP. STDP rules are specified by a plasticity kernel indicating when a pair of spikes evokes a weight change [Figs. 1(a)–1(b)]. When the plasticity kernel is monophasic, reflecting observations in *in vitro* slices from hippocampal area CA3 [4], all pairs of spikes whose absolute timing difference is less than a coincidence window, Δ_c, evoke potentiation events. When the kernel is biphasic, as observed in early STDP experiments in cultured hippocampal neurons [2], spike order matters: pre- followed by postsynaptic spikes lead to potentiation whereas post- followed by presynaptic spikes lead to depression [Fig. 1(b), bottom]. A variety of additional kernel shapes have been identified experimentally [3,34,35] and explored in modeling studies [12,36,37]. Here to illustrate the core phenomena at play and to simplify the subsequent mathematical analysis I consider rectangular symmetric monophasic kernels and antisymmetric biphasic kernels [Fig. 1(b)]. I first analyze symmetric monophasic STDP.

To determine the weight change distribution produced by STDP one must compute the distribution over the count of coincidence events occurring during the plasticity period. Geometrically, the number of coincidence events for symmetric monophasic STDP is equal to the number of intersections falling within the shaded diagonal band in Fig. 1(c). Vertical and horizontal lines represent pre- and postsynaptic spikes, respectively, and their intersections within the shaded region spike pairs whose absolute timing difference is less than Δ_c. When coincidence events are sparse, one can model them independently, such that the total coincidence count is Poisson distributed with mean

$$\mu = r_i r_j A = r_i r_j [L^2 - (L-\Delta_c)^2], \tag{2}$$

where A is the area of the shaded band, L is the plasticity period duration, and r_j and r_i are the pre- and postsynaptic spike rates. I refer to the Poisson distribution with mean μ given by Eq. (2) as the sparse approximation.

When coincidence events are not sparse, the critical mathematical challenge is handling correlations: the presence of a coincidence event at one time influences the probability of coincidence events at nearby times. To estimate the distribution of coincidence event counts, $P(c)$, while accounting for such correlations, instead of counting intersections within the diagonal shaded region, we approximate this region as composed of a series of K square blocks, each of width L^{block}, as shown in blue in Fig. 1(c). The intersection counts are then independent across blocks, and within each block this count $c^{\mathrm{block}} = n_i^{\mathrm{block}} n_j^{\mathrm{block}}$ is distributed as the product of two independent Poisson variables,

$$n_i^{\mathrm{block}} \sim \mathrm{Poisson}(r_i L^{\mathrm{block}}), \quad n_j^{\mathrm{block}} \sim \mathrm{Poisson}(r_j L^{\mathrm{block}}), \tag{3}$$

i.e., of the spike counts of each neuron in the block. To determine L^{block} I use $K = \mathrm{Round}[L/(2\Delta_c)]$ blocks of duration $L^{\mathrm{block}} = 2\Delta_c(1+\epsilon)$. The quantity ϵ represents a correction to ensure the total area of the blocks is equal to the area of the gray band and is given by $\epsilon = \sqrt{(L-\Delta_c/2)/(2K\Delta_c)} - 1$, which ensures the correct expectation value of the total number of coincidence events (Appendix A1 in the Supplemental Material [38]). The distribution over c^{block} is computed from the joint spike count distribution $P(n_i^{\mathrm{block}}, n_j^{\mathrm{block}}) = P(n_i^{\mathrm{block}})P(n_j^{\mathrm{block}})$:

$$P(c^{\mathrm{block}}) = \sum_{n_i^{\mathrm{block}}, n_j^{\mathrm{block}}} P\left(n_i^{\mathrm{block}}\right) P\left(n_j^{\mathrm{block}}\right) \mathbb{1}\left[c^{\mathrm{block}} = n_i^{\mathrm{block}} n_j^{\mathrm{block}}\right], \tag{4}$$

where $\mathbb{1}$ is the indicator function, which equals 1 when the condition in its argument is met and 0 otherwise. While the right-hand side of Eq. (4) does not admit a closed-form expression, the sum can nonetheless be directly computed if one

truncates the support of the Poisson distributions, i.e., assumes $P(n_i^{\text{block}}) \approx 0$ for sufficiently large n_i^{block}.

The total number of coincidence events across the whole plasticity period is then

$$c = \sum_{k=1}^{K} c_k^{\text{block}}, \tag{5}$$

where $c_k^{\text{block}} \sim P(c^{\text{block}})$ is the coincidence count in the kth block. Because the counts in each block are independent, the distribution of the sum of events across all blocks is

$$P(c) = \underbrace{P(c^{\text{block}}) * P(c^{\text{block}}) * \cdots * P(c^{\text{block}})}_{K \text{ times}}, \tag{6}$$

i.e., the (K–1)-fold convolution of $P(c^{\text{block}})$ with itself. From $P(c)$ one now arrives at the block approximation of the entropy produced by monophasic STDP at a single synapse:

$$h_1^{\text{STDP}} = H[c] = -\sum_c P(c) \log P(c). \tag{7}$$

This quantity depends on Δ_c and the pre- and postsynaptic firing rates, r_j and r_i, respectively. Before proceeding, I introduce the dimensionless parameters $\lambda_i \equiv r_i L$, $\lambda_j \equiv r_j L$, and $\Delta_c^* \equiv \Delta_c / L$. This reflects the observation that given these variables, the coincidence count is independent of the units of time. I will often also use the variable $\lambda \equiv \lambda_i = \lambda_j$ when the pre- and postsynaptic neurons have the same firing rate.

Comparison of the block approximation with simulations indicates that it is accurate over a wide range of parameters [Figs. 1(d)–1(e)]. Computing the Kullback-Leibler divergence (D_{KL}) between the approximations and the entropy estimated from direct simulations of coincidence counts reveals that the sparse approximation succeeds only when coincidence events are rare, whereas the block approximation remains accurate over a much wider range of parameters. In particular, although one expects a Gaussian approximation (Appendix A3 [38]) to also succeed for large K (i.e., small Δ_c^*) due to the central limit theorem, the block approximation strongly outperforms the former at small and medium K [Fig. 1(f)]. For a range of λ, the block approximation begins to fail only around $\Delta_c^* \approx 1/3$. Inspecting $P(c)$ computed analytically via the block approximation vs simulations reveals that part of the failure arises from the jagged nature of the block approximation for small K, due in part to the nature of the support of a product of Poisson variables, which is affected, for instance, by the distribution of prime numbers [Fig. S2(a) [38]]. Interestingly, the numerical simulations appear to smooth over this jaggedness.

One can compute $P(c)$ similarly for a biphasic STDP rule in which the order of the spikes determines whether potentiation or depression is triggered [Fig. 1(b), bottom]. Consistent with canonical experimental work [2,5] I allow pre- followed by postsynaptic spikes (causal) to evoke positive, potentiation events, and post- followed by presynaptic spikes (acausal) to evoke negative, depression events. The total plasticity event count that determines the final synaptic weight is then the number of potentiation minus the number of depression events. The analogous block approximation for the biphasic STDP rule is more complex than for the monophasic STDP rule, as detailed in Appendix A4 and Fig. S1(b) [38] but still accurately reproduces the simulated event count histograms over a reasonable range of Δ_c^* and λ [Fig. S2(b)–S2(d) [38]], although note that the accuracy appears to break down for large λ (e.g., $\lambda \gtrapprox 5$), unlike in the case of the monophasic rule.

A crucial point of comparison for the STDP rules are simpler rules that count spikes, rather than coincidence events. I consider two such additional rules. First, I consider a one-factor rule in which the event count is the count of presynaptic spikes (or equivalently postsynaptic spikes [39,40]) alone, as suggested by the operation of synapses in the hippocampal mossy fiber pathway, which to first order appear to have little dependence on postsynaptic spiking [41,42]. The plasticity event count is thus equal to the number of presynaptic spikes, hence Poisson-distributed with mean λ_j [Fig. 1(g)]. Its entropy h_1^{1f} is in turn given by

$$h_1^{1f} = \lambda_j[1 - \log \lambda_j] + e^{-\lambda_j} \sum_{a=0}^{\infty} \frac{\lambda_j^a \log(a!)}{a!}, \tag{8}$$

which for large λ_j can be approximated as

$$h_1^{1f} \approx \frac{1}{2} \log 2\pi e \lambda_j - \frac{1}{12\lambda_j} - \frac{1}{24\lambda_j^2} - \frac{19}{360\lambda_j^3} + \mathcal{O}\left(\frac{1}{\lambda_j^4}\right). \tag{9}$$

In computing this quantity I use the first expression truncated at $a = 20$ for $\lambda_j \leqslant 4$ and the second expression for $\lambda_j > 4$, ignoring the last term.

