Fast adaptation of feedforward connections through BTSP supports similarity search and explains the repulsion effect

Yujie Wu and Wolfgang Maass†

Institute of Theoretical Computer Science, Graz University of Technology, Graz, Austria
†Corresponding author. E-mail: yujie.wu@igi.tugraz.at; maass@igi.tugraz.at

ABSTRACT

Recent experimental studies in the awake brain have identified the rule for synaptic plasticity that is instrumental for the creation of memory traces for episodic and conjunctive memories in area CA1 of the mammalian brain: behavioral time scale synaptic plasticity (BTSP). This rule differs in essential aspects from previously studied plasticity rules. But a learning theory for BTSP and a principled understanding of its implications on the ensemble of created memory traces are still missing. We introduce a simple mathematical model for BTSP and a theory that enables a principled understanding of the relations between memory traces it creates. In particular, we show that it surpasses random projections, which formed the basis for a previously considered biological algorithm, in the creation of memory traces that support similarity search. Furthermore, BTSP induces the repulsion effect for closely related memory items, a well-known fingerprint of memory organization in the human brain that could not be explained by previous models.

Introduction

The brain has to solve a really challenging algorithmic problem when it stores episodic memories: It has to allocate on-the-fly neurons that form a trace or tag for a new experience that can subsequently be reactivated with partial cues for that experience, i.e., which support similarity search. Recent experimental work has elucidated an essential ingredient of the algorithm which the brain uses for that: Behavioral time scale synaptic plasticity (BTSP) (Bittner et al., 2017; Magee and Grienberger, 2020; Milstein et al., 2021). In contrast to plasticity rules that have previously been studied in memory models, such as the Hebbian plasticity and STDP (spike-timing-dependent plasticity), BTSP does not depend on the firing activity of the postsynaptic neuron. Instead, traces for memories represented through firing activity in area CA3 are created in area CA1 through the synaptic connections from CA3 to CA1 and the plasticity processes that are gated by plateau potentials from another brain area, the entorhinal cortex (EC) (Grienberger and Magee, 2022). Furthermore, unlike most other plasticity rules used in memory models that have to rely on in-vitro studies or studies in the anaesthetized brain, BTSP has been validated through experiments in awake and behaving animals.

Other converging experimental data suggest that area CA1 of the hippocampus is the site where tags for episodic and conjunctive memories are first created in the mammalian brain (Sugar and Moser, 2019).

More precisely, BTSP differs in the following aspects from traditionally studied plasticity rules:

- BTSP does not depend on the firing of the postsynaptic neuron.

- It acts on the time-scale of seconds, a time-scale that is suitable for creating episodic and other forms of conjunctive memories that need to integrate temporally dispersed information.
- BTSP does not require dozens of repetitions of a protocol for the induction of synaptic plasticity, but is effective in a single or few trials, i.e., it provides a mechanism for one-shot learning.
- It is gated by synaptic inputs which the CA1 neurons receive from EC. This gating input appears to be largely stochastic (Grienberger and Magee, 2022).
- The amount by which BTSP changes a synaptic weight depends primarily on the preceding value of this weight (Milstein et al., 2021).
- BTSP contains also a rather peculiar form of LTD, whereby especially large weights are subject to LTD (long-term depression).

A first mathematical model for BTSP and its impact on the induction of place cells was provided by (Milstein et al., 2021). There, and in the subsequent modelling studies of (Cone and Shouval, 2021; Li and Roxin, 2023) for the induction of place cells, BTSP was modelled through a differential equation with continuous time and weights. Similarly as for neuron models, where the strongly simplified discrete model of (McCulloch and Pitts, 1943) enabled important analyses on a larger scale, a simplified rule for BTSP with binary weights and discrete time is also desirable for BTSP for analyses of its impact on larger memory systems. A first such rule had been proposed, based on experimental data for connections to CA1, in a preliminary version of the present study in (Wu and Maass, 2023). Subsequently, a similar rule was also proposed based on new experimental data on BTSP for recurrent connections within area CA3 (Li et al., 2023).

We analyze the impact of the simplified version of the BTSP rule from (Wu and Maass, 2023) in a 2-layer neural network model that consists of a population of input neurons which are randomly connected to a population of output neurons, that models the population of pyramidal cells in area CA1. Since anatomical data show that lateral connections between pyramidal cells in CA1 are sparse (Andersen et al., 2007), we did not include lateral connections in the output layer of the model. Note that this model only addresses the system of memory traces that are created through the application of BTSP to the feedforward synaptic connections from area CA3 to area CA1(Bittner et al., 2017; Grienberger and Magee, 2022), not to synaptic connections within a recurrently connected network such as CA3 (Li et al., 2023). We analyze the relation between memory traces, also referred to as memory tags, that BTSP creates in the output population for a large number of memory items that are sequentially presented in the input population. We focus on the question of to what extent the resulting system of memory traces supports similarity search. The experimental data on BTSP (Bittner et al., 2017) suggest that the memory traces that are created through BTSP in CA1 support similarity search, since places that are close to the site of the induction of a memory trace in CA1 tend to activate similar sets of CA1 neurons. Therefore, we examine in our neural network model the question of to what extent a partial cue, where a substantial fraction of the original memory item in the input population has been masked, activates a set of neurons in the output population that is similar to the memory trace for the original memory item. Whereas most prior analytical studies of memory models focused on the case where the memory items are orthogonal to each other, we also analyze the case where different memory items have substantial overlap, since this appears to be the case for different episodic memories that share common components. The theory that we have developed predicts that BTSP supports similarity search, also for large numbers of non-orthogonal memory items. Results of numerical simulations for a smaller scale agree very well with the predictions of this theory.

A biological algorithm for solving a somewhat related problem, the creation of succinct representations of odors by Kenyon cells in the fly brain provides a useful reference point (Dasgupta et al., 2017). Like BTSP it relies on a stochastic component, which defines in the fly brain a random projection of sensory inputs through a fixed random assignment of synaptic weights.

However, the mammalian brain cannot use the same projection method, since it requires that input patterns are projected into a much higher dimensional space. This condition is fulfilled by the olfactory system in the fly brain, since there are approximately 40 times more Kenyon cells than projection neurons which project odor information to Kenyon cells. In contrast, the number of neurons in the mammalian area CA1 is at most by a factor of 1.5 larger than the number of neurons (in area CA3 and EC) that project synaptic input into CA1. This holds true for both the rodent and the human brain (see Table 3-1 in (Andersen et al., 2007)). Furthermore, the algorithm of (Dasgupta et al., 2017) requires a winner-take-all (WTA) computation within the pool of Kenyon cells, in addition to a fixed random projection. However, such global WTA computations are not plausible for CA1, since it is so much larger than the pool of Kenyon cells in the fly: It consists of 0.39 to 14 million neurons in the mammalian brain (Andersen et al., 2007). Hence, evolution had to find a different algorithm for storing memory traces in area CA1.

We show that BTSP provides an essential new component to the algorithm in the mammalian hippocampus: It combines the stochasticity of weight assignments, that is implemented in the hippocampus through random plateau potentials according to the data from (Grienberger and Magee, 2022), with data-dependent fast synaptic plasticity. It turns out that on this basis, no WTA computations are required in the projection target. Furthermore, this online biological learning algorithm surpasses the performance of another prominent method for the locality-sensitive hashing goal that is discussed in (Dasgupta et al., 2017): the use of fixed random projections (RP) without WTA computations.

Finally, we wondered whether this biologically more realistic rule for synaptic plasticity in the hippocampus could also explain another salient functional property of the assignment of memory traces that the mammalian brain uses: The repulsion effect. This effect has been demonstrated in numerous studies of the human memory system (Chanales et al., 2017; Wanjia et al., 2021; Zheng et al., 2021). It denotes the effect that the human brain creates more dissimilar traces for related memory items than for unrelated memory items. This effect has been conjectured to support differential handling of closely related memories, a cornerstone of higher cognitive function. We show that BTSP automatically produces the repulsion effect.

Results

A transparent mathematical formula captures core features of experimental data on BTSP. Memory tags, also referred to as memory traces, are assigned through synaptic connections to pyramidal cells in areas CA1 to neural representations of salient experiences that are encoded through neuronal firing in areas CA3 and EC, see Fig. 1A for a schematic drawing of such a pyramidal cell. Experimental data from (Bittner et al., 2015, 2017; Grienberger and Magee, 2022) show that the weights of these synaptic connections are governed by BTSP. In particular, weight updates do not depend on the firing of the CA1 cell, but on the presence of a plateau potential in its apical dendrites that lasts for several seconds, and is triggered by synaptic input from layer 3 of the entorhinal cortex (EC3). Such plateau potential opens the plasticity window for several seconds before and after its onset (= time 0 in Fig. 1B). Experimental data on the resulting weight change, in particular on the decision between LTP (long-term potentiation) and LTD are summarized in Fig. 3E of (Milstein et al., 2021). These data are redrawn here in Fig. 1B for the convenience of the reader. One sees that the resulting weight change depends on the temporal difference between the time when the synaptic input arrives and the onset of the plateau potential, and the prior strength of this synapse. The black curve shows that LTP occurs for synapses whose prior strength was very small, provided that the synaptic input arrives within the interval from -3 to 2s relative to the onset of the plateau potential. Outside of this time interval the weights of these synapses remain largely unchanged. The orange curve shows that LTD is applied to synapses that previously had maximal synaptic strength, provided that the synaptic input arrives within the windows from -6 to -3s or from 2 to 4s relative to the onset of the plateau potential. If the synaptic input arrives in between, the resulting weight change is negligible. Thus the dependence of the decision between LTP and LTD on the relative timing between the arrival of the synaptic input pattern and the onset of the plateau potential can be summarized by

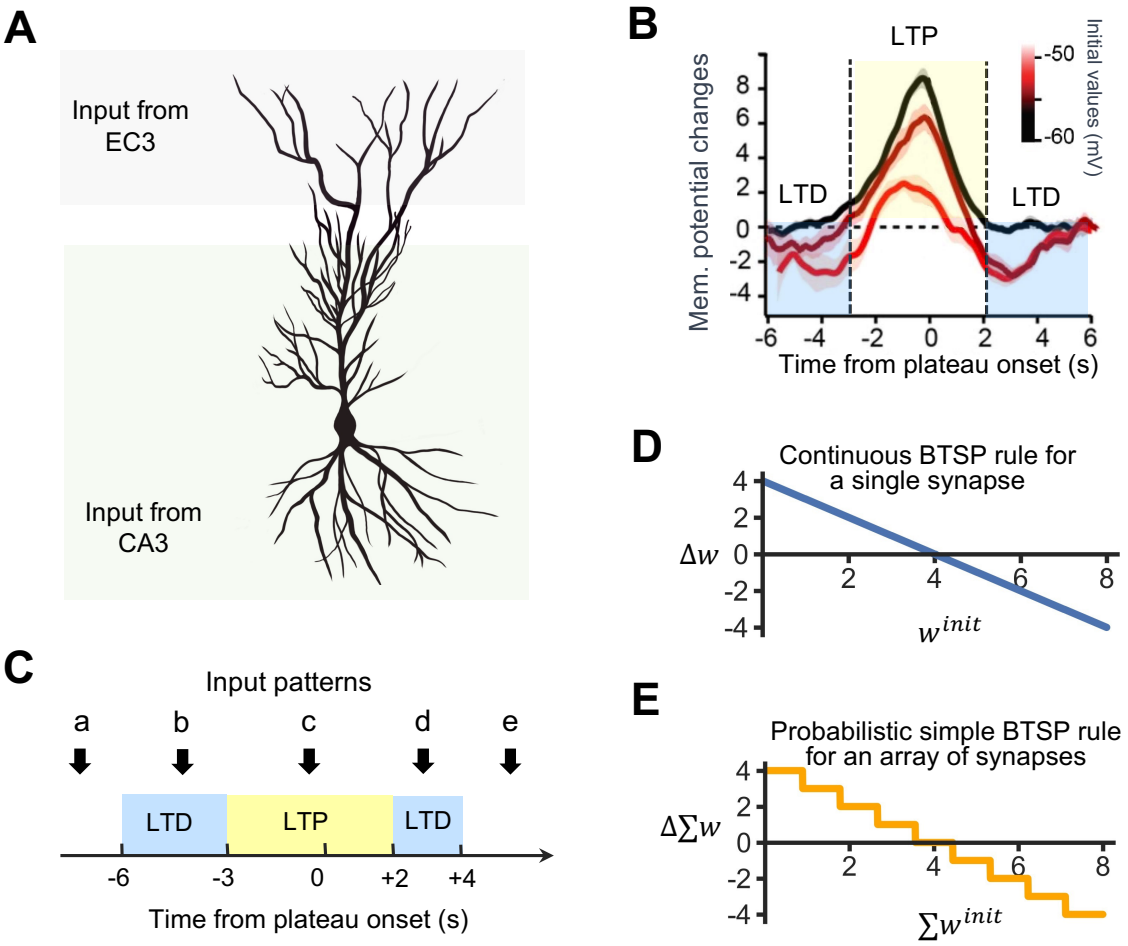

Figure 1. Experimental data and simplified rules for BTSP. (**A**) A CA1 pyramidal cell gets synaptic inputs from two different sources: One stream of synaptic inputs arrives in distal parts of apical dendrites from layer 3 of the entorhinal cortex (EC3) and induces plateau potentials according to (Grienberger and Magee, 2022). Another stream arrives from area CA3 at more proximal synapses, see Fig.5-1 in (Andersen et al., 2007). (**B**) According to the experimental data from Fig.3E in (Milstein et al., 2021) the change of synaptic weights through BTSP depends on the initial weight value and the time of presynaptic firing relative to the onset of a plateau potential (= time 0), Resulting weight changes are shown for 3 different ranges of the initial synaptic strength (black: maximal initial strength, orange: minimal initial strength, dark red: intermediate initial strength). Synaptic strength was measured in terms of the impact of synaptic input on the postsynaptic membrane potential. This panel is redrawn from Fig.3E of (Milstein et al., 2021). (**C**) Resulting scheme for the decision between LTP and LTD. According to the experimental data reported in Fig. 3 E of (Milstein et al., 2021) only input patterns that arrive up to 6s before or at most 4s after the initiation of a plateau potential in the CA1 neuron trigger synaptic plasticity through BTSP. Hence BTSP would apply in this drawing to patterns b, c and d, but not to patterns a or e. (**D**) Weight changes in dependence of the initial weight for a more complex model of BTSP according to Eq.1 in (Milstein et al., 2021), where the initial weight can assume continuous values. We consider here the case where the input patterns arrive at the same time as the plateau potential. (**E**) Demonstration that a simpler version of the BTSP rule for binary weights according to Eq.1 produces similar changes to the sum of binary weights when applied to an array of 8 synaptic connections between a presynaptic neuron and the CA1 neuron.

