Causal Spike Timing-Dependent Plasticity Prevents Assembly Fusion in Recurrent Networks

Xinruo Yang [*]
Department of Neurobiology, University of Chicago, Chicago, Illinois 60637, USA
and Grossman Center for Quantitative Biology and Human Behavior, University of Chicago, Chicago, Illinois 60637, USA

Brent Doiron [†]
Departments of Neurobiology and Statistics, University of Chicago, Illinois 60637, USA
and Grossman Center for Quantitative Biology and Human Behavior, University of Chicago, Chicago, Illinois 60637, USA

(Received 16 January 2025; accepted 5 December 2025; published 27 January 2026)

The organization of neurons into functionally related assemblies is a fundamental feature of cortical networks, yet our understanding of how these assemblies maintain distinct identities while sharing members remains limited. Here we analyze how spike timing-dependent plasticity (STDP) shapes the formation and stability of overlapping neuronal assemblies in recurrently coupled networks of spiking neuron models. Using numerical simulations and an associated mean-field theory, we demonstrate that the temporal structure of the STDP rule, specifically its degree of causality, critically determines whether assemblies that share neurons maintain segregation or merge together after training is completed. We find that causal STDP rules, where potentiation or depression occurs strictly when presynaptic spikes precede or follow postsynaptic spikes, allow assemblies to remain distinct even with substantial overlap in membership. This stability arises because causal STDP effectively cancels the symmetric correlations introduced by common inputs from shared neurons. In contrast, acausal STDP rules lead to assembly fusion when overlap exceeds a critical threshold, due to unchecked growth of common input correlations. Our results provide theoretical insight into how spike timing-dependent learning rules can support distributed representations where individual neurons participate in multiple assemblies while maintaining functional specificity.

DOI: 10.1103/9nv2-y63v

I. INTRODUCTION

The synaptic connections between neurons are not arbitrarily distributed, but rather they reflect an organization that supports brain network function. The pioneering work of Donald Hebb [1] proposed that a population of neurons that repeatedly or persistently fire together will form a strongly interconnected group, often termed a neuronal assembly [2,3]. The strong recurrent connectivity within an assembly acts to reinforce neuronal response, buffering against the inherent unreliable nature of single neuronal activity [4]. This is one reason why neuronal assemblies serve as essential building blocks of many theories of neural computation [2,5,6] and associative memory [7,8]. Since the mid-2000s, significant experimental evidence has surfaced to support the existence of assembly structure in cortical networks [9–12]. In particular, in rodent visual cortex there is clear evidence of stronger and more probable connections between neurons with similar orientation tuning [13–15], indicating that assembly structures and neuron functions are related. A longstanding goal in neuroscience is to identify the synaptic learning mechanisms that underlie assembly formation.

Many early modeling studies focused on firing rate–based plasticity rules for synaptic wiring [16–18]. Later experimental studies revealed that the spike time correlations between pre- and postsynaptic neurons over fine timescales (tens of milliseconds) also play an important role in synaptic learning [19–23]. These recordings of spike timing-dependent plasticity (STDP) uncovered a temporally causal rule for synaptic learning (often termed Hebbian), where if a presynaptic neuron spikes before a postsynaptic neuron spikes, then the synapse is potentiated (strengthened); otherwise, the synapse is depressed (weakened). Theoretical work first explored a natural consequence of this causal rule and established that feedforward chains of synaptically wired neurons would easily develop [24–27]. However, a feedforward chain is an architecture that is quite distinct from a recurrently coupled neuronal assembly.

Later theoretical work considered recurrently coupled networks with causal STDP learning. When pairs of neurons are considered then if the STDP rule is potentiation dominated then strong reciprocal connections can develop [28]. In large networks, strong recurrently coupled assemblies can be trained even with depression-dominated STDP, due to synchronous common inputs to pairs of neurons within an assembly driving temporally selective synaptic potentiation [29]. Associated mean-field theory shows how assembly structure can be self-reinforcing, so that trained synaptic structures are stable post training [29–31]. These studies show

[*]Contact author: xinruoyang@uchicago.edu
[†]Contact author: bdoiron@uchicago.edu

how temporally precise STDP can be used to not only embed feedforward wiring but also can support the formation of recurrently coupled neuronal assemblies.

However, many of these studies only consider assembly learning with no overlap in membership—meaning that a neuron is a part of only one assembly. This assumption is at odds with the clear evidence that individual neurons participate in the representation of multiple distinct stimuli and exhibit a mixed or varied selectivity in their responses [32–35]. Thus, if each stimulus has an associated assembly and neurons are tuned for multiple stimuli, then it is expected that assemblies will share members. This overlap is essential for networks to be able to store a large number of distinct stimuli in its recurrent synaptic coupling [36,37]. Understanding how STDP affects stable assembly formation when there is significant overlap is an understudied problem and much remains to be explored.

The temporal selectivity of STDP is quite diverse, with many biological components controlling how pre- and postsynaptic spike timing translate into potentiation and depression [38,39]. Indeed, there is a broad class of STDP learning rules, with some being of a classic causal form, while others show a symmetric rule where the ordering of pre- and postsynaptic spike times is unimportant and only their temporal coincidence matters [40]. Historically, causal STDP have been the focus of many theoretical studies [24,26,28–30,41]. However, recently Manz and colleagues [31] studied the formation of stable assemblies with membership overlap in networks with acausal (symmetric) excitatory STDP. They showed how assemblies could remain segregated with limited assembly overlap. Our study explores how the degree of causality in STDP affects the tolerance for assembly overlap. Through the derivation of self-consistent mean-field equations for weight dynamics, we show how networks with causal or asymmetric STDP are far more tolerant of assembly overlap than networks with acausal or symmetric STDP. In this way, our work then extends the functionality of causal STDP to beyond simply allowing feedforward structure to develop, also greatly enhancing the assembly capacity in recurrent networks.

II. RESULTS

A. Model framework

1. Neuronal dynamics

We consider a network of N spiking neuron models, with $N_E = 4N/5$ of the neurons being excitatory (E) and the remaining $N_I = N/5$ of them being inhibitory (I). Let $\gamma_E = N_E/N = 4/5$ and $\gamma_I = N_I/N = 1/5$ denote the fractions of excitatory and inhibitory neurons, respectively. The membrane potential of neuron i in population $\alpha \in \{E, I\}$, denoted as V_α^i, obeys an exponential integrate-and-fire model formalism [42]:

$$\tau_\alpha \frac{dV_\alpha^i}{dt} = \left(E_L - V_\alpha^i\right) + \Delta_T \exp\left(\frac{V_\alpha^i - V_T}{\Delta_T}\right) + I_\alpha^i(t) + \sum_{\beta=E,I}\sum_{j=1}^{N_\beta} W_{\alpha\beta}^{ij}(t)\cdot\left(J_{\text{syn}} * y_\beta^j(t)\right). \quad (1)$$

The passive properties of the neuron are given by the leak reversal potential E_L and membrane time constant τ_α. Action potential initiation is modeled by the exponential term in Eq. (1), where parameter V_T sets the membrane value where the action potential upstroke is approximately triggered, and Δ_T is a scaling parameter. Spike and reset dynamics are given by the discontinuity $V_\alpha^i(t_\alpha^{ik}) = V_s \to V_\alpha^i([t_\alpha^{ik}]_+) = V_r < V_s$, where t_α^{ik} is the kth time neuron i from population α crosses the spike threshold V_s and is subsequently reset to V_r and held there for a refractory period τ_{ref} [Fig. 1(c)]. The spike train for neuron i in population α is given by $y_\alpha^i(t) = \sum_k \delta(t - t_\alpha^{ik})$. The third term in Eq. (1), $I_\alpha^i(t)$, denotes a background external input to neuron i, which is modeled by an associated diffusion approximation [43–47]:

$$I_\alpha^i(t) = \mu_\alpha^{\text{ext}} + \sigma_\alpha^{\text{ext}}\xi_\alpha^i(t). \quad (2)$$

Here $\xi_\alpha^i(t)$ is a Gaussian white noise stochastic process characterized by expectation $\langle\xi_\alpha^i(t)\rangle = 0$ and correlation function $\langle\xi_\alpha^i(t)\xi_\beta^j(t')\rangle = \delta_{\alpha\beta}\delta^{ij}\delta(t-t')$, where δ^{ij} and $\delta_{\alpha\beta}$ are Kronecker delta functions ($\delta_{\alpha\beta} = 1 \leftrightarrow \alpha = \beta$; $\delta^{ij} = 1 \leftrightarrow i = j$) and $\delta(t-t')$ is a Dirac delta function. Finally, the fourth term in Eq. (1) is the recurrent input from the other neurons in the network, where $J_{\text{syn}}(t) = \frac{1}{\tau_{\text{syn}}}\exp(-t/\tau_{\text{syn}})H(t)$ is an exponential synaptic filter describing the dynamics of the postsynaptic current, with $*$ denoting temporal convolution and $H(t)$ being the Heaviside function [$H(t) = 1$ for $t > 0$ and $H(t) = 0$ for $t \leqslant 0$]. The coefficient $W_{\alpha\beta}^{ij}$ is the synaptic strength from neuron j in population β to neuron i in population α.

All neuronal parameters used in our study are given in Table I–IV (see Appendix C). A glossary of neuronal terms used in our network model is given in Table V.

2. Excitatory synaptic weight dynamics

In this network we consider the well-studied phenomenological model of STDP, where the synapse strength $W_{\alpha\beta}^{ij}(t)$ changes depending on the difference between pre- and postsynaptic spike times [29,30,44,48,49]. We only consider the $E \to E$ and the $I \to E$ synapses as plastic.

For each pair of excitatory pre- and postsynaptic spike times ($t_{\text{pre}} = t^{jk}, t_{\text{post}} = t^{il}$) with time lag $s = t_{\text{post}} - t_{\text{pre}}$, where t^{jk} indicates the time of kth spike of neuron j, the synapse evolves as:

$$W_{EE}^{ij} \longrightarrow W_{EE}^{ij} + L(s).$$

Here $L(s)$ is the STDP learning rule. We omit the subscript $\alpha\beta$ notation in W^{ij} for the convenience of presentation. Pairwise STDP rules of this form are inherently unstable, often leading to unbounded weight growth or decay [29,30]. To ensure stability, we therefore constrain the synaptic weights to the range $[0, W_{EE}^{\max}]$.

We focus on two distinct $E \to E$ learning rules. The first [Fig. 2(c)] is a temporally causal $E \to E$ rule (sometimes referred to as Hebbian), $L^{\text{c}}(s)$, which is defined as follows:

$$L^{\text{c}}(s) = \begin{cases} f_+^{\text{c}}\exp\left(-\frac{s}{\tau_+^{\text{c}}}\right), & \text{if } s > 0. \\ -f_-^{\text{c}}\exp\left(\frac{s}{\tau_-^{\text{c}}}\right), & \text{if } s \leqslant 0. \end{cases}$$

The parameters $\tau_+^{\text{c}} > 0$ and $\tau_-^{\text{c}} > 0$ represent the timescales of the forward and backward decay of the causal STDP rule,

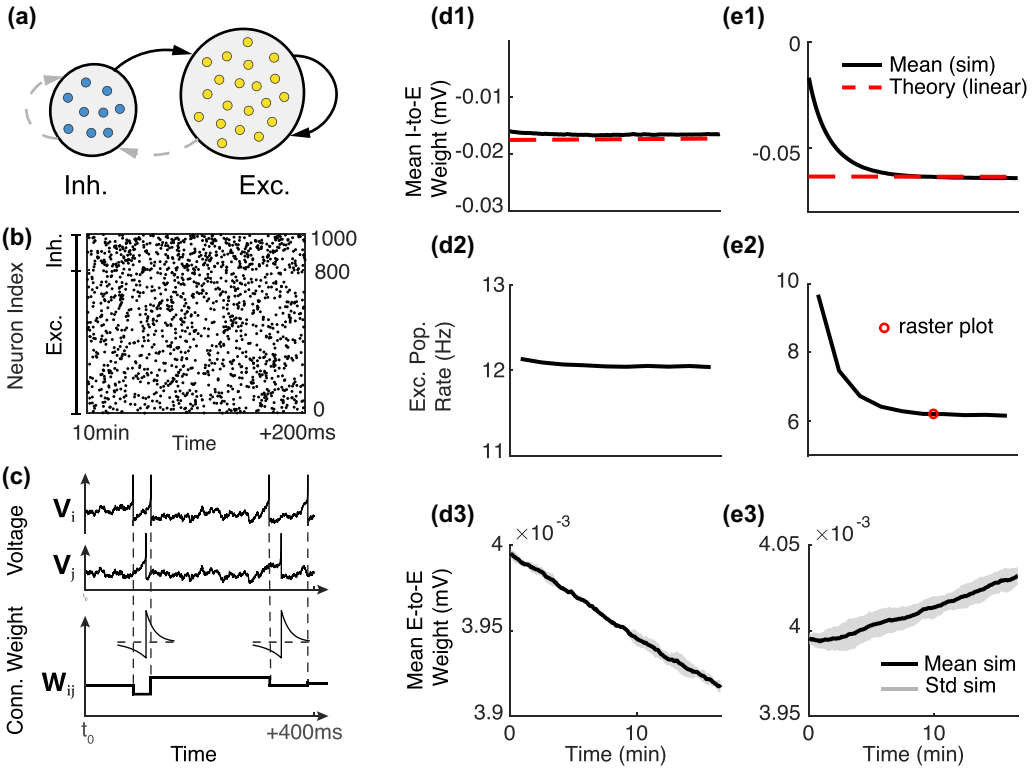

FIG. 1. Illustration of the firing rate dependency of synaptic weight dynamics. (a) Schematic of a recurrently coupled network of spiking neuron models, composed of excitatory (E) and inhibitory (I) populations. The solid black lines indicate plastic synapses and the dashed gray lines indicate static (nonplastic) connections. (b) A representative raster plot showing asynchronous spiking activity from a network simulation where the target excitatory firing rate was 6 Hz. (c) Synaptic weight evolution under the pairwise STDP rule; W_{ij} is updated based on the relative timing of presynaptic (neuron j) and postsynaptic (neuron i) spike times. [(d) and (e)] Comparison of network dynamics for two different target E neuron firing rates: $r_E^{\text{target}} = 12$ Hz [(d1), (d2), (d3)] and $r_E^{\text{target}} = 6$ Hz [(e1), (e2), and (e3)]. [(d1) and (e1)] Evolution of the mean I-to-E synaptic weight: $W_{EI} = (N_E N_I)^{-1} \sum_{ij} W_{EI}^{ij}$. Simulation results (solid lines) are well predicted by our linear theory (dashed red lines). [(d2) and (e2)] The mean E population firing rate successfully converges to its respective target value, demonstrating effective homeostatic control. [(d3) and (e3)] Evolution of the mean E-to-E synaptic weight: $W_{EE} = N_E^{-2} \sum_{ij} W_{EE}^{ij}$. In this example we use a more realistic STDP rule $L^{\text{c}}(s)$ with $\tau_+^{\text{c}} = 15$ ms $< \tau_-^{\text{c}} = 30$ ms, $S_0 = -1.36 \times 10^{-6}$ mV, $f_-^{\text{c}} = 10^{-4}$ mV and $W_{EE}^{\max} = 0.06$ mV. For the inhibitory plasticity rule, we have $\tau^{\text{h}} = 30$ ms, $f^{\text{h}} = -6 \times 10^{-4}$ mV, and $W_{EI}^{\min} = -0.48$ mV. Other STDP parameters f_-^{c} and d^{h} could be solved consistently.

while $f_+^{\text{c}} > 0$ and $f_-^{\text{c}} > 0$ denote the maximal amplitudes of each synaptic change, respectively. This rule is termed causal because presynaptic spikes that precede postsynaptic spikes ($s > 0$) result in synaptic potentiation [$L^{\text{c}}(s) > 0$], whereas if they follow postsynaptic spikes ($s < 0$), then synaptic depression occurs [$L^{\text{c}}(s) < 0$].

The second learning rule, $L^{\text{ac}}(s)$, is temporally acausal and obeys:

$$L^{\text{ac}}(s) = \begin{cases} f_+^{\text{ac}} \exp\left(-\frac{s}{\tau_+^{\text{ac}}}\right) \cdot (T_+ - s), & \text{if } s > 0 \\ f_-^{\text{ac}} \exp\left(\frac{s}{\tau_-^{\text{ac}}}\right) \cdot (T_- + s), & \text{if } s \leqslant 0. \end{cases}$$

Here the parameters $f_+^{\text{ac}} > 0$, $f_-^{\text{ac}} > 0$, $\tau_+^{\text{ac}} > 0$, and $\tau_-^{\text{ac}} > 0$ serve the same roles as in $L^{\text{c}}(s)$, and the $(T_\pm \mp s)$ terms biases the plasticity to depression for large $|s|$. The learning rule is acausal because the relative timing of the pre- and postsynaptic spike times is less important, and synaptic potentiation occurs for small $|s|$ and depression for larger $|s|$.

In this study, we operate in a weak-coupling regime where $W_{\alpha\beta}^{ij}$ scales as $1/N \equiv \epsilon$. We assume that the integral of the STDP function $L^{\text{c/ac}}(s)$ is negative, and its magnitude is small compared to its L_1 norm:

$$0 > \int L^{\text{c/ac}}(s) ds / ||L^{\text{c/ac}}(s)||_1 \sim \mathcal{O}(\epsilon).$$

This condition implies that the potentiation and depression components of $L^{\text{c/ac}}(s)$ are nearly balanced, such that they roughly cancel each other out upon integration. This assumption ensures that the plasticity will not be dominated by chance spike coincidences of pre- and postsynaptic neurons [8,29], which would otherwise lead to trivial dynamics characterized by uniform potentiation or depression. Furthermore, the small negative integral of $L^{\text{c/ac}}(s)$ introduces a depression-dominated drift that counteracts the tendency for weights to grow in the absence of training due to positive spike-train cross-correlations, thereby preventing runaway potentiation in spontaneous conditions.

Note that the overall scaling of $L(s)$ defines the synaptic update size ΔW_{ij} and thus determines the learning rate. The weight dynamics depend linearly on $L(s)$, given the fixed pre- and postsynaptic spike trains. This scaling can then be adjusted to control the learning speed without altering the qualitative outcome. Here we assume the scaling is

comparable to that of the weights, i.e., $L(s) \sim W \sim \mathcal{O}(\epsilon)$. When we derive the mean-field equations (Sec. II D), we use this fact to introduce a normalized function $L(s) \to L(s)/\epsilon$ which simplifies the notation in the mean-field equations. A glossary of synaptic terms used in our network model is given in Table V.

3. Inhibitory synaptic weight dynamics

Additionally, we assume the connections from $I \to E$ are plastic. Their plasticity is described by the homeostatic learning rule $L^{\mathrm{h}}(s)$ [50,51]:

$$L^{\mathrm{h}}(s) = f^{\mathrm{h}} \exp\left(-\frac{|s|}{\tau^{\mathrm{h}}}\right), \tag{3}$$

with $f^{\mathrm{h}} < 0$ represents the maximal amplitude change of the inhibitory synapse corresponding to pre- and postsynaptic spikes with time lag s. In addition to the synaptic update corresponding to spike pairs, each presynaptic (inhibitory) spike drives depression of the inhibitory synapses by $d^{\mathrm{h}} > 0$ [52,53]:

$$W_{EI}^{ij} \to W_{EI}^{ij} + d^{\mathrm{h}}.$$

We note that for $I \to E$ connections we have $W_{EI}^{ij} < 0$ so that adding $d^{\mathrm{h}} > 0$ leads to a weakening of the synapse. Analogous to the excitatory synapses, we bound the inhibitory weights to the range $[W_{EI}^{\min}, 0]$.

All plasticity parameters used in our study are given in Table II (see Appendix C).

B. The fate of assembly structure

The base network connectivity in our model is unstructured before training. Specifically, the adjacency matrix (existence of a connection) for the connectivity from presynaptic population β to postsynaptic population α is generated via the simple rule:

$$G_{\alpha\beta}^{ij} \sim \text{Bernoulli}(p_{\alpha\beta}). \tag{4}$$

Here $p_{\alpha\beta}$ is the connection probability from a neuron in population β to a neuron in population α. Further, we initialize the synaptic weights to be homogeneous within a connection class:

$$W_{\alpha\beta}^{ij}(t=0) = \epsilon \cdot G_{\alpha\beta}^{ij} J_{\alpha\beta}. \tag{5}$$

The combination of Eqs. (4) and (5) is such that there is no initial assembly structure prewired within the network.

1. The impact of firing rates on synaptic dynamics

To begin our analysis we first consider synaptic dynamics in the absence of any structured external input given to the network, so that the E and I neuron spiking activity is only weakly correlated through recurrent interactions [Figs. 1(a) and 1(b)]. A change in synaptic strength occurs via the spike timing relationships between pre- and postsynaptic neurons and how they relate to the STDP rule $L(s)$ [Fig. 1(c)]. Without external correlations we do not predict assemblies to emerge, and rather statistically homogeneous synaptic weight dynamics are expected [30].