Second, I consider the "sum rule" in which the final synaptic weight is equal to the sum of the pre- and postsynaptic spike counts. As the sum of two Poisson variables is also Poisson, the distribution of final synaptic weights is Poisson with mean $\lambda_i + \lambda_j$, and the entropy h_1^{Sum} can be computed via Eqs. (8)–(9) by letting $\lambda_j \to \lambda_i + \lambda_j$ [Fig. 1(g)].

IV. INFORMATION RETENTION BY A SINGLE SYNAPSE

Using the above formulas I directly computed the weight entropy of a single synapse produced by each rule for equal pre- and postsynaptic firing rates λ. For sufficiently large λ and Δ_c^*, monophasic STDP rules outperform both the one-factor and sum rule [Fig. 2(a)]. For small λ or Δ_c^*, however, one-factor rules outperform STDP, and the one-factor rule is additionally outperformed by the sum rule. The minimum Δ_c^* needed by STDP to outperform the one-factor rule grows as λ decreases, i.e., at low firing rates a larger coincidence-detection window is required [Fig. 2(a)]. Thus, across a sizable parameter regime the spike-counting one-factor and sum rules store more information than STDP and generally become more favored at lower firing rates (i.e., as λ decreases for fixed L). Intuitively, spikes are more frequent than coincidence events in low firing-rate regimes, hence are able to transmit more information into synaptic weights. Similar results were obtained for the biphasic STDP rule, with the total entropy similar or slightly less than that of the monophasic STDP rule (Fig. S3 [38]). For the rest of this study I thus focus on the monophasic STDP rule, as it represents equal or larger information retention than biphasic STDP, and is moreover approximately the rule reported from *in vitro* measurements of the recurrent neuronal network of hippocampal CA3 [4].

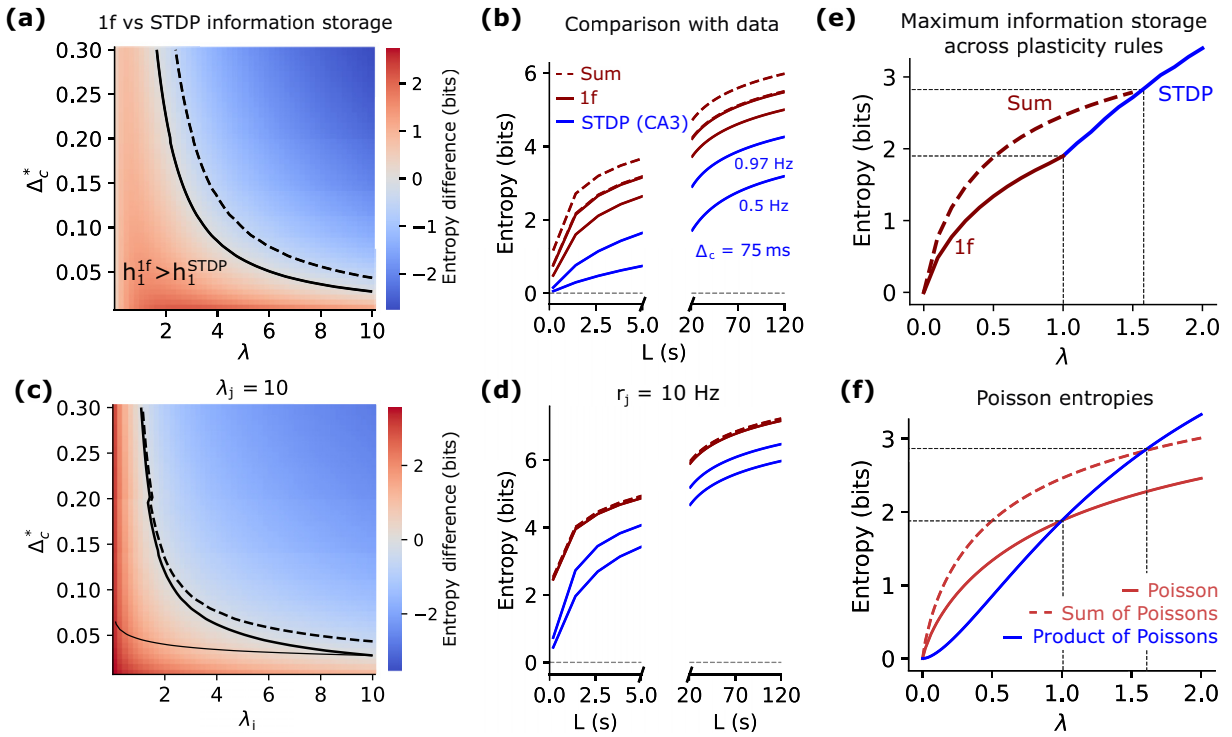

FIG. 2. Information retention by STDP, one-factor, and sum rules at a single synapse. (a) Entropy difference, $h_1^{1f}(\lambda) - h_1^{STDP}(\lambda, \Delta_c^*)$, between weight change distributions of the one-factor and STDP rule vs λ and Δ_c^*. The solid (dashed) line shows the boundary below which the one-factor (sum) rule retains more information than STDP. (b) Entropy of weight changes for the three rules, using experimental values for $r = 0.5$ Hz [24] or 0.97 Hz [23] and $\Delta_c = 75$ ms [4], vs the plasticity period duration L. For the sum and one-factor rules, the lower (upper) lines correspond to $r = 0.5$ Hz ($r = 0.97$ Hz). (c) As in (a), except here the presynaptic neuron has a firing rate of $\lambda_j = 10$. The thick (thin) solid curve shows the boundary below which a one-factor rule counting presynaptic (postsynaptic) spikes outperforms STDP. Below the dashed curve the sum rule outperforms STDP. (d) As in (b) except when the presynaptic neuron fires at $r_j = 10$ Hz. (e) Maximum weight change entropy by any rule (one-factor, sum, and the best STDP rule) vs λ. For each λ the optimal Δ_c^* for the STDP rule was determined by estimating entropy via direct simulations (Fig. S3(a) [38]), to avoid the deviation between the block approximation and sample estimation for large Δ_c^* [Fig. 1(e)]. (f) Analytically computed entropies of a Poisson random variable, the sum of two independent Poisson variables with equal means, and the product of two independent Poisson variables with equal means.

How do the predictions of the theory compare to empirical activity and plasticity measurements? Area CA3 is a subfield of hippocampus considered central to memory formation, characterized by recurrent excitatory connectivity and long modeled as an attractor or autoassociative neural network due to such networks' possible emergence from coincidence- or correlation-based plasticity rules [6,7,43,44]. Indeed, a symmetric, monophasic STDP rule has been reported in CA3 slices *in vitro*, with $\Delta_c \approx 75$ ms [4]. *In vivo*, however, empirically observed mean firing rates in CA3 are quite low, even during locomotion, with time-averaged firing rates often no larger than 0.5–1 Hz [23,24]. As shown in Fig. 2(b), the information stored by our STDP model with these parameters is less than that stored by the one-factor and sum rules, independent of the duration L of the plasticity period.

An important observation, however, is that while time-averaged firing rates in CA3 neurons are very low, many of these neurons are thought to be place cells, whose firing rates substantially increase when an animal is in a specific physical location within an environment, called the neuron's place field [45]. In CA3, mean within-field firing rates reach values of up to around 10 Hz [24]. Accordingly, I computed information retention by one-factor and sum rules vs STDP when the presynaptic neuron operates at $\lambda_j = 10$, as a function of the postsynaptic firing rate and coincidence detection window [Fig. 2(c)]. As shown, while STDP is now more advantaged than when both neurons have low firing rates, the one-factor and sum rules still retain more information than STDP when the postsynaptic firing rate is low. Figure 2(d) shows a similar analysis as Fig. 2(b) except when the presynaptic neuron fires at 10 Hz, demonstrating that spike-counting rules outperform the empirical STDP rule observed in CA3 even if the presynaptic cell has an elevated within-place-field-like firing rate. Thus, in *in vivo* operating regimes, STDP rules estimated from *in vitro* experiments may store less information than plasticity rules that simply count spikes.