the simple scheme of Fig. 1C. We focus on the case of binary weights, which is defined by the black and orange curves in Fig. 1B, to make an analytical understanding substantially simpler. On the side we would like to point out that the dark red 3rd curve in that figure, which shows weight changes for the case of intermediate initial weight values, indicates a similar dependence of the decision between LTP and LTD on time relative to the onset of the plateau potential. These considerations motivate the simple rule for BTSP, which we will use as the default rule in this paper:

$$\Delta w_i = \begin{cases} +1 & \text{if } x_i = 1 \text{ and } w_i = 0, \\ -1 & \text{if } x_i = 1 \text{ and } w_i = 1, \\ 0 & \text{else.} \end{cases} \left.\vphantom{\begin{matrix}1\\1\\1\end{matrix}}\right\} \begin{matrix}\text{with probability 0.5, provided there is a plateau potential with} \\ \text{onset between -6 and 4s relative to the arrival time of } \boldsymbol{x},\end{matrix} \tag{1}$$

The probability 0.5 occurs in this rule because the experimental data show that if the arrival time of the input pattern $\boldsymbol{x}$ falls into a time interval from -6 to 4s relative to the onset of a plateau potential, it falls with probability 0.5 into the interior window from -3 to 2s where LTP is applied, and with the same probability into the outer time windows from -6 to -3s or from 2 to 4s, where LTD is applied. One can assume here for simplicity that all synaptic inputs from the input pattern $\boldsymbol{x}$ arrive at the same time at the CA1 neuron. But the results remain virtually unchanged if one allows that the arrival times of the synaptic inputs of the pattern $\boldsymbol{x}$ are spread out uniformly and randomly over the salient interval from -6 to 4s so that the decision between LTP and LTD is drawn independently for each synapse i with probability 0.5. This is the standard case that we consider in our theoretical analysis and numerical simulations.

One can further simplify this BTSP rule by observing that the previously described probabilistic version of Eq.1 yields practically the same learning results when one deletes the dependence of a weight change on a drawing with probability 0.5, and instead reduces the probability of a plateau potential by this factor. In this way, one arrives at the following deterministic version of the simple BTSP rule. That had first been proposed in (Wu and Maass, 2023):

$$\Delta w_i = \begin{cases} +1 & \text{if } x_i = 1 \text{ and } w_i = 0, \\ -1 & \text{if } x_i = 1 \text{ and } w_i = 1, \\ 0 & \text{else.} \end{cases} \left.\vphantom{\begin{matrix}1\\1\\1\end{matrix}}\right\} \begin{matrix}\text{provided there is a plateau potential with} \\ \text{onset between -6 and 4s relative to the arrival time of } \boldsymbol{x},\end{matrix} \tag{2}$$

We show in Methods and Fig. S2 that this deterministic version of the BTSP rule yields practically the same results, both theoretically and in numerical experiments as the probabilistic version of Eq.1.

We show in Fig. 1D and E that the impact of a more complex rule for BTSP with continuous weight values, such as the rule proposed in (Milstein et al., 2021), can also be modelled by the previously described rules with binary synaptic weights: One just had to assume that the two neurons are connected by several synapses, each of which has just binary weights. One should note that experimental data provide in general no information about the number of synaptic connections between two neurons, and leave open whether seemingly continuously-valued synaptic weights are caused by multiple synapses with binary weights as indicated in Fig. 1E. But in any case, a restriction to the case of binary weights tremendously simplifies the theoretical analysis of BTSP.

Predictions of a BTSP learning theory for the organization of memory traces in area CA1. Converging experimental data suggest that memory traces for complex memory items, including episodic memories, are created in area CA1 (Sugar and Moser, 2019) through the application of BTSP to synaptic connections from areas CA3 and EC to CA1. We want to understand the fundamental properties of the resulting ensemble of memory traces in CA1 on the basis of the previously described simple mathematical model for BTSP. Since lateral connections between pyramidal cells in area CA1 are very sparse, we model this pool of target neurons by a pool of n disconnected McCulloch-Pitts neurons. Each neuron compares its weighted sum of inputs to a firing threshold and outputs a value of 1 if the sum exceeds the threshold, otherwise it outputs 0. We consider a setting where during the learning

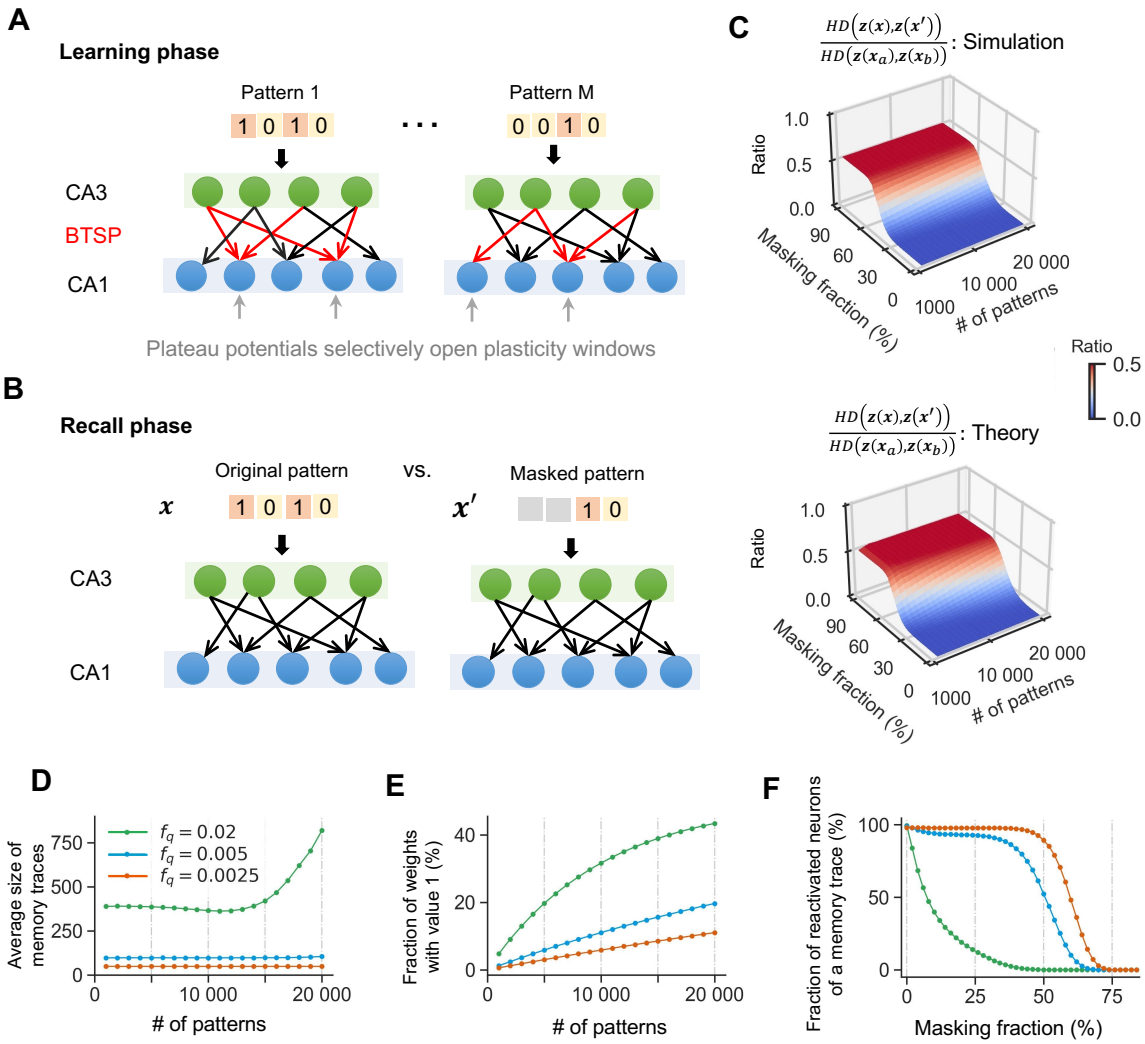

Figure 2. Structure and functional properties of memory traces created through BTSP in area CA1. (**A**) Schematic illustration of the model for the creation of memory traces (tags) in area CA1 through BTSP. During the learning phase, synaptic connections between neurons in area CA3 and CA1 are subject to BTSP for a sequence of M memory items, which are modelled as binary activity patterns in CA3 that are each presented once (one-shot learning). For each memory item in area CA3, a random set of CA1 neurons receives a plateau potential, which opens for this memory item the plasticity window for synaptic connections from CA3 to CA1. (**B**) During the recall phase, where no further learning takes place, partially masked versions $\boldsymbol{x}'$ of previously learned memory items $\boldsymbol{x}$ are presented where some input bits are masked, i.e., set to 0. (**C**) The HD between memory traces for original and partially masked memory items is divided by the average HD between memory traces for different memory items $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$ from the M memory items that had been presented during learning. We refer to this ratio as the relative dissimilarity of the traces for $\boldsymbol{x}$ and $\boldsymbol{x}'$. Relative dissimilarity remains small for masking fractions up to about 33%. This relative dissimilarity depends in the range of up to 20,000 very little on the total number M of memory items that had been presented during learning. Results of numerical simulations agree very well with theoretical results in the lower panel, which were obtained from Eqs 21 and 22. (**D**) Average size of memory traces (= number of CA1 neurons that fire during recall with the full memory item) as a function of the total number M of memory items that have been learnt. The resulting size depends substantially on the assumed probability f_q of a plateau potential in CA1 neurons during the learning of a memory item. (**E**) Fraction of weight value of 1 as a function of the total number M of memory items that have been learnt. (**F**) Fraction of a memory trace that gets reactivated with a partial cue, plotted as a function of the masking fraction.

phase M memory items $\boldsymbol{x}$, each represented by a string of m bits, are sequentially presented to the input layer of the network, and BTSP is applied once for each memory item to the synaptic connections from the pool of input neurons to the pool of target neurons that models area CA1, see Fig. 2A. This sequence of memory items could not only represent a sequence of spatial locations as in the rodent experiments, but also a sequence of salient experiences that arise during a day in the life of a human.

Our objective is to understand essential properties of the map from binary input patterns $\boldsymbol{x}$ to binary output patterns $\boldsymbol{z}(\boldsymbol{x})$ that is created by BTSP during the learning phase, see Fig. 2B. We usually refer to these input patterns as memory items, and to the output patterns as traces for these memory items. Of particular interest is the similarity of $\boldsymbol{z}(\boldsymbol{x}')$ to $\boldsymbol{z}(\boldsymbol{x})$ for the case where $\boldsymbol{x}'$ is a variation of one of the M input patterns $\boldsymbol{x}$, to which BTSP had been applied during the learning phase. We focus on the case where a substantial fraction of the bits of $\boldsymbol{x}$ has been masked in $\boldsymbol{x}'$, i.e., replaced by 0. We measure the dissimilarity between the resulting output patterns through the Hamming distance (HD) between $\boldsymbol{z}(\boldsymbol{x}')$ and $\boldsymbol{z}(\boldsymbol{x})$, i.e., through the number of neurons in CA1 that belongs to one of the two memory traces but not to the other. Since the absolute value of this HD depends on various other parameters, we consider instead the relative dissimilarity between $\boldsymbol{z}(\boldsymbol{x}')$ and $\boldsymbol{z}(\boldsymbol{x})$. This relative dissimilarity is calculated by dividing the absolute HD by the average HD between the traces $\boldsymbol{z}(\boldsymbol{x}_a)$ and $\boldsymbol{z}(\boldsymbol{x}_b)$, where $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$ represent different memory items from the sequence of M input patterns. We propose that this relative dissimilarity provides a suitable notion of the dissimilarity of two memory traces in CA1 from the perspective of downstream neural systems.

Of particular importance is the dependence of this relative dissimilarity on the fraction of bits of $\boldsymbol{x}$ that have been masked in $\boldsymbol{x}'$. It is worth noting that both HD values need to be analyzed during the recall phase after all input patterns in the sequence have been learned. This consideration is essential since these HD values are prone to continuous changes during the processing of that sequence through BTSP, making the theoretical analysis quite non-trivial.

Fig. 2C depicts the resulting relative dissimilarity obtained after completing the learning phase for different numbers of memory items, represented by the variable M, in the input sequence to the network. This is shown both as results of numerical simulations and as results of the learning theory for BTSP that we have developed. One sees that both results agree very well and that the relative dissimilarity is very small across the range of values for M when no more than one-third of the bits are masked in the recall cue.

Note that 0.5 is a critical value for the relative dissimilarity, as exceeding this threshold makes it unlikely to determine the original memory item from which the masked recall cue $\boldsymbol{x}'$ was created based on the memory traces.