The inhibitory rule $L^{\mathrm{h}}(s)$ [Eq. (3)] is designed to homeostatically regulate E network activity [50]. Indeed, the firing rates of E neurons, $r_E^i = \langle y_E^i(t) \rangle$, under the inhibitory STDP rule obey $r_E^i \to r_E^{\text{target}}$ over time, where r_E^{target} is a desired "target" firing rate. For our form of $L^{\mathrm{h}}(s)$ we have that $r_E^{\text{target}} = -d^{\mathrm{h}}/(2f^{\mathrm{h}}\tau^{\mathrm{h}})$, with $r_E^{\text{target}} > 0$ being ensured by $d^{\mathrm{h}} > 0$ and $f^{\mathrm{h}} < 0$. To illustrate, we show simulations for two different target rates [Figs. 1(d1)–1(d3) with $r_E^i \to r_E^{\text{target}} = 12$ Hz and Figs. 1(e1)–1(e3) with $r_E^i \to r_E^{\text{target}} = 6$ Hz]. Homeostasis is accomplished by $L^{\mathrm{h}}(s)$ adjusting the $I \to E$ weights [Figs. 1(d1) and 1(e1)], causing the E population to fire at the correct target rate [Figs. 1(d2) and 1(e2)]. This mechanism is further explained in the Sec. IV C. We note that other forms of inhibitory plasticity have also been proposed to ensure network stability. For instance, recent work [54,55] has shown that symmetric inhibitory plasticity can strengthen feedback inhibition onto highly active excitatory neurons, providing an alternative stabilization mechanism.

We next explore how r_E^{target}, and hence inhibitory STDP, affects the evolution of the recurrent excitatory weights W_{EE}^{ij}. The pair-based learning rule $L(s)$ produces a weight change ΔW_{EE}^{ij} over an interval $(t, t+T)$ following [41]:

$$\Delta W_{EE}^{ij} = \int_t^{t+T} L(s)\left[y_E^i(s) y_E^j(t+s)\right] ds. \tag{6}$$

If we take the weight update to be small (i.e., $f_+, f_- \ll \mathcal{O}(\epsilon)$), then $|\Delta W_{EE}^{ij}| \ll |W_{EE}^{ij}|$ and the evolution of W_{EE}^{ij} occurs on a slow timescale, typically minutes to hours. On this slow timescale, and allowing $\langle y_E^i(t) \rangle = r_E^i \to r_E^{\text{target}}$, the mean synaptic weight $W_{EE} = N_E^{-2} \sum_{ij} W_{EE}^{ij}$ obeys the following expression [30,41]:

$$\frac{dW_{EE}}{dt} = p_{EE}\left[r_E^{\text{target}}\right]^2 S_0 + \text{effects due to spiketime correlations}, \tag{7}$$

where

$$S_0 = \int_{-\infty}^{\infty} L(s) ds.$$

The first term models how chance coincidences in pre- and postsynaptic neuron spiking activity affects the evolution of W_{EE}, with the likelihood of chance coincident spiking at lag s being $[r_E^{\text{target}}]^2$ and the probability of a connection from a pre- to postsynaptic E neurons being p_{EE}. The coefficient S_0 is the net effect of the plasticity rule $L(s)$, considering both synaptic potentiation $[L(s) > 0]$ and depression $[L(s) < 0]$ over all time lags s. The second term considers temporally structured spiketime correlations, and we defer its treatment to later sections.

To ensure that networks without structured correlated spiking activity do not spuriously grow synaptic connections, we take throughout this study that $S_0 < 0$ [i.e., a depression-dominated $L(s)$]. For this reason, we intentionally employ an STDP function $L(s)$ that departs from a perfectly antisymmetric (odd) or symmetric (even) form, which is incidentally the more biologically realistic scenario [38–40]. Since $S_0 < 0$, then for a large r_E^{target} we see that the mean excitatory weight, W_{EE}, slowly depresses owing to a dominate $p_{EE}[r_E^{\text{target}}]^2 S_0 < 0$ in Eq. (7) [Fig. 1(d3)]. By contrast, for a smaller r_E^{target} then the influence of the weak (but nonzero) spiketime correlations throughout the network can reverse the baseline tendency for

synapses to depress, and rather the mean W_{EE} potentiates [Fig. 1(e3)]. We will give a complete theory for how spike-time correlations affect the plasticity of synaptic weights in later sections. This section simply illustrates how inhibitory plasticity determines (through r_E^{target}) the baseline depression of W_{EE} when $S_0 < 0$.

Finally, we remark that under these spontaneous conditions, we do not observe the emergence of significant structure, such as the feedforward motifs reported in other work [24–27]. Instead, weights tend to potentiate or depress based on their initial conditions relative to a separatrix [29,30]. We confirmed this qualitative long-term behavior in Appendix B, where a modified parameter set was used to show that weights saturate at their boundaries on a computationally tractable timescale.

2. Training assembly structure

To embed assembly structure in the network we will follow [29,30] and use a common dynamic "training" signal given to subgroups of neurons with the aim of strengthening within group synaptic connectivity. Throughout our study we consider two distinct training signals, $s_A(t)$ and $s_B(t)$, and assign E neurons to receive signal A, signal B, or both signals. We accomplish this by augmenting the external input during training to be

$$I_E^i(t) = \mu_E^{\text{ext}} + \sigma_E \xi_E^i(t) + M_A^i s_A(t) + M_B^i s_B(t).$$

Here μ_E^{ext} and $\sigma_E \xi_E^i(t)$ set the background state as defined in Eq. (2). The two signals $s_A(t)$ and $s_B(t)$ are also modeled as white noise fluctuations yet are a common source of fluctuations across all neurons that are members of an intended assembly. During training, if the ith E neuron is intended to be a member of assembly A (or B), then M_A^i (or M_B^i) is set to 1.

The central question of our study is as follow: What is the fate of assembly structure after training has finished? Before training, connection strengths are small and not organized into any group structure [Fig. 2(a), left]. During training, the STDP rules will promote strong within group connectivity, leaving between group connectivity weak [Fig. 2(a), middle]. After training has stopped, there are three possible fates for assembly structure [Fig. 2(a), right]. First, the assemblies may dissolve and all connections will depress to match the strengths before training. Second, the assemblies will remain segregated, with the within assembly connectivity persisting at high levels, while the between assembly connections remain weak. Third, the assemblies may fuse together and all connections, within and between, potentiate to high levels. The segregated case represents stable memory formation as the structure embedded by the training phase persists post training. The faded or fused assemblies represent failed learning, since the end state of the network shows no evidence of the training inputs.

To investigate assembly fate we will consider two control parameters of our network. The first is the degree of assembly overlap [Fig. 2(b)]. When assigning neurons to assembly membership a fraction of neurons will have both $M_A^i = M_B^i = 1$. Let the number of E neurons that are members of both assemblies be N_O, and each assembly have N_X neurons that are exclusive members (we assume equally sized assemblies).

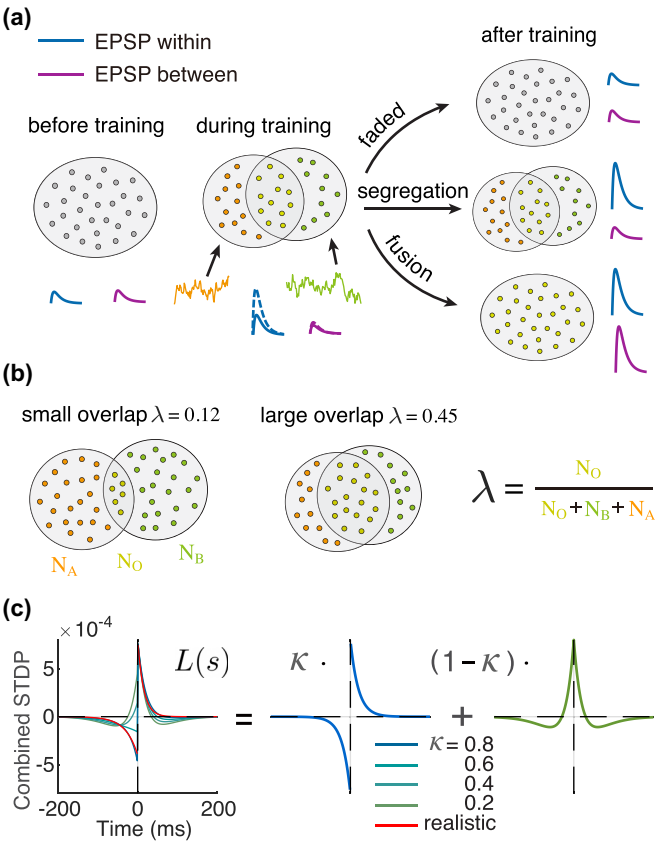

FIG. 2. (a) Illustration of three possible assembly fates after training. We create two assemblies with overlap using two uncorrelated signals, and there are three possible dynamics after training. (1) The assembly structure fades away and all synaptic connections depress to their original pretraining values. (2) The assemblies remain segregated with strong within assembly connections and weak between assembly connections. (3) The assemblies fuse into one large assembly with strong connections throughout. (b) Definition of the overlap ratio λ. (c) Illustration of the $E \to E$ STDP rule with causality parameter κ. For comparison, the red curve models a more biologically realistic STDP rule where the potentiation time constant is shorter than the depression time constant ($\tau_+^c < \tau_-^c$). This realistic case is closely approximated by our model with $\kappa = 0.8$ (blue curve).

With each neuron assigned to at least one assembly we have $N_E = 2N_X + N_O$. The overlap parameter is then:

$$\lambda = \frac{N_O}{N},$$

and λ ranges from no overlap ($\lambda = 0$) to complete overlap ($\lambda = \gamma_E = 0.8$, where γ_E is the fraction of excitatory neurons). We consider a symmetric configuration where the two assemblies are identical: They comprise the same number of neurons ($N_X + N_O$), share the same initial weight strengths, and are driven by training signals of equal magnitude. This symmetry allows us to simplify the analysis by focusing on the weight dynamics of a single, representative assembly.

The second parameter κ controls the degree of causality in the STDP learning rule [Fig. 2(c)]. Specifically, $E \to E$ plasticity obeys the following:

$$L(s) = \kappa L^{\text{c}}(s) + (1 - \kappa) L^{\text{ac}}(s).$$

For the chosen STDP parameters $\{f_+^{ac/c}, f_-^{ac/c}, \tau_+^{ac/c}, \tau_-^{ac/c}\}$, when $\kappa = 0$ the rule is roughly acausal, while for $\kappa = 1$ the rule is roughly causal. While the causal $L^c(s)$ and acausal $L^{ac}(s)$ synaptic plasticity rules have very distinct temporal shapes (close to odd or even functions, respectively), they are both constructed to possess the same small, negative integral [i.e., $S_0 \equiv \int L^c(s)ds = \int L^{ac}(s)ds < 0$]. This ensures that the net integral of our combined STDP function, $L(s)$, is independent of the causality parameter κ, allowing for a fair comparison across different degrees of causality (Table II). Consequently, any change in weight dynamics when varying κ can be attributed purely to how the temporal symmetry of the learning rule interacts with spike-time correlations, rather than to a shift in the overall balance of potentiation and depression. Although our study utilizes specific STDP functions, this framework provides general insights. Since any learning rule can be decomposed into its even and odd components, our approach of varying κ systematically explores how the degree of temporal symmetry—a fundamental property of any STDP rule—governs assembly stability. For instance, the classic asymmetric STDP curve observed in experiments [22] is well approximated by our model with $\kappa \approx 0.8$ [Fig. 2(c)].

It is natural to expect that as the overlap parameter λ grows the possibility of assembly fusion increases, since for large λ the assemblies will share many members and consequently interact more strongly. Thus, by varying λ we will have an expected influence on assembly fate. However, it is less obvious how assembly fusion versus segregation will be affected by the degree of causality κ in the STDP rule $L(s)$. Classically, causality in $L(s)$ promotes unidirectional connections [49,56,57]: If neuron j tends to lead neuron i in spiking ($s > 0$), then potentiation in W^{ij} occurs, but this necessarily causes any backward connection W^{ji} to depress since i must lag j ($s < 0$). This feature of causal STDP is ideal to prevent assembly fusion, since unwanted reciprocal connections between assemblies will be suppressed. However, this same logic is at odds with the formation of strong reciprocal connections within an assembly in the first place. The central goal of our study is to characterize the boundary between assembly segregation and assembly fusion defined in (κ, λ) parameter space. Experimental evidence favors the decaying time constant for potentiation being smaller than depression, making the STDP function $L(s)$ deviate from an almost odd form $L^c(s)$. However, our linear combination of causal-acausal rule gives very good approximation for such an STDP function $L(s)$ [as is shown in Fig. 2(c), red curve].

C. Numerical exploration of assembly fusion or segregation

To begin, we divide our network into four sub-populations defined by the training phase [Fig. 3(a)]. Neurons that receive only stimulus A or B are identified as population A or B, respectively. The group of neurons that receive both stimuli is identified as the overlap population O. Finally, the inhibitory neurons (which do not receive a training signal) form the last population I. We separate E neurons in the overlap from the remaining neurons in the assemblies since their evolution is distinct from that of nonoverlap E neurons. After sufficient training [Fig. 3(b)] the E network organizes itself into clearly demarcated A, B, and O populations [Fig. 3(c)].

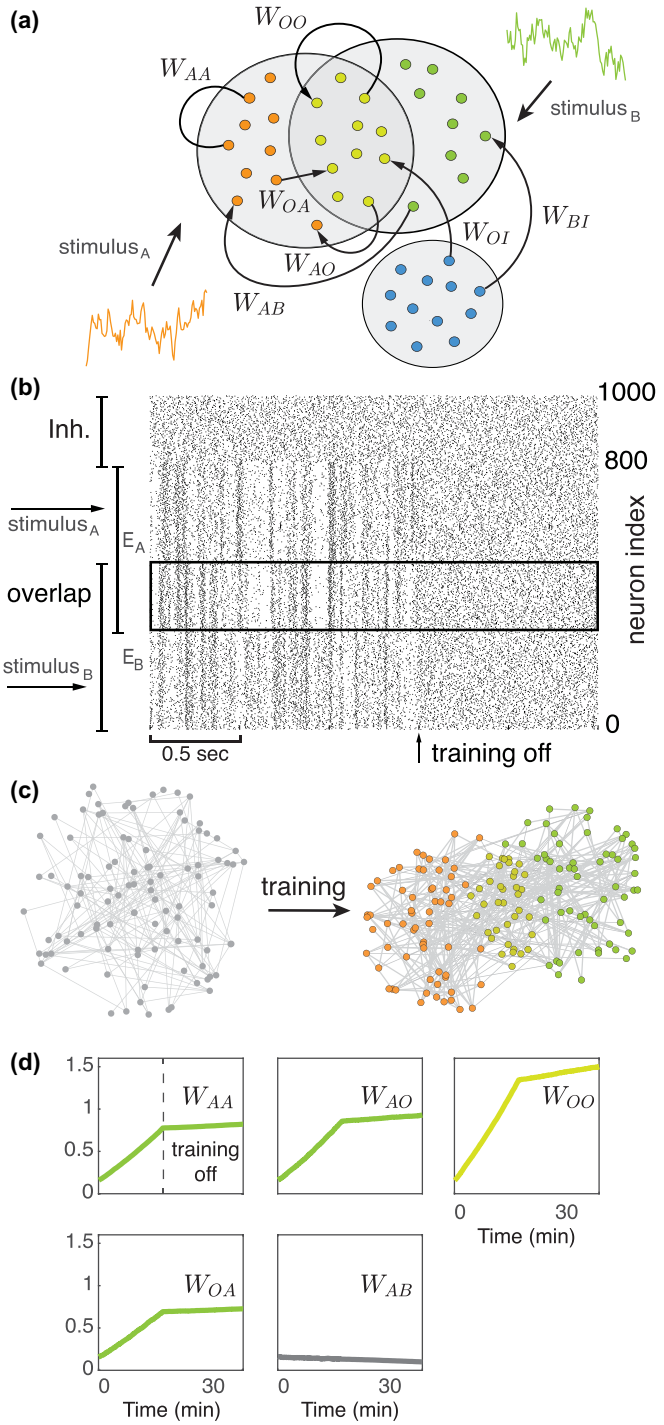

FIG. 3. Illustration of the network model and training protocol. (a) We consider an $E - I$ network model with two assemblies (A and B) with overlap (O) in the excitatory population due to common training signals. The set $\{W_{AA}, W_{AB}, W_{AO}, W_{OA}, W_{AI}\}$ of the key plastic weights are marked. (b) A raster plot of spike times of all neurons during and after training. (c) Visualization of the connectivity between a subset of the excitatory neurons. (d) Dynamics of the mean averaged weights during and after training. Unless otherwise specified, parameters are as listed in Table II.

The network dissection into $\{A, B, O, I\}$ prompts us to define normalized mean variables for the dynamic weights in the

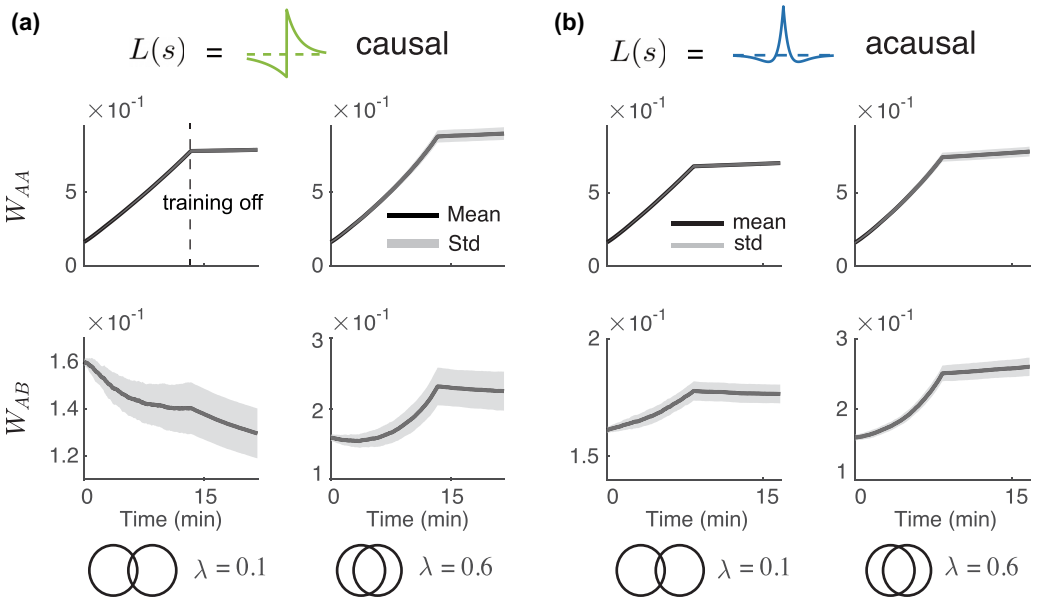

FIG. 4. The dynamics of the mean weights between neuron pairs in the same assembly, W_{AA}, and neuron pairs from different assemblies, W_{AB}, during and after training for small overlap $\lambda = 0.1$ and large overlap $\lambda = 0.6$. (a) For the causal STDP rule ($\kappa = 1$), the assemblies remain segregated post training (W_{AA} increases while W_{AB} decreases) for both small and large overlap. (b) For the acausal STDP rule ($\kappa = 0$) the assemblies fuse (both W_{AA} and W_{AB} increase) for large overlap. The solid black line plots the mean weight evolution averaged over 20 independent noise realizations, while the shaded gray area indicates the standard deviation. Due to the significant odd component in the causal STDP rule ($\tau_+^c = 15$ ms $< \tau_-^c = 30$ ms), this network required a stronger training protocol. It was trained with a signal of $\sigma_{\text{train}} = 1.5$ mV for 800 s. In contrast, the network with the acausal rule was trained using a weaker signal ($\sigma_{\text{train}} = 0.8$ mV) for 500 s. Other simulation parameters can be found in Table II.

network:

$$W_{\alpha\beta} = \frac{1}{N_\alpha N_\beta} \sum_{i\in\alpha, j\in\beta} \frac{W_{\alpha\beta}^{ij}}{\epsilon} \sim \mathcal{O}(1), \tag{8}$$

where $\alpha, \beta \in \{A, B, O, I\}$. Because of the symmetry between the A and B populations only the mean weight vector $\mathbf{W_m} = [W_{AA}, W_{AB}, W_{AO}, W_{OA}, W_{OO}]$ need be considered [Fig. 3(b)]. During training the recurrent synapses within the assemblies show rapid growth [Fig. 3(d); W_{AA}, W_{OA}, W_{AO}, W_{OO}], while the between assembly weights remain small [Fig. 3(d); W_{AB}]. The rapid synaptic growth for within assembly coupling is due to the strong correlating force of the training signals. After training the spike train correlations become much smaller, and consequently the evolution of the synaptic weights slows significantly. However, the basic structure of a strong within (W_{AA}) and weak between (W_{AB}) assembly connectivity persists. This is the key signature of stable assembly segregation.