I next quantified when the best possible STDP rule would outperform the one-factor or sum rules. I define the best STDP rule via the Δ_c^* that maximizes the final weight entropy, i.e., information stored, which I determine either analytically via the above equations, or numerically in the large Δ_c^* regime in which the block approximation breaks down. Two clear phase transitions appeared, with the best STDP rule outperforming the one-factor rule above $\lambda \approx 1$, and outperforming the sum rule above $\lambda \approx 1.6$, corresponding to final weight entropies of slightly less than 2 and 3 bits,

respectively [Fig. 2(e)]. This means that if the expected number of pre- or postsynaptic spikes in the plasticity period is less than 1 or 1.6, no possible STDP rule as I have defined it can outperform a spike-counting rule at storing information. Equivalently, the one-factor and sum rules outperform STDP if the synapse can store less than about 2 or 3 bits, respectively (i.e., STDP would require a longer plasticity period to store the same amount of information). This result can also be understood by comparing the entropies of (i) a Poisson variable (one-factor rule), (ii) the sum of two Poisson variables with equal λ (sum rule), and (iii) the product of two Poisson variables with equal λ (STDP), with the latter representing the coincidence count distribution when the detection window $\Delta_c^* = 1$. As shown in Fig. 2(f), these entropies produce similar phase transitions as those computed via the plasticity analysis. Although the STDP entropy is not strictly maximized for $\Delta_c^* = 1$ [Fig. S3(a) [38]], the entropy for the best possible STDP rule is nonetheless quite close to the entropy of the product of two Poisson variables, revealing that the comparison of the plasticity rules undertaken so far is quite similar to the entropies of these simple forms of random variables.

V. PLASTICITY EVENTS IN A NETWORK

I now estimate the information, h^R_{Ntwk}, stored in a network of neurons in which multiple synapses are plastic simultaneously according to learning rule R. I assume all neurons follow independent Poisson processes. Let N be the number of neurons and M the number of plastic synapses during the plasticity period of length L [Fig. 3(a)]. Each synapse is directional, i.e., defined by its pre- and postsynaptic neurons. For simplicity I assume the set of plastic synapses are chosen uniformly at random, without replacement, from the set of all possible synapses, excluding self-connections. The connection probability is thus $p_{\text{syn}} \approx M/(N^2 - N) \approx M/N^2$ for large networks, and the mean in-degree and out-degree both given by $\text{E}[\text{in-degree}] = \text{E}[\text{out-degree}] \approx p_{\text{syn}}(N-1) = M/N$.

In the simplest approximation, one can treat all synapses as independent, such that for any rule,

$$P(\{w_{ij}\}) = \prod_{(i,j)} P(w_{ij}) \quad \text{hence} \quad h^R_{\text{Ntwk}} \equiv H[\{w_{ij}\}] = M h^R_1, \tag{10}$$

where the product is over all plastic synapses, and h^R_1 is the entropy of a single synaptic weight change given learning rule R. In this approximation the single-synapse entropy is simply scaled by M so the relative information retention by the different rules follows immediately from the single-synapse case.

More generally, there will be correlations among synaptic weight changes in the network. In our model all neurons are independent, hence weight changes can only be correlated if there is overlap between their pre- and postsynaptic neurons. If two synapses share no neurons they will be independent. For the one-factor (presynaptic-only) rule the correlations are straightforward. The weight changes in two synapses will be identical if they share a presynaptic neuron and otherwise independent, hence

$$h^{1f}_{\text{Ntwk}} = N_{\text{pre}} h^{1f}_1, \tag{11}$$

where N_{pre} is the total number of unique presynaptic neurons across the M synapses. If all M synapses emanate from the same presynaptic neuron, $N_{\text{pre}} = 1$ whereas if all M synapses emanate from different presynaptic neurons, then $N_{\text{pre}} = M$. One can approximate the expectation of N_{pre} as the average number of unique outcomes from M samples from N options, which yields

$$\text{E}\left[h^{1f}_{\text{Ntwk}}\right] = \text{E}[N_{\text{pre}}] h^{1f}_1 \approx N\left(1 - \left(1 - \frac{1}{N}\right)^M\right) h^{1f}_1. \tag{12}$$

Synaptic weight correlations for the (monophasic) STDP and sum rules are more complex. For these rules, the weight updates w_{ij}, w_{pq} will only be identical if $p = j$ and $i = q$, i.e., if the synapses are reciprocal (if $i = p$, $j = q$ this would mean they are the same synapse, and we do not allow duplicate synapses). They will be correlated but not identical if they share a single neuron, i.e., if either $i = p$, $i = q$, $j = p$, or $j = q$. In the subsequent analyses I will assume that all neurons have the same firing rate λ, and that all synapses use the same Δ_c^* for the STDP rule. Then the pairwise entropy $H[w_{ij}, w_{jk}]$ is equal for every pair of synapses sharing one neuron, as is their mutual information: $\text{MI}[w_{ij}; w_{jk}] = H[w_{ij}] + H[w_{jk}] - H[w_{ij}, w_{jk}]$. This quantity can be computed either exactly (sum rule) or via the block approximation (STDP), using a straightforward extension of the single-synapse method (Appendix A5 [38]). The difference between the joint entropy and sum of independent entropies for the STDP rule vs Δ_c^*, and for the sum rule, for several λ, is shown in Fig. 3(b), indicating that for a single pair of synapses sharing a neuron the loss of information due to correlations is small.

To estimate information retention stored in the weights of a network by the STDP and sum rules I employ a recursive method, adding synapses one-by-one and updating the estimate of the total entropy in turn. The entropy added by the first synapse is simply the single-synapse entropy, h_1^{STDP} or h_1^{Sum}. To add a synapse and update the entropy estimation, suppose that m synapses have already been added with the total entropy estimated at h^R_m, using rule R. We wish to compute the entropy h^R_{m+1}. Let Ω_B be the set of synapses neighboring synapse $m+1$, i.e., sharing one neuron with the synapse I are adding [Fig. 3(c)]; let Ω_A be all synapses already accounted for, excluding Ω_B, i.e., $|\Omega_A| = m - |\Omega_B|$ and $H[\Omega_A, \Omega_B] = h^R_m$; and let Ω_C be synapse $m+1$. Then, $h^R_{m+1} = H[\Omega_A, \Omega_B, \Omega_C]$. To proceed analytically, I leverage two approximations. In Approximation 1, I treat any two synapses that do not share a neuron as conditionally independent given all other synapses [in the model two synapses sharing no neurons are only marginally independent: $P(w_{ij}, w_{kl}) = P(w_{ij})P(w_{kl})$, so this approximation imposes further structure]. Following from this approximation, we have that Ω_C and Ω_A are conditionally independent given Ω_B. Therefore

$$\begin{aligned} h^R_{m+1} &= H[\Omega_A, \Omega_B, \Omega_C] = H[\Omega_A, \Omega_B] + H[\Omega_C|\Omega_A, \Omega_B] \\ &\approx H[\Omega_A, \Omega_B] + H[\Omega_C|\Omega_B] \\ &= H[\Omega_A, \Omega_B] + H[\Omega_B, \Omega_C] - H[\Omega_B] \\ &= h^R_m + H[\Omega_B, \Omega_C] - H[\Omega_B], \end{aligned} \tag{13}$$

