Anti-Hebbian plasticity protects against memory interference

Hebbian learning strengthens synaptic weights between correlated neurons and is believed to support assembly formation and pattern completion. Anti-Hebbian learning, on the other hand, strengthens connections between negatively correlated neurons, while its contribution to memory processing is unclear. Here we show in theory and experiment how anti-Hebbian plasticity at excitatory synapses onto inhibitory neurons (E-to-I synapses) protects against interference between competing memories. In the hippocampus, anti-Hebbian plasticity at E-to-I synapses is mediated by calcium-permeable AMPA receptors (CP-AMPARs) that allow calcium influx due to presynaptic activity despite post-synaptic hyperpolarization. Using a viral approach targeting hippocampal parvalbumin-positive interneurons, we replaced CP-AMPARs with calcium-impermeable receptors and observed a sharp decrease in anti-Hebbian plasticity. We developed an excitatory-inhibitory network model, where anti-Hebbian plasticity at E-to-I synapses allowed the storage of the negative part of a pattern covariance matrix, while the positive part was stored via Hebbian plasticity. When presented with a recall cue, the network converged to the cued pattern and maintained high accuracy despite strong cue noise or large pattern overlaps. Interestingly, when the strength of anti-Hebbian synapses was reduced, recall performance remained high for low cue noise and small pattern overlaps, but deteriorated quickly when noise or overlaps increased. These predictions were confirmed in CP-AMPAR-deficient mice that performed indistinguishably from controls in a spatial memory task when delays between an encoding and a recall phase were short (corresponding to low cue noise in the model) but showed decreased performance for longer delays (high cue noise). Similarly, CP-AMPAR-deficient mice learned the location of an escape platform in the Morris water maze but performed poorly when the platform's location was changed; indicating interference between highly similar memory patterns. Together, these results suggest that CP-AMPAR-mediated anti-Hebbian plasticity at E-to-I synapses protects against interference between competing memory patterns by facilitating the storage of negative correlations.

Hebbian plasticity of E-to-E synapses has been suggested to be key for storing *positive* correlations between elements of a memory.[1] Hebbian plasticity in I-to-E and E-to-I synapses has been suggested to restore E-I balance during memory formation.[2] Here, we study *anti*-Hebbian plasticity in E-to-I synapses and show that it can store *negative* correlations between elements of a memory, and thereby support memory recall from ambiguous cues. Previous memory networks considered inhibitory interactions to maintain a global balance between excitation and inhibition. In these models, inhibitory neurons provide stabilizing feedback but do not directly contribute to memory storage.[3–5] In contrast, experimental evidence suggests a role for inhibition beyond balancing excitation.[6–8]

We consider E-I rate networks with dynamics

$$\tau \dot{\mathbf{r}} \propto -\mathbf{r} + [\mathbf{W}\mathbf{r}]_+^{\circ n}, \quad [x]_+ = \max(x, 0), \tag{1}$$

$$\mathbf{r} = \begin{pmatrix} \mathbf{r}_E \\ \mathbf{r}_I \end{pmatrix}, \ \mathbf{W} = \begin{pmatrix} \mathbf{W}_{EE} & -\mathbf{W}_{EI} \\ \mathbf{W}_{IE} & -\mathbf{W}_{II} \end{pmatrix}, \tag{2}$$

where bold symbols are matrices and vectors, and $[\cdot]^{\circ n}$ denotes an element-wise power. For synapses w_{ij}^{+AB} connecting neurons i, j of type $A, B \in \{E, I\}$, we assume a synapse-type-specific competitive Hebbian learning rule[9] with a BCM-like plasticity threshold[10]

$$\dot{w}_{ij}^{+AB} \propto \left(r_i^A - \bar{r}_i^A\right) r_j^B, \quad \sum\nolimits_j w_{ij}^{+AB} = \omega_{AB}^{+} \tag{3}$$

where r are firing rates with means $\bar{r}$ and synapses compete for type-specific resource pools such that total synaptic weights ω_{AB}^{+} remain constant.

We assume that during memory storage, the network is dominated by memory patterns, $\mathbf{r}_A = \mathbf{p}$. Then recurrent weights become proportional to the row-normalized positive part of the pattern covariance matrix $\bar{\mathbf{C}}^{+}$, while synapses that connect negatively correlated neurons decay to zero, maintaining Dale's law. Negative correlations are instead captured by anti-Hebbian learning, corresponding to CP-AMPA-mediated plasticity at E-to-I synapses[11,12] (Fig. 1*A*):

$$\dot{w}_{ij}^{-IE} \propto -\left(r_i^I - \bar{r}_i^I\right) r_j^E, \quad \sum\nolimits_j w_{ij}^{-IE} = \omega_{IE}^{-} \tag{4}$$