The dissimilarity measurements depend also on the bit-length m of input patterns and the bit-length n of output patterns. The parameter m models the number of pyramidal cells in area CA3, which is estimated to be around 0.25 million for rats and around 2.83 million in the human brain. The parameter n models the number of pyramidal cells in area CA1, which is estimated to be around 0.39 million for rats and around 14 million in the human brain. Since the size of these biological values makes extensive numerical analyses cumbersome, we have focused on values that are down-scaled roughly by a factor of 1/10 from the size of these structures in the rat brain, and by a factor of 1/100 from the sizes in the human brain. These yields are default values in the next analyses. We will discuss in a later section theoretical estimates of the functional impact when these default values are scaled up to those of the brain. Another important parameter is the probability that a neuron in the input layer is connected to a neuron in the output layer. Since CA1 pyramidal cells are estimated to around 30,000 dendritic spines according to chapter 5 of (Andersen et al., 2007), we assume in our reduced-size model that a neuron in the output layer receives on average synaptic inputs from a random set of 15,000 neurons in the input layer, indicating a connection probability of $f_w = 0.6$. Also the fraction f_p of bits that assume value 1 in an input pattern (corresponding to the fraction of neurons in CA3 that fire during the presentation of a memory item) is relevant. This fraction was estimated by (Guzman et al., 2016) to have a value $f_p = 0.005$, which we use as the default value. Since we are not aware of reliable experimental data on the firing threshold of pyramidal cells in CA1, we used grid search to set the firing

thresholds of CA1 neurons at a value that optimally supports recall from partial cues (see *Methods*).

Of particular interest is the impact of parameters that govern the generation of plateau potentials, since they have no analogue in previously considered rules for synaptic plasticity. The experimental data of (Grienberger and Magee, 2022) suggest that one can model the generation of plateau potentials in CA1 neurons by a stochastic process, whereby CA1 neurons receive independently with probability f_q a plateau potential during the presentation of a memory item $\boldsymbol{x}$ during learning. Furthermore, the experimental data shown in their Fig.4G suggest that 0.005 is a reasonable estimate for f_q. We show in Fig. 2D-F that the value of f_q has a significant impact on the size of memory traces that are created during learning, and also on the fraction of weights with value 1 that emerge during the learning phase. Fig. 2F shows that the value of f_q also has a significant impact on the recall of memory traces with partial cues. This impact hinges on the fact that plateau potentials not only cause LTP, but also LTD if the current weight is already large. Therefore a larger value of f_q has the effect that more weights that were first set to 1 when BTSP was applied to a memory item during learning become with higher probability subject to LTD during the application of BTSP to subsequent memory items in the learned sequence. This effect makes the memory trace less stable, and the neurons in $\boldsymbol{z}(\boldsymbol{x})$ become harder to reactivate with partial cues. But the relatively low value 0.005 that is suggested by the data of (Grienberger and Magee, 2022) entails that most neurons in the output layer can be reactivated even when a third of the bits have been masked, as already shown in terms of relative dissimilarity in Fig. 2C.

The impact of the application of BTSP to other memory items that occur later in the sequence of M items makes a theoretical analysis of the memory trace for a given memory item $\boldsymbol{x}$ nontrivial. In particular, one needs to take into account that weights which have been increased during the creation of a memory trace for $\boldsymbol{x}$ through BTSP may become later subject to LTD when a subsequently processed memory item shares presynaptic neurons and some of the neurons in its memory trace receive another plateau potential. Therefore previously developed analytical tools for the analysis of the resulting ensemble of memory traces with synaptic plasticity rules that do not have the same strong dependence of the weight update on the preceding weight value, such as those in (Hopfield, 1982, 2007; Widrich et al., 2020), can not be used in an analysis of the impact of BTSP. Instead, tools from combinatorial mathematics and probability theory were needed to create a learning theory for BTSP. The two panels of Fig. 2C show that the resulting theory is very successful in predicting the outcome of numerical simulations. An additional demand on a learning theory for BTSP arises from the content of the memory items for which traces are created through BTSP in area CA1: Episodic memories, or more generally conjunctive memories (Bittner et al., 2015), where heterogeneous pieces of information are integrated into a single memory trace. For recall of an episodic memory, not just a few percent, but substantial parts of the original input pattern may be missing in the recall cue. Furthermore, the content of different episodic memories is not necessarily orthogonal because they may contain common components, such as a common spatial position. Both of these features, which are not covered by traditional memory theories in neural networks, need to be addressed in a learning theory for BTSP.

Distinguishing fundamental properties of the system of memory traces that BTSP creates with those of other memory models. Since classical Hopfield networks (HFNs) and RP are arguably the most commonly considered models for memory organization, we compare the results of BTSP from Fig. 2C with corresponding measurements for these other memory models. We constructed the HFN and its synaptic weights by following a recent study(Sharma et al., 2022), using the same connection probability $f_w = 0.6$ as for the BTSP and the same input patterns. For the HFN we write $\boldsymbol{z}(\boldsymbol{x})$ for the set of neurons that are activated by a cue $\boldsymbol{x}$ after 100 synchronous update steps of the HFN after the presentation of the cue. In order to arrive at a fair comparison, we used for the RP the same fraction of weights with value 1 in the RP as that is created by BTSP after learning 20,000 patterns (about 20%).

The first row of Fig. 3, panels A-C, shows the resulting Hamming distance $HD(\boldsymbol{z}(\boldsymbol{x}), \boldsymbol{z}(\boldsymbol{x}'))$ between the trace for a memory item $\boldsymbol{x}$ that occurred during learning, and the trace for a variation $\boldsymbol{x}'$ where a random subset of 33% of

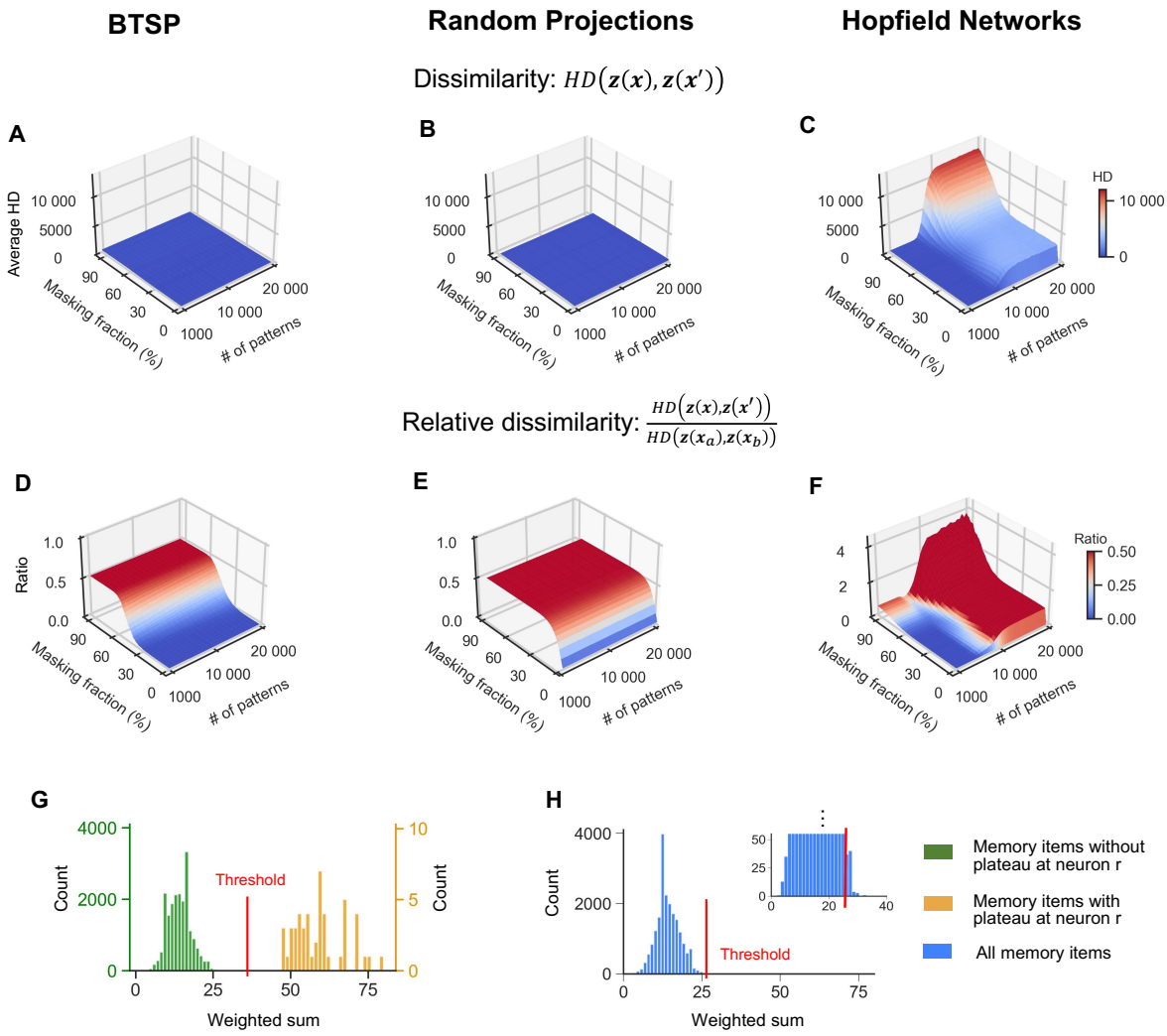

Figure 3. Recall with partial cues from the memory structure created through BTSP, a random projection (RP) and a classical Hopfield network(HFN). Panels (**A**), (**B**) and (**C**) show that the dissimilarity between the original memory trace and the memory trace recalled by a partial cue is significantly smaller for BTSP and RP compared to a HFN, particularly when the number of stored patterns is above 10,000. Panels (**D**), (**E**) and (**F**) support the same conclusion for relative dissimilarity. Furthermore, panels D and E show that the relative dissimilarity remains in the case of BTSP small for masking fractions up to 33%, whereas it approaches the RP already for small masking fractions the value 0.5, where one can no longer distinguish to which memory item the trace for the masked cue belongs. Shown are the results of simulations, averaged over 10 runs with new random generation of memory items and plateau potentials. Grid search was used to optimize the firing threshold of neurons in each of the three models (see *Methods*). (**G**) The distribution of inputs (= weighted sums) to a sample CA1 neuron for all 20,000 memory items that occurred during learning with BTSP. Memory items to which the neuron belongs and BTSP was applied are highlighted in yellow, while other memory items are represented in green. Since the threshold can maintain a significant distance from values of the weighted sum for memory items to which it belongs (for the case of a complete cue), the neuron still fires when a substantial fraction of the bits in the memory items are masked. The neuron belongs to the trace for 98 memory items. (**H**) The distribution of inputs to a sample CA1 neuron for all 20,000 memory items by the RP. The red line marks the optimal value of the neuron threshold, as determined through grid search. One sees both in the full-scale plot and in the inset that moving the threshold to a lower value would result in the inclusion of this neuron in a significantly larger number of memory traces. At the shown value, the neuron belongs to the trace for 102 memory items.

1's has been masked. One sees that this HD is substantially larger in the HFN for larger numbers of patterns. The relative dissimilarities, depicted in the second row of Fig. 3, show that this also holds for the relative similarity measure. In addition, a comparison of panels D and E shows that the memory traces that are created through BTSP tolerate a substantially larger masking fraction during recall. This holds in spite of the fact that the firing thresholds of neurons in the output layer had been optimized so that RP and BTSP provide optimal performance for recall with partial cues (see *Methods*).

This performance difference can be explained through the analysis of the distribution of weighted sums of inputs to a sample CA1 neuron for 20,000 memory items that occurred during learning, shown in panels G and H of Fig. 3. Since BTSP creates, in contrast to RP, memory traces for a particular set of memory items, it is able to make the synaptic input at a CA1 neuron substantially larger for those memory items where it belongs to their memory traces than for other memory items. Therefore, the firing threshold of a CA1 neuron can be chosen so that it clearly separates these two sets of synaptic inputs. In particular, it can be chosen to be substantially smaller than the weighted sums for those memory items where it belongs to their memory trace. This enables the neuron to fire for these memory items also when a substantial number of 1's of the memory item has been masked, i.e., if a fraction of the terms in the weighted sum is missing. In contrast, for the case of a RP there is no clear separation of the sizes of synaptic inputs for memory items where this neuron belongs to their trace versus synaptic inputs that are caused by other memory items, see Fig. 3H. In particular, many of the latter cause a synaptic input that lies just below the threshold, no matter how it is chosen.

Hence there is no safety margin that would allow us to lower the threshold, in order to achieve that the neuron also fires for partially masked versions of memory items to whose trace it belongs, without making the neuron part of the memory trace for many further memory items (but again not for masked versions of them). To demonstrate the ubiquity of this fundamental difference between the distribution of inputs to CA1 neurons in the case of BTSP and a random projection, we show in Figure S1 these distributions for three further randomly chosen CA1 neurons.

BTSP enables similarity search also for non-orthogonal memory items. Episodic and conjunctive memory involve the storage of patterns that may have common components, and are therefore in general not orthogonal. However, the orthogonality of memory items is assumed in the theory of most memory models. Our theory also works for non-orthogonal memory items, and predicts, as depicted in the right column of Fig. 4, that BTSP is capable of creating memory structures that support similarity search also in this case. This prediction is further confirmed by numerical simulations, as shown in the left column of the figure. Increasing the number of common bits only leads to a moderate increase in the ratio $HD(\boldsymbol{z}(\boldsymbol{x}),\boldsymbol{z}(\boldsymbol{x}'))/HD(\boldsymbol{z}(\boldsymbol{x}_a),\boldsymbol{z}(\boldsymbol{x}_b))$, which is crucial for evaluating the ability to recall using partial cues. Note that in our default setup, memory items have an average of 125 bits with a value of 1.

Properties of memory traces created by BTSP at the scale of the brain. We have so far considered an input dimension of $m = 25,000$ and a dimension of $n = 39,000$ for the network output. According to (Andersen et al., 2007) the real size of area CA3 (the input structure) and of the output structure CA1 are in the rat brain 10 times larger, and in the human brain 100 times larger. Hence, the question arises of how the previously discussed results on the memory traces created through BTSP scale up to these larger sizes. Unfortunately, it is challenging to directly probe this scale with numerical experiments. The theory that we have developed can be applied to these larger scales. We show in Fig. 5 that the theory fits the results of numerical experiments very well for scaling factors between 0.05 and 1.5 of the previously considered network sizes: Compare the results in the left and middle columns. The right column shows absolute (first row) and relative dissimilarity (last row) for scales that are up to 100 times larger than in our default setting. We always assume here a masking fraction of 33% for partial cues. One sees that the relative dissimilarity stays small for up to 40,000 memory items in the largest model.