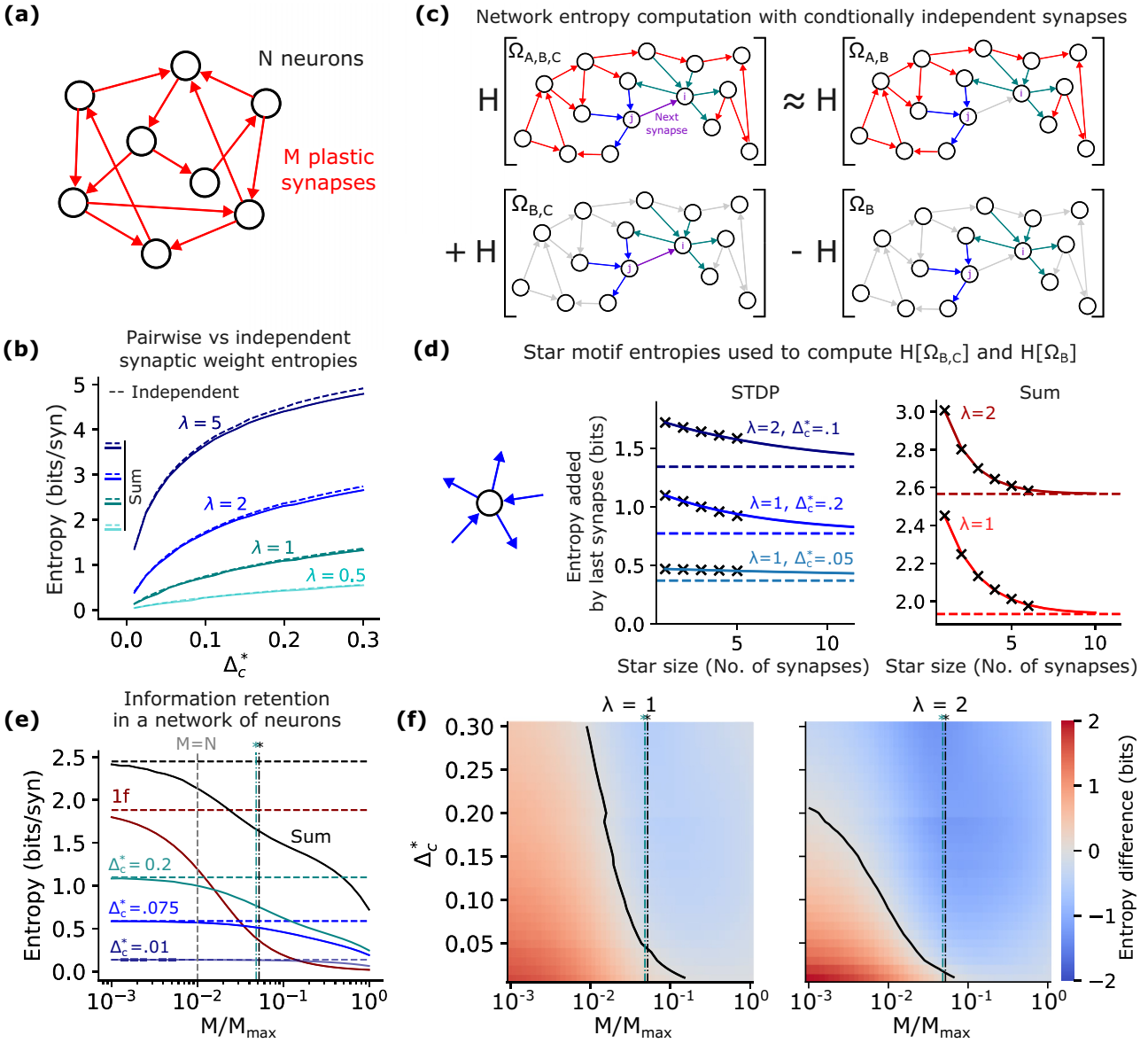

FIG. 3. Information retained in a network. (a) Network schematic depicting that M synaptic connections in a network of N neurons are plastic for L seconds. (b) Joint entropy of a pair of synapses sharing one neuron (solid lines) vs the sum of entropies of two independent synapses (dashed lines), for the STDP rule (block approximation) vs Δ_c^* and λ; and for the sum rule for several λ. (c) Schematic depicting how the entropy estimate of a set of synaptic weights is updated when a new synapse is added. Here $\Omega_{A,B,C} \equiv \Omega_A \bigcup \Omega_B \bigcup \Omega_C$, $\Omega_{A,B} \equiv \Omega_A \bigcup \Omega_B$, etc. (d) Computation and extrapolation of entropies of star motifs (a set of synapses sharing one neuron; left) for the STDP and sum rules. Crosses represent directly computed entropies, colored lines the extrapolation, and dashed lines the baseline term of the extrapolation. (e) Entropy of weight distributions in a network vs number of plastic synapses, using Eq. (12) for the one-factor rule or the approximation of (c) [Eqs. (13)–(14)] for the STDP and sum rules, for $\lambda = 1$ and three different Δ_c^* for the STDP rule, with $N = 100$ ($M_{\max} \approx 10^5$). For the STDP and sum rules, averages over 30 instantiations of M random synaptic connections are shown, since different instantiations of the network can yield different connectivity patterns. Dashed horizontal lines show corresponding entropies when all synapses are independent. The two dot-dashed vertical lines with stars above indicate when the mean star motif size used in the total entropy computation exceeded 5 (teal) or 6 (black), corresponding to (d). (f) Heat map of the entropy difference, $h^{1f}_{\text{Ntwk}} - h^{\text{STDP}}_{\text{Ntwk}}$ vs Δ_c^* and $M/M_{\max}$, for two λ. In both plots the sum rule exceeds STDP across the entire parameter space shown. [The slight kink at $\Delta_c^* = 0.2$ (left) arises from the Round function in the STDP block approximation.]

where I have used the chain rule for entropy and Approximation 1.

To compute $H[\Omega_B]$ and $H[\Omega_B, \Omega_C]$ I introduce Approximation 2. Let i be the postsynaptic neuron of synapse $m+1$ and Ω_i be all existing synapses (excluding the new synapse, $m+1$) that touch neuron i [Fig. 3(c); cyan], and similarly for the presynaptic neuron j and its neighboring synapses Ω_j [Fig. 3(c); blue]. In Approximation 2 I treat Ω_i and Ω_j as sharing no neurons, such that Ω_i and Ω_j are marginally independent and also conditionally independent given Ω_C (via

Approximation 1). Then,

$$H[\Omega_B] \approx H[\Omega_i] + H[\Omega_j] \quad \text{and}$$
$$H[\Omega_B, \Omega_C] \approx H[\Omega_i, \Omega_C] + H[\Omega_C, \Omega_j] - H[\Omega_C]. \quad (14)$$

The terms Ω_i and Ω_j, as well as $\Omega_i \bigcup \Omega_C$ and $\Omega_C \bigcup \Omega_j$, constitute star motifs (a collection of m^* synapses sharing one neuron), whose entropy can be computed directly for small m^* (up to $m^* = 5$ for the STDP rule and $m^* = 6$ for the sum rule) and extrapolated with a simple formula for larger star sizes [Fig. 3(d) and Appendix A5 [38]]. Figure S4 [38] depicts the extrapolations used for large star motif sizes and how often these extrapolations and Approximation 2 are used as a function of m.

Equations (13)–(14), together with Approximations 1 and 2, thus enable a direct calculation of h^R_{m+1}. To compute the entropy of the entire network I repeat this procedure until all M synapses are added. The final entropy determined through this procedure, while an approximation, accounts for a large portion of the correlation structure in $\{w_{ij}\}$ produced through the STDP and sum rules.

VI. INFORMATION RETENTION IN A NETWORK

Using the above approximations for the entropy of synaptic weight updates across the network, I computed this entropy as a function of $M/M_{\max}$, where $M_{\max} = N(N-1) \approx N^2$ is the maximum possible number of plastic synapses in the network. As shown in Fig. 3(e), for $\lambda = 1$ and $M \ll N$, one-factor rules outperform STDP for several Δ_c^*, which follows largely from the single-synapse analysis, since correlations are weak in this regime hence the total information stored by any rule is approximately the single synapse information multiplied by M. One-factor rules, however, are affected strongly by correlations as $M/M_{\max}$ increases, since the total information is constrained by the number of neurons $N \geqslant N_{\text{pre}}$ in the network, and any set of synapses that share a common presynaptic neuron store only as much information as a single synapse; hence the total information stored by the one-factor rule decreases quickly as M surpasses N. As expected from the single-synapse analysis (Fig. 2), STDP rules also gain increasing performance over one-factor rules as λ increases [Fig. 3(f)]. When λ is small, however, STDP rules may also be able to outperform one-factor rules if plasticity becomes sufficiently dense, i.e., as M increases [Figs. 3(e)–3(f)], with the analysis suggesting—to the extent that the approximations remain accurate in this regime—that STDP is less limited by network correlations than the one-factor rule hence may overtake the latter in storing information at a network level.

Similar to the single-synapse analysis, at low firing rates the sum rule outperforms both the one-factor and STDP rules [Figs. 3(e)–3(f)]. As more synapses become plastic, however, the sum rule may also be less limited by correlations than the one-factor rule, since collections of synapses sharing the same presynaptic neuron can measure different postsynaptic activities, although accurate comparisons become difficult in the dense plasticity regime $M/M_{\max} \to 1$ where Approximation 2 becomes more significantly invoked [Figs. 3(e)–3(f), S4 [38]]. Together, these results suggest the following in the regime of low firing rates (small λ): (i) when plasticity is sparse, STDP is outperformed by the one-factor rule, which is in turn outperformed by the sum rule; (ii) as plasticity becomes dense, STDP may outperform the one-factor rule at a network level; and (iii) the sum rule may outperform both STDP and the one-factor rule regardless of how many synapses are plastic.

VII. INFORMATION RETENTION UNDER ENERGETIC CONSTRAINTS

Performance optimization has proven a powerful principle for explaining diverse biological phenomena [46]. In the nervous system, such analyses have focused predominantly on peripheral, e.g., sensory systems, and it is unknown whether a similar approach may be useful for understanding brain areas such as hippocampus, which are deeper in the brain. Given our observations above, is there a relevant quantity that may be optimized by a hippocampal CA3-like network? Here I hypothesize that such a network may be well posed for maximizing information storage about spike trains per unit energy.