After memory storage, recurrent weights become

$$\mathbf{W} = \mathbf{W}^{+} + \beta\, \mathbf{W}^{-} = \boldsymbol{\mathcal{W}}^{+} \otimes \bar{\mathbf{C}}^{+} + \beta\, \boldsymbol{\mathcal{W}}^{-} \otimes \bar{\mathbf{C}}^{-}, \tag{5}$$

where '$\otimes$' is the Kronecker product, and $\boldsymbol{\mathcal{W}}^{+}, \boldsymbol{\mathcal{W}}^{-}$ hold type-specific weight norms $\omega_{AB}^{+}, \omega_{IE}^{-}$. To simulate a change in the strength of anti-Hebbian synapses, we introduced the scalar factor β.

Under some simplifying assumptions, we can show analytically that stored patterns become autonomous fixed points of the E-I network. Numerically, the network converges to the closest pattern when initialized with a noisy memory cue (Fig. 1*B*; *C* & *D*, top). When the strength of anti-Hebbian weights is decreased, the model predicts decreased recall performance when cue noise is large (Fig. 1*C* & *D*, bottom; *E* & *F*), but maintained performance when cue noise is small (Fig. 1*E*). We tested these predictions in PV-Cre mice deprived of CP-AMPAR via bilateral (dorsal and ventral) injection of a GluA2 virus, which decreases anti-Hebbian plasticity (Fig. 1*G*). In a delay-dependent memory task (Fig. 1*H*), animal performance was only decreased when the delay between memory encoding and recall was long, while performance for shorter delays was similar