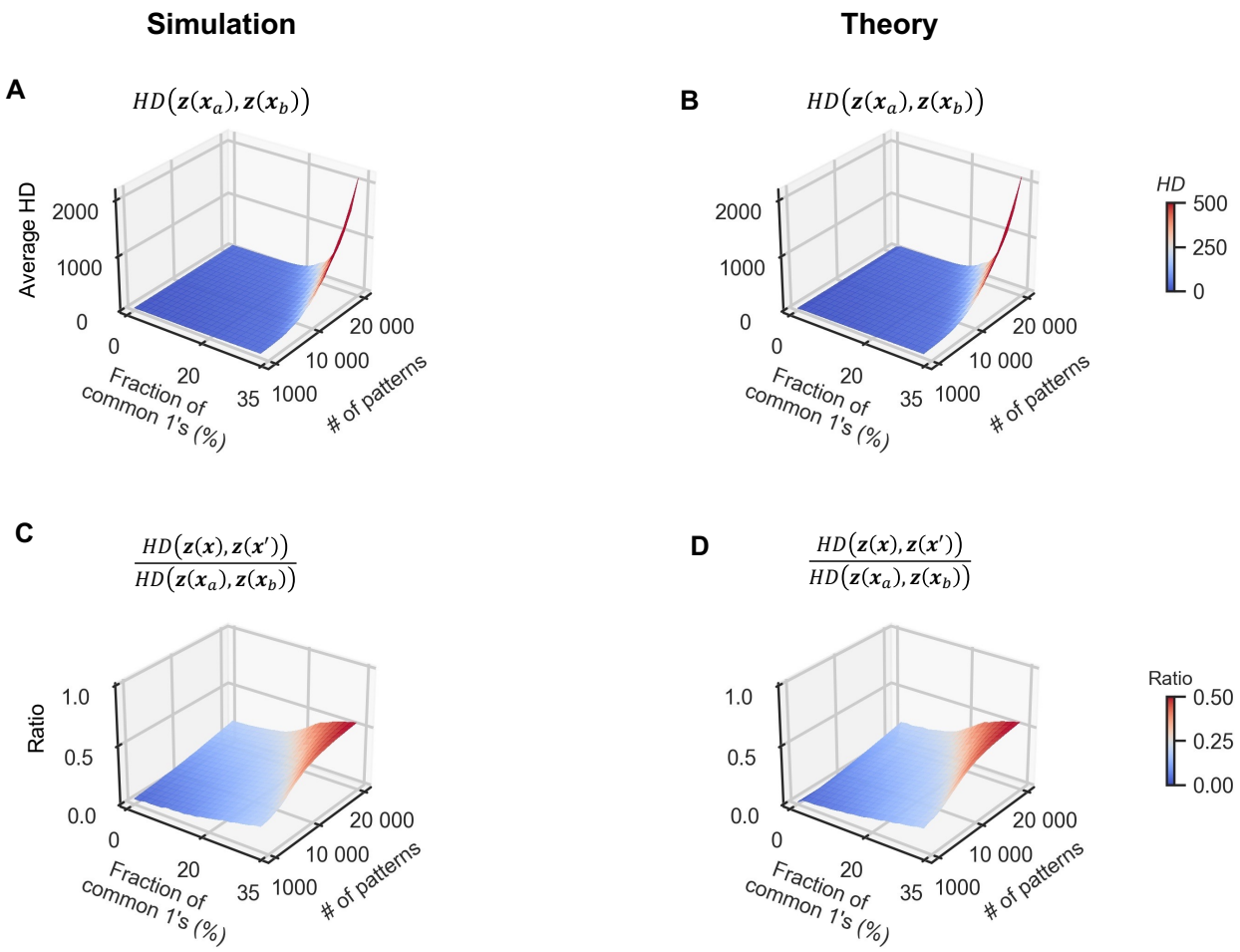

Figure 4. Recall with partial cues from memory traces created by BTSP for non-orthogonal memory items. Dissimilarity of memory traces for non-orthogonal memory items, analysed through numerical simulations (**A**) and theory (**B**). The fraction of common 1's refers to the average number of common 1's in any two input patterns divided by the average number of 1's in an input pattern. The theory predicts the results of numerical simulations very well. Relative dissimilarity for recall with partial cues (masking fraction 33%), analysed through numerical simulations (**C**) and theory (**D**). Theoretical results were obtained from Eqs.30 and 31. Both simulations and theory show that a fraction of 10 to 20% common 1's in different memory items has only a mild impact on recall with partial cues.

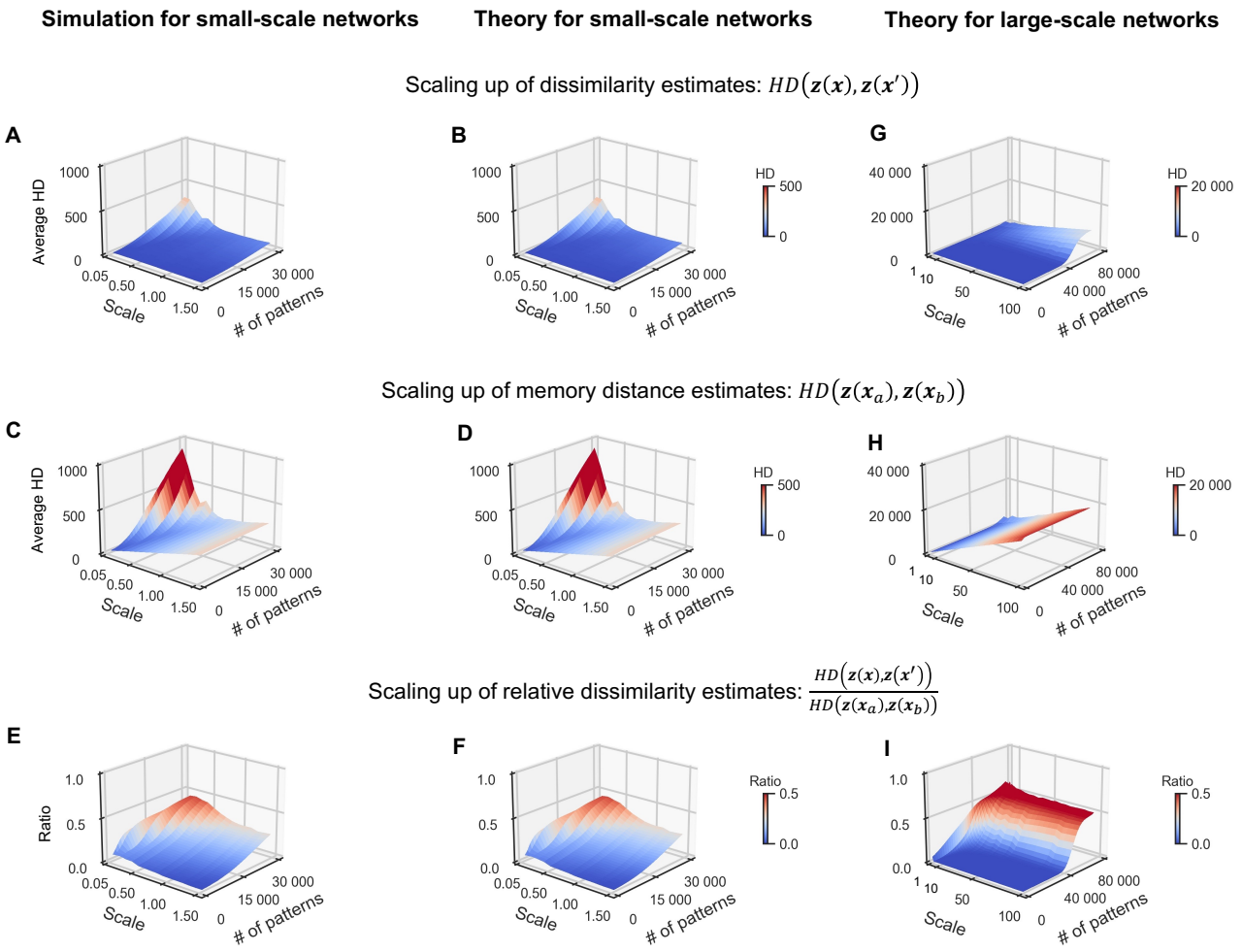

Figure 5. Fundamental properties of memory traces created through BTSP for neural networks at the scale of the brain. $HD(\boldsymbol{z}(\boldsymbol{x}), \boldsymbol{z}(\boldsymbol{x}'))$ refers to the average HD between the memory traces recalled by any complete pattern $\boldsymbol{x}$ or the corresponding masked pattern $\boldsymbol{x}'$ with 33% masking fraction. We scaled the input size m and output size n according to the scaling factor defined on the x-axis. Simulation (**A**) and theoretical prediction (**B**) of $HD(\boldsymbol{z}(\boldsymbol{x}), \boldsymbol{z}(\boldsymbol{x}'))$ at various scales that are still accessible for numerical simulations. Simulation (**C**) and theoretical prediction (**D**) of the dissimilarity between traces for different memory items at different scales. Simulation (**E**) and theoretical prediction (**F**) of relative dissimilarity for recall with partial cues at different scales. One sees in panels A-F that the theory predicts very well the outcome of numerical experiments at these smaller scales. (**G**) - (**I**): Predictions of the same measurements as in panels A-F at larger scales, up to the scale of the human brain (scaling factor 100).

BTSP induces a repulsion effect for closely related memory items. Experimental studies (Chanales et al., 2017; Wammes et al., 2022; Wanjia et al., 2021; Zheng et al., 2021) of the human brain, in particular of area CA1, have exhibited a somewhat surprising effect: Memory traces for closely related memory items are less similar than traces for unrelated memory items. This repulsion effect is thought to be important for the encoding of episodic memories, as it enables the separation of events that differ just in a few details, e.g. two scenarios that look alike but occurred on different days or lead to different outcomes. Previous memory models have not been able to explain this repulsion effect. In particular, a RP produces an opposite effect, where traces for related memory items have a larger overlap than for unrelated memory items, see Fig. 6A. Next, we demonstrate that BTSP, more specifically the LTD component of BTSP, is able to explain the repulsion effect in human memory.

Fig. 6B shows results of numerical simulations for a setup that reflects the experimental design in the studies of (Chanales et al., 2017; Wammes et al., 2022; Wanjia et al., 2021; Zheng et al., 2021): A sequence of M memory items are learned through BTSP, and these are pairwise unrelated except for two that have an overlap ratio of 40%. These two items may appear at any position in the sequence. We then study the overlap of their memory traces after learning has been completed for the sequence. We use here a definition of overlap that corresponds to the definition that was used in the experimental measurements from the human brain, where the Pearson correlation coefficient was applied to voxels of fMRI experiments: We calculate the overlap ratio of two memory items as the fraction of the number of overlapping 1's in the two memory traces, divided by the average number of 1's in a memory trace. Memory items without overlap were chosen as memory items without common 1's. The result for the BTSP model is very similar to the results from the human brain, see Fig. 3C of (Chanales et al., 2017).

The repulsion effect for memory traces that are generated by BTSP can also be explained theoretically. If a CA1 neuron receives plateau potentials for both memory items during learning, most of its weights from neurons that are active in both memory items will have a value 0 after learning: They are first increased and then decreased by BTSP. However, if the two memory items have largely non-overlapping sets of 1's, the LTD part of BTSP will only affect weights from those input bits that have value 1 in both input patterns. The neuron is therefore more likely to fire after learning for both input patterns. This effect increases the overlap ratio of traces for unrelated memory items.

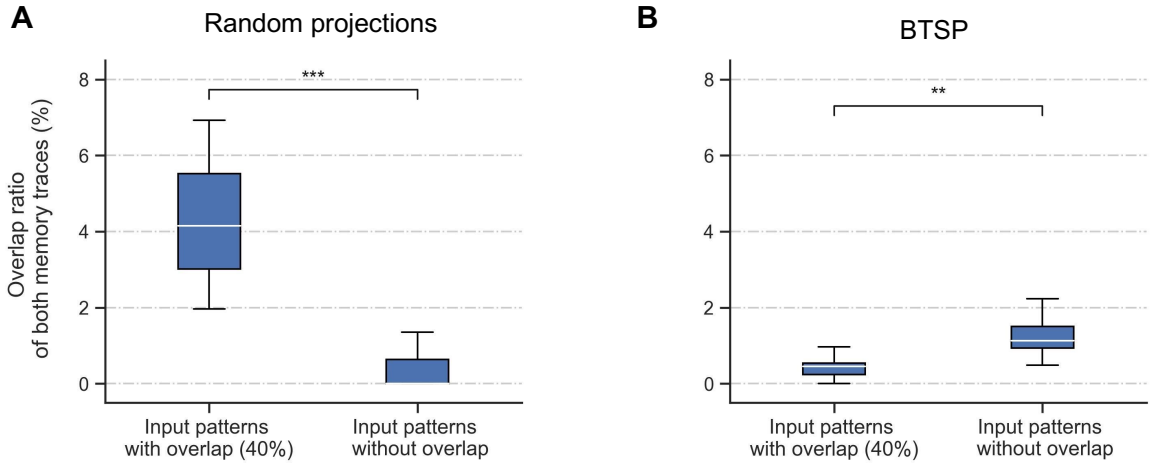

Figure 6. In contrast to a random projection, BTSP can explain the repulsion effect in human memory. (**A**) The overlap ratio of the memory traces for a pair of memory items with 40% overlap, when memory traces are created by a RP ($***p < 0.001$, two-tailed one-sample t-test), and for a pair of unrelated memory items. The overlap ratio is larger for the more similar memory items. (**B**) The same analysis is performed for the memory traces created by BTSP. The overlap ratio is smaller for more similar memory items ($**p = 0.002 < 0.01$, two-tailed one-sample t-test), as in experimental data from the human brain. The error bars represent in both panels the standard deviation over 20 trials.