Generating spikes, as well as learning via synaptic plasticity, are both thought to require energy [30,47–51]. The energy needed for the former one expects to be minimized by operating at as low firing rates as possible, which appears largely consistent with *in vivo* data [22–24]. Here I focus instead on the energy required to learn via synaptic plasticity. I model this quantity, E_{learn}, as composed of two terms:

$$E_{\text{learn}} = ML + \gamma M. \quad (15)$$

The first term accounts for the total amount of time L each synapse is plastic (the plasticity period), aggregated over all M synapses. That is, I assume it requires energy for the synapse to remain in a plastic state, e.g., to power the biophysical processes needed to record information. The second term reflects a fixed cost γ for each plastic synapse, for instance to initiate the plasticity period or persist the synapse from degradation after learning. The overall goal of this equation is to capture the assumption that plasticity is energetically costly due to the physical processes needed to instantiate it, with the general hypothesis being that energy cost is minimized by writing to as few synapses as possible, with each such synapse plastic for as short a period as possible. For simplicity I will also assume that Eq. (15) is general across rules, i.e., E_{learn} does not depend on the biophysical processes specific to each rule, although such a scenario will be important to investigate in future work. While more complex models are possible, Eq. (15) represents a parsimonious model of energy expenditure that increases with both M and L.

Crucially, although I will assume for simplicity that all synapses have identical parameters, I will allow variability in when their individual plasticity periods occur. Formally, I let each synapse be subject to a binary plasticity gating signal that equals 1 when the synapse is plastic and 0 otherwise, such that the synaptic weight can only change in response to pre- and/or postsynaptic spikes when this gating signal equals 1. In turn, I will assume that only synapses that experience a gating signal and hence a plasticity period at some point during the learning session will expend energy.

We now embark on the task of understanding the energy costs needed to store I bits of information in the network

under different learning rules and patterns of gating signals. In the remainder of my analyses I will assume that the network is sufficiently large relative to I that only a sparse fraction of synapses need to be plastic at any time [i.e., the $M/M_{\max} \to 0$ regime of Figs. 3(e)–3(f)] in order to store the desired information. This in turn enables modeling the total information stored in the network as simply the sum of the information stored in the individual synapses (although in my very last analysis I will consider synapses sharing common postsynaptic neurons). I will also assume that the gating signals are sufficiently sparse that it is very unlikely that any individual synapse should experience more than one gating signal during a single learning session. The number M of total plastic synapses is then simply equal to the total number of gating signals that occur. If I is too large or the network too small to achieve successful storage with only a sparse set of synapses plastic at any time, then the results that follow may not completely hold.

In this setting, what is the most energy-efficient way to store I bits in a network over a learning session of length T, which may be substantially larger than any plasticity period duration L? That is, what learning rules and plasticity gating signals minimize E_{learn}? To gain intuition I first consider deterministic plasticity gating signals. Figure 4(a) shows three example simulations each tuned to store approximately the same amount of total information [$I \approx 4$ bits; Fig. 4(b)], but with different gating signal schedules. In Simulation A, all M synapses are plastic over the whole learning session. This leads to both a large ML and γM [Fig. 4(d)]. In Simulation B, each synapse is plastic for the minimal possible L needed to store I/M bits [i.e., using the best plasticity rule; Figs. 2(e) and 4(c)]. This leads to a smaller ML, but equal γM as in Simulation A [Fig. 4(d)]. In Simulation C the information content is distributed over fewer synapses, leading to a smaller γM than Simulations A and B [Fig. 4(d)]. However, in C, each synapse must be plastic for longer than in B, and disproportionately so: in the regime where the sum rule is optimal, for instance, one synapse plastic for 1 s stores less information than two synapses plastic for 0.5 s [Fig. 2(e)]. Simulation C thus has a lower γM than Simulation B but a higher ML [Fig. 4(d)]. As Simulation A has the highest E_{learn}, one expects the most energy-efficient plasticity process to lie somewhere between B and C. The final quantity E_{learn} as a function of M thus reflects competition between a decreasing ML and increasing γM, with optima depending on γ [Fig. 4(e)]. For large γ, using as few synapses as possible can be optimal. As γ decreases, however, it becomes favorable to distribute storage over more than the minimal number of synapses to reduce the ML term. Intuitively, to minimize E_{learn} one wishes to write to as few synapses as possible as fast as possible.

Simulations B and C, in which only one or a few synapses are plastic at any moment, also have the advantage of distributing information storage more evenly over the total learning session than in Simulation A [Fig. 4(f)]. Simulation A stores information faster at first and slower later on, due to the sublinear increase of the entropy of sums of random variables. Cycling through synapses, each plastic for a short period, diminishes this effect at the population level, such that the overall rate at which information is written into the network remains more stable over time.

Second, I consider the case of stochastic gating signals. Indeed, in hippocampus the gating signals in BTSP (plateau potentials) are thought to be both sparse but also largely stochastic [8,27]. In this setting the weights are no longer a deterministic function of the spikes, and one must thus take care in computing information retention since the right-hand equality in Eq. (1) no longer holds. Instead, I compute the mutual information by directly calculating how much the entropy of the spike trains decreases upon observation of the weights (Appendix A9 [38]):

$$\mathrm{MI}\left[\{w_{ij}\};\left\{t_k^l\right\}\right] = H\left[\left\{t_k^l\right\}\right] - H\left[\left\{t_k^l\right\}|\{w_{ij}\}\right]. \tag{16}$$

For simplicity and to illustrate our key points, for the remainder of this study I will consider binary synapses $w_{ij} \in \{0, 1\}$. Past work modeled BTSP synapses as binary [29,53] to provide conservative estimates of synaptic weight resolution; one-bit synapses are also relevant to neuromorphic devices, since the latter are far easier to implement with memristors than multibit synapses [29,54]. In this setting the sum rule is optimal, followed by the one-factor rule, followed by STDP [Fig. 2(e)]. As a final useful simplification, I will consider learning sessions discretized into integer multiples of L, containing $Q \equiv T/L$ distinct periods in which each synapse could potentially be plastic [Fig. 4(g)]. The energy specified by Eq. (15) is determined entirely by L and the number of gating signals. How does stochasticity of the gating signals change the total information stored at fixed energy cost?

Formally, I implement plasticity gating as follows. In each period q in each synapse (i, j), spiking produces an eligibility trace $e_{ij}^q \in \{0, 1\}$ via one of the original three learning rules, with probability $p_e \equiv p(e_{ij}^q = 1)$. For instance, if the presynaptic one-factor rule is used, and there is approximately a 0.5 probability of a spike occurring in a period q, then the eligibility traces roughly indicate which synapses experienced a presynaptic spike in which periods [Fig. 4(g)]. If the gating signal $g_{ij}^q = 1$ during period q, the eligibility trace is measured by the synapse and converted to a weight change $w_{ij} \leftarrow e_{ij}^q$; if $g_{ij}^q = 0$ the eligibility trace is not measured, hence forgotten, and the synapse is left unchanged. Since all information about the spikes flows through the eligibility traces, which are a deterministic function of the spikes, the mutual information between the weights and spikes equals that between the weights and eligibility traces; i.e., given e_{ij}^q, the precise timing of the spike(s) within period q can have no additional influence on the weights.

Given this model I examined two forms of gating stochasticity [Fig. 4(g)]. First, I considered fully random gating signals independent across all synapses and periods. For multiple p_e fully random gating led to a dramatic degradation of the information stored that became worse for longer learning sessions [Fig. 4(h)]. Intuitively, a weight $w_{ij} = 0$ reveals almost nothing about the spike trains surrounding synapse (i, j), since most probably a gating signal simply did not occur; on the other hand, for the potentiated synapses ($w_{ij} = 1$), one learns only that these synapses each experienced at least one positive eligiblility trace during the learning session, which occurs for typical spike trains (as long as p_e is not too small), hence the entropy of the spike trains decreases very little. Thus, while sparse gating reduces energy through Eq. (15),