These initial results prompt us to investigate the (κ, λ) dependence of the dynamics of $(W_{AA}(t), W_{AB}(t))$ post training. To begin, we consider the fully causal STDP learning rule ($\kappa = 1$). For both small ($\lambda = 0.1$) and large ($\lambda = 0.6$) assembly overlap the mean weight dynamics post training preserve the assembly structure embedded during training [Fig. 4(a)]. This is clear since for both small and large λ the mean within assembly synaptic weight W_{AA} continues to grow after training has ceased, while the mean between assembly weight W_{AB} continues to decay. By contrast, a fully acausal STDP learning rule ($\kappa = 0$) and large λ lead to assembly fusion [Fig. 4(b)]. These numerical results show that the temporal structure of the learning rule $L(s)$ determines the assembly fate post training. To deepen our understanding of this dependence we next develop a mean-field theory for the mean synaptic weights $\mathbf{W_m}$.

D. Mean-field theory of synaptic weight dynamics

Following past work [28–30,41,58], we separate the slow timescale of weight dynamics from the fast timescale of the cross-covariance of neuron spike trains to derive a set of self-consistent differential equations for the slow weight evolution.

We rewrite Eq. (6) that gives the update of $\Delta W_{\alpha\beta}^{ij}$ over an interval $(t, t+T)$ as follows [41]:

$$\Delta W_{\alpha\beta}^{ij} = \int_t^{t+T} L(s)\big[y_\alpha^i(s) y_\beta^j(t+s)\big] ds, \tag{9}$$

where $y_\alpha^i(t)$ is the spike train of neuron i in population α.

Small weight updates, $|\Delta W_{\alpha\beta}^{ij}| \ll |W_{\alpha\beta}^{ij}|$, ensure that the evolution of $W_{\alpha\beta}^{ij}$ occurs on a slow timescale, so that the weight dynamics obey [41]:

$$\frac{dW_{\alpha\beta}^{ij}}{dt} = \int_{-\infty}^{\infty} L(s)\big[r_\alpha^i r_\beta^j + C_{\alpha\beta}^{ij}(s)\big] ds. \tag{10}$$

Here we have expanded upon Eq. (7) and included the cross-covariance function between spike trains $y_\alpha^i(t)$ and $y_\beta^j(t)$ given by $C_{\alpha\beta}^{ij}(s) = \langle (y_\alpha^i(t) - r_\alpha^i)(y_\beta^j(t+s) - r_\beta^j)\rangle$. The function $C_{\alpha\beta}^{ij}(s)$ describes the mean corrected expectation of neuron j spiking at a time interval s after ($s > 0$) or before ($s < 0$) a spike from neuron i. Equation (10) describes how the slow dynamics of weight dynamics depend on the statistics of fast-timescale spiking activity.

On the fast timescale of neuronal spiking, we consider the recurrent synaptic connectivity $\mathbf{W}$ to be fixed; $\mathbf{W}$ is the full $N \times N$ connectivity matrix. To determine spike train correlations from network connectivity, we employ linear response theory in the weakly correlated firing regime [29,30,59–61]. In this regime, the cross-covariance is computed in the frequency domain as:

$$\widehat{\mathbf{C}} = (\mathbf{I} - \widehat{\mathbf{W}} \circ \widehat{\mathbf{K}})^{-1} \widehat{\mathbf{C}_0} (\mathbf{I} - \widehat{\mathbf{W}}^* \circ \widehat{\mathbf{K}}^*)^{-1},$$

where $\mathbf{C}$ is the cross-covariance matrix with element $C^{ij}(s)$ and $\mathbf{C}_0$ is the autocovariance matrix of neurons in the absence of coupling, and here we use their Fourier transforms $\widehat{\mathbf{C}}(\omega)$ and $\widehat{\mathbf{C}_0}(\omega)$. The kernel $\mathbf{K}$ describes neural response properties corresponding to the input spikes from presynaptic neurons, which could be computed as $K_{ij}(s) = \chi^i * J_{\text{syn}}(s)$. The neuronal response kernel $\chi^i(s)$ describes how the neural response changes due to a change or modulation in the input current, and $J_{\text{syn}}(s)$ is the synaptic filter [62]. Finally, $\mathbf{I}$ is the identity matrix and $\widehat{\mathbf{K}}^*$ and $\widehat{\mathbf{W}}^*$ denotes the complex conjugate of corresponding matrices $\widehat{\mathbf{K}}$ and $\widehat{\mathbf{W}}$, respectively. A complete derivation is given in Sec. IV A.

In our parameter regime of weak connections, the spectral radius ρ of the effective interaction matrix satisfies $\rho(\mathbf{W} \circ \mathbf{K}) < 1$. This allows us to expand the inverse matrix (and its complex conjugate) as a Neumann series:

$$(\mathbf{I} - \mathbf{W} \circ \mathbf{K})^{-1} = \sum_{k=0}^{\infty} (\mathbf{W} \circ \mathbf{K})^k. \tag{11}$$

Physically, this expansion decomposes the inverse matrix into a sum over all possible signaling paths (motifs) of length k between neurons. Inserting this expansion into Eq. (10) and grouping terms by connectivity motifs yields the mean-field equations for the population-averaged weights:

$$\frac{dW_{\alpha\beta}}{dt} = \sum_{\text{motif}} S^{\text{motif}}_{\alpha\beta} q^{\text{motif}}_{\alpha\beta}(\mathbf{W_m}). \tag{12}$$

Each term represents the contribution of a specific motif, factorized into a structural component and a temporal component. The polynomial $q^{\text{motif}}_{\alpha\beta}(\mathbf{W_m})$ characterizes the collective strength of the weights involved in the motif. With the exception of the zeroth-order firing rate term, these polynomials typically scale as $\mathcal{O}(\epsilon)$, reflecting the influence of weak synaptic weights on cross-covariance (e.g., for a direct feedback motif $\alpha \to \beta$, $q^b_{\alpha\beta} = \epsilon W_{\beta\alpha}$).

The coefficient $S^{\text{motif}}_{\alpha\beta}$ quantifies the net effect of the plasticity rule on that motif. To facilitate the mean-field analysis, we apply a consistent normalization scheme. Since the weight updates scale linearly with the learning rule, we introduce a renormalized learning rule $L(s) \to L(s)/\epsilon$. This ensures that $L(s) \sim \mathcal{O}(1)$, consistent with our normalized mean weights $W_{\alpha\beta} \sim \mathcal{O}(1)$. Consequently, the motif coefficient is defined compactly as:

$$S^{\text{motif}}_{\alpha\beta} = \int_{-\infty}^{\infty} L(s) M^{\text{motif}}_{\alpha\beta}(s) ds. \tag{13}$$

This definition ensures that the coefficient itself remains of order $\mathcal{O}(1)$. The kernel $M^{\text{motif}}_{\alpha\beta}(s)$ represents the temporal structure of the correlations induced by the motif. The motif kernels depend on the cellular response K_α and the autocovariance C^0_α (see Sec. IV A). The homeostatic rule for inhibitory plasticity forces the excitatory neurons to fire at the target firing rate $r_E = r_E^{\text{target}}$ [29,50]. This implies that the response kernels $K_\alpha \approx K_E$ and the autocovariance $C^0_\alpha \approx C^0_E$ are approximately uniform across all excitatory neurons, irrespective of their assembly association. Thus, the motif kernels $M^{\text{motif}}_{\alpha\beta}(s)$ do not depend on population indices $\alpha, \beta \in \{A, B, O\}$ and can be described by a single kernel for the excitatory population. A glossary of terms used in our mean field theory is given in Table VI.

E. Circuit motif derivation of the evolution of synaptic weights

To derive a closed form dynamical system for $\mathbf{W_m}$ we only consider up to the second-order motifs in Eq. (12).

1. Zeroth-order motif

The zeroth-order motif is defined by $q^0_{EE} = p_0 r_E^2$ and $M^0 = 1$. Thus, the coefficient S^0 is determined only by the integration of the STDP curve: $S^0 = \int L(s) ds$ (as discussed earlier). This term captures the synaptic plasticity owing to chance and unstructured pre- and postsynaptic spiking activity. Since the higher-order terms involve synaptic connections then they scale with $\epsilon = 1/N \ll 1$. Without any constraints on $L(s)$ the zeroth-order term $p_0 r^2 S^0 \sim \mathcal{O}(1)$ [since $p_0 r^2 \sim \mathcal{O}(1)$] and would dominate the weight dynamics in Eq. (12) so that $W_{AB} \approx p_0 r^2 \epsilon^{-1} S^0 t$ [30]. This would lead to uninteresting synaptic dynamics, where only rapid assembly fusion or assembly decay are possible solutions. As a consequence, throughout our work we enforce the following constraint on $L(s)$:

$$p_0 r^2 S^0 \sim \mathcal{O}(\epsilon) \Longrightarrow S^0 = \int_{-\infty}^{\infty} L(s) ds \sim \mathcal{O}(\epsilon). \tag{14}$$

This permits higher-order terms in Eq. (12) to contribute to the growth or decay of the synaptic weights. This mathematical constraint has a clear biological interpretation: For synaptic weights to remain stable, the overall tendency for potentiation (strengthening) must be almost perfectly offset by the tendency for depression (weakening).

2. First-order motifs

The first-order motifs include direct feedforward (f) connections $\beta \to \alpha$ and feedback (b) paths $\alpha \to \beta$ with corresponding response kernels $M^{\text{motif=f}}(s)$ and $M^{\text{motif=b}}(s)$ [denoted as $M^{\text{f}}(s)$ and $M^{\text{b}}(s)$ for simplicity]:

$$\begin{aligned} q^{\text{f}}_{\alpha\beta} &= \epsilon W_{\alpha\beta}; \quad M^{\text{f}}(s) = K_E * C^0_E(s), \\ q^{\text{b}}_{\alpha\beta} &= \epsilon p_0 W_{\beta\alpha}; \quad M^{\text{b}}(s) = C^0_E * K^-_E(s). \end{aligned} \tag{15}$$

The negative superscript indicates reversed time $K^-(s) = K(-s)$ coming from the inverse Fourier transform of the complex conjugate. The p_0 in $q^{\text{b}}_{\alpha\beta}$ captures the probability of a backwards connection from populations β to α. The forwards and backwards first-order coefficients S^{f}_E and S^{b}_E are then

given by:

$$S^{\mathrm{f}} = \int_{-\infty}^{\infty} L(s)\big(K_E * C_E^0\big)(s)ds,$$
$$S^{\mathrm{b}} = \int_{-\infty}^{\infty} L(s)(C_E^0 * K_E^-)(s)ds. \tag{16}$$

3. Second-order motifs

The second-order motifs include common inputs (cc) $\alpha \leftarrow \upsilon \rightarrow \beta$ from another population υ and feedforward chains (fc) $\beta \rightarrow \upsilon \rightarrow \alpha$ and backward chains (bc) $\alpha \rightarrow \upsilon \rightarrow \beta$ through an intermediary population υ. The corresponding response kernels are respectively:

$$M_\upsilon^{\mathrm{cc}}(s) = \big(K_E * C_\upsilon^0 * K_E^-\big)(s),$$
$$M_\upsilon^{\mathrm{fc}}(s) = \big(K_E * K_\upsilon * C_E^0\big)(s),$$
$$M_\upsilon^{\mathrm{bc}}(s) = \big(C_E^0 * K_\upsilon^- * K_E^-\big)(s).$$

We distinguish the kernels by the intermediate population $\upsilon \in \{E, I\}$. This is because while the homeostatic inhibitory plasticity makes it that excitatory populations $\alpha, \beta \in \{A, B, O\}$ all have the same response properties, those of the inhibitory and excitatory populations may differ [i.e., in general we have $K_E(s) \neq K_I(s)$ and $C_E^0(s) \neq C_I^0(s)$]. From these kernels, we compute the second-order motif coefficients as:

$$S_\upsilon^{\mathrm{cc}} = \int_{-\infty}^{\infty} L(s)\big(K_E * C_\upsilon^0 * K_E^-\big)(s)ds,$$
$$S_\upsilon^{\mathrm{fc}} = \int_{-\infty}^{\infty} L(s)\big(K_E * K_\upsilon * C_E^0\big)(s)ds, \tag{17}$$
$$S_\upsilon^{\mathrm{bc}} = \int_{-\infty}^{\infty} L(s)\big(C_E^0 * K_\upsilon^- * K_E^-\big)(s)ds.$$

To simplify the analysis of the mean-field equations, we further assume $W_{OA} \approx W_{AO}$. While this property is not strictly guaranteed by symmetry, as populations A and O may receive inputs with distinct correlation structures (e.g., through different values of W_{AA} and W_{OO}), this approximation is nonetheless justified. Through the analysis of the full system (Sec. IV B), we show that the divergence between W_{OA} and W_{AO} remains bounded, constrained by the initial difference and the excitatory weight cap $W_{EE}^{\max}$. Furthermore, our numerical simulations confirm the validity of this approximation. With this assumption, the coefficients for forward chain and backward chains always appear as a sum $S^{\mathrm{fb}} = S^{\mathrm{fc}} + S^{\mathrm{bc}}$ in our mean-field equations, and we use S^{fb} for simplicity of exposition. Consequently, in what follows, we will consider the reduced four-dimensional equations of $\mathbf{W_m} = [W_{AA}, W_{AB}, W_{AO}, W_{OO}]$.

Substituting these simplified coefficients, the full expressions for the second-order motifs are as follows:

$$\sum_{\mathrm{cc,fb}} S_{\alpha\beta}^{\mathrm{motif}} q_{\alpha\beta}^{\mathrm{motif}}(\mathbf{W_m})$$
$$= \epsilon \cdot p_0(S^{\mathrm{cc}} + S^{\mathrm{fb}})\left[\left(\frac{\gamma_E - \lambda}{2}\right)F_{\alpha\beta}(\mathbf{W_m}) + \lambda W_{\alpha O} W_{O\alpha}\right]$$
$$+ \epsilon \cdot p_0\gamma_I\big(S_I^{\mathrm{cc}} W_{\alpha I} W_{\beta I} + S_I^{\mathrm{fb}} W_{\alpha I} W_{IE}\big). \tag{18}$$

The second-order terms are distinguished by the function $F_{\alpha\beta}(\mathbf{W_m})$ which captures the two pathway common inputs to the postsynaptic population α that originate in the presynaptic population β:

$$F_{AA}(\mathbf{W_m}) = W_{AA}^2 + W_{AB}^2;$$
$$A \rightarrow A \rightarrow A \text{ and } A \rightarrow B \rightarrow A,$$
$$F_{AB}(\mathbf{W_m}) = 2W_{AA}W_{AB};$$
$$B \rightarrow A \rightarrow A \ \ (\times 2),$$
$$F_{AO}(\mathbf{W_m}) = W_{AA}W_{AO} + W_{AB}W_{AO};$$
$$O \rightarrow A \rightarrow A \text{ and } O \rightarrow B \rightarrow A,$$
$$F_{OO}(\mathbf{W_m}) = 2W_{AO}^2;$$
$$O \rightarrow A \rightarrow O \ \ (\times 2).$$

In the above we have assumed symmetry in the weight evolution with $W_{BA} = W_{AB}$, $W_{OB} = W_{OA}$, and $W_{BO} = W_{AO}$. We remark that the term $\lambda W_{\alpha O} W_{O\alpha}$ results from the $\alpha \rightarrow O \rightarrow \beta$ and $\beta \rightarrow O \rightarrow \alpha$ pathways and uses the above-mentioned symmetries.

The final term in Eq. (18) captures how correlated activity induced by common inhibition drives $W_{\alpha\beta}$ dynamics. In principle, we should have other two dynamical weights W_{AI} and W_{OI} and analyze a full six-dimensional mean field. However, as in Ref. [29], we take the timescale of homeostatic inhibition to be much faster than that of excitatory learning ($1/\epsilon$). This allows inhibition to adiabatically track the excitatory weights and maintain control of the excitatory firing rate r_E. We can write $W_{AI} = W_{BI}$ by symmetry and W_{OI} in terms of the excitatory weights and the steady state firing rates r_E and r_I:

$$W_{AI} = \mu_A^{\mathrm{diff}} - \frac{r_E}{r_I\gamma_I}\left[\left(\frac{\gamma_E - \lambda}{2}\right)(W_{AA} + W_{AB}) + \lambda W_{AO}\right],$$
$$W_{OI} = \mu_O^{\mathrm{diff}} - \frac{r_E}{r_I\gamma_I}[(\gamma_E - \lambda)W_{AO} + \lambda W_{OO}]. \tag{19}$$

Here $\mu_\alpha^{\mathrm{diff}}$ is a constant bias that determines the inhibitory weight strength needed for neuron population α to maintain its target firing rate r_α^{diff}, as specified by the inhibitory plasticity rule, when receiving only external background input r_α^{ext} without excitatory input. A detailed derivation of inhibitory connection strength $W_{\alpha I}$ is provided in Sec. IV C.

4. Self-consistent mean-field theory for $\mathbf{W_m}$

Truncating Eq. (12) at second order and inserting Eqs. (14), (15), and (18) gives the evolution equation for the generic weight $W_{\alpha\beta}$ ($\alpha, \beta \in \{A, B, O\}$) as:

$$\frac{dW_{\alpha\beta}}{dt} = p_0 r_E^2 S^0 + \epsilon \cdot (S^{\mathrm{f}} W_{\alpha\beta} + p_0 S^{\mathrm{b}} \cdot W_{\beta\alpha})$$
$$+ \epsilon \cdot p_0(S^{\mathrm{cc}} + S^{\mathrm{fb}})$$
$$\times \left[\left(\frac{\gamma_E - \lambda}{2}\right)F_{\alpha\beta}(\mathbf{W_m}) + \lambda W_{\alpha O} W_{O\alpha}\right]$$
$$+ \epsilon \cdot p_0\gamma_I\big(S_I^{\mathrm{cc}} W_{\alpha I} W_{\beta I} + S_I^{\mathrm{fb}} W_{\alpha I} W_{IE}\big). \tag{20}$$

Recall that $S^0 \sim \mathcal{O}(\epsilon)$ [see Eq. (14)] so that, in principle, all terms on the right-hand side of Eq. (20) can contribute to the dynamics of $\mathbf{W_m}$.

Equations (19) and (20) constitute a self-consistent mean-field theory for the evolution of mean synaptic weights $\mathbf{W_m}$.

The influence of overlap parameter λ is explicit in the second-order terms in Eq. (20). However, the influence of the causality of the learning rule $L(s)$ given by κ is more subtle. The derivation of Eq. (20) is based on the separation of the fast timescale of spiking activity from the slow timescale of synaptic plasticity. As in many fast-slow systems analyses, the fast system is connected to the slow system via a temporal averaging of the fast system. This averaging is expressed in the calculations of the motif coefficients S in Eqs. (14), (16), and (17) where $L(s)$ (and hence κ) appear.

We remark that our mean-field theory ignores that the operating point of the network about which we linearize in principle depends on the weight matrix $\mathbf{W_m}$, and thus the kernel $K_E(s)$ and the autocovariance function $C_E^0(s)$ would also depend on $L(s)$. However, in practice the inhibitory plasticity ensures that the E neurons firing rates remain at their target value irrespective of $\mathbf{W}$ (Fig. 1, Sec. IV C; see also Refs. [29,50]). Because of this homeostasis the linearization point is roughly independent of the recurrent weights, and thus $K_E(s)$ and $C_E^0(s)$ are also independent of $L(s)$ [29]. This simplifies the analysis of assembly fate considerably.

In the next section we use our mean-field theory to study the (κ, λ) dependence of assembly dynamics post training.

F. How motif coefficients depend on the learning rule

We begin our analysis by considering the forward motif coefficient S^{f}. The corresponding motif kernel $M^{\mathrm{f}}(s)$ is weighted towards positive lag s, where the presynaptic neuron j fires before the postsynaptic neuron i [Fig. 5(a), top]. The causal ($\kappa = 1$) STDP rule $L(s)$ strictly potentiates for $s > 0$, so that the combination of $M^{\mathrm{f}}(s)$ and $L(s)$ leads to a large potentiation component [large yellow box in Fig. 5(a), top]. The backwards coefficient S^{b} is the same as S^{f} except that it is biased towards the spike train correlations where postsynaptic neuron i fires before presynaptic neuron j, given by the kernel $K_E^-(s)$. This occurs for $s < 0$ and results in a large depression component for the causal $L(s)$ rule [Fig. 5(a), bottom].

A similar argument applies for S^{f} and S^{b} with the acausal ($\kappa = 0$) learning rule $L(s)$ [Fig. 5(b)]. The one distinction is that both S^{f} and S^{b} show potentiation (i.e., >0) for the acausal rule. This is due to the temporal overlap of spike train correlations and the narrow potentation component of the acausal $L(s)$ for small $|s|$. However, the key distinction between assembly dynamics with causal or acausal learning occurs with the second-order terms of our theory.