Discussion

We have analyzed a biological algorithm for creating functionally useful memory tags that support similarity search. It is especially important for applications where data can not be projected into a much higher dimensional target space, and global computational operations in the target space, such as WTA operations, are not feasible. Both of these conditions were required for the fly algorithm of (Dasgupta et al., 2017), the arguably most prominent previously studied biological paradigm for the creation of memory tags. We have investigated the creation of memory tags in area CA1 of the mammalian hippocampus, where neither a projection of memory items from CA3 into a much higher dimensional space, nor global WTA operations in CA1 are available. We propose that the brain overcomes this hurdle by using an adaptive mechanism, rather than fixed synaptic weights as in the fly algorithm. It is implemented through an experimentally very well-supported rule for synaptic plasticity, BTSP, that differs in several essential properties from previously studied rules for synaptic plasticity in biological or artificial neural networks. We have proposed a simple mathematical model for BTSP that captures the core features of BTSP while being at the same time sufficiently simple to support the theoretical analysis. On this basis, we could establish a previously missing framework for understanding essential functional properties of the ensemble of memory traces that are created through BTSP in area CA1. We found that the predictions of this theoretical model agree very well with the results of numerical simulations. In particular, our theoretical analysis provides a principled understanding of an essential advantage of BTSP over previously studied synaptic plasticity rules: It creates traces for memory items so that cues where a substantial fraction of the content of a stored memory item is missing still lead to activation of a very similar set of neurons as the complete memory item. In addition, BTSP also supports the creation of well-separated tags for non-orthogonal memory items. Both of these properties appear to be essential for the function of memory traces in the brain, but have rarely been addressed in previous memory models.

We have shown that BTSP also induces another characteristic property of memory traces in the brain: the repulsion effect. This effect, which has been reported in numerous studies of the human brain such as (Chanales et al., 2017; Wammes et al., 2022; Wanjia et al., 2021; Zheng et al., 2021), refers to the peculiar property that traces for rather similar memory items are more dissimilar than for unrelated memory items. We have shown that the repulsion effect can be explained on the basis of a specific property of BTSP: The decision between potentiation and depression of a synaptic weight through BTSP depends on the current value of the synaptic weight (Milstein et al., 2021).

This is not the first theoretical study on BTSP. The role of BTSP in mouse experiments had been studied analytically in (Cone and Shouval, 2021). Their analysis focused on the assumption that the same CA1 neuron is subject to plateaus in repetitions of the same path (memory item) during treadmill experiments. In contrast, we have focused on the role of BTSP in the creation of memory traces through one-shot learning, where each memory item, such as the content of an episodic memory, is presented only once. In addition, our analysis is based on a scenario where plateau potentials that trigger BTSP occur very infrequently and highly randomly, in accordance with the data of (Grienberger and Magee, 2022). Another recent study (Li and Roxin, 2023), developed concurrently with our model but independently, focused on a more detailed continuous model for BTSP. Their research analyze the role of BTSP in recurrent networks, such as CA3, which is not addressed in our study. Altogether, it is clear that BTSP opens a new chapter in the study of memory models that requires besides further experimental investigations also a new generation of models for the creation of memory traces through synaptic plasticity.

Table of content for Methods

Table 1: Definitions and explanations of key parameters used in the theoretical analysis. 16

A learning theory for BTSP 17

Background and preparations for memory trace estimates 17

Estimates of properties of memory traces for orthogonal memory items 17

Estimates of firing probability of CA1 neurons after learning M orthogonal memory items 17

Estimates of expected HD between memory traces for the original and masked memory items 19

Estimates of expected HD between memory traces for different orthogonal memory items 20

Estimates of properties of memory traces for non-orthogonal memory items 21

Estimates of firing probability of CA1 neurons after learning M non-orthogonal memory items 21

Estimates of expected HD between memory traces for the original and masked memory items 23

Estimates of expected HD between memory traces for different non-orthogonal memory items 25

Details of experimental setup and simulations 25

Details of simulation of continuous weight changes by binary BTSP rule (Fig. 1D and E). 25

Details of analysis of the masking fraction (Figs. 2C-F) and common bits (Fig. 4) 26

Details of comparison with other memory models (Fig. 3). 26

Details of experimental setup of scale-up estimates (Fig. 5). 26

Details of analysis of the repulsion effect (Fig. 6) 27

Table 1. Definitions and explanations of key parameters used in the theoretical analysis.

Variables	Category	Explanations	Definitions
$f(*)$	General notations	The probability mass function of binomial distribution	
$F(*)$		The cumulative distribution function of binomial distribution	
$Q(\boldsymbol{x})$		The set, which collects the index of CA1 neurons that receives plateau potential for $\boldsymbol{x}$ during learning	
$supp(\boldsymbol{x})$		The set, which collects positions of 1's in $\boldsymbol{x}$	
$\iota(\boldsymbol{x})$		The size of $supp(\boldsymbol{x})$	
$S(j)$		The set, which collects the index of patterns for which the CA1 neuron j receives plateau potentials during learning	
$\boldsymbol{y}(j)$		The unit vector with only j-th bit position being 1	
c		The number of common 1's in different non-orthogonal memory items	
f_q		The probability of plateau potential	
f_p		The input density	
f_w		The probability of connections	
f_d		The fraction of 1's being masked to zeros	
p_{f1}(or p_{f2})	Estimates for orthogonal memory items	The conditional probability of a CA1 neuron firing in response to $\boldsymbol{x}$ in the recall phase, given that the CA1 neuron receives (or does not receive) plateau for $\boldsymbol{x}$ during learning	Eqs.15 and 16
p_e(or p_o)		The conditional probability that a synapse connecting 1's of $\boldsymbol{x}$ and a CA1 neuron remains in an active state during the recall phase, given that the CA1 neuron receives (or does not receive) a plateau potential for $\boldsymbol{x}$ during the learning phase.	Eqs.13 and 14
$p_{f_{\text{avg}}}$		The probability of a CA1 neuron firing in response to $\boldsymbol{x}$ after learning	Eq.17
$\hat{f}_p$		The probability of the remaining bits (excluding the common 1's) being 1	
c_w	The estimates for non-orthogonal memory items	The total number of active synapses connecting common 1's to a CA1 neuron in the recall phase	
$\hat{p}_e$(or $\hat{p}_o$)		The conditional probability that a synapse connecting 1's of $\boldsymbol{x}$ and a CA1 neuron keeps in an active state in the recall phase, given that the CA1 neuron receives (or does not receive) plateau for $\boldsymbol{x}$ during learning	Eqs.23 and 24
$\hat{p}_{f1}$(or $\hat{p}_{f2}$)		The conditional probability of a CA1 neuron firing in response to $\boldsymbol{x}$ in the recall phase, given that the CA1 neuron receives (or does not receive) plateau for $\boldsymbol{x}$ during learning	Eqs.25 and 26
$\hat{p}_{f_{\text{avg}}}$		The probability of a CA1 neuron firing in response to $\boldsymbol{x}$ in the recall phase	Eq.27
$\hat{p}_{c1}$(or $\hat{p}_{c2}$)		The conditional probability that CA1 neurons fire differently for $\boldsymbol{x}$ and $\boldsymbol{x}'$ in the recall phase, given that the CA1 neuron receives (or does not) plateau for $\boldsymbol{x}$ during learning	Eqs.28 and 29
$\hat{s}_j$		The number of weight updates of a synapse connecting common 1's to the CA1 neuron j during learning	

Methods

1 A learning theory for BTSP

1.1 Background and preparations for memory trace estimates

We will employ the properties of the binomial distribution for the memory trace estimates. A summary of the key parameters used in the following derivations is provided in Table 1. Additionally, we introduce the following Lemma which is useful in handling the parity of weight updates used in Eqs.13, 14, 23 and 24.

Lemma 1. *Let $f(k_x, N_x, p_x)$ be the probability of binomial distribution that a random variable X gets k_x success in N_x Bernoulli trials. Then, the probability that X is even can be given by*

$$P(X \text{ is even}) = \frac{1}{2}\big(1 + (1-2p_x)^{N_x}\big). \tag{3}$$

Similarly, the probability that X is odd can be given by

$$P(X \text{ is odd}) = \frac{1}{2}\big(1 - (1-2p_x)^{N_x}\big). \tag{4}$$

Proof. Recall that $f(k_x, N_x, p_x)$ represents the probability of a binomial distribution that a random variable X gets k_x successes in N_x Bernoulli trials. Note that

$$((1-p_x)+p_x)^{N_x} = \sum_{k_b=0}^{N_x} C_{N_x}^{k_b} p_x^{k_b} (1-p_x)^{N_x-k_b} = P(X=0) + P(X=1) + \cdots + P(X=N_x), \tag{5}$$

$$((1-p_x)-p_x)^{N_x} = \sum_{k_b=0}^{N_x} C_{N_x}^{k_b} (-p_x)^{k_b} (1-p_x)^{N_x-k_b} = P(X=0) - P(X=1) + \cdots + (-1)^{N_x} P(X=N_x). \tag{6}$$

Combining the two equations then gives

$$P(X \text{ is even}) = \frac{1}{2}\big(1 + (1-2p_x)^{N_x}\big), \tag{7}$$

$$P(X \text{ is odd}) = \frac{1}{2}\big(1 - (1-2p_x)^{N_x}\big). \tag{8}$$

□

1.2 Estimates of properties of memory traces for orthogonal memory items

Estimates of firing probability of CA1 neurons after learning M orthogonal memory items. To estimate the HD of memory traces in CA1, we need to estimate the probability that a CA1 neuron fires differently for different memory items. In this section, we derive a basic estimate called the firing probability, which represents the probability that a CA1 neuron fires in response to the memory item $\boldsymbol{x}$ after learning M patterns.

Given that the firing behaviour of a CA1 neuron is determined by whether the weighted sum of binary inputs is greater than a firing threshold v_{th}, the firing probability depends on the number of weights that are in the active status (i.e., $w_{ji} = 1$) after the learning is completed. On this basis, we start with the modelling of the effect of BTSP on the active state of w_{ji} after learning.

To model the binary plateau pattern, we introduce the unit vector $\boldsymbol{y}(j) \in \mathbb{R}^{n\times 1}$ with only the j bit position being 1. For simplicity, in our following analysis, we assume that the initial weights are set to a zero matrix and use the term $\boldsymbol{W}$ to refer to the weights after learning M patterns. The weight matrix $\boldsymbol{W}$ can be written as the summation of multiple weight updates, organized in the order of the bit position j, as follows:

$$\boldsymbol{W} = \sum_{j=1}^{n} \boldsymbol{y}(j) \times \hat{1}(\sum_{k\in S(j)} \boldsymbol{x}_k^T), \tag{9}$$

where the set $S(j)$ collects the index of patterns for which the CA1 neuron j receives the plateau potentials during the learning process. $\hat{1}(x)$ is an indicator function which models the impact of the parity of weight update:

$$\hat{1}(x) = \begin{cases} 1, & \text{if } x \text{ is odd,} \\ 0, & \text{if } x \text{ is even.} \end{cases} \tag{10}$$

Next, we calculate the weighted sum of inputs:

$$\boldsymbol{u}(\boldsymbol{x}) = \boldsymbol{W} \times \boldsymbol{x} = \sum_{j=1}^{n} \left(\boldsymbol{y}(j) \times \hat{1}(\sum_{k\in S(j)} \boldsymbol{x}_k^T) \right) \times \boldsymbol{x} = \sum_{j=1}^{n} \boldsymbol{y}(j) \times \left(\hat{1}(\sum_{k\in S(j)} \boldsymbol{x}_k^T) \times \boldsymbol{x} \right). \tag{11}$$

where $\boldsymbol{x}_k$ represents the memory item for which the CA1 neuron j receives a plateau potential during the learning phase.

Let u_j be the j_{th} element of $\boldsymbol{u}$. According to the definition of $\boldsymbol{y}(j)$, we have

$$u_j(\boldsymbol{x}) = \hat{1}(\sum_{k\in S(j)} \boldsymbol{x}_k^T) \times \boldsymbol{x} = \sum_{i\in supp(\boldsymbol{x})} \hat{1}(\sum_{k\in S(j)} x_{ki}), \tag{12}$$

where x_{ki} is the element i in $\boldsymbol{x}_k$.

Note that $\hat{1}(\sum_{k\in S(j)} x_{ki})$ is a binary variable that represents the active state of w_{ji}. Its value is determined by the parity of weight updates during the learning phase. If we consider the probability that a memory pattern induces the update of weight w_{ji} during the learning phase, this probability can be modeled by Bernoulli distribution with a probability of $f_p f_q/2$ due to the independence of the input density f_p, plateau probability f_q, and the weight update probability 0.5. On this basis, we can estimate the probability $P(\hat{1}(\sum_{k\in S(j)} x_{ki}) = 1)$ in the following two cases:

- For any $j_1 \in Q(\boldsymbol{x})$ and any $i \in supp(\boldsymbol{x})$, we estimate the probability p_e that $w_{j_1 i}$ is active after learning M patterns:

$$p_e := P\big(\hat{1}(\sum_{k\in S(j_1)} x_{ki}) = 1|\boldsymbol{x}\big) = P\big(\sum_{k\in S(j_1)} x_{ki} \text{ is odd}|\boldsymbol{x}\big) = P\big(\sum_{k\in S(j_1)} x_{ki} - 1 \text{ is even}\big) = \frac{1}{2}\big(1 + (1 - f_p f_q)^{M-1}\big), \tag{13}$$

  where the estimate of the parity in the last equality uses Lemma 1.