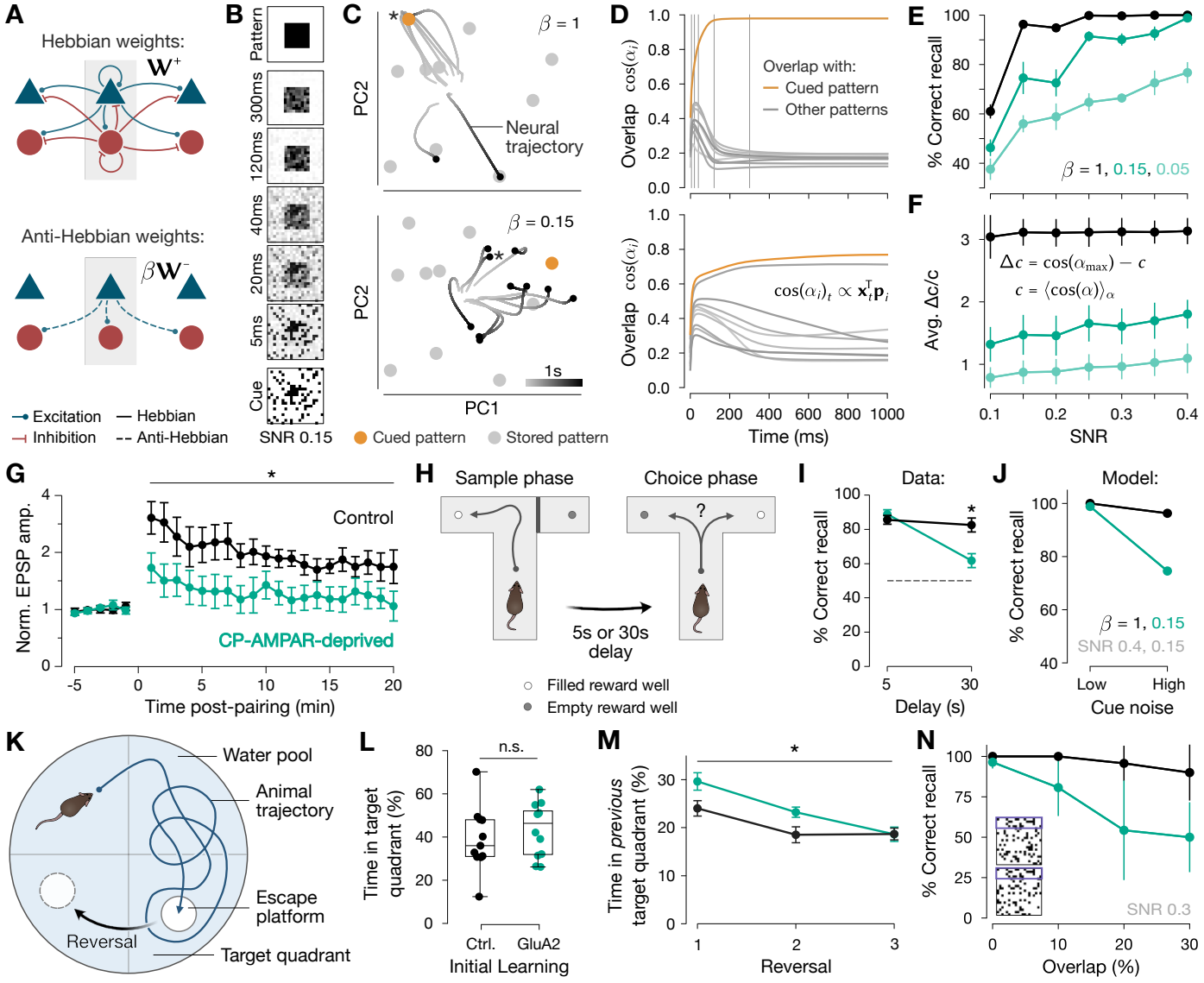

Figure 1. **(A)** Construction of E-I network. Only outgoing synapses from center E-I pair shown. **(B)** Neural activity $\mathbf{x}_E$ converging from noisy cue with low signal-to-noise ratio (SNR, bottom) to stored binary pattern (top) (cf. *D*, top). Pixels sorted to form a square for better visibility. **(C)** Neural trajectories $\mathbf{x}_E(t)$ in the principal pattern subspace. For large β, the network converged to stored patterns (top). For smaller β, recall accuracy decreased (bottom). Stars '$*$' indicate example trajectories in *D*. **(D)** Overlap (cosine similarity) of excitatory neural activities with cued (orange) and non-cued patterns (gray). Top: large β, bottom: small β (cf. *C*). Vertical lines mark sample points in *B*. **(E)** Average correct recall for different SNRs and β. Correct recall means the network has the largest overlap with the cued pattern after 1*s* (cf. *D*). Error bars indicate standard deviations over 7 randomly initialized networks, each storing 10 binary patterns. Each pattern was probed with 10 cues. **(F)** Average recall accuracy, quantified as the normalized distance between the largest and the average neural overlap with stored patterns after 1*s*. Error bars are standard deviations over recall cues ($n = 700$). **(G)** Anti-Hebbian plasticity recorded in CP-AMPAR-deprived parvalbumin-positive interneurons and controls ($p = 0.03$, mixed ANOVA). Error bars are ±SEM. **(H)** Spatial memory task. To get a reward in the choice phase, animals had to recall and choose the T-maze arm that was blocked in the preceding sampling phase. **(I)** Animal performance depended on the delay between sampling and choice phase ($p = 0.001$, mixed ANOVA). Dashed line (gray) indicates performance for random policy. Error bars are ±SEM. **(J)** Recall performance of a model network reflecting experimental results in *I*. Data replotted from *E*. Error bars are ±SEM. **(K)** In the Morris water maze mice had to find a platform that was submerged in opaque water. After some time, the platform location was reversed to a different quadrant. **(L)** Time spent in target quadrant after learning the initial platform location does not differ between CP-AMPAR-deprived (GluA2) and control animals ($p = 0.43$, one-way ANOVA). **(M)** Time spent in previous target quadrant ($p = 0.025$, mixed ANOVA). Error bars are ±SEM. **(N)** Network recall performance as a function of overlap between 2 out of 10 stored memory patterns. Average over 7 networks, 2 patterns per network, and 10 cues per pattern. Error bars are standard deviations over networks. Inset: two example patterns with 20% overlap.

to controls (Fig. 1*I*), matching our model's predictions (Fig. 1*J*). Similarly, in the water maze, CP-AMPAR-deprived mice performed indistinguishably from controls when learning the initial platform location (Fig. 1*L*). However, when the platform was repositioned, animals spent more time in the *previous* target quadrant compared to controls (Fig. 1*M*), indicating difficulty in separating similar memory patterns. This was reflected in our networks with decreased anti-Hebbian synaptic strength, where recall performance dropped with increasing overlap between memory patterns (Fig. 1*N*).

References: ***1***. J. J. Hopfield, *PNAS* (1982). ***2***. T. P. Vogels *et al.*, *Science* (2011). ***3***. P. E. Latham, S. Nirenberg, *Neural Comp.* (2004). ***4***. Y. Roudi, P. E. Latham, *PLoS CB* (2007). ***5***. W. Gerstner *et al.* (Cam. Univ. Press, 2014). ***6***. D. Dupret *et al.*, *Neuron* (2013). ***7***. S. McKenzie, *Hippocampus* (2018). ***8***. L. Topolnik, S. Tamboli, *Nature Rev. Neuro.* (2022). ***9***. S. Eckmann *et al.*, *PNAS* (2024). ***10***. E. L. Bienenstock *et al.*, *JoN* (1982). ***11***. K. P. Lamsa *et al.*, *Science* (2007). ***12***. F. Laezza *et al.*, *Science* (1999).