Unlike the correlations through monosynaptic forward and backward connections, the correlations in the spike trains from neurons i and j that are induced by common inputs are symmetric across lag s. This is true for the direct common input term S^{cc} [Figs. 5(c) and 5(d), top] and when we combine the forward and backward chain terms $S^{\mathrm{fb}} = S^{\mathrm{fc}} + S^{\mathrm{bc}}$, the effective correlations are also symmetric [Figs. 5(c) and 5(d), bottom].

The causal STDP rule ($\kappa = 1$) is such that $L(s)$ is almost an odd function (in s). In this case, if $S^0 = \int_{-\infty}^{\infty} L(s; \kappa = 1) ds \sim \mathcal{O}(\epsilon)$, then when $L(s)$ are integrated against the even motif kernel functions $M^{\mathrm{cc}}(s) = K_E * C_E^0 * K_E^-(s)$ we obtain:

$$S^{\mathrm{cc}} = \int_{-\infty}^{\infty} M^{\mathrm{cc}}(s) L(s; \kappa = 1) ds \sim \mathcal{O}(\epsilon).$$

This assumes that the timescales of $M^{\mathrm{cc}}(s)$ are not much faster than those of $L(s)$ (i.e., $\tau_s \sim \tau_+ \sim \tau_-$) so that $\mathcal{O}(1)$ differences in $L(s)$ for $\pm s \to 0$ do not dominate S^{cc}. The same is true for the chain term S^{fb}. Consequently, the coefficients S^{cc} and S^{fb} are much smaller than S^{f} and S^{b} [very small yellow boxes in Fig. 5(c)]. These small S^{cc} coefficients imply that the dynamics of $W_{\alpha\beta}$ in Eq. (20) is effectively linear, since the quadratic terms are negligible for reasonable values of $W_{\alpha\beta}$.

By contrast, when the STDP learning rule is fully acausal ($\kappa = 0$) then $L(s)$ is an even function (over s). Consequently, when $L(s)$ is integrated against the even and narrow spike train correlations from common inputs we have:

$$S^{\mathrm{cc}} = \int_{-\infty}^{\infty} M^{\mathrm{cc}}(s) L(s; \kappa = 0) ds \sim \mathcal{O}(1).$$

This requires that $M^{\mathrm{cc}}(s)$ primarily overlaps the potentiation component of $L(s)$; if that is the case, then $S^{\mathrm{cc}} \sim \mathcal{O}(1)$ even when $S^0 \sim \mathcal{O}(\epsilon)$. The result is sizable S^{cc} and S^{fb} coefficients [purple boxes in Fig. 5(d)], especially for S^{cc}. The same argument and conclusion holds for the coefficients S_I^{cc} and S_I^{fb}. Thus, for acausal learning the quadratic terms in Eq. (20) may contribute to the overall dynamics of assembly structure. We explore this possibility in the next section.

G. Causal or acausal learning allows assembly segregation or fusion for large overlap

We have established that for the causal STDP rule the quadratic terms in Eq. (20) are negligible (due to small coefficients S^{cc} and S^{fb}). Since the overlap parameter λ only appears in the quadratic terms, then for networks with causal STDP assembly overlap is expected to have minimal impact on post-training assembly dynamics. In particular, if during training we embed segregated assemblies where $W_{AB} < W_{AA}$ and they persist post training for $\lambda = 0$, then they should persist across a broad range of $\lambda > 0$ as well. This prediction is validated through simulations of the full spiking model network, as well as the associated mean-field theory in Eqs. (19) and (20) and across a range of λ [Fig. 6(a)]. We note the excellent agreement between the trial averaged simulations of our spiking network with plastic synapses ($N = 1000$ and $p_0 = 0.1$ so approximately 1×10^5 synapses) and our five-dimensional mean-field theory in Eq. (20) [compare red and black curves in Fig. 6(a)].

By contrast to networks with causal learning, networks with acausal STDP have non-negligible coefficients for the quadratic (disynaptic) terms in the mean-field theory in Eq. (20). For small λ and if sufficient training is given so that W_{AA} is significantly larger that W_{AB} at the conclusion of training, then the assemblies will remain segregated post training [Fig. 6(b), top]. However, for larger overlap λ the common term $\lambda W_{AO} W_{OA} \approx \lambda W_{AO}^2$ in Eq. (20) can dominate the disynaptic interactions, and this term is symmetric across the evolution of both W_{AA} and W_{AB}. Consequently, this will promote growth for both within and between assembly weights, so that assembly fusion occurs [Fig. 6(b), bottom].

In sum, we see that the influence of common inputs can cause assembly fusion if the learning rule $L(s)$ is acausal and the assembly overlap is large. We explore this result in the next section.

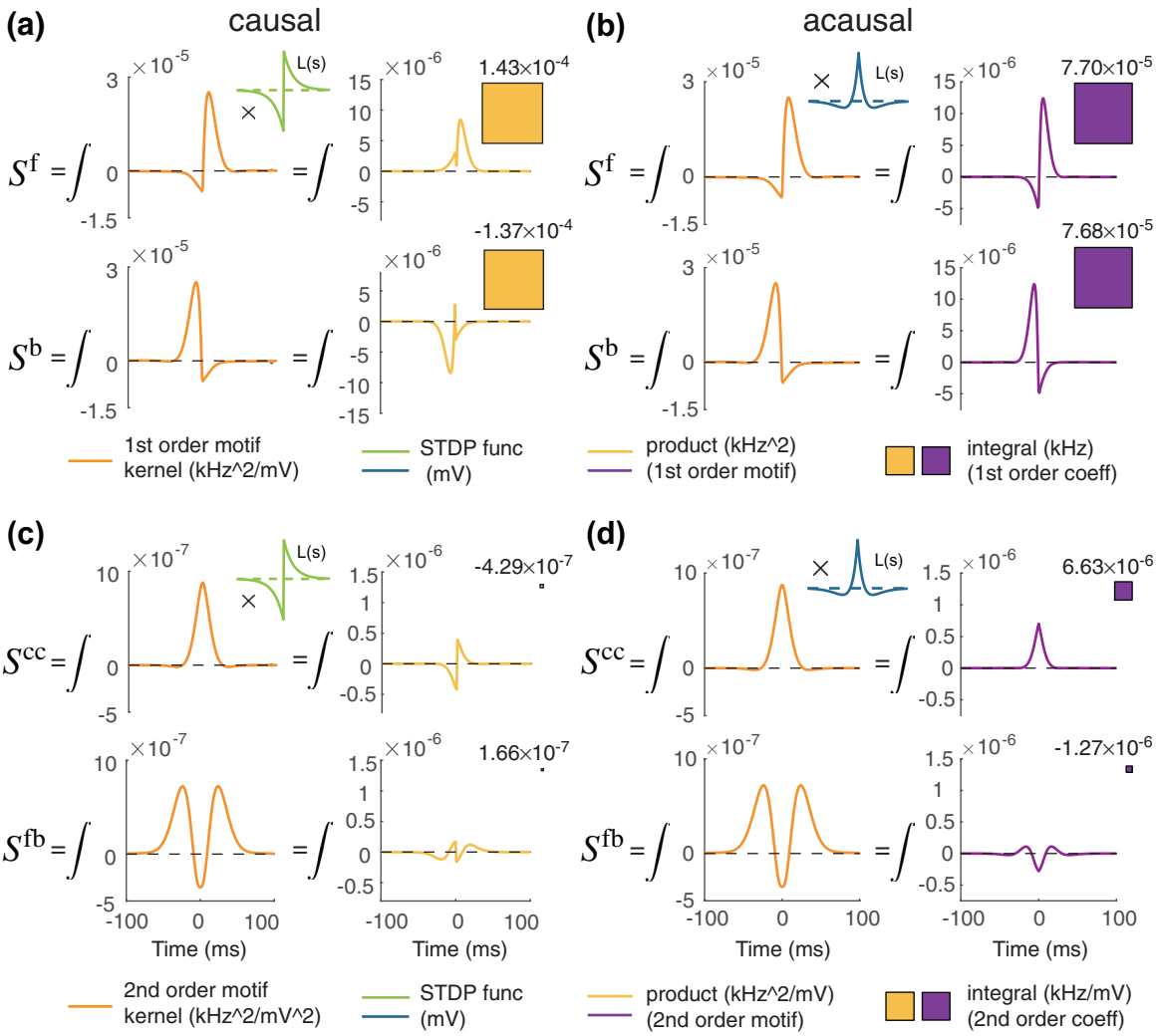

FIG. 5. Motif coefficients for assembly weight dynamics for the causal ($\kappa = 1$) and acausal ($\kappa = 0$) STDP rules. [(a) and (c)] Calculation of coefficients for causal STDP. Panel (a) displays the first-order coefficients (S^{f} and S^{b}) and (c) displays the second-order coefficients (S^{cc} and S^{fb}). In each row, the left plots show the STDP rule $L(s) = L^{\mathrm{c}}(s)$ (green) and the corresponding motif kernels (orange); the right plots show the product of these two functions, the integral of which yields the coefficient. The area of each yellow square represents the absolute magnitude of the corresponding coefficient. Note that for the causal rule, the second-order coefficients are negligible compared to the first-order coefficients because the product of the motif kernel and the STDP rule exhibits approximate odd symmetry. [(b) and (d)] Same as in (a) and (c) but for acausal STDP $L(s) = L^{\mathrm{ac}}(s)$. For the acausal rule, the second-order coefficients are comparable in magnitude to the first-order coefficients. The coefficient values appear small because they are computed on the fast timescale (milliseconds). To isolate the influence of the STDP rule's odd component, the causal rule is implemented here with symmetric time constants ($\tau_-^{\mathrm{c}} = \tau_+^{\mathrm{c}} = 20$ ms) and no synaptic weight cap. Unless otherwise specified, these STDP rules and parameters are used in all subsequent figures.

H. The boundary between assembly segregation and fusion

The fate of assembly training (segregation versus fusion) is determined from the growth or decay of the between assembly mean weight W_{AB}. To give deeper insight into the mechanics of assembly segregation or fusion we consider the approximated dynamical system where $\mathbf{W_m} = [W_{AA}, W_{AB}, W_{AO}, W_{OO}] = [W_{\max}, W_{AB}, W_{\max}, W_{\max}]$. This amounts to assuming that the within and overlap assembly weights evolve to a large (maximal) value $W_{\max}$ and we are left considering only the one-dimensional dynamics of the mean between assembly weights W_{AB}. Under this reduction we can write the following dynamics for W_{AB}:

$$\frac{1}{\epsilon}\frac{dW_{AB}}{dt} = k_0 + k_1 W_{AB} + k_2 W_{AB}^2. \tag{21}$$

Here the coefficients k_0, k_1, and k_2 are given by:

$$\begin{aligned} k_0 &= p_0 r^2 S^0/\epsilon + p_0\gamma_I S_I^{\mathrm{cc}}(D_0^2) + p_0\gamma_I S_I^{\mathrm{fb}}(D_0 W_{IE}) \\ &\quad + p_0(S^{\mathrm{cc}} + S^{\mathrm{fb}})\lambda W_{\max}^2, \\ k_1 &= (S^{\mathrm{f}} + p_0 S^{\mathrm{b}}) + p_0\gamma_I S_I^{\mathrm{cc}}(-2 D_0 D_1) \\ &\quad + p_0\gamma_I S_I^{\mathrm{fb}}(-D_1 W_{IE}) \end{aligned}$$

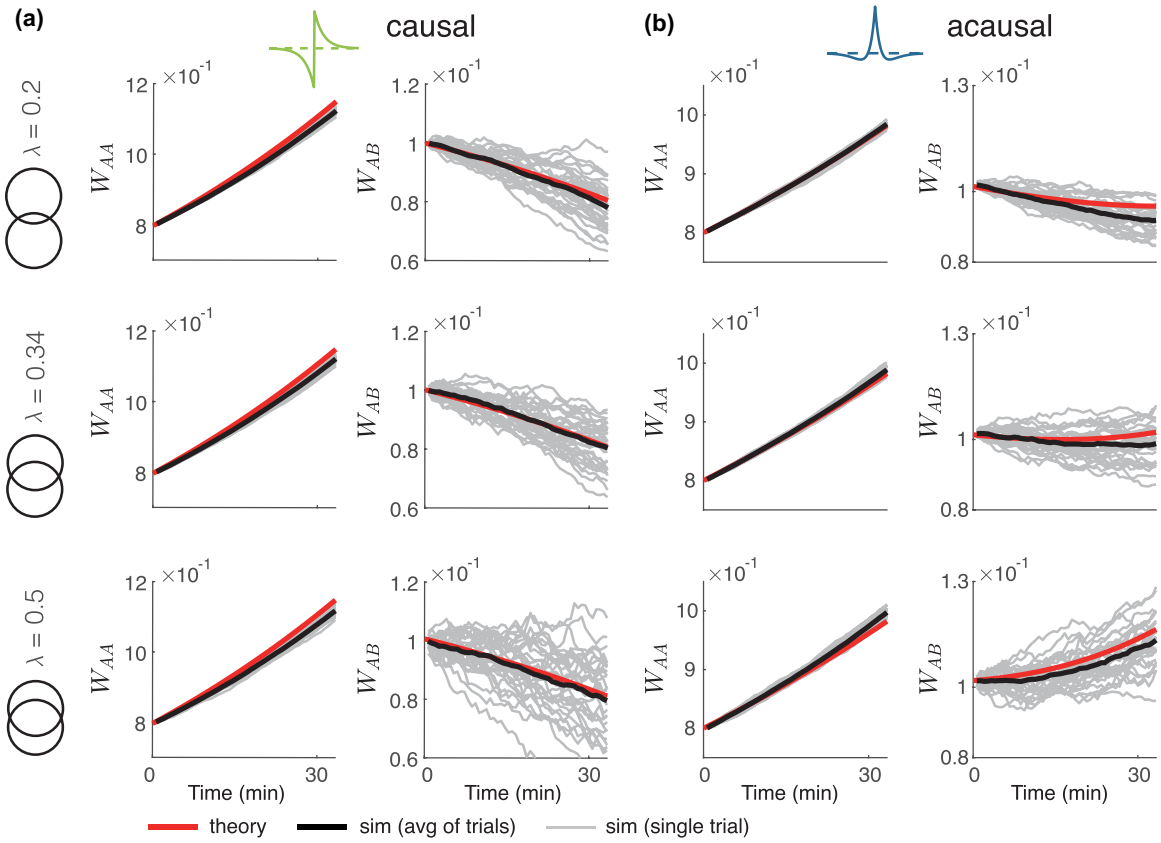

FIG. 6. Almost linear assembly weight dynamics for the causal STDP rule ($\kappa = 1$) and the nonlinear (quadratic) dynamics for acausal ($\kappa = 0$) STDP rule. (a) Comparison between theory and simulation for the within assembly weight W_{AA} and between assembly weight W_{AB} post training for causal STDP rule $L(s) = L^{c}(s)$. The gray curves are multiple realizations from the network of spiking neuron models; the black curve is the average of these realizations; and the red curve is the theoretical mean-field prediction. We perform this for three overlap λ values (top to bottom). (b) Same as (a), but for acausal STDP rule $L(s) = L^{ac}(s)$. The STDP rules and all other simulation parameters are identical to those used in Fig. 5 (see Table II for details), with the exception that synaptic weights were uncapped to facilitate a direct comparison with theory.

$$+ p_0(S^{cc} + S^{fb})(\gamma_E - \lambda)W_{max},$$

$$k_2 = p_0\gamma_I S_I^{cc} D_1^2.$$

Here we introduce two hyperparameters D_0 and D_1 for simplicity:

$$D_0 = \mu_A^{diff} - \frac{r_E^{target}(\gamma_E + \lambda)}{2r_I\gamma_I}W_{max},$$

$$D_1 = \frac{r_E^{target}(\gamma_E - \lambda)}{2r_I\gamma_I}.$$

where D_1 is a positive constant and D_0 is an intermediate variable that depends on W_{max}. Both coefficients exhibit clear dependence on the overlap ratio λ. The $\mathcal{O}(1)$ scaling of these coefficients, combined with $W_{AB} \sim \mathcal{O}(1)$, ensures that all the right-hand-side terms in Eq. (21) maintain $\mathcal{O}(1)$ scaling. We assume that when training terminates (at $t = 0$), the initial condition for the between-assembly weight is set to $W_{AB}(0) < W^{max}$.

When the STDP learning $L(s)$ is causal ($\kappa = 1$) then all second-order motif can be ignored (i.e., S^{cc}, S^{fb}, S_I^{cc}, and S_I^{fb} are all approximately 0). This makes $k_2 \approx 0$, $k_1 = S^{f} + p_0 S^{b} > 0$, and $k_0 = p_0 S^0 (r_E^{target})^2/\epsilon < 0$ [since $S^0 < 0$ for our choice of $L(s)$]. For this system there is a unstable (threshold) point $W_{AB}^t = -k_0/k_1$, where for initial condition $W_{AB}(0) > W_{AB}^t$ then $W_{AB}(t)$ grows, while for $W_{AB}(0) < W_{AB}^t$ then $W_{AB}(t)$ decays. Importantly, for $\kappa = 1$ we have that k_0 and k_1 are independent of the overlap parameter λ, and consequently so is the threshold W_{AB}^t [Fig. 7(a), green curve]. Thus, for training that drives $W_{AB}(0) < W_{AB}^t$, then after training we will have assembly segregation independent of λ [Fig. 7(b); green curves].

For acausal STDP learning [$L(s)$ with $\kappa = 0$], the coefficients k_0 and k_1 take more complex forms and $k_2 > 0$. Specifically, all coefficients k_0, k_1, and k_2 depend on the overlap parameter λ and second-order motif coefficients S^{motif} and by extension also on the causality parameter κ. Notably, the zeroth-order motif coefficient $S^0 < 0$ remains largely independent of other terms in k_0, allowing us to select parameter regimes where $k_0 < 0$. Under these conditions, the quadratic equation $k_2 W_{AB}^2 + k_1 W_{AB} + k_0 = 0$ yields two real roots: one positive and one negative. Since $k_2 > 0$, the positive root is an unstable fixed point W_{AB}^t, again establishing a threshold between potentiation and depression of W_{AB}. In our model, W_{AB}^t decreases with increasing λ, as the overlap population facilitates the growth of W_{AB} [Fig. 7(a), blue curve]. Consequently, at large overlap values λ, when $W_{AB}(0) > W_{AB}^t$,

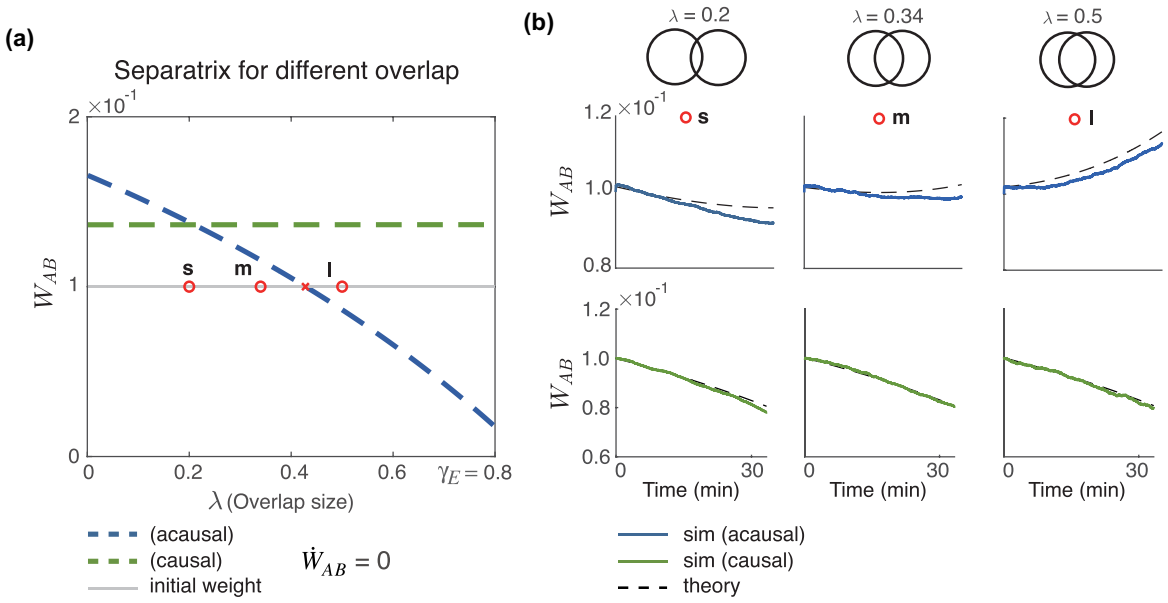

FIG. 7. One-dimensional dynamical system for the evolution of the between assembly weight W_{AB}. (a) The threshold weight W^t_{AB} that separates growth and decay of $W_{AB}(t)$ as the overlap ratio λ varies. The green dashed line is for the causal rule $L(s;\kappa = 1)$ where the threshold is independent of λ. The blue dashed line is for the acausal rule $L(s;\kappa = 0)$ where as overlap λ increases W^t_{AB} is lowered. The gray line is the initial condition $W_{AB}(0)$ immediately after training. The three red dots with label s (small), m (medium), l (large) are the three overlaps considered in Fig. 6. (b) The after training dynamics for W_{AB} in our reduced system with the three overlaps considered in panel (a). The dashed curves are from our reduced theory in Eq. (21) and the solid lines are from simulations of our full spiking network. The top row is for the acausal rule for $L(s)$, while the bottom row is for the causal rule.

assembly fusion occurs [Fig. 7(b), blue curves]. This relationship between overlap and fusion is robust across a broad range of parameters and is not specific to our parameter choice.