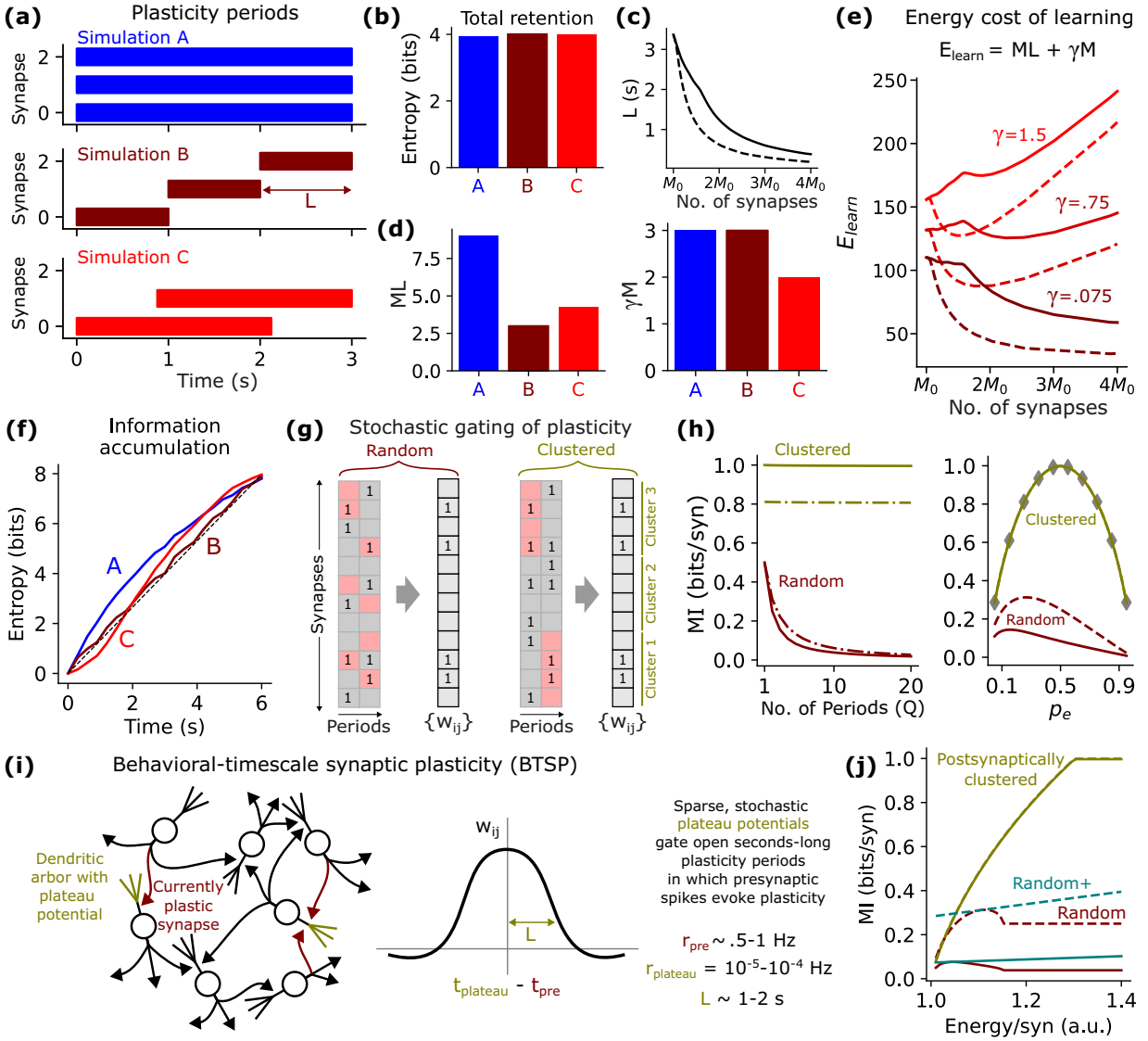

FIG. 4. Information storage under energetic constraints. (a) Three example simulations with deterministic gating signals, each using up to three synapses to store information about spike trains. (b) The total entropy of each simulation's synaptic weights after 3 s, estimated from 10^5 trials using a firing rate of 0.5 Hz, assuming independent synapses (sparse plasticity). (c) Optimal L (the shortest plasticity period using any rule [Fig. 2(c)]) vs the number of plastic synapses used to store the same total information. Here 96 bits are stored in at least $M \geqslant M_0 = 32$ independent synapses (3 bits per synapse). The solid (dashed) curve shows the best L between the one-factor (sum) and STDP rules. (d) Learning cost terms for the simulations in (a). (e) Total learning costs vs M for a simulation in which at least $M \geqslant M_0 = 32$ synapses store 96 bits, for different γ. $L = r\lambda$ is computed by selecting the λ that optimizes information retention between the one-factor and STDP rules (solid lines) or between the sum and STDP rules (dashed lines). (f) Total entropy, measured for simulations running from 0, 1, ..., 6 total seconds in which six (A, B) or four (C) synapses were used. While the total entropy or information is equal after the 6 s for all simulations, the simulations differ in the timecourse of how this information is accumulated. (g) Two forms of stochastic gating of plasticity. Inscribed 1's represent positive eligibility traces and pink shading gating signals. (h) Information about spike trains stored in one-bit synapses vs number of periods (left) for $p_e = .5$ (solid) or $p_e = .25$ (dot-dashed), or vs p_e (right) for $Q = 2$ (dashed) or $Q = 5$ (solid) periods. Diamonds: information stored by single synapses given deterministic gating. (i) Schematic of experimentally observed BTSP rule [8,25,26,52]. Spike rates are from area CA3 reported in Refs. [23,24] in behaving rodents. The plateau rate is estimated from [8] (Appendix A10 [38]). (j) Information about spike trains stored by the two schemes in (g) when gating signals are postsynaptically clustered, vs energy budget. Dark red lines show direct replacement of clustered with random gating with M fixed. Cyan lines ("Random+") show information retained after additionally optimizing M at fixed energy [Eq. (15); $\gamma = 1$] for $Q = 10$ (solid) or $Q = 2$ (dashed).

when this gating is maximally random information storage significantly suffers.

I therefore considered an alternative case in which the gating signals are random but clustered, as in biological BTSP in which all synapses presynaptic to a postsynaptic neuron experiencing a plateau potential become plastic simultaneously [Fig. 4(i)]. Strikingly, one finds that when gating signals are clustered, the information retention increases to a value close to that expected in the deterministic case [Fig. 4(h)]. Intuitively, if any $w_{ij} = 1$ within a cluster, this implies that

the entire cluster received a gating signal during some period q' within the learning session; during this period the collection of eligibility traces in that cluster were all converted into weights, such that the resulting weights in the cluster are essentially a perfect snapshot of the eligibility traces at q'. Since typical spike trains will not produce a snapshot of eligibility traces exactly equal to the $\{w_{ij}\}$ in the cluster (if the cluster has more than just a few synapses), the entropy of the spike trains thus dramatically decreases upon observation of these weights. Indeed, the essential information lost in the clustered gating scenario is only which period q' received the gating signal, which is only $\log_2 Q$ bits per cluster. Thus, randomizing the gating signals destroys information, but clustering them largely restores it.

Finally, in biological BTSP plasticity is indeed clustered, but via a shared postsynaptic cell [Fig. 4(i)]. This means that the best plasticity rule—the sum rule—loses its advantage and approaches performance equivalent to the one-factor (presynaptic only) rule due to the correlated postsynaptic spikes, decreasing information storage (Appendix A9 [38]). I therefore sought to quantitatively compare postsynaptically clustered gating to random gating using the sum rule. For each possible amount of information stored per synapse (up to a maximum of 1 bit) in the postsynaptically clustered case, I determined the effective spike rates and energy cost of learning [Eq. (15)] in the network then computed how much information would be stored per synapse at this same energy cost if gating were made fully random and the sum rule were used instead. In this case the loss of information due to postsynaptic clustering was much smaller than the loss of information due to randomizing the gating signals, such that postsynaptically clustered gating (effectively using the one-factor rule to produce eligibility traces) still stored much more information per synapse than randomized gating but using the sum rule, with a more pronounced effect for longer learning sessions (larger Q) [Fig. 4(j)]. Thus, when gating signals are stochastic, the information gained by postsynaptically clustering these gating signals, as in BTSP, outweighs the information lost due to shared postsynaptic neurons. The BTSP-like rule therefore decreases energy usage by only allowing a few synapses to be plastic at a time, yet retains substantial information about past spike trains by counting presynaptic spikes during the plasticity periods and clustering the gating signals that determine which synapses are plastic when.

VIII. DISCUSSION

Physical learning systems, whether biological or artificial, must store information and generally use energy to do so. The above results suggest that in low firing rate activity regimes, it may be advantageous for only a few synapses to be plastic at any time, and for these synapses to measure spike counts during their plasticity periods, i.e., time averages or integrals of spiking activity rather than correlations and spike-time differences, at least for the first few bits of information to be stored. Intuitively, spikes are more frequent than spike-pair coincidences for Poisson processes at low firing rates, thus have higher capacity to transmit information. While in this study I considered an implementation of STDP using rectangular temporal kernels [Fig. 1(b)], in future work it will be useful to investigate nonrectangular kernels observed *in vitro*, in which trial-averaged weight updates are continuous valued and depend more smoothly on the pre-post time difference. Notable variants of STDP have also been proposed, for instance in which postsynaptic voltage prior to spiking influences the sign of the weight change [55], or in which plasticity is evoked by triplets, rather than pairs of spikes [56]. In future work it will be interesting to compute information retention in these rules as well, although the intuition we have developed here suggests that at low firing rates any rule depending on precise spike coincidences will generally store less information than spike-counting rules unless spike statistics are significantly non-Poissonian. While I have assumed an objective function of storing as much total information about the spike trains as possible, it may also be the case that certain features of the spike trains are more relevant than others [57,58], and it will be important to explore how different plasticity rules might optimize the storage of spike trains in which different features are differently important. Curiously, for the one-factor, sum, and BTSP-like plasticity rules discussed in this study, precise spike times are forgotten, and only information in the neurons' spike counts can be stored, although spike timing could support other functions.