- For any $j_2 \notin Q(\boldsymbol{x})$ and any $i \in supp(\boldsymbol{x})$, we estimate the probability p_o that $w_{j_2 i}$ is active after learning M patterns:

$$p_o := P\big(\hat{1}(\sum_{k\in S(j_2)} x_{ki}) = 1|\boldsymbol{x}\big) = P(\sum_{k\in S(j_2)} x_{ki} \text{ is odd}|\boldsymbol{x}) = \frac{1}{2}\big(1 - (1 - f_p f_q)^{M-1}\big). \tag{14}$$

The above equations provide estimates of the probability that any given connection is in an active state after learning M patterns. In order to incorporate the effects of the inherited connection probability f_w, we directly multiply f_w by p_o (or p_e) since they are independent. We then employ the cumulative binomial probability F to estimate the firing probability of CA1 neurons given the size of support set $\iota(\boldsymbol{x})$ as I:

- For any $j_1 \in Q(\boldsymbol{x})$, we have:

$$p_{f1}(j_1, \boldsymbol{x}, I) := P(u_{j_1} > v_{th} | \boldsymbol{x}, I, j_1 \in Q(\boldsymbol{x})) = 1 - F(v_{th}, I, f_w p_e). \tag{15}$$

- For any $j_2 \notin Q(\boldsymbol{x})$, we have:

$$p_{f2}(j_2, \boldsymbol{x}, I) := P(u_{j_2} > v_{th} | \boldsymbol{x}, I, j_2 \notin Q(\boldsymbol{x})) = 1 - F(v_{th}, I, f_w p_o). \tag{16}$$

The above Eqs.15 and 16 show that the formulas of firing probability do not depend on the specific positions of j_1 and j_2. For brevity, we will omit the indices j_1 and j_2 of the brackets in the following derivations.

Given the estimates of p_{f1} and p_{f2}, we can finally obtain the firing probability of a CA1 neuron by applying the total probability rule, which allows us to relax the condition of the given size $\iota(\boldsymbol{x})$ of the support set:

$$\begin{aligned}
p_{f_{avg}}(\boldsymbol{x}) &= P(u_j > v_{th} | \boldsymbol{x}) \\
&= \sum_{I=0}^{m} P(u_j > v_{th}, \iota(\boldsymbol{x}) = I | \boldsymbol{x}) \\
&= \sum_{I=0}^{m} P(\iota(\boldsymbol{x}) = I) P(u_j > v_{th} | \iota(\boldsymbol{x}) = I, \boldsymbol{x}) \\
&= \sum_{I=0}^{m} P(\iota(\boldsymbol{x}) = I) \Big(P(j \in Q(\boldsymbol{x})) P(u_j > v_{th} | j \in Q(\boldsymbol{x})) + P(j \notin Q(\boldsymbol{x})) P(u_j > v_{th} | j \notin Q(\boldsymbol{x})) \Big) \\
&= \sum_{I=0}^{m} C_m^I f_p^I (1 - f_p)^{m-I} \Big(f_q p_{f1}(\boldsymbol{x}, I) + (1 - f_q) p_{f2}(\boldsymbol{x}, I) \Big).
\end{aligned} \tag{17}$$

Estimates of expected HD between memory traces for the original and masked memory items. In this section, we utilize the properties of p_e and p_o to estimate the expected HD between memory traces for the original and masked memory items.

Suppose $\boldsymbol{x}$ is a random binary memory item with the input density of f_d and $\boldsymbol{x'}$ is the masked pattern of $\boldsymbol{x}$ with the fraction f_d of 1's being masked to zeros. We formulate the expected HD in terms of the probability that individual neurons fire differently for any complete memory item and its partially masked pattern:

$$\begin{aligned}
&E(HD(\boldsymbol{z}(\boldsymbol{x}), \boldsymbol{z}(\boldsymbol{x'}))) \\
&= \sum_{j=1}^{n} P(z_j(\boldsymbol{x}) \neq z_j(\boldsymbol{x'})) \\
&= \sum_{j=1}^{n} P(j \in Q(\boldsymbol{x})) P\big(z_j(\boldsymbol{x}) \neq z_j(\boldsymbol{x'}) | j \in Q(\boldsymbol{x})\big) + P(j \notin Q(\boldsymbol{x})) P\big(z_j(\boldsymbol{x}) \neq z_j(\boldsymbol{x'}) | j \notin Q(\boldsymbol{x})\big) \\
&= n f_q P\big(z_j(\boldsymbol{x}) \neq z_j(\boldsymbol{x'}) | j \in Q(\boldsymbol{x})\big) + n(1 - f_q) P\big(z_j(\boldsymbol{x}) \neq z_j(\boldsymbol{x'}) | j \notin Q(\boldsymbol{x})\big)
\end{aligned} \tag{18}$$

where $Q(\boldsymbol{x})$ refers to the set of indices of the CA1 neurons that got plateau potentials for $\boldsymbol{x}$ during the learning process. Eq.18 indicates that to estimate this expression $E\big(HD(\boldsymbol{z}(\boldsymbol{x}), \boldsymbol{z}(\boldsymbol{x'}))\big)$, we only need to estimate the probability that a CA1 neuron fires differently for $\boldsymbol{x}$ and $\boldsymbol{x'}$ for both the set $Q(\boldsymbol{x})$ and its complement.

By using p_{f_1} and p_{f_2}, we can estimate $P(z_j(\boldsymbol{x}) \neq z_j(\boldsymbol{x'}))$ in the following two separate cases:

- For any $j_1 \in Q(\boldsymbol{x})$, we have:

$$\begin{aligned}
&P\big(z_{j_1}(\boldsymbol{x}) \neq z_{j_1}(\boldsymbol{x'}) | j_1 \in Q(\boldsymbol{x}), \imath(\boldsymbol{x}) = I\big) \\
&= P\big(z_{j_1}(\boldsymbol{x}) = 1, z_{j_1}(\boldsymbol{x'}) = 0 | j_1 \in Q(\boldsymbol{x}), \imath(\boldsymbol{x}) = I\big) + P\big(z_{j_1}(\boldsymbol{x}) = 0, z_{j_1}(\boldsymbol{x'}) = 1 | j_1 \in Q(\boldsymbol{x}), \imath(\boldsymbol{x}) = I\big) \\
&= P\big(z_{j_1}(\boldsymbol{x}) = 1, z_{j_1}(\boldsymbol{x'}) = 0 | j_1 \in Q(\boldsymbol{x}), \imath(\boldsymbol{x}) = I\big) \quad \text{(Note that } P(z_{j_1}(\boldsymbol{x}) = 0, z_{j_1}(\boldsymbol{x'}) = 1) = 0) \\
&\approx P\big(z_{j_1}(\boldsymbol{x}) = 1 | j_1 \in Q(\boldsymbol{x}), \imath(\boldsymbol{x}) = I\big) P\big(z_{j_1}(\boldsymbol{x'}) = 0 | j_1 \in Q(\boldsymbol{x}), \imath(\boldsymbol{x}) = I\big) \\
&= \big(1 - F(v_{th}, I, f_w p_e)\big) F\big(v_{th}, I(1 - f_d), f_w p_e\big).
\end{aligned} \tag{19}$$

  The above estimate approximates $P(z_{j_1}(\boldsymbol{x}) = 1, z_{j_1}(\boldsymbol{x'}) = 0)$ as two independent terms, $P(z_{j_1}(\boldsymbol{x}) = 1)$ and $P(z_{j_1}(\boldsymbol{x'}) = 1)$. It ignores the dependence on the same parity of the weight updates indicated by Eqs.15 and 16. Nevertheless, our numerical estimates show good consistency using the simple strategy. A more precise but complicated estimate is provided in the following derivation for the non-orthogonal memory items (Eqs.28 and 29), where we further employ the total probability rule to deal with this dependency and make this estimate applicable for the common 1's.

- For any $j_2 \notin Q(\boldsymbol{x})$, similarly, we have

$$\begin{aligned}
&P\big(z_{j_2}(\boldsymbol{x}) \neq z_{j_2}(\boldsymbol{x'}) | j_2 \notin Q(\boldsymbol{x}), \imath(\boldsymbol{x}) = I\big) \\
&\qquad = P\big(z_{j_2}(\boldsymbol{x}) = 1, z_{j_2}(\boldsymbol{x'}) = 0 | j_2 \notin Q(\boldsymbol{x}), \imath(\boldsymbol{x}) = I\big) \\
&\qquad \approx F(v_{th}, I(1 - f_d), f_w p_o)\big(1 - F(v_{th}, I, f_w p_o)\big).
\end{aligned} \tag{20}$$

By combining Eqs.19 and 20, we can obtain estimates for the expected HD,

$$\begin{aligned}
&E(HD(\boldsymbol{z}(\boldsymbol{x}), \boldsymbol{z}(\boldsymbol{x'}))) \\
&= \sum_{I=0}^{m} Pr(\imath(\boldsymbol{x}) = I) E\big(HD(\boldsymbol{z}(\boldsymbol{x}), \boldsymbol{z}(\boldsymbol{x'})) | \imath(\boldsymbol{x}) = I\big) \\
&= \sum_{I=0}^{m} n Pr(\imath(\boldsymbol{x}) = I) \Big(f_q P\big(z_{j_1}(\boldsymbol{x}) \neq z_{j_1}(\boldsymbol{x'}) | j_1 \in Q(\boldsymbol{x}), \imath(\boldsymbol{x}) = I\big) + (1 - f_q) P\big((z_{j_2}(\boldsymbol{x})) \neq z_{j_2}(\boldsymbol{x'}) | j_2 \notin Q(\boldsymbol{x}), \imath(\boldsymbol{x}) = I\big) \Big) \\
&\approx n \sum_{I=0}^{m} C_m^I f_p^I (1 - f_p)^{m-I} \Big(f_q F(v_{th}, I(1 - f_d), f_w p_e)\big(1 - F(v_{th}, I, f_w p_e)\big) + (1 - f_q) F(v_{th}, I(1 - f_d), f_w p_o)\big(1 - F(v_{th}, I, f_w p_o)\big) \Big).
\end{aligned} \tag{21}$$

$E\big(HD(\boldsymbol{z}(\boldsymbol{x}), \boldsymbol{z}(\boldsymbol{x'}))\big)$ provides an estimate of dissimilarity between the memory traces for the original and masked memory items, which was used in Fig. 2C and Fig. 5.

Estimates of expected HD between memory traces for different orthogonal memory items. To estimate the average HD between memory traces of two different memory items, we use $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$ to represent two different orthogonal patterns. Owing to the assumption of orthogonal memory items, we can directly utilize $p_{f_{avg}}$ to estimate

$HD(\boldsymbol{z}(\boldsymbol{x}_a), \boldsymbol{z}(\boldsymbol{x}_b))$ as follows,

$$
\begin{aligned}
& E\big(HD(\boldsymbol{z}(\boldsymbol{x}_a), \boldsymbol{z}(\boldsymbol{x}_b))\big) \\
&= \sum_{j=1}^{n} P\big(z_j(\boldsymbol{x}_a) \neq z_j(\boldsymbol{x}_b)\big) \\
&= \sum_{j=1}^{n} P\big(z_j(\boldsymbol{x}_a) = 1, z_j(\boldsymbol{x}_b) = 0\big) + P\big(z_j(\boldsymbol{x}_a) = 0, z_j(\boldsymbol{x}_b) = 1\big) \\
&= \sum_{j=1}^{n} P(z_j(\boldsymbol{x}_a) = 1) P(z_j(\boldsymbol{x}_b) = 0) + P(z_j(\boldsymbol{x}_a) = 0) P(z_j(\boldsymbol{x}_b) = 1) \\
&= 2n p_{f_{avg}} (1 - p_{f_{avg}}).
\end{aligned}
\tag{22}
$$

Eq.22 provides the estimates of the memory traces for different original memory items. By combining the estimates in Eq.21 and Eq.22, we can estimate the ratio of HD used in Fig. 2C and Fig. 5.

1.3 Estimates of properties of memory traces for non-orthogonal memory items

Estimates of firing probability of CA1 neurons after learning M non-orthogonal memory items. The firing probability serves as a foundation for the subsequent estimates of HD of memory traces. Our estimates for the firing probability follow a similar strategy to that used for orthogonal memory items. However, a key difference is that we need to consider the impact of weight updates by BTSP on the common 1's. To address this, we analyze and model the number of patterns for which a CA1 neuron receives the plateau potential during the learning process. This allows us to incorporate the effects of weight updates on the common 1's and refine our estimates accordingly.

Consider a random binary memory item $\boldsymbol{x}$ with c common 1's that are shared with all other non-orthogonal memory items. The remaining elements in $\boldsymbol{x}$ have an input density of $\hat{f}_p$. To estimate the firing probability, we need to model the effects of the common bits on the firing probability. Note that the dependency of the active state of weight w_{ji} on the parity of updates still holds for non-orthogonal memory items, but with a different probability of weight update for synaptic weights that connect the common 1's. On this basis, we follow the derivation of Eq.13 and derive estimates for the conditional probability $P\big(\hat{1}(\sum_{k \in S(j)} x_{ki}) = 1 | S_j\big)$ given the size s_j of $S(j)$ in the following two separate cases:

- For any $j_1 \in Q(\boldsymbol{x})$ and any $i \in supp(\boldsymbol{x})$, we estimate the probability $\hat{p}_e(s_{j_1})$ that $w_{j_1 i}$ is active given the size s_{j_1} of $S(j_1)$:

$$
\hat{p}_e(s_{j_1}) := P\big(\hat{1}(\sum_{k \in S(j_1)} x_{ki}) = 1 | \boldsymbol{x}, s_{j_1}\big) = P(\sum_{k \in S(j_1)} x_{ki} - 1 \text{ is even} | s_{j_1}) = \frac{1}{2}\big(1 + (1 - \hat{f}_p)^{s_{j_1} - 1}\big). \tag{23}
$$

- For any $j_2 \notin Q(\boldsymbol{x})$ and any $i \in supp(\boldsymbol{x})$, we estimate the probability $\hat{p}_o(s_{j_2})$ that $w_{j_2 i}$ is active given the size s_{j_2} of $S(j_2)$:

$$
\hat{p}_o(s_{j_2}) := P\big(\hat{1}(\sum_{k \in S(j_2)} x_{ki}) = 1 | \boldsymbol{x}, s_{j_2}\big) = P(\sum_{k \in S(j_2)} x_{ki} \text{ is odd} | s_{j_2}) = \frac{1}{2}\big(1 - (1 - \hat{f}_p)^{s_{j_2}}\big). \tag{24}
$$

Having obtained the estimates for the probability that the synaptic weight remains active after learning, we can proceed to estimate the firing probability. In order to model the effect of common bits, we introduce the following two variables, c_w and $\hat{s}_j$:

First, to incorporate the impact of the connection probability f_w on the common bits, we introduce c_w, which represents the number of all active connections from the input neurons receiving common 1's to a given CA1 neuron at the end of learning. This variable captures the level of connectivity between the common bits and the CA1 neuron.