Generally, for κ not equal to 0 or 1, the STDP rule $L(s) = \kappa L^{\rm c}(s) + (1-\kappa)L^{\rm ac}(s)$ would be a combination of these two cases. We notice that all the S coefficients are linear in $L(s)$, and hence also linear in κ. Consequently, the mean weights dynamics are merely linear combinations of the dynamics for the weights when $\kappa = 1$ and $\kappa = 0$. Thus, for $\kappa \in (0, 1)$ we do not expect new dynamics, and expect only the threshold W^t_{AB} to be of interest. For arbitrary (κ, λ) we can compute the threshold $W^t_{AB}(\kappa, \lambda)$ and compare it to the initial condition: $W_{AB}(0) - W^t_{AB}(\kappa, \lambda)$. This gives us a boundary in (κ, λ) parameter space that separates the assembly fates of fusion and segregation (Fig. 8, red curve). We see that for large overlap λ then for sufficiently acausal $L(s)$ rules we have the fusion state. By contrast, for sufficiently causal $L(s)$ rules we will never have a fusion state even for $\lambda \to \gamma_E$ (of course for $\lambda = \gamma_E = 0.8$ the notion of segregation versus fusion is moot).

Our simulation and theoretical analysis was performed on network models with uncapped plasticity (i.e., we ignore the requirement $W^{ij}_{EE} \in W[0, W^{\max}_{EE}]$). This was done to achieve a good quantitative match between our network simulations and theoretical estimates. Although our theory does not take capped synaptic weights into consideration, the simulation results with capping shows qualitatively consistent results with our theoretical approach [Figs. 3(d) and 4]. Indeed, numerical simulations indicate that our theory provides an accurate approximation of the weight dynamics, provided the bulk of the weight distribution remains far from the boundaries.

I. Same conclusions apply to more general, asymmetric networks

Our analysis has primarily focused on a simplified case with two symmetric assemblies (same number of neurons per assembly). This choice reduces the complexity of the mean-field equations and aids in a clearer interpretation. However, the core theoretical argument—that the approximate oddness of a causal STDP rule effectively cancels the correlation

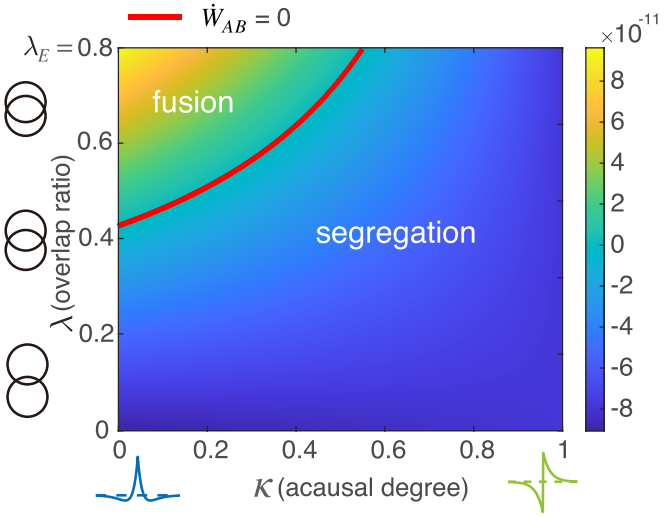

FIG. 8. Heat map of the value of the derivative dW_{AB}/dt corresponding to different overlap parameter λ and different causality parameter κ with the initial weight $W_{AB}(0)$. The red curve specifies $dW_{AB}/dt = 0$.

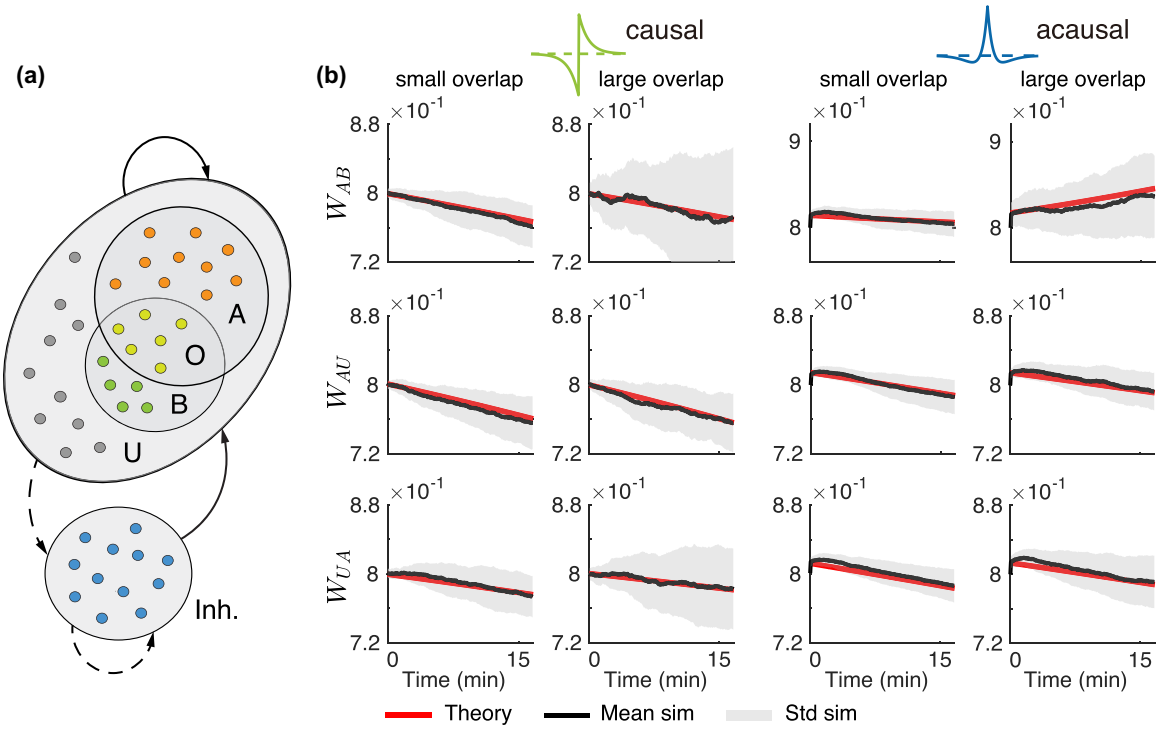

FIG. 9. Causal STDP maintains assembly segregation even in networks with asymmetric structure. (a) Schematic of the network with asymmetric assemblies (A, B) with overlap (O) and a population of unassigned neurons (U). (b) Each column represents a specific condition, comparing causal and acausal STDP for large and small assembly overlap. The rows display the evolution of the mean synaptic weight for different between-group connections. Top row: Between assemblies (W_{AB}). Middle row: From unassigned neurons to assembly A (W_{AU}). Bottom row: From assembly A to unassigned neurons (W_{UA}). The analysis employs the same fundamental causal and acausal STDP rules as in previous figures. The parameters are listed in Table IV.

structure generated by second-order motifs—does not depend on network symmetry. Therefore, our conclusions should generalize to more complex architectures.

To explicitly test this, we analyzed a more general, asymmetric network model [Fig. 9(a)]. In this extended framework, the excitatory population is partitioned into two assemblies, A and B (with an overlapping population, O), and a population of unassigned neurons, U (which did not receive training), all of which having different sizes. Applying our mean-field analysis to this network yielded a system of 16 dynamical equations for the $E \to E$ weights. For causal STDP rules, assemblies remain segregated even for large overlap, with between-assembly connections weakening over time [Fig. 9(b), left columns]. Here we only show the cross connections W_{AB} between two assemblies and the connections to and from the unassigned assembly, W_{AU}, and W_{UA}. The remaining connection weights also evolved as predicted; Fig. 11 shows a representative example of a network with large overlap and causal STDP. In contrast, for acausal STDP rules, assembly fate is again dependent on the overlap size; large overlaps lead to assembly fusion, whereas smaller overlaps permit segregation [Fig. 9(b), right columns]. Additionally, all the connections across assemblies and unassigned neurons remain weak. This demonstrates that the relationship between STDP causality and assembly stability is a robust principle that is not limited to symmetric network configurations. A more detailed description of this asymmetric model is provided in Appendix A.

III. DISCUSSION

Following previous work [26,29,30], we derived a low-dimensional dynamical system describing the mean weight dynamics of a weakly coupled network of spiking neuron models whose synapses obey a STDP rule. We used this low-dimensional mean-field theory to analyze how the stability of trained assembly structure depends on the combination of assembly overlap and the degree of causality in the STDP learning rule. A causal STDP rule suppresses fusion dynamics and allows assemblies to stay segregated regardless of the overlap size. This was not the case for acausal STDP rules, where for sufficient assembly overlap a fused assembly would develop after training.

The key difference between assembly stability with causal versus acausal STDP rules is how they treat common input projections to neuron pairs between assemblies. The coefficients for the weight dynamics characterizing how the common input motifs affect the dynamics is computed as an integral of the product of the STDP rule and filtered spike train autocovariance, of which the latter is an even function. For causal STDP rules, which are close to odd functions, this product is an odd function so the integral is near zero. This makes the weight dynamics unaffected by the common inputs coming from the overlap group. By contrast, an acausal STDP curve is close to an even function, so that its product with autocovariance is also even, and hence common input effects cannot be ignored. As a result

fused assemblies can occur for sufficiently large assembly overlap.

However, purely causal STDP rules make training from correlated noise difficult, as their odd symmetry cancels the net effect of symmetric spike-time correlations, thus preventing synaptic potentiation. While this limitation can be overcome in high-firing-rate regimes [39,63–67], here we explore an alternative solution effective at lower rates, rooted in the learning rule itself. We propose that a biologically plausible approach is to use an STDP rule that incorporates an acausal (even) component. This addition allows the network to strengthen connections based on correlated activity (Fig. 4). Indeed, STDP curves measured experimentally [20,22,23] are typically asymmetric but not strictly odd and are well approximated by our model with $\kappa \approx 0.8$ [Fig. 2(c), red curve]. This mixed causality is functionally advantageous: It allows the network to reinforce assembly structure in response to strong correlations during training, while maintaining overall stability when such correlations are weak.

A. Constraints and limitations of our theory

In this paper we built upon previous work [29,68] and developed a mean-field theory for synaptic weight dynamics in networks of spiking neuron models. Our primary extension is the inclusion of second-order motifs, which allows the theory to account for spike-time cross-covariance generated by common inputs. For this theory to be mathematically tractable, we made several simplifying assumptions, which we discuss below.

Our theory assumes weak synaptic connections, a common simplification in mean-field approaches that offers several advantages [8,29,30,60]. First, it sidesteps the complexities of network stability. Networks with strong connections often require a delicate balance of excitation and inhibition to prevent runaway activity, which complicates the theoretical treatment of multiple interacting assemblies [48,64]. Weak coupling ensures the convergence of mathematical approximations used to derive the cross-covariance, such as the Neumann series expansion of the cross-covariance. In strongly coupled balanced networks, there is no guarantee that such expansions are valid. Second, weak coupling facilitates the derivation of dynamics at the level of individual synapses. In contrast, theories for strongly coupled networks often describe the bulk dynamics of neural populations, making a straightforward theoretical extension to multiple distinct assemblies more challenging [48].

Another key assumption is a moment-closure approximation, where correlations between synaptic weights are considered negligible (e.g., $\langle W_{ik}W_{kj}\rangle \approx \langle W_{ik}\rangle\langle W_{kj}\rangle$). This allows us to close the mean-field equations at the level of quadratic terms in the average synaptic weights. This condition holds for our initial setup but may be violated in biological circuits possessing preexisting structure from development or prior learning. While a full investigation of this scenario is beyond our current scope, we hypothesize that the strong, common input from our training protocol would dominate plasticity dynamics. We expect this powerful training signal to effectively override or reshape any weaker, preexisting network configurations, imprinting the desired assembly structure. Extending our mean-field theory to networks with preexisting structure could likely be achieved by incorporating methods from previous work on calculating connection motifs [30,60,61].

Our theory treats synaptic weights as uncapped, meaning they are not constrained to a finite range (e.g., $[0, W_{EE}^{\max}]$). This mathematical idealization is crucial for achieving a quantitative match between theory and simulations. Although the theory does not explicitly account for weight bounds, simulations that incorporate them show results that are qualitatively consistent with our theoretical framework (Figs. 3 and 4). The theory provides an accurate approximation of the weight dynamics as long as the bulk of the weight distribution remains far from the saturation boundaries. Indeed, even when weights are capped, their long-term qualitative evolution—approaching and saturating at the boundaries—aligns with our expectations (Appendix B).

Finally, we assumed homogeneous firing rates across excitatory neurons for simplicity. While biological networks exhibit significant rate heterogeneity [69–71], this simplification is not critical to our core conclusions. Since homeostatic inhibitory plasticity controls excitatory firing rates [29,50], rate heterogeneity can be readily introduced by varying the homeostatic targets across the inhibitory population. Although this would quantitatively alter the weight dynamics, the fundamental symmetry properties of the motif kernels, which are central to our argument about STDP causality, are expected to be preserved on average. Therefore, we are confident that our main conclusions regarding the relationship between STDP rules and assembly fate would hold in a more heterogeneous network.

B. Rate-based versus timing-based synaptic learning

While STDP learning is inherently temporal, the degree of synaptic potentiation compared to depression is well known to depend on the firing rates of the pre- and postsynaptic neurons [18,63,72,73]. Biologically realistic STDP learning rules based on voltage or calcium dynamics capture the rate dependence of STDP [40,74]. When these realistic learning rules are used then the strong reciprocal coupling between neuron pairs that is the basis of assembly structure can be embedded in networks of spiking neuron models [64,75]. However, the complexity of these plasticity rules make a comprehensive theory of assembly formation in spiking systems difficult to formulate. Alternatively, models that consider simple, phenomenological STDP rules which depend only on the relative timing of pre- and postsynaptic spiking are amenable to mean-field treatments [26,30,58].

The mechanisms behind the formation and stability of assembly structure with rate-based [64,75] and timing-based plasticity are distinct. In a rate-based scenario, coordinated high firing rates for neurons within an assembly is a necessary condition for potentiation of within assembly synapses. However, due to balanced recurrent inhibition only a subset of assemblies can have access to high firing rates at a given time. A consequence of this is that training can only be achieved through giving signals sequentially, so that coordinated high firing rates can drive potentiation for synapses within the same assembly and low firing rates can drive depression for synapses between neuron pairs in different assemblies. In our

temporal framework, the formation of the assembly structure depends on the strong spike time covariance within assemblies due to correlated stimuli (or feedforward inputs). Because of this the training of different assemblies can occur simultaneously, as opposed to sequentially.

The stability of the assembly structure during spontaneous activity after training also differs between rate-based and timing-based learning. In the rate-based scenario, the trained structure is a multiattractor state, and the assembly structure maintains itself through random activation of different assemblies (attractor states), ultimately reinforcing the learned structure [64]. Recent work has suggested that such metastability can also be achieved by a BCM-like, local plasticity rule alone, without homeostatic mechanisms such as inhibitory plasticity or nonlocal synaptic renormalization [76]. In the spike timing–based scenario, it is the strong spike train correlations within assemblies due to strong recurrent connections that reinforce the learned structure. Hence, the assembly structure reinforces itself even when all neurons fire at approximately the same firing rate.

A common criticism of purely timing-based learning is that for nontrivial weight dynamics to occur we require the STDP curve to be roughly balanced between potentiation and depression (i.e., $S^0 = \int_{-\infty}^{\infty} L(s)ds \sim \mathcal{O}(\epsilon)$). This may occur for a restricted range of pre- and postsynaptic neuron firing rates [63,64], yet in general is not guaranteed. Rate-based or the recently coined behavioral timescale plasticity mechanisms [17] are more robust and do not require fine-tuning of the plasticity rule. Nonetheless, in all cases network learning requires access to both potentiation and depression mechanics, else all synaptic connections will become homogeneous in strength. The requirement that $S^0 \sim \mathcal{O}(\epsilon)$ for STDP learning simply has this fact being satisfied for statistically stationary spike trains.

Finally, the STDP models employed in our work are phenomenological ones where ΔW is additive; see Eq. (9), meaning that it does not depend on W. In general for such models weight dynamics either increases to the upper bound or depresses to zero [77,78], which is inconsistent with unimodal and long-tailed weight distributions typically observed in experiments [79,80]. Multiplicative STDP learning does not have this failure [78]; however, such a rule does not permit the motif-based mean-field theory we use in our study [30]. Further modifications could be done by introducing axonal or dendritic delays or weight dependence of plasticity which can yield weight distributions more closely resembling those observed in neural tissue [77,78,81–83]. Additionally, our conclusion is made for the phenomenological shape of the STDP rule, the comparison to more realistic synaptic plasticity rules, such as triple STDP or calcium-based models [40,72] remains a topic for future analysis.

C. Overlap capacity

The stability of overlapping assemblies has been recently explored by Manz and colleagues [31] in networks with acausal excitatory STDP. Their work suggests that stable assembly structures with overlap can emerge under specific conditions: either when neurons maintain appropriate spontaneous firing rates or when most neurons participate in multiple assemblies with similar saturation levels. Alternatively, assembly segregation can be achieved through the plasticity of inhibitory synapses. For instance, Ref. [84] showed that with acausal plasticity rules, assembly fusion is prevented by introducing two subpopulations of inhibitory neurons with opposing plasticity rules (Hebbian and anti-Hebbian) that respectively provide feedback control and competitive lateral inhibition. Distinct inhibitory plasticity rules has also been explored as a means to generate functionally relevant connectivity structures [54,55]. Our study is complementary to these works, focusing on how the temporal structure of the STDP learning rule affects the maximal tolerated overlap. In particular, we have shown that overlap is much more tolerated when the STDP learning rule is causal, as opposed to the acausal rule used in Ref. [31].

In our study, we only considered two symmetric assemblies with a common overlap, which is the simplest case. Applying the same derivation to the case of multiple assemblies with different population sizes would result in a larger-dimensional mean-field theory. We would need to allow for the case that $W_{AB} \neq W_{BA}$, and for a three population network we would have four overlap populations to model: O_{AB}, O_{BC}, O_{AC}, and O_{ABC}. However, all the dynamical equations for $W_{\alpha\beta}$ would have a similar form as our two assemblies case yet with extra terms corresponding to common input through weak connections from other assemblies. Nevertheless, the motif coefficients S^{cc}, S^{fc}, and S^{bc} would still be the same. Hence our theoretical framework could be extended to multiple assemblies and our general conclusion likely still holds. We verified our conclusion in a network with two unsymmetrical overlapping assemblies and a group of unassigned excitatory neurons.

Considering multiple assemblies then begs the question of assembly capacity: How many assemblies of size $M < N$ can be embedded into a network of N neurons and remain segregated post training? This question is distinct, but related, to more classic studies of memory capacity in recurrent networks [36,37]. For instance, in Hopfield networks capacity is measured by many input patterns can be stored in the recurrent weight matrix and reliably retrieved [85,86]. In standard Hopfield networks all neurons contribute to the storage of each input pattern and hence the pattern "overlap" is significant (or complete). In our network model capacity could be measured by an assembly population's ability to perform pattern completion and pattern segregation of inputs that are coherent with the training set [64,87]. Assembly fusion would certainly compromise pattern segregation, as now inputs to assembly α would also activate assembly β through the strong $\alpha \to \beta$ wiring. This research avenue will be pursued in future studies.