The model suggests a normative account for the recently discovered learning rule BTSP [8,25,26,29] as a rule that stores information in as few synapses as possible as fast as possible (Fig. 4) to implement an energetically efficient learning system. The final analysis also showed that if gating of plasticity is stochastic, clustering plastic synapses increases information retention. As shown in Fig. 4(j), the information gain from clustering the gating signals overcomes the information loss from correlated postsynaptic spikes when synapses share a common postsynaptic neuron. In principle, further information could be retained if the gating signal clusters comprised synapses with no shared pre- or postsynaptic neurons, since then the sum rule could be fully exploited, but this may be more difficult to achieve biologically than clustering on the postsynaptic neuron, as in BTSP. Further notable features of BTSP include a depression component activated when synapses are already strong before the onset of the plasticity period [26] and the co-occurrence of plateau potentials with bursts of postsynaptic spikes [8,25], which will be interesting to explore in future work; the latter, for instance, suggests that postsynaptic spikes may still play some role in plasticity even though they are not required for induction. In future work it will also be important to consider the role of more realistic spatiotemporally heterogeneous firing rates [59], non-Poissonian or non-independent activity patterns such as stereotyped spike sequences [62], modulation of spiking activity by inputs, e.g., encoding the position of the animal [45,60,61], nonspiking activity patterns such as sharp-wave ripples [62], and synaptic forgetting [63,64]. It will also be important to consider in more detail the energetics of the specific biophysical processes that may underlie different learning rules.

Our theory, although idealized, makes specific predictions about learning in the brain. First, it is not unrealistic that spike-counting and STDP rules both play some role *in vivo*, given the diversity of plasticity rules reported across *in vitro* experiments [3,5]. I predict, however, that STDP-like rules

will be found preferentially in neural circuits or subnetworks operating at higher firing rates, since this regime produces more frequent coincidence events that are in turn better able to transmit information into weights. Second, I predict that in brain areas exhibiting BTSP [8,25,26,52], the duration of the plasticity period should be inversely related to the firing rate of the presynaptic neuron, since lower firing rates necessitate longer plasticity windows to store the same amount of information. This prediction moreover suggests that the seconds-long plasticity periods in BTSP, which are thought to reflect a behavioral timescale (compared with STDP's tens of milliseconds), may additionally reflect the time needed for synapses to store information at low hippocampal firing rates.

While information is a natural quantity for studying theoretical bounds on learning [63–69], in the living brain this information must be stored in such a way that it can influence subsequent neural activity dynamics and computation. A longstanding appeal of correlation-based and STDP-like rules is their suitability for creating attractor networks that can perform computations via fixed-point or metastable dynamics [12,43,44,70,71]. How spike-counting rules could produce similarly complex and computationally useful dynamics is less clear given the relative simplicity of their induction mechanism. Recent work, however, showed that such rules may be well suited for encoding information in representations resembling virtual odor trails, which could produce flexible dynamics and computation through integration into a feedback controller [31,72].

ACKNOWLEDGMENTS

I would like to thank Aaron Milstein for helpful comments on an early version of the manuscript, Yiqin Gao and Luca di Carlo for insightful discussions throughout the development of the ideas presented in this study, as well as my reviewers for valuable feedback.

DATA AVAILABILITY

The software and data that support the findings of this article are openly available [73].

[1] H. Markram, J. Lübke, M. Frotscher, and B. Sakmann, Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs, Science **275**, 213 (1997).

[2] G.-q. Bi and M.-m. Poo, Synaptic modifications in cultured hippocampal neurons: Dependence on spike timing, synaptic strength, and postsynaptic cell type, J. Neurosci. **18**, 10464 (1998).

[3] D. E. Feldman, The spike-timing dependence of plasticity, Neuron **75**, 556 (2012).

[4] R. K. Mishra, S. Kim, S. J. Guzman, and P. Jonas, Symmetric spike timing-dependent plasticity at CA3–CA3 synapses optimizes storage and recall in autoassociative networks, Nat. Commun. **7**, 11552 (2016).

[5] J. C. Magee and C. Grienberger, Synaptic plasticity forms and functions, Annu. Rev. Neurosci. **43**, 95 (2020).

[6] R. P. Kesner, Behavioral functions of the CA3 subregion of the hippocampus, Learning Memory **14**, 771 (2007).

[7] E. T. Rolls, A quantitative theory of the functions of the hippocampal CA3 network in memory, Front. Cell. Neurosci. **7**, 98 (2013).

[8] Y. Li, J. J. Briguglio, S. Romani, and J. C. Magee, Mechanisms of memory-supporting neuronal dynamics in hippocampal area CA3, Cell **187**, 6804 (2024).

[9] T. Masquelier, R. Guyonneau, and S. J. Thorpe, Competitive STDP-based spike pattern learning, Neural Comput. **21**, 1259 (2009).

[10] I. R. Fiete, W. Senn, C. Z. Wang, and R. H. Hahnloser, Spike-time-dependent plasticity and heterosynaptic competition organize networks to produce long scale-free sequences of neural activity, Neuron **65**, 563 (2010).

[11] S. Klampfl and W. Maass, Emergence of dynamic memory traces in cortical microcircuit models through STDP, J. Neurosci. **33**, 11515 (2013).

[12] A. Litwin-Kumar and B. Doiron, Formation and maintenance of neuronal assemblies through synaptic plasticity, Nat. Commun. **5**, 5319 (2014).

[13] T. M. George, W. de Cothi, K. L. Stachenfeld, and C. Barry, Rapid learning of predictive maps with STDP and theta phase precession, eLife **12**, e80663 (2023).

[14] J. Stubenrauch, N. Auer, R. Kempter, and B. Lindner, Stochastic synaptic dynamics under learning, PRX Life **4**, 023006 (2026).

[15] D. H. Perkel, G. L. Gerstein, and G. P. Moore, Neuronal spike trains and stochastic point processes: I. The single spike train, Biophys. J. **7**, 391 (1967).

[16] M. N. Shadlen and W. T. Newsome, The variable discharge of cortical neurons: Implications for connectivity, computation, and information coding, J. Neurosci. **18**, 3870 (1998).

[17] F. Rieke, D. Warland, R. (d) R. Van Steveninck, and W. Bialek, *Spikes: Exploring the Neural Code* (MIT Press, Cambridge, 1999).

[18] P. Dayan and L. F. Abbott, *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems* (MIT Press, Cambridge, 2005).

[19] J. W. Pillow, J. Shlens, L. Paninski, A. Sher, A. M. Litke, E. Chichilnisky, and E. P. Simoncelli, Spatio-temporal correlations and visual signalling in a complete neuronal population, Nature (London) **454**, 995 (2008).

[20] C. Aghamohammadi, C. Chandrasekaran, and T. A. Engel, A doubly stochastic renewal framework for partitioning spiking variability, Nat. Commun. **16**, 8656 (2025).

[21] R. Pang, Balanced state of networks of winner-take-all units, PLoS Comput. Biol. **21**, e1013081 (2025).

[22] P. Lennie, The cost of cortical computation, Curr. Biol. **13**, 493 (2003).

[23] M. Frerking, J. Schulte, S. Wiebe, and U. Staubli, Spike timing in CA3 pyramidal cells during behavior: Implications for synaptic transmission, J. Neurophysiol. **94**, 1528 (2005).

[24] K. Mizuseki, S. Royer, K. Diba, and G. Buzsáki, Activity dynamics and behavioral correlates of CA3 and CA1 hippocampal pyramidal neurons, Hippocampus **22**, 1659 (2012).

[25] K. C. Bittner, A. D. Milstein, C. Grienberger, S. Romani, and J. C. Magee, Behavioral time scale synaptic plasticity underlies CA1 place fields, Science **357**, 1033 (2017).

[26] A. D. Milstein, Y. Li, K. C. Bittner, C. Grienberger, I. Soltesz, J. C. Magee, and S. Romani, Bidirectional synaptic plasticity rapidly modifies hippocampal representations, eLife **10**, e73046 (2021).

[27] C. Grienberger and J. C. Magee, Entorhinal cortex directs learning-related changes in CA1 representations, Nature (London) **611**, 554 (2022).

[28] A. D. Madar, A. D. Milstein, T. J. O'Dell, A. Jain, C. Clopath, and M. E. Sheffield, Behavioral timescale synaptic plasticity: A burst in the field of learning and memory, J. Neurosci. **45**, e1332252025 (2025).