Second, to capture the impact of probabilistic updates on synapses connecting input neurons that receives common 1's input, we introduce $\hat{s}_j$ to represent the number of updates of a synapse connecting the input neuron that receives common 1's during the learning of M patterns. A noteworthy observation is that if $\hat{s}_j$ is odd, then the synapse will be in the active state after the completion of learning M patterns and thereby the partial weighted sum of inputs from the common 1's is equal to c_w; otherwise, the synapse will be in the inactive state and the weighted sum is equal to 0.

Utilizing these two variables, we can evaluate the conditional firing probability of CA1 neurons as follows,

- For any $j_1 \in Q(\boldsymbol{x})$, we have

$$
\begin{aligned}
&\hat{p}_{f1}(\boldsymbol{x}, I, c_w, c) \\
&= P\big(z_{j_1}(\boldsymbol{x}) = 1 | j_1 \in Q(\boldsymbol{x}), \imath(\boldsymbol{x}) = I\big) \\
&= \sum_{l=0}^{M} P\big(z_{j_1}(\boldsymbol{x}) = 1, s_{j_1} = l | j_1 \in Q(\boldsymbol{x}), \imath(\boldsymbol{x}) = I\big) \quad \text{(by the total probability rule)} \\
&= \sum_{l=0}^{M} P(s_{j_1} = l) P\big(z_{j_1}(\boldsymbol{x}) = 1 | j_1 \in Q(\boldsymbol{x}), \imath(\boldsymbol{x}) = I, s_{j_1} = l\big) \quad \text{(by considering the parity of } \hat{s}_{j_1}\text{)} \\
&= \sum_{l=0}^{M} P(s_{j_1} = l) \Big(P\big(z_{j_1}(\boldsymbol{x}) = 1, \hat{s}_{j_1} \text{is even} | j_1 \in Q(\boldsymbol{x}), \imath(\boldsymbol{x}) = I, s_{j_1} = l\big) + P\big(z_{j_1}(\boldsymbol{x}) = 1, \hat{s}_{j_1} \text{is odd} | j_1 \in Q(\boldsymbol{x}), \imath(\boldsymbol{x}) = I, s_{j_1} = l\big) \Big) \\
&= \sum_{l=0}^{M} P(s_{j_1} = l) P(\hat{s}_{j_1} \text{is odd}) P\big(z_{j_1}(\boldsymbol{x}) = 1 | j_1 \in Q(\boldsymbol{x}), \imath(\boldsymbol{x}) = I, s_{j_1} = l, \hat{s}_{j_1} \text{is odd}\big) \\
&+ \sum_{l=0}^{M} P(s_{j_1} = l) P(\hat{s}_{j_1} \text{is even}) P\big(z_{j_1}(\boldsymbol{x}) = 1 | j_1 \in Q(\boldsymbol{x}), \imath(\boldsymbol{x}) = I, s_{j_1} = l, \hat{s}_{j_1} \text{is even}\big) \\
&= \sum_{l=0}^{M} P(s_{j_1} = l) \Big(\frac{1}{2}\big(1 - F(v_{th} - c_w, I - c, f_w \hat{p}_e(l))\big) + \frac{1}{2}\big(1 - F(v_{th}, I - c, f_w \hat{p}_e(l))\big) \Big) \\
&= \sum_{l=0}^{M} C_M^l f_q^l (1 - f_q)^{M-l} \Big(\frac{1}{2}\big(1 - F(v_{th} - c_w, I - c, f_w \hat{p}_e(l))\big) + \frac{1}{2}\big(1 - F(v_{th}, I - c, f_w \hat{p}_e(l))\big) \Big) \\
&\approx \sum_{l=0}^{M} \frac{(M f_q)^l}{2 e^{M f_q} l!} \Big(2 - F\big(v_{th} - c_w, I - c, f_w \hat{p}_e(l)\big) - F\big(v_{th}, I - c, f_w \hat{p}_e(l)\big) \Big).
\end{aligned}
\tag{25}
$$

The last equation uses the Poisson approximation to the binomial probability for simplifying analysis.

- For any $j_2 \notin Q(\boldsymbol{x})$, we can obtain the estimate in a similar manner

$$
\begin{aligned}
&\hat{p}_{f2}(\boldsymbol{x}, I, c_w, c) \\
&= P(z_{j_2}(\boldsymbol{x}) = 1 | j_2 \notin Q(\boldsymbol{x}), \imath(\boldsymbol{x}) = I) \\
&= \sum_{l=0}^{M} P\big(z_{j_2}(\boldsymbol{x}) = 1, s_{j_2} = l, j_2 \notin Q(\boldsymbol{x}), \imath(\boldsymbol{x}) = I\big) \\
&= \sum_{l=0}^{M} P(s_{j_2} = l) P\big(z_{j_2}(\boldsymbol{x}) = 1 | j_2 \notin Q(\boldsymbol{x}), \imath(\boldsymbol{x}) = I, s_{j_2} = l\big) \quad \text{(by considering the parity of } \hat{s}_{j_2}) \\
&= \sum_{l=0}^{M} P(s_{j_2} = l) \Big(P\big(z_{j_2}(\boldsymbol{x}) = 1, \hat{s}_{j_2} \text{is even} | j_2 \notin Q(\boldsymbol{x}), \imath(\boldsymbol{x}) = I, s_{j_2} = l\big) + P\big(z_{j_2}(\boldsymbol{x}) = 1, \hat{s}_{j_2} \text{is odd} | j_2 \notin Q(\boldsymbol{x}), \imath(\boldsymbol{x}) = I, s_{j_2} = l\big) \Big) \\
&= \sum_{l=0}^{M} C_M^l f_q^l (1 - f_q)^{M-l} \Big(\frac{1}{2} \big(1 - F(v_{th}, I - c, f_w \hat{p}_o(l))\big) + \frac{1}{2} \big(1 - F(v_{th} - c_w, I - c, f_w \hat{p}_o(l))\big) \Big) \\
&\approx \sum_{l=0}^{M} \frac{(M f_q)^l}{2 e^{M f_q} l!} \Big(2 - F\big(v_{th}, I - c, f_w \hat{p}_o(l)\big) - F\big(v_{th} - c_w, I - c, f_w \hat{p}_o(l)\big) \Big).
\end{aligned}
\tag{26}
$$

Combining Eqs.25 and 26 gives the estimate of the average firing probability of CA1 neurons

$$
\begin{aligned}
&\hat{p}_{f_{avg}}(\boldsymbol{x}, c) \\
&= \sum_{I'=0}^{m-c} \sum_{c_w=0}^{c} P\big(\imath(\boldsymbol{x}) = I' + c, c_w\big) P\big(u_j > v_{th} | \imath(\boldsymbol{x}) = I' + c, c_w\big) \\
&= \sum_{I'=0}^{m-c} \sum_{c_w=0}^{c} P(\imath(\boldsymbol{x}) = I' + c) P(c_w) \Big(f_q \hat{p}_{f1}\big(\boldsymbol{x}, I' + c, c_w, c\big) + (1 - f_q) \hat{p}_{f2}(\boldsymbol{x}, I' + c, c_w, c) \Big) \\
&= \sum_{I'=0}^{m-c} \sum_{c_w=0}^{c} C_{m-c}^{I'} C_c^{c_w} \hat{f}_p^{I'} (1 - \hat{f}_p)^{m-c-I'} f_w^{c_w} (1 - f_w)^{c-c_w} \big(f_q \hat{p}_{f1}(\boldsymbol{x}, I' + c, c_w, c) + (1 - f_q) \hat{p}_{f2}(\boldsymbol{x}, I' + c, c_w, c) \big),
\end{aligned}
\tag{27}
$$

Estimates of expected HD between memory traces for the original and masked memory items. In this section, we use the firing probability of CA1 neurons to estimate the expected HD between memory traces for the original and masked memory items. As indicated by Eq.19, to estimate $E\big(HD(\boldsymbol{z}(\boldsymbol{x}), \boldsymbol{z}(\boldsymbol{x'}))\big)$, we only need to estimate $P(z_j(\boldsymbol{x}) = 1, z_j(\boldsymbol{x'}) = 0 | \imath(\boldsymbol{x}) = I)$.

Unlike the estimate of orthogonal memory items, the interaction of common bits and masking bits complicates the estimates. For simplification, we use the average number of masking common 1's, $c f_d$, to approximate the effects of the masking fraction on common 1's. On this basis, we evaluate $P(z_j(\boldsymbol{x}) = 1, z_j(\boldsymbol{x'}) = 0 | \imath(\boldsymbol{x}) = I)$ in two different cases, $j_1 \in Q(\boldsymbol{x})$ and $j_2 \notin Q(\boldsymbol{x})$.

- For any $j_1 \in Q(\boldsymbol{x})$, we follow the derivation of Eq.25 and have

$$
\begin{aligned}
&\hat{p}_{c1}(j_1, \boldsymbol{x}, \boldsymbol{x'}, I, c_w) \\
&= P\big(z_{j_1}(\boldsymbol{x}) = 1, z_{j_1}(\boldsymbol{x'}) = 0 | j_1 \in Q(\boldsymbol{x}), \imath(\boldsymbol{x}) = I\big) \\
&= \sum_{l=0}^{M} P\big(z_{j_1}(\boldsymbol{x}) = 1, z_{j_1}(\boldsymbol{x'}) = 0, s_{j_1} = l | j_1 \in Q(\boldsymbol{x}), \imath(\boldsymbol{x}) = I\big) \\
&= \sum_{l=0}^{M} P(s_{j_1} = l) P\big(z_{j_1}(\boldsymbol{x}) = 1, z_{j_1}(\boldsymbol{x'}) = 0 | j_1 \in Q(\boldsymbol{x}), \imath(\boldsymbol{x}) = I, s_{j_1} = l\big) \\
&= \sum_{l=0}^{M} P(s_{j_1} = l) P(\hat{s}_{j_1} \text{is even}) P\big(z_{j_1}(\boldsymbol{x}) = 1, z_{j_1}(\boldsymbol{x'}) = 0 | j_1 \in Q(\boldsymbol{x}), \imath(\boldsymbol{x}) = I, s_{j_1} = l, \hat{s}_{j_1} \text{is even}\big) \\
&+ \sum_{l=0}^{M} P(s_{j_1} = l) P(\hat{s}_{j_1} \text{is odd}) P\big(z_{j_1}(\boldsymbol{x}) = 1, z_{j_1}(\boldsymbol{x'}) = 0 | j_1 \in Q(\boldsymbol{x}), \imath(\boldsymbol{x}) = I, s_{j_1} = l, \hat{s}_{j_1} \text{is odd}\big) \\
&\approx \sum_{l=0}^{M} \frac{(Mf_q)^l}{2e^{Mf_q} l!} \Big(\big(1 - F(v_{th} - c_w, I - c, f_w \hat{p}_e(l))\big) F\big(v_{th} - c_w(1 - f_d), (I - c)(1 - f_d), f_w \hat{p}_e(l)\big) \\
&+ \big(1 - F(v_{th}, I - c, f_w \hat{p}_e(l))\big) F\big(v_{th}, (I - c)(1 - f_d), f_w \hat{p}_e(l)\big) \Big),
\end{aligned}
\tag{28}
$$

where the handling of parity of s_{j_1} follows the same as Eq.25 and thus we simplify the derivation here.

- For any $j_2 \notin Q(\boldsymbol{x})$, we can derive the estimate in a similar manner:

$$
\begin{aligned}
&\hat{p}_{c2}(j_2, \boldsymbol{x}, \boldsymbol{x'}, I, c_w) \\
&= P\big(z_{j_2}(\boldsymbol{x}) = 1, z_{j_2}(\boldsymbol{x'}) = 0 | j_2 \notin Q(\boldsymbol{x}), \imath(\boldsymbol{x}) = I\big) \\
&= \sum_{l=0}^{M} P\big(z_{j_2}(\boldsymbol{x}) = 1, z_{j_2}(\boldsymbol{x'}) = 0, s_{j_2} = l | j_2 \notin Q(\boldsymbol{x}), \imath(\boldsymbol{x}) = I\big) \\
&= \sum_{l=0}^{M} P(s_{j_2} = l) P\big(z_{j_2}(\boldsymbol{x}) = 1, z_{j_2}(\boldsymbol{x'}) = 0 | j_2 \notin Q(\boldsymbol{x}), \imath(\boldsymbol{x}) = I, s_{j_2} = l\big) \\
&= \sum_{l=0}^{M} P(s_{j_2} = l) P(\hat{s}_{j_2} \text{is even}) P\big(z_{j_2}(\boldsymbol{x}) = 1, z_{j_2}(\boldsymbol{x'}) = 0 | j_2 \notin Q(\boldsymbol{x}), \imath(\boldsymbol{x}) = I, s_{j_2} = l, \hat{s}_{j_2} \text{is even}\big) \\
&+ \sum_{l=0}^{M} P(s_{j_2} = l) P(\hat{s}_{j_2} \text{is odd}) P\big(z_{j_2}(\boldsymbol{x}) = 1, z_{j_1}(\boldsymbol{x'}) = 0 | j_2 \notin Q(\boldsymbol{x}), \imath(\boldsymbol{x}) = I, s_{j_2} = l, \hat{s}_{j_2} \text{is odd}\big) \\
&\approx \sum_{l=0}^{M} \frac{(Mf_q)^l}{2e^{Mf_q} l!} \Big(\big(1 - F(v_{th}, I - c, f_w \hat{p}_o(l))\big) F\big(v_{th}, (I - c)(1 - f_d), f_w \hat{p}_o(l)\big) \\
&+ \big(1 - F(v_{th} - c_w, I - c, f_w \hat{p}_o(l))\big) F\big(v_{th} - c_w(1 - f_d), (I - c)(1 - f_d), f_w \hat{p}_o(l)\big) \Big).
\end{aligned}
\tag{29}
$$

As the above Eqs.28 and 29 do not depend on the specific positions of j_1 and j_2, we omit the indices j_1 and j_2 in the following derivations.