IV. METHODS

A. Diffusion-based theory for first- and second-order spike train statistics

We present the method used to derive spike train firing rates and cross-covariance in our network model. For large networks, the recurrent input to neuron i in population α can be treated using a diffusion approximation:

$$I_\alpha^i(t) = \sum_{\beta=E,I} \sum_{j=1}^{N^\beta} W_{\alpha\beta}^{ij}(t) \cdot \left(J_{\text{syn}} * y_\beta^j(t)\right)$$

$$\approx \mu_{\text{rec}}^i(t) + \sigma_{\text{rec}}^i(t)\xi^i(t).$$

Here we drop the population α notation for ease of exposition. In brief, the recurrent input is replaced by a white noise process with time-varying noise strength (time-changed Brownian motion). The external input I^i_{ext} is similarly modeled as white noise (Wiener process) $I^i_{\text{ext}} = \mu^{\text{ext}}_\alpha + \sigma^{\text{ext}}_\alpha \xi^i_\alpha(t)$ [see Eq. (2)]. The sum of these two independent stochastic processes yields a white noise with time-varying strength [88], resulting in the voltage equation for neuron i in population α [modified from Eq. (1)]:

$$\tau \frac{dV^i}{dt} = (E_L - V^i) + \Delta_T \exp\left(\frac{V^i - V_T}{\Delta_T}\right) + \mu^i(t) + \sigma^i(t)\sqrt{2\tau}\xi^i(t). \tag{22}$$

The mean input $\mu(t)$ and noise strength $\sigma(t)$ are given by:

$$\mu^i(t) = \mu_{\text{ext}} + \mu^i_{\text{rec}}(t),$$
$$\sigma^i(t) = \sqrt{\frac{\sigma^2_{\text{ext}} + \left(\sigma^i_{\text{rec}}\right)^2(t)}{2\tau}} \approx \frac{\sigma_{\text{ext}}}{\sqrt{2\tau}}.$$

The approximation for $\sigma^i(t)$ holds because $\sigma^i_{\text{rec}} \sim \mathcal{O}(\epsilon)$ when recurrent connections are weak.

The voltage equation is a Langevin equation, whose probability distribution $P(V, t)$ obeys a corresponding Fokker-Planck equation:

$$\frac{\partial P}{\partial t}(V^i, t) = \frac{\partial P}{\partial V^i}\left(V^i - \mu(t) + \frac{\sigma^2(t)}{2}\frac{\partial P}{\partial V^i}\right),$$

with boundary conditions:

$$P(V_{\text{th}-}, t) = P(V_{\text{re}+}, t) - P(V_{\text{re}-}, t) = 0,$$
$$J(V_{\text{th}-}, t) = J(V_{\text{re}+}, t) - J(V_{\text{re}-}, t) = r^i(t).$$

These conditions arise from the threshold and reset mechanism, with $r^i(t)$ being the firing rate of neuron i at time t.

To compute temporal mean firing rates and cross-covariance, we assume the network operates near a steady state with small fluctuations. This approach separates the input to each neuron into static and dynamic components:

$$\mu^i(t) = \overline{\mu}^i + \widetilde{\mu}^i(t),$$

where $\overline{\mu}^i$ represents the time-averaged input and $\widetilde{\mu}^i(t)$ satisfies $\int_0^\infty \widetilde{\mu}^i(s)ds = 0$. We again assume the variance remains approximately static, $\sigma^i \approx \sigma_{\text{ext}}$.

Following Refs. [62,89] we compute the steady-state firing rates r^i_0 given the mean input $\overline{\mu}^i$ and diffusion parameter σ_{ext}:

$$r^i_0 = f(\overline{\mu}^i, \sigma_{\text{ext}}),$$

by numerically solving the Fokker-Planck equation. Since neurons within each subpopulation α share identical parameters except for μ^i, their transfer functions are equivalent:

$$r^i_0 = f_\alpha(\overline{\mu}^i, \sigma_{\text{ext}}),$$

for neuron i in subpopulation $\alpha \in \{A, B, O, I\}$.

The mean input current $\overline{\mu}^i$ depends on both external and recurrent inputs, which in turn depend on the recurrent weights $\mathbf{W}$ and firing rates $\mathbf{r}_0$. This leads to a self-consistent system of equations:

$$\mathbf{r}_0 = \mathbf{f}(\mathbf{r}_0; \mathbf{W}, \sigma_{\text{ext}}),$$

where the weights $\mathbf{W}$ evolve slowly enough to be treated as static parameters.

The firing rates can be obtained by integrating the auxiliary differential equations [90]:

$$\dot{r}^i = -r^i + f_\alpha(\mu^i, \sigma_{\text{ext}}),$$

until reaching a fixed point $\dot{r}_i = 0$, starting from an initial guess r_{initial}. This method effectively estimates individual neuron firing rates under weak connections embedded in white noise. Our theoretical predictions align with simulation results, particularly in showing that homeostatic PV plasticity drives excitatory neurons toward the target rate r_E^{target}.

To compute the spike train cross-covariance we employ a linear response analysis of the Fokker-Planck equation. In response to a small input perturbation $I_\epsilon(t)$, the firing rate response can be approximated as $I_\epsilon * \chi_r(t)$, where $\chi_r(t)$ is the response kernel computed numerically from the Fokker-Planck equation [62,89].

In the frequency domain this becomes $\hat{I}_\epsilon(\omega)\hat{\chi}_r(\omega)$, for frequency variable ω. For small fluctuations $\mu^i_{\text{small}}(t)$, we can apply this to mean-corrected spike trains [29,60]:

$$\hat{y}^i(\omega) = \int_{-\infty}^{\infty} \left(y^i(t) - r^i_0\right) e^{-2\pi i\omega t} dt.$$

Following Refs. [59,60] we decompose the spike train as:

$$\begin{aligned}\hat{y}^i(\omega) &= \hat{y}^i_0(\omega) + \hat{\chi}^i(\omega)\widehat{\mu}^i_{\text{small}}(\omega) \\ &= \hat{y}^i_0(\omega) + \hat{\chi}^i(\omega)\sum_{j=1}^{N} W^{ij}\hat{J}_{\text{syn}}(\omega)\hat{y}^j(\omega),\end{aligned} \tag{23}$$

where $\chi^i(\omega)$ represents the Fourier transform of the response kernel and $J_{\text{syn}}(\omega)$ is the Fourier transform of the synaptic filter.

To simplify the matrix notation, we define the effective interaction matrix [29,60]:

$$\widehat{\mathbf{K}}(\omega) = \left[\hat{\chi}^i(\omega)\hat{J}^{ij}_{\text{syn}}(\omega)\right]. \tag{24}$$

This definition allows us to rewrite Eq. (23) as:

$$\widehat{\mathbf{y}} = \widehat{\mathbf{y}}_0 + \widehat{\mathbf{W}} \circ \widehat{\mathbf{K}}(\omega) \cdot \widehat{\mathbf{y}},$$
$$\widehat{\mathbf{y}} = (\mathbf{I} - \widehat{\mathbf{W}} \circ \widehat{\mathbf{K}}(\omega))^{-1} \cdot \widehat{\mathbf{y}}_0.$$

The cross-covariance in the frequency domain is therefore given by:

$$\begin{aligned}\widehat{\mathbf{C}}(\omega) &= \langle \widehat{\mathbf{y}}(\omega) \cdot \widehat{\mathbf{y}}^*(\omega) \rangle \\ &= [\mathbf{I} - \widehat{\mathbf{W}} \circ \widehat{\mathbf{K}}(\omega)]^{-1} \widehat{\mathbf{C}}_0(\omega) [\mathbf{I} - (\widehat{\mathbf{W}} \circ \widehat{\mathbf{K}}(\omega))^*]^{-1},\end{aligned} \tag{25}$$

where $\widehat{\mathbf{C}}_0(\omega)$ denotes the Fourier transform of the mean-subtracted autocovariance matrix. Given that weights $\mathbf{W}$ evolve on a much slower timescale than spiking activity, we can treat them as static when computing the cross-covariance $\mathbf{C}(s; \mathbf{W})$, which depends on the connection matrix $\mathbf{W}$. Our simulation results confirm that this linear response approach provides accurate estimates of cross-covariance in our network model.

B. Mean-field dynamics of population averaged synaptic weights

In this section we derive the mean-field equations of all the averaged weights $\mathbf{W_m} = [W_{AA}, W_{AB}, W_{AO}, W_{OA}, W_{OO}]$. After training, the network is dissected into four populations $\{A, B, O, I\}$. We recall that we have introduced the overlap ratio $\lambda = \frac{N_O}{N}$, and excitatory and inhibitory ratios $\gamma_E = \frac{N_E}{N}$, $\gamma_I = \frac{N_I}{N}$. Notice here population O contains all the neurons in the overlap between two assemblies, and population A or B contain only the neurons in the assembly A or B respectively, but not in the overlap part. Hence the number of neurons in these two populations equals to $N_A = N_B = \frac{\gamma_E - \lambda}{2} N$.

We define the normalized mean synaptic weights from population β to α as:

$$W_{\alpha\beta} = \frac{1}{N_\alpha N_\beta} \sum_{i\in\alpha} \sum_{j\in\beta} \frac{W_{\alpha\beta}^{ij}}{\epsilon} \sim \mathcal{O}(1).$$

Applying the methods introduced in Sec. II D, we expand the cross-covariance using a Neumann series and truncate at the second-order motifs. The resulting weight dynamics for a generic connection $W_{\alpha\beta}$ is given by:

$$\begin{aligned}
\frac{dW_{\alpha\beta}}{dt} &= p_{\alpha\beta} S_0 r_E^2 + \epsilon \cdot (S^{\mathrm{f}} W_{\alpha\beta} + p_{\alpha\beta} S^{\mathrm{b}} W_{\beta\alpha}) \\
&\quad + \epsilon \cdot p_{\alpha\beta} \Bigg[S^{\mathrm{cc}} \left(\sum_\upsilon \epsilon N_\upsilon \cdot W_{\alpha\upsilon} W_{\beta\upsilon} \right) \\
&\quad + S^{\mathrm{fc}} \left(\sum_\upsilon \epsilon N_\upsilon \cdot W_{\alpha\upsilon} W_{\upsilon\beta} \right) \\
&\quad + S^{\mathrm{bc}} \left(\sum_\upsilon \epsilon N_\upsilon \cdot W_{\upsilon\alpha} W_{\beta\upsilon} \right) \\
&\quad + \gamma_I \cdot \left(S_I^{\mathrm{fc}} W_{\alpha I} W_{I\beta} + S_I^{\mathrm{cc}} W_{\alpha I} W_{\beta I} + S_I^{\mathrm{bc}} W_{I\alpha} W_{\beta I} \right) \Bigg].
\end{aligned}$$

Crucially, all terms on the right-hand side are $\mathcal{O}(\epsilon)$. To be specific, for the zeroth-order motif, consistency requires that $S_0 \sim \mathcal{O}(\epsilon)$, as discussed in Sec. II E. For the second-order terms, the explicit factor of ϵ outside the brackets arises from the normalization of the mean weights $W_{\alpha\beta} \sim \mathcal{O}(1)$. Since the unscaled synaptic weights are proportional to ϵ, the second-order interaction involves a product of two weights (ϵ^2). By grouping one ϵ with the population size N_υ [using the relation $\epsilon N_\upsilon \sim \mathcal{O}(1)$], we are left with a single net factor of ϵ, confirming the order of the approximation.

To simplify this system, we exploit the symmetry of the network architecture (e.g., $W_{AB} = W_{BA}$). Furthermore, we employ the approximation $W_{AO} \approx W_{OA}$. This approximation is justified on two grounds. First, by applying Grönwall's inequality, we establish that the divergence between W_{OA} and W_{AO} is bounded by their initial difference after training (up to a constant factor). Since the training signals are symmetric, this initial difference is typically small, as confirmed by our simulations [Fig. 3(d)]. Second, and crucially, our analysis focuses on the regime where the fate of weak cross-assembly connections (W_{AB}) is determined by their proximity to the potentiation-depression separatrix. In contrast, the strong within-assembly and overlap connections ($W_{AA}, W_{AO}, W_{OA}, W_{OO}$) lie far from this threshold and consistently potentiate. Consequently, as W_{AO} and W_{OA} approach the hard bound $W_{EE}^{\max}$, the assumption $W_{AO} \approx W_{OA}$ becomes asymptotically exact. Incorporating these symmetries allows us to reduce the system to the following four-dimensional mean-field equations for $\mathbf{W_m}$:

$$\begin{aligned}
\frac{dW_{AB}}{dt} &= p_0 r^2 S^0 + \epsilon \cdot W_{AB}(S^{\mathrm{f}} + p_0 S^{\mathrm{b}}) \\
&\quad + \epsilon \cdot p_0 \gamma_I \left(S_I^{\mathrm{cc}} W_{AI}^2 + S_I^{\mathrm{fb}} W_{AI} W_{IE} \right) \\
&\quad + \epsilon \cdot p_0 (S^{\mathrm{cc}} + S^{\mathrm{fb}}) \left[(\gamma_E - \lambda) W_{AA} W_{AB} + \lambda W_{AO}^2 \right], \\
\frac{dW_{AA}}{dt} &= p_0 r^2 S^0 + \epsilon \cdot W_{AA}(S^{\mathrm{f}} + p_0 S^{\mathrm{b}}) \\
&\quad + \epsilon \cdot p_0 \gamma_I \left(S_I^{\mathrm{cc}} W_{AI}^2 + S_I^{\mathrm{fb}} W_{AI} W_{IE} \right) \\
&\quad + \epsilon \cdot p_0 (S^{\mathrm{cc}} + S^{\mathrm{fb}}) \\
&\quad \times \left[\left(\frac{\gamma_E - \lambda}{2} \right) \left(W_{AA}^2 + W_{AB}^2 \right) + \lambda W_{AO}^2 \right], \\
\frac{dW_{OO}}{dt} &= p_0 r^2 S^0 + \epsilon \cdot W_{OO}(S^{\mathrm{f}} + p_0 S^{\mathrm{b}}) \\
&\quad + \epsilon \cdot p_0 \gamma_I \left(S_I^{\mathrm{cc}} W_{OI}^2 + S_I^{\mathrm{fb}} W_{OI} W_{IE} \right) \\
&\quad + \epsilon \cdot p_0 (S^{\mathrm{cc}} + S^{\mathrm{fb}}) \left[(\gamma_E - \lambda) W_{AO}^2 + \lambda W_{OO}^2 \right], \\
\frac{dW_{AO}}{dt} &= p_0 r^2 S^0 + \epsilon \cdot W_{AO}(S^{\mathrm{f}} + p_0 S^{\mathrm{b}}) \\
&\quad + \epsilon \cdot p_0 \gamma_I \left(S_I^{\mathrm{cc}} W_{AI} W_{OI} + S_I^{\mathrm{fb}} W_{AI} W_{IE} \right) \\
&\quad + \epsilon \cdot p_0 (S^{\mathrm{cc}} + S^{\mathrm{fb}}) \\
&\quad \times \left[\left(\frac{\gamma_E - \lambda}{2} \right) (W_{AA} + W_{AB}) W_{AO} + \lambda W_{AO} W_{OO} \right].
\end{aligned}$$

Finally, the system is closed by the inhibitory synaptic weights, which follow the linear relationship described in Sec. IV C:

$$\begin{aligned}
W_{AI} &= \left(\mu_E^{\mathrm{target}} - \mu_A^{\mathrm{ext}} \right) / (r_I \gamma_I) \\
&\quad - r_E / (r_I \gamma_I) \cdot [(\gamma_E - \lambda)/2 \cdot (W_{AA} + W_{AB}) + \lambda W_{AO}], \\
W_{OI} &= \left(\mu_E^{\mathrm{target}} - \mu_O^{\mathrm{ext}} \right) / (r_I \gamma_I) \\
&\quad - r_E / (r_I \gamma_I) \cdot [(\gamma_E - \lambda) W_{AO} + \lambda W_{OO}].
\end{aligned}$$

Here the first term $(\mu_E^{\mathrm{target}} - \mu_A^{\mathrm{ext}})/(r_I \gamma_I)$, denoted as μ_E^{diff} for brevity before, determines the inhibitory weight strength required for excitatory neurons to maintain their target firing rate r_α^{diff} when receiving only external background input r_α^{ext} without excitatory input. This relationship is maintained through homeostatic inhibitory plasticity [29,50]. A detailed derivation is provided in Sec. IV C.

C. Inhibitory STDP and homeostatic control of firing rates

In this section, we describe how inhibitory synaptic weights from I neurons to E neurons ($W_{EI}^{ij} < 0$) evolve under STDP. The inhibitory plasticity rule combines two components: a timing-dependent function and presynaptic spike-induced depression.

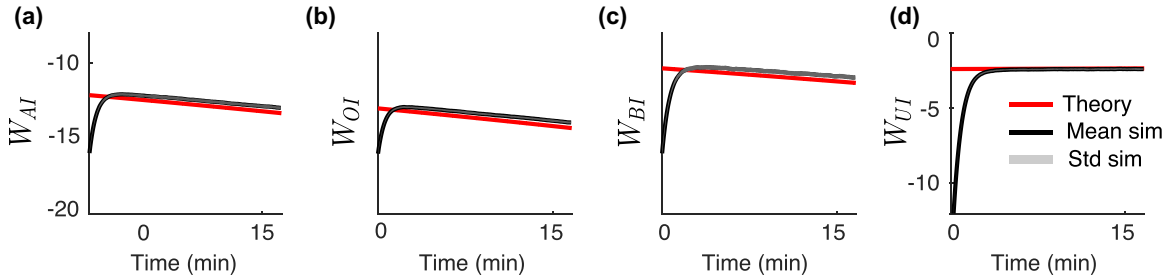

FIG. 10. Comparison of simulated and theoretically estimated inhibitory connection weights in an asymmetric network with four excitatory populations. The theoretical estimation, derived using the method in Sec. IV C, depends linearly on the excitatory weights. Panels (a)–(d) show the normalized mean weight from the inhibitory population to excitatory subpopulations (a)–(d), respectively.

The homeostatic STDP function is defined for lag $s = t_{\text{post}} - t_{\text{pre}}$ as:

$$L^{\text{h}}(s) = f^{\text{h}} \exp\left(-\frac{|s|}{\tau^{\text{h}}}\right).$$

Here $f^{\text{h}} < 0$ is the update step size and τ^{h} is the STDP decay time constant. Additionally, each presynaptic spike triggers synaptic depression with magnitude $d^{\text{h}} > 0$:

$$W_{EI}^{ij} \to W_{EI}^{ij} + d^{\text{h}}.$$

This plasticity model induces a target firing rate r_E^{target}, defined by the relationship between depression and potentiation:

$$r_E^{\text{target}} = -\frac{d^{\text{h}}}{2f^{\text{h}}\tau^{\text{h}}}.$$

$r_E^{\text{target}} > 0$ is ensured by $d^{\text{h}} > 0$ and $f^{\text{h}} < 0$. After averaging over neuronal populations, the mean-field dynamics for the inhibitory weights $W_{\alpha I}$ (from population I to excitatory population α) reduces to:

$$\frac{dW_{\alpha I}}{dt} = p_{EI} S_{EI}^0 \cdot r_I\left[r_E^{\text{target}} - r_\alpha\right] + \sum_{\text{motif}} S^{\text{motif}} \cdot q_{\alpha\beta}^{\text{motif}}(\mathbf{W_m}),$$

where $S_{EI}^0 = 2f^h\tau^h < 0 \sim \mathcal{O}(1)$. Since the motif terms are $\mathcal{O}(\epsilon)$, the leading-order dynamics depend primarily on the firing rates r_I and r_α. Given that the weights satisfy $W_{\alpha I} \sim \mathcal{O}(\epsilon)$, the stationarity of this equation implies:

$$r_E^{\text{target}} - r_\alpha \sim \mathcal{O}(\epsilon). \tag{26}$$

Thus, on the slow timescale of plasticity, the inhibitory rule effectively constrains the excitatory population α to track the target firing rate r_E^{target}.

The mean firing rate of excitatory neurons $\alpha \in \{A, B, O\}$ follows:

$$r_\alpha = f_E(\mu_\alpha, \sigma_\alpha),$$
$$\mu_\alpha = \mu_{\text{ext}} + \sum_\beta \epsilon N_\beta \cdot W_{\alpha\beta} r_\beta + \epsilon N_I \cdot W_{\alpha I} r_I,$$

where $\epsilon N_I = \gamma_I$ by definition and $f_E(\mu, \sigma)$ is a monotonically increasing transfer function. For a given noise intensity σ_{ext}, there exists an input μ_E^{target} such that $r_E^{\text{target}} = f_E(\mu_E^{\text{target}}, \sigma_{\text{ext}})$. Linearizing the transfer function around this target (or applying the mean value theorem) to Eq. (26) yields:

$$f'(\tilde{\mu}, \sigma) \cdot \left[\mu_E^{\text{target}} - \mu_\alpha\right] \sim \mathcal{O}(\epsilon).$$

Assuming the gain f' is nonzero, this implies that the total mean input must remain close to the target level:

$$\mu_\alpha \approx \mu_E^{\text{target}}.$$

This equilibrium is maintained because the inhibitory input term, $\gamma_I W_{\alpha I} r_I$, dynamically adjusts to counterbalance the recurrent excitatory drive, $\sum_\beta \epsilon N_\beta W_{\alpha\beta} r_\beta$.

The I neuron firing rate r_I remains static due to nonplastic connections $W_{I\alpha}$ and W_{II}:

$$\mu_I = \mu_{\text{ext}} + \sum_{\alpha\in E} \epsilon N_\alpha \cdot W_{I\alpha} r_E^{\text{target}} + \epsilon N_I \cdot W_{II} r_I,$$
$$r_I = f_I(\mu_I, \sigma_{\text{ext}}).$$

Consequently, the inhibitory connection weights can be directly computed:

$$W_{\alpha I} = \left(\mu_e^{\text{target}} - \mu_{\text{ext}} - \sum_{\beta\in E} \epsilon N_\beta \cdot W_{\alpha\beta} r_\beta\right) \Big/ (\gamma_I r_I).$$

This relationship demonstrates that inhibitory connections track excitatory weights linearly while maintaining target firing rates, a result we verified through numerical simulations (Fig. 10).