[29] Y. Wu and W. Maass, A simple model for behavioral time scale synaptic plasticity (BTSP) provides content addressable memory with binary synapses and one-shot learning, Nat. Commun. **16**, 342 (2025).

[30] H. L. Li and M. C. Van Rossum, Energy efficient synaptic plasticity, eLife **9**, e50804 (2020).

[31] R. Pang, J. Arbelaiz, and J. W. Pillow, Learning neural dynamics through instructive signals, bioRxiv (2025).

[32] C. E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. **27**, 379 (1948).

[33] A. Morrison, M. Diesmann, and W. Gerstner, Phenomenological models of synaptic plasticity based on spike timing, Biol. Cybern. **98**, 459 (2008).

[34] N. Caporale and Y. Dan, Spike timing–dependent plasticity: A Hebbian learning rule, Annu. Rev. Neurosci. **31**, 25 (2008).

[35] H. Z. Shouval, S. S.-H. Wang, and G. M. Wittenberg, Spike timing dependent plasticity: A consequence of more fundamental learning rules, Front. Comput. Neurosci. **4**, 19 (2010).

[36] T. P. Vogels, H. Sprekeler, F. Zenke, C. Clopath, and W. Gerstner, Inhibitory plasticity balances excitation and inhibition in sensory pathways and memory networks, Science **334**, 1569 (2011).

[37] B. Confavreux, Z. P. Harrington, M. Kania, P. Ramesh, A. N. Krouglova, P. A. Bozelos, J. H. Macke, A. M. Saxe, P. J. Gonçalves, and T. P. Vogels, Memory by a thousand rules: Automated discovery of multi-type plasticity rules reveals variety & degeneracy at the heart of learning, bioRxiv (2025).

[38] See Supplemental Material at http://link.aps.org/supplemental/10.1103/m8bw-9fzg for the appendices and supplemental figures, which includes Refs. [74–76].

[39] H. K. Titley, N. Brunel, and C. Hansel, Toward a neurocentric view of learning, Neuron **95**, 19 (2017).

[40] R. Pang and A. L. Fairhall, Fast and flexible sequence induction in spiking neural networks via rapid excitability changes, eLife **8**, e44324 (2019).

[41] D. Vandael, C. Borges-Merjane, X. Zhang, and P. Jonas, Short-term plasticity at hippocampal mossy fiber synapses is induced by natural activity patterns and associated with vesicle pool engram formation, Neuron **107**, 509 (2020).

[42] D. Vandael and P. Jonas, Structure, biophysics, and circuit function of a "giant" cortical presynaptic terminal, Science **383**, eadg6757 (2024).

[43] J. J. Hopfield, Neural networks and physical systems with emergent collective computational abilities. Proc. Natl. Acad. Sci. USA **79**, 2554 (1982).

[44] N. Brunel, Is cortical connectivity optimized for storing information? Nat. Neurosci. **19**, 749 (2016).

[45] J. O'Keefe and J. Dostrovsky, The hippocampus as a spatial map: Preliminary evidence from unit activity in the freely-moving rat, Brain research **34**, 171 (1971).

[46] M. Bauer, W. Bialek, C. Goddard, C. M. Holmes, K. Krishnamurthy, S. E. Palmer, R. Pang, D. J. Schwab, and L. Susman, Optimization and variability can coexist, arXiv:2505.23398.

[47] D. Attwell and S. B. Laughlin, An energy budget for signaling in the grey matter of the brain, J. Cereb. Blood Flow Metab. **21**, 1133 (2001).

[48] F. Mery and T. J. Kawecki, A cost of long-term memory in Drosophila, Science **308**, 1148 (2005).

[49] C. Howarth, P. Gleeson, and D. Attwell, Updated energy budgets for neural computation in the neocortex and cerebellum, J. Cereb. Blood Flow Metab. **32**, 1222 (2012).

[50] P.-Y. Plaçais and T. Preat, To favor survival under food shortage, the brain disables costly memory, Science **339**, 440 (2013).

[51] A. Pache and M. C. van Rossum, Energetically efficient learning in neuronal networks, Curr. Opin. Neurobiol. **83**, 102779 (2023).

[52] K. Xiao, Y. Li, B. J. Sullivan, G. Li, and J. C. Magee, Rapid neocortical network modifications via dendritic plateau potential induced plasticity, bioRxiv (2025).

[53] P. Y. Li and A. Roxin, Rapid memory encoding in a recurrent network model with behavioral time scale synaptic plasticity, PLoS Comput. Biol. **19**, e1011139 (2023).

[54] D. Kudithipudi, C. Schuman, C. M. Vineyard, T. Pandit, C. Merkel, R. Kubendran, J. B. Aimone, G. Orchard, C. Mayr, R. Benosman, *et al.*, Neuromorphic computing at scale, Nature (London) **637**, 801 (2025).

[55] P. J. Sjöström, G. G. Turrigiano, and S. B. Nelson, Rate, timing, and cooperativity jointly determine cortical synaptic plasticity, Neuron **32**, 1149 (2001).

[56] J.-P. Pfister and W. Gerstner, Triplets of spikes in a model of spike timing-dependent plasticity, J. Neurosci. **26**, 9673 (2006).

[57] L. Buesing and W. Maass, A spiking neuron as information bottleneck, Neural Comput. **22**, 1961 (2010).

[58] S. E. Palmer, O. Marre, M. J. Berry, and W. Bialek, Predictive information in a sensory population, Proc. Natl. Acad. Sci. USA **112**, 6908 (2015).

[59] A. A. Koulakov, T. Hromádka, and A. M. Zador, Correlated connectivity and the distribution of firing rates in the neocortex, J. Neurosci. **29**, 3685 (2009).

[60] J. O'Keefe, Place units in the hippocampus of the freely moving rat, Exper. Neurol. **51**, 78 (1976).

[61] E. H. Nieh, M. Schottdorf, N. W. Freeman, R. J. Low, S. Lewallen, S. A. Koay, L. Pinto, J. L. Gauthier, C. D. Brody, and D. W. Tank, Geometry of abstract learned knowledge in the hippocampus, Nature (London) **595**, 80 (2021).

[62] M. F. Carr, S. P. Jadhav, and L. M. Frank, Hippocampal replay in the awake state: A potential substrate for memory consolidation and retrieval, Nat. Neurosci. **14**, 147 (2011).

[63] S. Fusi, P. J. Drew, and L. F. Abbott, Cascade models of synaptically stored memories, Neuron **45**, 599 (2005).

[64] M. K. Benna and S. Fusi, Computational principles of synaptic memory consolidation, Nat. Neurosci. **19**, 1697 (2016).

[65] A. L. Fairhall, G. D. Lewen, W. Bialek, and R. R. de Ruyter van Steveninck, Efficiency and ambiguity in an adaptive neural code, Nature (London) **412**, 787 (2001).

[66] G. Chechik, Spike-timing-dependent plasticity and relevant mutual information maximization, Neural Comput. **15**, 1481 (2003).

[67] T. Toyoizumi, J.-P. Pfister, K. Aihara, and W. Gerstner, Spike-timing-dependent plasticity and mutual information maximization for a spiking neuron model, Adv. Neural Inf. Proc. Syst. **17** (2004).

[68] T. Toyoizumi, J.-P. Pfister, K. Aihara, and W. Gerstner, Generalized Bienenstock–Cooper–Munro rule for spiking neurons that maximizes information transmission, Proc. Natl. Acad. Sci. USA **102**, 5239 (2005).

[69] T. M. Bartol Jr., C. Bromer, J. Kinney, M. A. Chirillo, J. N. Bourne, K. M. Harris, and T. J. Sejnowski, Nanoconnectomic upper bound on the variability of synaptic plasticity, eLife **4**, e10778 (2015).

[70] H. Ramsauer, B. Schäfl, J. Lehner, P. Seidl, M. Widrich, T. Adler, L. Gruber, M. Holzleitner, M. Pavlović, G. K. Sandve, *et al.*, Hopfield networks is all you need, arXiv:2008.02217.

[71] D. Krotov, A new frontier for Hopfield networks, Nat. Rev. Phys. **5**, 366 (2023).

[72] R. Pang and S. Recanatesi, A non-Hebbian code for episodic memory, Sci. Adv. **11**, eado4112 (2025).

[73] https://github.com/rkp8000/original_synapse.

[74] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, *et al.*, SciPy 1.0: Fundamental algorithms for scientific computing in Python, Nat. Meth. **17**, 261 (2020).

[75] A. Treves and S. Panzeri, The upward bias in measures of information derived from limited data samples, Neural Comput. **7**, 399 (1995).

[76] W. Bialek, *Biophysics: Searching for Principles* (Princeton University Press, Princeton, 2012).