Combining Eqs.28 and 29 gives the estimate of $E\big(HD(\boldsymbol{z}(\boldsymbol{x}),\boldsymbol{z}(\boldsymbol{x'}))\big)$

$$
\begin{aligned}
&E\big(HD(\boldsymbol{z}(\boldsymbol{x}),\boldsymbol{z}(\boldsymbol{x'}))\big)\\
&=n\sum_{I=c}^{m}Pr(\iota(\boldsymbol{x})=I)P\big(z_j(\boldsymbol{x})=1,z_j(\boldsymbol{x'})=0,\forall j|\iota(\boldsymbol{x})=I\big)\\
&=n\sum_{I=c}^{m}Pr(\iota(\boldsymbol{x})=I)\Big(f_qP\big(z_{j_1}(\boldsymbol{x})=1,z_{j_1}(\boldsymbol{x'})=0|\iota(\boldsymbol{x})=I,j_1\in Q(\boldsymbol{x})\big)\\
&+(1-f_q)\big(P(z_{j_2}(\boldsymbol{x})=1,z_{j_2}(\boldsymbol{x'})=0|\iota(\boldsymbol{x})=I,j_2\notin Q(\boldsymbol{x}))\big)\Big)\\
&\approx n\sum_{I'=0}^{m-c}\sum_{c_w=0}^{c}C_{m-c}^{I'}C_c^{c_w}\hat{f}_p^{I'}(1-\hat{f}_p)^{m-c-I'}f_w^{c_w}(1-f_w)^{c-c_w}\Big(f_q\hat{p}_{c1}\big(\boldsymbol{x},\boldsymbol{x'},I'+c,c_w\big)+(1-f_q)\hat{p}_{c2}\big(\boldsymbol{x},\boldsymbol{x'},I'+c,c_w\big)\Big).
\end{aligned}
\tag{30}
$$

$E\big(HD(\boldsymbol{z}(\boldsymbol{x}),\boldsymbol{z}(\boldsymbol{x'}))\big)$ provides the estimate of dissimilarity between the memory traces for the original and masked memory items, which was used in Fig. 4B.

Estimates of expected HD between memory traces for different non-orthogonal memory items. To estimate the expected HD of memory traces between non-orthogonal memory items, we use $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$ to represent two different non-orthogonal memory items. By using the estimate of firing probability for non-orthogonal memory items, we can directly obtain an approximate estimate for this HD:

$$
\begin{aligned}
&E\big(HD(\boldsymbol{z}(\boldsymbol{x}_a),\boldsymbol{z}(\boldsymbol{x}_b))\big)\\
&=\sum_{j=1}^{n}P\big(z_j(\boldsymbol{x}_a)=1,z_j(\boldsymbol{x}_b)=0\big)+P\big(z_j(\boldsymbol{x}_a)=0,z_j(\boldsymbol{x}_b)=1\big)\\
&\approx\sum_{j=1}^{n}P(z_j(\boldsymbol{x}_a)=1)P(z_j(\boldsymbol{x}_b)=0)+P(z_j(\boldsymbol{x}_a)=0)P(z_j(\boldsymbol{x}_b)=1)\\
&=2n\hat{p}_{f_{avg}}(\boldsymbol{x}_a,c)\big(1-\hat{p}_{f_{avg}}(\boldsymbol{x}_a,c)\big).
\end{aligned}
\tag{31}
$$

In the above estimation we disregard the correlation arising from common bits for simplicity. This approximation is suitable for cases where the fraction of common 1's is small and the equal sign can be taken when the fraction of common 1's is 0. A more precise but intricate estimate can be obtained by expanding the estimation along $S(j)$, as exemplified by Eqs.25 and 26.

The estimated HD in Eq.31 provides an approximation of the memory traces for different original memory items. By combining this estimate with the estimate in Eq.30, we can directly obtain the estimate of the ratio of HD used in Fig. 4D.

2 Details of experimental setup and simulations

Details of simulation of continuous weight changes by binary BTSP rule (Fig. 1D and E). In Fig. 1D , we generated the curve by employing Eq.1 in (Milstein et al., 2021). We considered the case where the input patterns arrive at the same time and simulated the rule with the following parameter values: $w^{max}=8,\tau_{ET}=1000ms,\tau_{IS}=500ms,k^+=0.5,k^-=0.5,\alpha^+=0.01,\beta^+=35,\alpha^-=0.01,\beta^-=800$. For the simulation depicted in Fig. 1E, we considered a scenario where two neurons are connected by 8 synapses with binary synaptic weights. The initial binary weights were randomly sampled from a Bernoulli distribution with a probability of 0.5. We conducted 10,000 simulations and plotted the average change of weights based on the sum of initial weight values.

Details of analysis of the masking fraction (Figs. 2C-F) and common bits (Fig. 4). In our simulation, we initialized the weights as a zero matrix and used the default parameter setting of $m = 25,000$, $n = 39,000$, $f_p = 0.005$, $f_q = 0.005$, and $f_w = 0.6$. For each specific masking fraction f_d, we generate $\boldsymbol{x}'$ by randomly masking $\lfloor m f_p f_d \rfloor$ of the 1's of $\boldsymbol{x}$. We conducted each experiment 10 times and reported the average results. For the data points in the simulation, we sampled M at intervals of 1000 and sampled the masking fraction and the fraction of common 1's at intervals of 0.02.

For Figs. 2C-F and 4, we selected the optimal v_{th} by minimizing the predicted HD ratio with the constrain that the firing probability of CA1 neurons for the complete memory items falls within a reasonable range [0.05%,10%]. Throughout the optimization process, we used a fixed masking fraction of 33% and the maximum number of input patterns M (i.e., 20,000) for calculating the firing probability of CA1 neurons (Eqs. 17 and 27) and the predicted HD ratio. We searched for the optimal threshold for three different values of f_q used in Fig. 2C-F and searched for the optimal threshold once for Fig. 4.

For Fig. 4, we first generated the corresponding common 1's with the number of $c = \lfloor m f_c f_p \rfloor$, for all input patterns. We then generated the remaining bits with the density of $\hat{f}_p = (m f_p - c)/(m - c)$, ensuring that the average number of 1's in each pattern remains the same as the value $m f_p$ used for orthogonal memory items. We generated the corresponding data sets for each specific value of c and randomly masked 33% elements of each non-orthogonal memory item for testing. The predicted ratio was obtained from Eqs.30 and 31.

Details of comparison with other memory models (Fig. 3). We built the BTSP model with the default parameter configuration and build the HFN with the same input size and connection probability. To establish the HFN with sparse binary inputs, we followed the previous studies(Hopfield, 1982; Sharma et al., 2022) and formalized the weight matrix $\boldsymbol{W}^H$ of the HFN by

$$w_{ji}^H = \begin{cases} 0, & i = j, \\ \frac{1}{m}\sum_{h=1}^{M}(x_{hi} - f_p)(x_{hj} - f_p), & i \neq j. \end{cases} \tag{32}$$

To generate a HFN with the sparsely-connected structure, we first generated a fixed random binary matrix based on the connection probability f_w and multiplied it with w_{ji}^H after each weight update. We searched for the optimal threshold of the state update function from $[0, 0.02]$ with the minimum step 0.001. By referring to the memory capacity defined in section 17.2.4 of (Gerstner et al., 2014), we adopted an optimal criterion that maximizes the number of patterns that HFN can recall correctly within the error tolerance $m * 0.001$ for the dissimilarity (i.e., HD) between the memory traces of the original pattern and the masked pattern. We used the resulting optimal threshold 0.001 in Fig. 3. We used the synchronous update rule to iterate the state of all neurons 100 times for retrieval.

We constructed the RP using binary connection weights. The weights were initialized with the same fraction of weights with value 1 created by BTSP after learning 20,000 patterns. We binarized the projection results by applying a fixed threshold. To optimize this threshold in the binary representation, we used a grid search approach. The optimization criterion was to minimize the simulated HD ratio with a masking fraction of 33%. For numerical simulations, we sampled the number of memory items M at intervals of 1000 and the masking fraction at intervals of 0.02 to obtain the data points.

Details of experimental setup of scale-up estimates (Fig. 5). For each specific scaling factor γ defined on the x-axis, we conducted simulations with $m = \gamma * 25,000$, $n = \gamma * 39,000$, and other default hyperparameters: $f_p = 0.005$, $f_q = 0.005$, and $f_w = 0.6$. We searched for the optimal threshold v_{th} for each scale once. The optimization criterion follows the experimental setup used in Figs. 2 and 4. To obtain the data points for simulation, we sampled M at intervals of 1000 and sampled the simulated scale at intervals of 0.1. The reported results are the average of 10 simulation runs. Prediction of $HD(\boldsymbol{z}(\boldsymbol{x}), \boldsymbol{z}(\boldsymbol{x}')))$ in Fig. 5A was obtained from Eq.21. Prediction of $HD(\boldsymbol{z}(\boldsymbol{x}_a), \boldsymbol{z}(\boldsymbol{x}_b)))$

in Fig. 5B was obtained from Eqs.21 and 22. Prediction in Fig. 5C was obtained by calculating the quotient of the two HD values.

Details of analysis of the repulsion effect (Fig. 6). We used the following parameters in the experiment: $m = 25,000$, $n = 39,000$, $f_p = 0.005$, $f_q = 0.02$, $f_w = 0.6$. Specifically, we first generated 1000 orthogonal memory patterns with the input density f_p. We randomly selected one pattern and generated another pattern with a 40% overlap ratio, as well as a pattern that had no overlap 1's with the chosen pattern. The generated patterns, along with orthogonal memory items, were randomly presented to the network and BTSP was applied to each pattern once. After learning, we calculated the overlapping similarity between the selected patterns and the two specific patterns separately. The results were averaged over 20 runs with different randomly initialized parameters.

Acknowledgments We would like to thank Jeffrey C. Magee, Aaron D. Milstein, and Sandro Romani for their helpful comments on the first version of this manuscript from (Wu and Maass, 2023), and Alice Dauphin for a critical reading of the current version. This research was partially supported by the Human Brain Project (Grant Agreement number 785907) of the European Union, and a grant from Intel.

References

Andersen, P., Morris, R., Amaral, D., Bliss, T., and O'Keefe, J. (2007). *The hippocampus book*. Oxford university press.

Bittner, K. C., Grienberger, C., Vaidya, S. P., Milstein, A. D., Macklin, J. J., Suh, J., Tonegawa, S., and Magee, J. C. (2015). Conjunctive input processing drives feature selectivity in hippocampal ca1 neurons. *Nature neuroscience*, 18(8):1133–1142.

Bittner, K. C., Milstein, A. D., Grienberger, C., Romani, S., and Magee, J. C. (2017). Behavioral time scale synaptic plasticity underlies ca1 place fields. *Science*, 357(6355):1033–1036.

Chanales, A. J., Oza, A., Favila, S. E., and Kuhl, B. A. (2017). Overlap among spatial memories triggers repulsion of hippocampal representations. *Current Biology*, 27(15):2307–2317.

Cone, I. and Shouval, H. Z. (2021). Behavioral time scale plasticity of place fields: Mathematical analysis. *Frontiers in computational neuroscience*, 15:640235.

Dasgupta, S., Stevens, C. F., and Navlakha, S. (2017). A neural algorithm for a fundamental computing problem. *Science*, 358(6364):793–796.

Gerstner, W., Kistler, W. M., Naud, R., and Paninski, L. (2014). *Neuronal dynamics: From single neurons to networks and models of cognition*. Cambridge University Press.

Grienberger, C. and Magee, J. C. (2022). Entorhinal cortex directs learning-related changes in ca1 representations. *Nature*, pages 1–9.

Guzman, S. J., Schlögl, A., Frotscher, M., and Jonas, P. (2016). Synaptic mechanisms of pattern completion in the hippocampal ca3 network. *Science*, 353(6304):1117–1123.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558.

Hopfield, J. J. (2007). Hopfield network. *Scholarpedia*, 2(5):1977.

Li, P. Y. and Roxin, A. (2023). Rapid memory encoding in a recurrent network model with behavioral time scale synaptic plasticity. *bioRxiv*, pages 2023–05.

Li, Y., Briguglio, J., Romani, S., and Magee, J. C. (2023). Mechanisms of memory storage and retrieval in hippocampal area ca3. *bioRxiv*, pages 2023–05.

Magee, J. C. and Grienberger, C. (2020). Synaptic plasticity forms and functions. *Annual review of neuroscience*, 43:95–117.

McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5:115–133.

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

Sharma, S., Chandra, S., and Fiete, I. (2022). Content addressable memory without catastrophic forgetting by heteroassociation with a fixed scaffold. In *International Conference on Machine Learning*, pages 19658–19682. PMLR.

Sugar, J. and Moser, M.-B. (2019). Episodic memory: Neuronal codes for what, where, and when. *Hippocampus*, 29(12):1190–1205.

Wammes, J., Norman, K. A., and Turk-Browne, N. (2022). Increasing stimulus similarity drives nonmonotonic representational change in hippocampus. *ELife*, 11:e68344.

Wanjia, G., Favila, S. E., Kim, G., Molitor, R. J., and Kuhl, B. A. (2021). Abrupt hippocampal remapping signals resolution of memory interference. *Nature communications*, 12(1):1–11.

Widrich, M., Schäfl, B., Pavlović, M., Ramsauer, H., Gruber, L., Holzleitner, M., Brandstetter, J., Sandve, G. K., Greiff, V., Hochreiter, S., et al. (2020). Modern hopfield networks and attention for immune repertoire classification. *Advances in Neural Information Processing Systems*, 33:18832–18845.

Wu, Y. and Maass, W. (2023). Memory structure created through behavioral time scale synaptic plasticity. *bioRxiv*, pages 2023–04.

Zheng, L., Gao, Z., McAvan, A. S., Isham, E. A., and Ekstrom, A. D. (2021). Partially overlapping spatial environments trigger reinstatement in hippocampus and schema representations in prefrontal cortex. *Nature Communications*, 12(1):6231.