ACKNOWLEDGMENTS

B.D. is supported by the NIH Grants No. 1U19NS107613-01, No. R01NS133598, and No. CRCNS-R01EY034723; a Vannevar Bush faculty fellowship (N000141812002); and the Simons Foundation Collaboration on the Global Brain. This research benefited from Physics Frontier Center for Living Systems funded by the National Science Foundation (PHY-2317138). Financial support was provided via the National Institute from Mathematics and Theory in Biology (Simons Foundation Award No. MP-TMPS-00005320 and NSF Award No. DMS-2235451).

DATA AVAILABILITY

The data that support the findings of this article are openly available [91], embargo periods may apply.

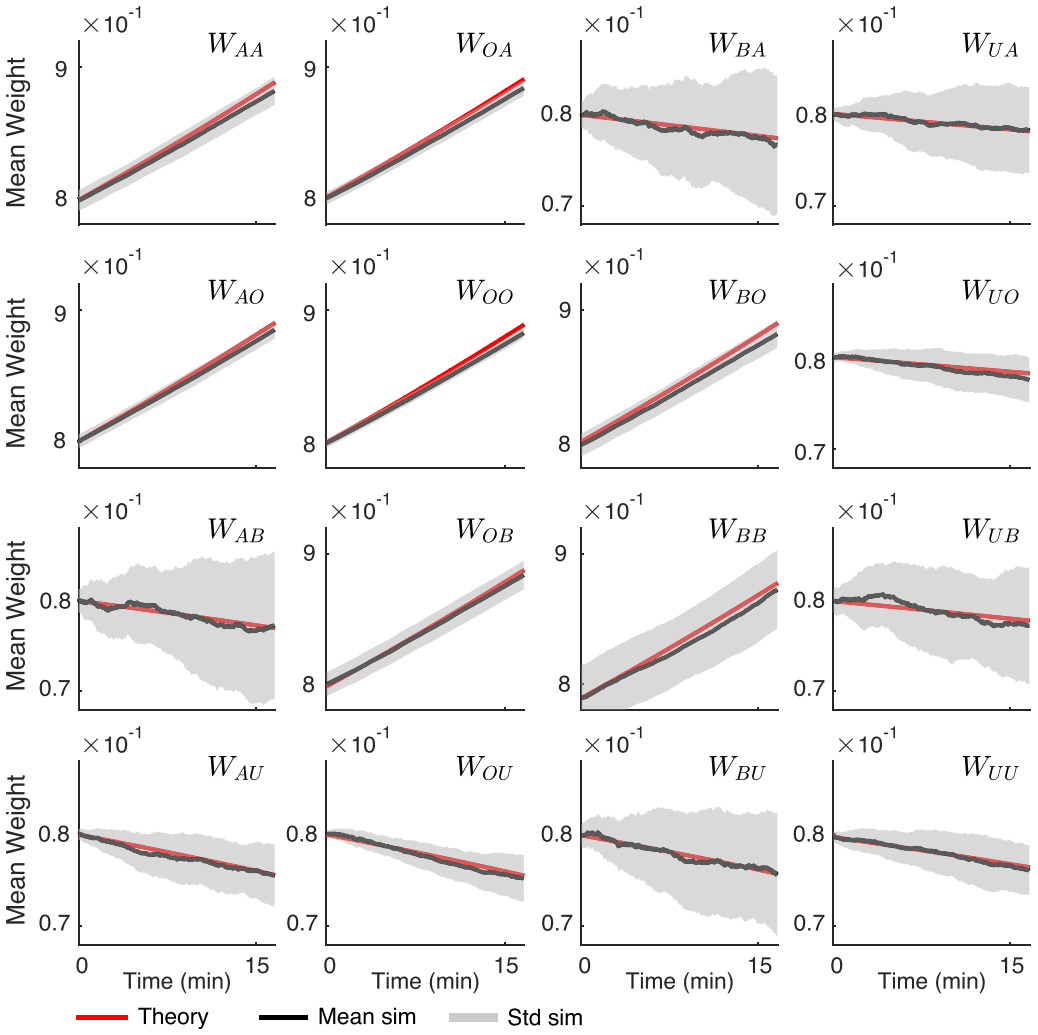

FIG. 11. Causal STDP leads to stable assembly segregation in an asymmetric network with large overlap. The figure displays the dynamics of all 16 classes of mean excitatory-to-excitatory ($E \to E$) weights in the network with an asymmetric assembly structure. The simulation confirms that under a causal STDP rule, the network self-organizes into a stable, segregated state: Strong within-assembly connections (e.g., W_{AA}, W_{OO}, and W_{AO}) consistently potentiate. Weak cross-assembly connections (e.g., W_{AB}, W_{AU}, and W_{UA}) consistently depress. These simulation results align with the predictions from the simplified theoretical model, confirming that causal STDP robustly supports segregated assembly structures even in more complex networks.

APPENDIX A: GENERALIZATION TO NETWORKS WITH MORE POPULATIONS

The mean-field framework developed in this paper can be readily extended to more general cases, such as networks with multiple, asymmetric assemblies of varying sizes (see Sec. II I). The fundamental argument for why causal STDP prevents fusion remains the same: It relies on the odd-even symmetry (parity) of the STDP function and the correlation kernels, not on the architectural symmetry of the network itself.

To validate this claim, we tested our conclusions on a more complex network model. This network consists of $N = 2000$ neurons, with $N_E = 1600$ Excitatory and $N_I = 400$ Inhibitory neurons. The E population is divided into four subpopulations: population A and B, containing neurons exclusive to each assembly; an overlap population O, containing neurons common to both assemblies; and an unassigned population U, with $N_U = 400$ neurons. We compared two scenarios, a large overlap case ($N_A = 300, N_O = 800, N_B = 100$) and a small overlap case ($N_A = 650, N_O = 100, N_B = 450$).

Relaxing the assumption of network symmetry increases the number of unique mean-field equations needed to describe the system. For instance, in a network with distinct populations A, B, O, and U, we can no longer assume that $W_{AB} = W_{BA}$. This results in a larger system of 16 coupled ordinary differential equations for the $E \to E$ weights and four equations for the homeostatic $I \to E$ weights. While the derivation for each equation follows the same process, the fundamental dynamics of each connection $W_{\alpha\beta}$ remain governed by the general form derived in Sec. IV B, with the summation index υ now extending over all E subpopulations $\{A, O, B, U\}$. To validate our approach while maintaining tractability, we analyze a simplified version of this high-dimensional system. In this approximation, we group connection types, representing all strong connections with a single value ($W_{\text{strong}} = W$) and

all weak connections with another ($W_{\text{weak}} = w$), and retain only the dominant term for each motif. Taking the connection W_{AB} as an example, the simplified equation of W_{AB} becomes

$$\frac{dW_{AB}}{dt} = p_{AB}S^0 r_E^2 + \epsilon \cdot \left(S^{\text{f}}_{AB} w + p_{AB} S^{\text{b}}_{AB} w\right) + \epsilon \cdot p_{AB}\Big[\left(S^{\text{fc}}_E + S^{\text{fb}}_E\right)\epsilon N_O \cdot W^2 + \epsilon N_I \cdot \left(S^{\text{fc}}_I W_{AI} W_{IB} + S^{\text{cc}}_I W_{AI} W_{BI} + S^{\text{bc}}_I W_{IA} W_{BI}\right)\Big].$$

The inhibitory connections are still computed from the linear estimation as:

$$W_{\alpha I} = \text{Const}^{\text{diff}}_{\alpha} - \frac{r_E}{N_I r_I}\left(\sum_{\beta} N_\beta W_{\alpha\beta}\right),$$

where the $\text{Const}^{\text{diff}}_{\alpha}$ are determined by solving $r_E^{\text{target}} = f(\mu_{\text{total}}, \sigma)$, as discussed in Sec. IV C. While this is a highly simplified model where W and w are fixed, we only expect it to provide a reasonable approximation when the connection weights $W_{\alpha\beta}$ do not change significantly. Nevertheless, this simplified model produces weight dynamics that show reasonable agreement with full network simulations (Fig. 11). This provides further evidence that our conclusions generalize to more complex and biologically realistic network structures.

APPENDIX B: LONG-TERM WEIGHT DYNAMICS AND DISTRIBUTION

The additive, pairwise STDP rule employed in our model (Sec. II A 2) can, in principle, lead to monotonic weight changes, where mean weights potentiate or depress until they saturate at the imposed upper or lower bounds. Our simulations implement these as soft boundaries (Figs. 3 and 4); a synapse at a bound cannot move further in that direction but is free to change in the opposite direction. This is a heuristic approach to account for the biological constraint that synaptic strengths are finite and must adhere to Dale's principle ($W_{EE} \geqslant 0$ and $W_{EI} \leqslant 0$).

Our mean-field theory describes the dynamics of the average synaptic weights, without requiring knowledge of the full distribution of synaptic weights. The validity of this simplification rests on two key assumptions. First, that the weight distribution does not destabilize the asynchronous firing activity of the network, preserving the validity of the linear response analysis. Second, that the influence of second-order connectivity motifs can be approximated by the product of mean weights. For example, for a second-order motif involving weight W_{ik} and W_{jk} we assume $\langle W_{ik} W_{jk}\rangle \approx \langle W_{ik}\rangle\langle W_{jk}\rangle$. Our simulations indicate that these assumptions generally hold for the parameter regimes and timescales considered in our main analysis.

In most cases we simulate the network for a duration sufficient to observe the potentiation or depression of mean synaptic weights and to validate our theoretical predictions. For instance, the weight dynamics under a causal STDP rule appear linear over these timescales, as the influence of spike-time correlations is largely canceled by the odd symmetry of the STDP function. In contrast, for acausal rules, the dynamics can be approximately quadratic (Figs. 6 and 7). We chose not to extend all simulations until the mean weights reached their bounds for two reasons: First, the weight dynamics is very slow near a separatrix, making such simulations computationally expensive, and, second, the agreement between our theory and simulations is already robust and nontrivial within the observed time frame.

For completeness, we performed an extended simulation of a network with one excitatory and one inhibitory population to examine the long-term evolution of its synaptic weights (Fig. 12). While a quantitative theory for the full weight distribution is beyond the scope of this work, these simulations provide key insights into its dynamics. As predicted by our mean-field analysis, the mean synaptic weights slowly drift toward their imposed bounds over long timescales. Concurrently, the shape of the weight distribution evolves significantly. For instance, an initially Gaussian-like distribution of $E \to E$ weights broadens over time. As the mean drifts toward the lower bound ($W_{EE} = 0$), a substantial fraction of weights accumulates at this boundary. This process skews the distribution, causing it to develop a heavy tail that resembles a log-normal distribution [Figs. 12(a3) and 12(c3)].

The evolution of the weight distribution is shaped by two distinct factors: interaction with synaptic bounds and the passage of time. To disentangle these effects, we first compare two networks that differ only in the zeroth-order term S_0 of the STDP rule. A more negative $S_0 < 0$ causes the mean weight to drift rapidly to the lower boundary, leading to a highly skewed distribution as weights accumulate at zero [Figs. 12(a1) and 12(a3)], with inhibition onto excitatory neurons depressing as a consequence [Fig. 12(a2) and 12(a4)]. In contrast, a less negative $S_0 < 0$ results in a much slower drift, and the distribution remains roughly symmetric and far from the boundary over the same time period [Figs. 12(b1)–12(b4)]. Given that their initial widths are comparable, this comparison demonstrates that the pronounced skewness is a direct consequence of the boundary interaction. Second, to illustrate the effect of time, the slow-drift simulation was continued for 10 times longer [Fig. 12(c)]. Although the mean weight eventually reaches a value similar to that in the fast-drift case, the final distribution is substantially broader [Fig. 12(a), blue, and 12(c), purple]. This result reveals that the variance of the weight distribution grows continuously over time, largely independent of the mean.

It is important to distinguish these generic drift dynamics observed in unstructured networks from the weight distributions in the structured assembly networks discussed in the main text (e.g., Fig. 11). In the simple *E-I* case (Fig. 12), the excitatory weights form a single homogeneous population that drifts collectively. However, in networks with trained assemblies, the synaptic population is segregated into distinct functional classes: within and across different assemblies. While the distribution of each individual class may resemble the broadened, boundary-distorted Gaussian observed in Fig. 12, their mean values drift in opposite directions. Consequently, the aggregate distribution of all excitatory weights in the network resembles a mixture of Gaussians (distorted by the boundaries), yet the evolution of the distribution for any single connection type follows the pattern shown in this subsection.

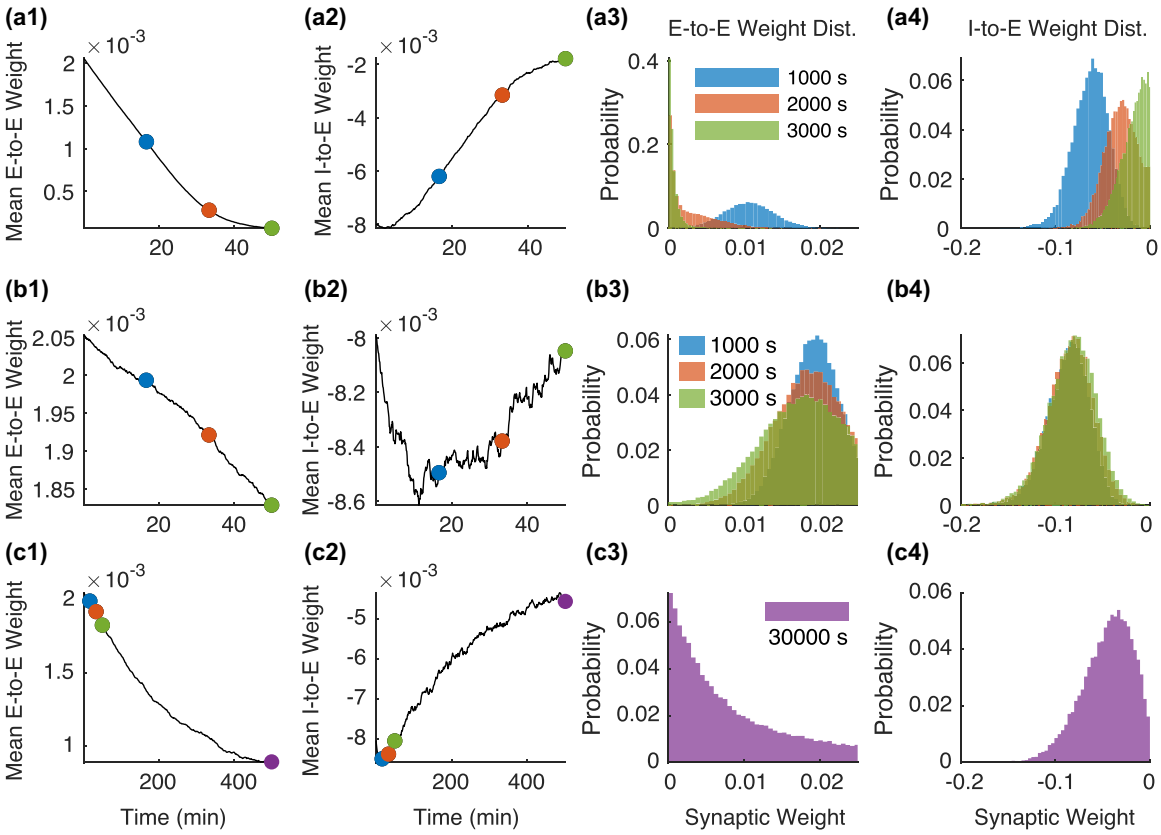

FIG. 12. Evolution of synaptic weights and their distributions over long timescales in a simple *E-I* network. We utilize a network without assembly structure here to isolate the drift dynamics imposed by the boundaries from the effects of assembly interactions. The figure illustrates how the weight distribution is shaped by both the drift speed of the mean weights and the total simulation time. The *E-I* network is the same as in Fig. 1, but a modified STDP rule and smaller synaptic cap were used to better visualize the dynamics (detailed parameters are in Table III). (a) Simulation over 50 minutes (3000 s) using an STDP rule with a large negative rate-dependent term (S_0). This condition drives the mean $E \to E$ weight rapidly towards its lower bound of zero. (b) Same simulation as (a), but with the magnitude of S_0 reduced by a factor of 10. Consequently, the mean weight drifts much more slowly over the same 50-minute duration. (c) The simulation from (b) continued for a 10-fold longer duration (500 min $\approx$ 8.3 h), allowing the slow-drifting weight distribution to eventually approach the lower bound. For each condition shown in rows (a), (b), and (c): the first and second columns show the time evolution of the mean synaptic weights for $E \to E$ and $I \to E$ connections, respectively (un-normalized). The third and fourth columns show snapshots of the corresponding weight distributions. The colors represent different time points: 1000 s (blue), 2000 s (orange), 3000 s (green), and 30 000 s (purple).

APPENDIX C: MODEL PARAMETERS

The parameters for the network models and plasticity rules used throughout this study are provided in the tables below. To clearly isolate the effects of STDP symmetry in our theoretical analysis (Figs. 5–8), we used a causal STDP rule with symmetric time constants ($\tau_-^c = \tau_+^c$) and uncapped synaptic weights, aligning with the assumptions of the mean-field theory. In contrast, simulations that included a training phase (Figs. 3 and 4) were designed to be more biologically realistic, employing an STDP rule with a faster potentiation constant ($\tau_-^c > \tau_+^c$) and weight bounds on all synaptic weights. The specific parameters for other figures, such as those demonstrating firing rate homeostasis (Fig. 1) or generalization to more complicated networks (Fig. 9), were chosen to best illustrate those results. Unless otherwise specified in the main text, all simulations used the default values listed here.

TABLE I. Neuronal and network model parameters.

Parameter	Description	Value
Single neuron properties		
τ_α	Membrane time constant	20 ms
E_L/V_L	Leak reversal potential	–55 mV
V_{th}	Spike threshold	–53 mV
V_{re}	Reset potential	–60 mV
Δ_T	Spike sharpness	3 mV
τ_{ref}	Refractory period	1 ms
τ_{syn}	Synaptic time constant	5 ms
Input properties		
I_i^α	Total external input	
μ_α^{ext}	Mean external input	0 mV
$\sigma_\alpha^{\text{ext}}$	Std. dev. of ext. input	4 mV
σ_{train}	Std. dev. of training signal	0.8/1.5 mV

TABLE II. $E \to E$ realistic STDP parameters (Figs. 3 and 4).

Parameter	Realistic	Odd
Causal STDP rule $[L^{c}(s)]$		
S_0^{c}	-8.32×10^{-5} mV	-2×10^{-5} mV
f_-^{c}	-4×10^{-4} mV	-8×10^{-4} mV
f_+^{c}	Solved for consistency	
τ_+^{c}	15 ms	20 ms [a]
τ_-^{c}	30 ms	20 ms [a]
Acausal STDP rule $[L^{ac}(s)]$		
S_0^{c}	-3.2×10^{-5}	-2×10^{-5}
f_+^{ac}	2.667×10^{-5} mV	2.667×10^{-5} mV [b]
f_-^{ac}	2.667×10^{-5} mV	2.667×10^{-5} mV [b]
T_+^{ac}, T_-^{ac}	30 ms	30 ms
τ_+^{ac}	30 ms	30 ms
τ_-^{ac}	Solved for consistency	
Inhibitory plasticity		
τ^{h}	30 ms	Same
f^{h}	–0.0024 mV	
r_E^{target}	12 Hz	
Static weights and bounds		
W_{EE}^{max}	0.24 mV	
W_{EI}^{min}	–0.96 mV	
W_{IE}	0.08 mV	
W_{II}	–0.32 mV	

[a]For the analytical framework, we set $\tau_+^{c} = \tau_-^{c}$ to maintain an approximately odd function. In simulations (Figs. 2 and 3), we used a faster potentiation ($\tau_+^{c} = 15$ ms) and slower depression ($\tau_-^{c} = 30$ ms) to accelerate training dynamics.
[b]These values, in combination with $T_\pm^{ac}$, were chosen to ensure comparable magnitudes between the causal and acausal STDP functions.

TABLE III. Parameters of simple *E-I* network (used in Figs. 1 and 12).

Parameter	Realistic	Long timescale
Causal STDP rule $[L^{c}(s)]$		
S_0^{c}	-2.72×10^{-5} mV	-6.8×10^{-6} mV [a]
f_-^{c}	-2×10^{-4} mV	-5×10^{-5} mV
f_+^{c}	Solved for consistency	
τ_+^{c}	15 ms	
τ_-^{c}	30 ms	
Inhibitory plasticity		
τ^{h}	30 ms	
f^{h}	–0.0012 mV	-3×10^{-4} mV
r_E^{target}	6/12 Hz [b]	
Static weights and bounds		
W_{EE}^{max}	0.12 mV	0.03 mV
W_{EI}^{min}	–0.96 mV	–0.24 mV
W_{IE}	0.04 mV	0.02 mV
W_{II}	–0.16 mV	–0.08 mV

[a]In Fig. 12(a), we set $S_0^{c} = -6.8 \times 10^{-5}$ to derive the weights approaching 0 faster.
[b]In Fig. 1(d), we have $r_E^{target} = 12$ Hz, in Fig. 1(e), we have $r_E^{target} = 6$ Hz.

TABLE IV. Parameters of network with unsymmetric assemblies (used in Figs. 9 and 11).

Parameter	Realistic
Causal STDP rule $(L^{c}(s))$	
S_0^{c}	-4.44×10^{-6} mV
f_-^{c}	-4×10^{-4} mV
f_+^{c}	Solved for consistency
τ_+^{c}	30 ms
τ_-^{c}	30 ms
Acausal STDP rule $[L^{ac}(s)]$	
S_0^{ac}	-2.963×10^{-6} mV
f_+^{ac}	1.33×10^{-5} mV
f_-^{ac}	1.33×10^{-5} mV
T_+^{ac}, T_-^{ac}	30 ms
τ_+^{ac}	30 ms
τ_-^{ac}	Solved for consistency
Inhibitory plasticity	
τ^{h}	30 ms
f^{h}	–0.0012 mV
r_E^{target}	12 Hz
Static weights and bounds	
W_{EE}^{max}	0.12 mV
W_{EI}^{min}	–0.48 mV
W_{IE}	0.04 mV
W_{II}	–0.16 mV

TABLE V. Glossary of network model terms.

Symbol	Description
Neuronal and network dynamics	
V_α^i	Membrane potential of neuron i in population α.
$y_\alpha^i(t)$	Spike train of neuron i in population α.
r_α^i	Mean firing rate of neuron i in population α.
$J_{\text{syn}}(t)$	Synaptic filter describing postsynaptic current dynamics.
$M_\alpha^i = 1$ or 0	Existence of training signal
$s_\alpha(t)$	Training signal common to population α
Network size	
N	Neural network size
N_α	Neural population size
$p_{\alpha\beta}$	Baseline connection probability
$\epsilon = 1/N$	Weight normalization constant
$J_{\alpha\beta}$	Initial connection weights
$W_{\alpha\beta}^{ij} \sim \mathcal{O}(\epsilon)$	Single connection weight
$W_{\alpha\beta} \sim \mathcal{O}(1)$	Normalized mean connection weight
λ	Fraction of excitatory neurons shared by both assemblies (overlap).
λ_E/λ_I	Fraction of exc./inh. neurons over network size
STDP	
$L(s)$	General STDP learning function, a mix of causal and acausal rules.
$L^{c}(s)$	Temporally causal (Hebbian) STDP learning rule.
$L^{ac}(s)$	Temporally acausal (symmetric) STDP learning rule.
$L^{h}(s)$	Homeostatic learning rule for inhibitory ($I \to E$) synapses.
κ	Parameter controlling the degree of causality in the STDP rule.

TABLE VI. Glossary of mean field terms.

Symbol	Description
Mean field theory	
$C^{ij}_{\alpha\beta}(s)$	Cross-covariance function of spike trains between neurons i and j.
$K^{ij}_{\alpha\beta}(s)$	Neural response kernel to presynaptic spikes
$\chi^{i}(s)$	Neural response kernel to postsynaptic current
$q^{\text{motif}}_{\alpha\beta}(\mathbf{W})$	Polynomial describing how weight evolution depends on a specific connectivity motif.
$M^{\text{motif}}_{\alpha\beta}(s)$	Response kernel of weight dynamics for a specific connectivity motif.
$S^{\text{motif}}_{\alpha\beta}$	Motif coefficient, quantifies the influence of a motif on plasticity.
k_0, k_1, k_2	Coefficients for the one-dimensional dynamics of W_{AB}.

[1] D. O. Hebb, *The Organization of Behavior: A Neuropsychological Theory* (Psychology Press, London, 2005).

[2] K. D. Harris and T. D. Mrsic-Flogel, Cortical connectivity and sensory coding, Nature (London) **503**, 51 (2013).

[3] H. Markram, W. Gerstner, and P. J. Sjöström, A history of spike-timing-dependent plasticity, Front. Synaptic Neurosci. **3**, 4 (2011).

[4] H. Tang, M. Schrimpf, W. Lotter, C. Moerman, A. Paredes, J. Ortega Caro, W. Hardesty, D. Cox, and G. Kreiman, Recurrent computations for visual pattern completion, Proc. Natl. Acad. Sci. USA **115**, 8835 (2018).

[5] G. Buzsáki, Neural syntax: Cell assemblies, synapsembles, and readers, Neuron **68**, 362 (2010).

[6] C. H. Papadimitriou, S. S. Vempala, D. Mitropolsky, M. Collins, and W. Maass, Brain computation by assemblies of neurons, Proc. Natl. Acad. Sci. USA **117**, 14464 (2020).

[7] G. Neves, S. F. Cooke, and T. V. Bliss, Synaptic plasticity, memory and the hippocampus: A neural network approach to causality, Nat. Rev. Neurosci. **9**, 65 (2008).

[8] C. Miehl, S. Onasch, D. Festa, and J. Gjorgjieva, Formation and computational implications of assemblies in neural circuits, J. Physiol. **601**, 3071 (2023).

[9] Y. Yoshimura, J. L. Dantzker, and E. M. Callaway, Excitatory cortical neurons form fine-scale functional networks, Nature (London) **433**, 868 (2005).

[10] R. Perin, T. K. Berger, and H. Markram, A synaptic organizing principle for cortical neuronal groups, Proc. Natl. Acad. Sci. USA **108**, 5419 (2011).

[11] A. Holtmaat and P. Caroni, Functional and structural underpinnings of neuronal assembly formation in learning, Nat. Neurosci. **19**, 1553 (2016).

[12] Y. Zaki, Z. T. Pennington, D. Morales-Rodriguez, M. E. Bacon, B. Ko, T. R. Francisco, A. R. LaBanca, P. Sompolpong, Z. Dong, S. Lamsifer, *et al.*, Offline ensemble co-reactivation links memories across days, Nature **637**, 145 (2025).

[13] H. Ko, S. B. Hofer, B. Pichler, K. A. Buchanan, P. J. Sjöström, and T. D. Mrsic-Flogel, Functional specificity of local synaptic connections in neocortical networks, Nature (London) **473**, 87 (2011).

[14] L. Cossell, M. F. Iacaruso, D. R. Muir, R. Houlton, E. N. Sader, H. Ko, S. B. Hofer, and T. D. Mrsic-Flogel, Functional organization of excitatory synaptic strength in primary visual cortex, Nature (London) **518**, 399 (2015).

[15] W.-C. A. Lee, V. Bonin, M. Reed, B. J. Graham, G. Hood, K. Glattfelder, and R. C. Reid, Anatomy and function of an excitatory network in the visual cortex, Nature (London) **532**, 370 (2016).

[16] E. L. Bienenstock, L. N. Cooper, and P. W. Munro, Theory for the development of neuron selectivity: Orientation specificity and binocular interaction in visual cortex, J. Neurosci. **2**, 32 (1982).

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

[18] M. Graupner, P. Wallisch, and S. Ostojic, Natural firing patterns imply low sensitivity of synaptic plasticity to spike timing compared with firing rate, J. Neurosci. **36**, 11238 (2016).

[19] C. C. Bell, V. Z. Han, Y. Sugawara, and K. Grant, Synaptic plasticity in a cerebellum-like structure depends on temporal order, Nature (London) **387**, 278 (1997).

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

[21] H. Markram, W. Gerstner, and P. J. Sjöström, Spike-timing-dependent plasticity: A comprehensive overview, Front. Synaptic Neurosci. **4**, 2 (2012).

[22] 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).

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

[24] N. Masuda and H. Kori, Formation of feedforward networks and frequency synchrony by spike-timing-dependent plasticity, J. Comput. Neurosci. **22**, 327 (2007).

[25] Y. K. Takahashi, H. Kori, and N. Masuda, Self-organization of feed-forward structure and entrainment in excitatory neural networks with spike-timing-dependent plasticity, Phys. Rev. E **79**, 051904 (2009).

[26] N. Ravid Tannenbaum and Y. Burak, Shaping neural circuits by high order synaptic interactions, PLoS Comput. Biol. **12**, e1005056 (2016).

[27] 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).

[28] B. Babadi and L. F. Abbott, Pairwise analysis can account for network structures arising from spike-timing dependent plasticity, PLoS Comput. Biol. **9**, e1002906 (2013).
[29] G. K. Ocker and B. Doiron, Training and spontaneous reinforcement of neuronal assemblies by spike timing plasticity, Cereb. Cortex **29**, 937 (2019).
[30] G. K. Ocker, A. Litwin-Kumar, and B. Doiron, Self-organization of microcircuits in networks of spiking neurons with plastic synapses, PLoS Comput. Biol. **11**, e1004458 (2015).
[31] P. Manz and R.-M. Memmesheimer, Purely stdp-based assembly dynamics: Stability, learning, overlaps, drift and aging, PLoS Comput. Biol. **19**, e1011006 (2023).
[32] M. Rigotti, O. Barak, M. R. Warden, X.-J. Wang, N. D. Daw, E. K. Miller, and S. Fusi, The importance of mixed selectivity in complex cognitive tasks, Nature (London) **497**, 585 (2013).
[33] D. J. Cai, D. Aharoni, T. Shuman, J. Shobe, J. Biane, W. Song, B. Wei, M. Veshkini, M. La-Vu, J. Lou, *et al.*, A shared neural ensemble links distinct contextual memories encoded close in time, Nature (London) **534**, 115 (2016).
[34] M. Wenzel and J. P. Hamm, Identification and quantification of neuronal ensembles in optical imaging experiments, J. Neurosci. Methods **351**, 109046 (2021).
[35] C. Gastaldi, T. Schwalger, E. De Falco, R. Q. Quiroga, and W. Gerstner, When shared concept cells support associations: Theory of overlapping memory engrams, PLoS Comput. Biol. **17**, e1009691 (2021).
[36] J. Aljadeff, M. Gillett, U. P. Obilinovic, and N. Brunel, From synapse to network: Models of information storage and retrieval in neural circuits, Curr. Opin. Neurobiol. **70**, 24 (2021).
[37] S. Fusi, Memory capacity of neural network models, arXiv:2108.07839.
[38] Z. Brzosko, S. B. Mierau, and O. Paulsen, Neuromodulation of spike-timing-dependent plasticity: Past, present, and future, Neuron **103**, 563 (2019).
[39] D. E. Feldman, The spike-timing dependence of plasticity, Neuron **75**, 556 (2012).
[40] M. Graupner and N. Brunel, Calcium-based plasticity model explains sensitivity of synaptic changes to spike pattern, rate, and dendritic location, Proc. Natl. Acad. Sci. USA **109**, 3991 (2012).
[41] R. Kempter, W. Gerstner, and J. L. van Hemmen, Hebbian learning and spiking neurons, Phys. Rev. E **59**, 4498 (1999).
[42] N. Fourcaud-Trocmé, D. Hansel, C. Van Vreeswijk, and N. Brunel, How spike generation mechanisms determine the neuronal response to fluctuating inputs, J. Neurosci. **23**, 11628 (2003).
[43] N. Brunel, Dynamics of sparsely connected networks of excitatory and inhibitory spiking neurons, J. Comput. Neurosci. **8**, 183 (2000).
[44] W. Gerstner, W. M. Kistler, R. Naud, and L. Paninski, *Neuronal Dynamics: From Single Neurons to Networks and Models of Cognition* (Cambridge University Press, Cambridge, UK, 2014).
[45] B. Gluss, A model for neuron firing with exponential decay of potential resulting in diffusion equations for probability density, Bull. Math. Biophys. **29**, 233 (1967).
[46] P. Johannesma, Diffusion models for the stochastic activity of neurons, in *Neural Networks: Proceedings of the-School on Neural Networks-June 1967 in Ravello* (Springer, Berlin, 1968), pp. 116–144.
[47] R. Capocelli and L. Ricciardi, Diffusion approximation and first passage time problem for a model neuron, Kybernetik **8**, 214 (1971).
[48] A. E. Akil, R. Rosenbaum, and K. Josić, Balanced networks under spike-time dependent plasticity, PLoS Comput. Biol. **17**, e1008958 (2021).
[49] L. F. Abbott and S. B. Nelson, Synaptic plasticity: Taming the beast, Nat. Neurosci. **3**, 1178 (2000).
[50] 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).
[51] R. C. Froemke, M. M. Merzenich, and C. E. Schreiner, A synaptic memory trace for cortical receptive field plasticity, Nature (London) **450**, 425 (2007).
[52] M. A. Woodin, K. Ganguly, and M.-m. Poo, Coincident pre-and postsynaptic activity modifies gabaergic synapses by postsynaptic changes in Cl^- transporter activity, Neuron **39**, 807 (2003).
[53] V. Kilman, M. C. Van Rossum, and G. G. Turrigiano, Activity deprivation reduces miniature ipsc amplitude by decreasing the number of postsynaptic $GABA_A$ receptors clustered at neocortical synapses, J. Neurosci. **22**, 1328 (2002).
[54] F. Lagzi, M. C. Bustos, A.-M. Oswald, and B. Doiron, Assembly formation is stabilized by parvalbumin neurons and accelerated by somatostatin neurons, bioRxiv (2021), https://doi.org/10.1101/2021.09.06.459211.
[55] D. Festa, C. Cusseddu, and J. Gjorgjieva, Structured stabilization in recurrent neural circuits through inhibitory synaptic plasticity, bioRxiv (2024), https://doi.org/10.1101/2024.10.12.618014.
[56] S. Song, K. D. Miller, and L. F. Abbott, Competitive Hebbian learning through spike-timing-dependent synaptic plasticity, Nat. Neurosci. **3**, 919 (2000).
[57] S. Song and L. F. Abbott, Cortical development and remapping through spike timing-dependent plasticity, Neuron **32**, 339 (2001).
[58] M. Gilson, A. N. Burkitt, D. B. Grayden, D. A. Thomas, and J. L. van Hemmen, Emergence of network structure due to spike-timing-dependent plasticity in recurrent neuronal networks v: Self-organization schemes and weight dependence, Biol. Cybern. **103**, 365 (2010).
[59] B. Lindner, B. Doiron, and A. Longtin, Theory of oscillatory firing induced by spatially correlated noise and delayed inhibitory feedback, Phys. Rev. E **72**, 061919 (2005).
[60] J. Trousdale, Y. Hu, E. Shea-Brown, and K. Josić, Impact of network structure and cellular response on spike time correlations, PLoS Comput. Biol. **8**, e1002408 (2012).
[61] Y. Hu, J. Trousdale, K. Josić, and E. Shea-Brown, Motif statistics and spike correlations in neuronal networks, J. Stat. Mech.: Theory Exp. (2013) P03012.
[62] M. J. E. Richardson, Firing-rate response of linear and nonlinear integrate-and-fire neurons to modulated current-based and conductance-based synaptic drive, Phys. Rev. E **76**, 021919 (2007).
[63] P. J. Sjöström, G. G. Turrigiano, and S. B. Nelson, Rate, timing, and cooperativity jointly determine cortical synaptic plasticity, Neuron **32**, 1149 (2001).

[64] A. Litwin-Kumar and B. Doiron, Formation and maintenance of neuronal assemblies through synaptic plasticity, Nat. Commun. **5**, 5319 (2014).
[65] O. Mackwood, L. B. Naumann, and H. Sprekeler, Learning excitatory-inhibitory neuronal assemblies in recurrent networks, Elife **10**, e59715 (2021).
[66] S. Sadeh and C. Clopath, Excitatory-inhibitory balance modulates the formation and dynamics of neuronal assemblies in cortical networks, Sci. Adv. **7**, eabg8411 (2021).
[67] C. Miehl and J. Gjorgjieva, Stability and learning in excitatory synapses by nonlinear inhibitory plasticity, PLoS Comput. Biol. **18**, e1010682 (2022).
[68] G. K. Ocker, Y. Hu, M. A. Buice, B. Doiron, K. Josić, R. Rosenbaum, and E. Shea-Brown, From the statistics of connectivity to the statistics of spike times in neuronal networks, Curr. Opin. Neurobiol. **46**, 109 (2017).
[69] A. Treves, S. Panzeri, E. T. Rolls, M. Booth, and E. A. Wakeman, Firing rate distributions and efficiency of information transmission of inferior temporal cortex neurons to natural visual stimuli, Neural Comput. **11**, 601 (1999).
[70] K. Mizuseki and G. Buzsáki, Preconfigured, skewed distribution of firing rates in the hippocampus and entorhinal cortex, Cell Rep. **4**, 1010 (2013).
[71] M. N. Insanally, B. F. Albanna, J. Toth, B. DePasquale, S. S. Fadaei, T. Gupta, O. Lombardi, K. Kuchibhotla, K. Rajan, and R. C. Froemke, Contributions of cortical neuron firing patterns, synaptic connectivity, and plasticity to task performance, Nat. Commun. **15**, 6023 (2024).
[72] J.-P. Pfister and W. Gerstner, Triplets of spikes in a model of spike timing-dependent plasticity, J. Neurosci. **26**, 9673 (2006).
[73] L. Montangie, C. Miehl, and J. Gjorgjieva, Autonomous emergence of connectivity assemblies via spike triplet interactions, PLoS Comput. Biol. **16**, e1007835 (2020).
[74] C. Clopath, L. Büsing, E. Vasilaki, and W. Gerstner, Connectivity reflects coding: A model of voltage-based stdp with homeostasis, Nat. Neurosci. **13**, 344 (2010).
[75] F. Zenke, E. J. Agnes, and W. Gerstner, Diverse synaptic plasticity mechanisms orchestrated to form and retrieve memories in spiking neural networks, Nat. Commun. **6**, 6922 (2015).
[76] X. Yang and G. La Camera, Co-existence of synaptic plasticity and metastable dynamics in a spiking model of cortical circuits, PLoS Comput. Biol. **20**, e1012220 (2024).
[77] J. Rubin, D. D. Lee, and H. Sompolinsky, Equilibrium properties of temporally asymmetric Hebbian plasticity, Phys. Rev. Lett. **86**, 364 (2001).
[78] R. Gütig, R. Aharonov, S. Rotter, and H. Sompolinsky, Learning input correlations through nonlinear temporally asymmetric Hebbian plasticity, J. Neurosci. **23**, 3697 (2003).
[79] S. Song, P. J. Sjöström, M. Reigl, S. Nelson, and D. B. Chklovskii, Highly nonrandom features of synaptic connectivity in local cortical circuits, PLoS Biol. **3**, e68 (2005).
[80] S. Lefort, C. Tomm, J.-C. F. Sarria, and C. C. Petersen, The excitatory neuronal network of the C2 barrel column in mouse primary somatosensory cortex, Neuron **61**, 301 (2009).
[81] B. Babadi and L. F. Abbott, Intrinsic stability of temporally shifted spike-timing dependent plasticity, PLoS Comput. Biol. **6**, e1000961 (2010).
[82] J. E. Rubin, Steady states in an iterative model for multiplicative spike-timing-dependent plasticity, Netw., Comput. Neural Syst. **12**, 131 (2001).
[83] M. Gilson and T. Fukai, Stability versus neuronal specialization for stdp: Long-tail weight distributions solve the dilemma, PLoS ONE **6**, e25339 (2011).
[84] R. Bergoin, A. Torcini, G. Deco, M. Quoy, and G. Zamora-López, Emergence and maintenance of modularity in neural networks with Hebbian and anti-Hebbian inhibitory stdp, PLoS Comput. Biol. **21**, e1012973 (2025).
[85] D. J. Amit, H. Gutfreund, and H. Sompolinsky, Spin-glass models of neural networks, Phys. Rev. A **32**, 1007 (1985).
[86] D. Krotov, A new frontier for Hopfield networks, Nat. Rev. Phys. **5**, 366 (2023).
[87] J. J. Knierim and J. P. Neunuebel, Tracking the flow of hippocampal computation: Pattern separation, pattern completion, and attractor dynamics, Neurobiol. Learn. Mem. **129**, 38 (2016).
[88] B. Oksendal, *Stochastic Differential Equations: An Introduction with Applications* (Springer Science & Business Media, New York, 2013).
[89] M. J. Richardson, Spike-train spectra and network response functions for non-linear integrate-and-fire neurons, Biol. Cybern. **99**, 381 (2008).
[90] M. Layer, J. Senk, S. Essink, A. van Meegen, H. Bos, and M. Helias, Nnmt: Mean-field based analysis tools for neuronal network models, Front. Neuroinform. **16**, 835657 (2022).
[91] https://github.com/brain-math/spknet_overlap.