A computational framework for memory engrams

Chiara Gastaldi and Wulfram Gerstner, EPFL, CH

This is a preprint version before editing. The final version has appeared in the book:
Engrams. Advances in Neurobiology, vol 38
Gräff, J., Ramirez, S. (eds). (Springer) pp 237-257 *(2024)*

Table of Contents

Abstract

Memory engrams in mice brains are potentially related to groups of concept cells in human brains. A single concept cell in human hippocampus responds, for example, not only to different images of the same object or person, but also to its name written down in characters. Importantly, a single mental concept (object or person) is represented by several concept cells and each concept cell can respond to more than one concept. Computational work shows how mental concepts can be embedded in recurrent artificial neural networks as memory engrams and how neurons that are shared between different engrams can lead to associations between concepts. Therefore observations at the level of neurons can be linked to cognitive notions of memory recall and association chains between memory items.

Vocabulary box

Association - Spontaneous transition between different mental concepts, or in the case of auto-associative memory, between different aspects of the same mental concept.

Assembly - A set of neurons participating in an memory engram with relatively strong connectivity amongst themselves form an assembly, also called Hebbian assembly.

Activity pattern - A configuration of neurons firing at high rate embedded in a sea of neurons firing at low rate.

Attractor network - Biological or artificial neural networks with properties such that the network activity converges to certain preferred activity patterns.

Concept - an object such as a person, a place, or an animal with individuality. Example: concepts are “my mother”, “the Sidney Opera house”, and “my dog Max”, not the generic classes such as “mothers = women who have one or more children”, “famous buildings”, and “dogs”.

Concept cell - A single neuron, often in hippocampus or more generally the Medium Temporal Lobe, that responds to the retrieval of a mental concept, independently of the specifics of sensory input

Memory engram - A set of cells that responds to a specific memory item. In this chapter we consider an assembly of concept cells as equivalent to an engram.

Hopfield model - The Hopfield network is an example of an attractor network where engrams correspond to memory items represented by preferred activity patterns.

Memory recall or concept retrieval - high activity in assembly of concept cells.

Overlap of engrams - fraction of concept cells shared across engrams.

Shared neuron - A neuron that participates in two or more different engrams.

Similarity measure, m^{μ} - measure of the correlation between the current state of activity of the memory network and the state of memory recall of a specific concept.

Sparse activity - Fraction of neurons in a given brain area that are active during the activation of one engram or one memory item

ξ_i^3 **=1** - neuron i is part of engram number 3.

Introduction: Associations and Auto-associations in memory

Humans can memorize thousands of concepts, where the term ‘concept’ is taken in a broad sense encompassing facts (‘who is the president of the United States?’), objects of daily life (‘imagine a spoon!’), naming of food items (‘this is a banana’), meanings of words, episodes of life, persons, family relations, or the layout of your home town To check whether your discussion partner remembers a ‘concept’, you may want to give a cue, for example by asking a question, providing a keyword, or showing a picture.

In computational research, the recall of memories of concepts has been formalized and categorized as either “associative” or “auto-associative” memory. To understand the idea of of “auto-associative” memory, let us consider the example of Fig. 1A, where we see an animal partially hidden behind a tree. We can easily guess that the hidden animal is a horse, even if the visual information is incomplete. Whenever we are able to recall the full memory of a concept starting from a piece of partial information on the *same* concept we speak of “auto-associative” memory recall. A rough black-and-white drawing of an apple may enable you to think of the taste and color of an apple and thus helps you to recall the rich concept of “apple” from partial information.

In contrast to auto-associative memory, associative memory (also called association memory) links two separate concepts. The example of Fig. 2B shows Laurel's famous comic character and reminds us of his equally famous comedy partner Hardy. Even though Laurel and Hardy are separate persons and, hence, separate concepts, their repeated and simultaneous appearances on TV have, for most fans of their TV comedies, consolidated the association between the two actors.

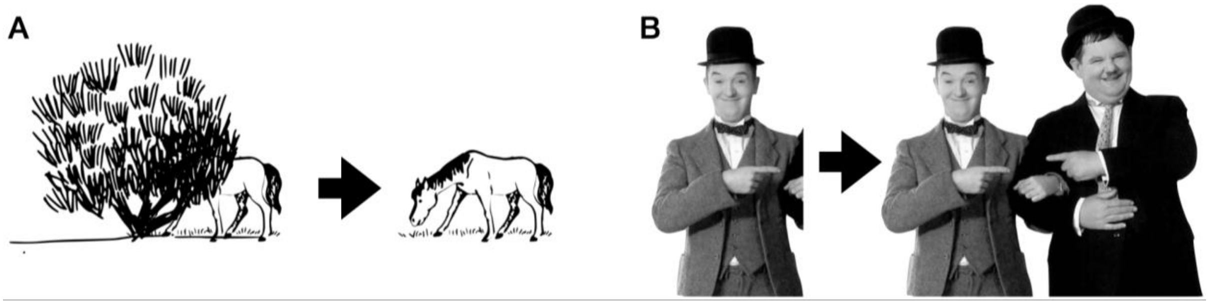


Figure 1 A) Example of auto-associative memory recall. B) Example of associative or episodic memory recall.

Associations are not restricted to concepts of the same type such as two persons. If you visited a famous place together with a good friend, a picture of that place will remind you of the friend. The personally experienced event (the trip in this example) acts as the glue that associates two otherwise unrelated concepts (the friend and the place). This type of learned association, if related to personal experience, is also called “episodic” memory. Note that, typically, before the trip you could already recognize your friend and had seen pictures of the famous place as separate entities, implying that the memory of the individual concepts was already stored somewhere in your brain; the association between the two concepts was built later. Associative (across mental concepts) and auto-associative (within mental concepts) memory work in parallel and possibly collaborate for concept recall.

Concept cells

Memory engrams can be induced in the hippocampal formation of mice by appropriate experimental stimulation paradigms (see Chapters xx and xx of this edited book). In a typical experiment, a memory engram is the ensemble of neurons (”assembly”) that respond selectively to a concept, such as a specific cage or stimulus. While obtaining experimental data in the human brain has ethical and practical limitations, a stream of experimental results hints at a mechanism in the human Medium Temporal Lobe (MTL) that is analogous to the formation of engrams in mice.

Experimentalists found neurons in the human MTL which selectively and consistently respond to stimuli representing specific individuals or places [1–3]. The experiments were conducted in patients suffering from severe treatment-resistant epilepsy, requiring surgical intervention. To determine the location of the epileptic focus in relation to crucial brain areas like those responsible for speech or motor control, electrophysiological recordings are conducted while the patient engages in various tasks. Unlike neurons in the visual cortex that respond specifically to visual stimuli, single neurons in the medial temporal lobe of the human cortex, particularly in the hippocampus, exhibit broader responsiveness to a range of stimuli associated with the same mental concept. For instance, a neuron may respond to both the written word "Sydney Opera" and a picture of the Sydney Opera House (Fig. 2A). This finding allows us to interpret such neurons as being part of an assembly that encodes the mental concept of "Sydney Opera". Other neurons in the same area may respond to the “Tower of Pisa” and yet others to a famous actor. These neurons have been named "concept cells” [1].

Coding for single concepts

Each mental concept is believed to be represented by a group of concept cells [1-3] which have been hypothesized to form a memory engram [43]. The group of concept cells simultaneously increase their firing rates when one of the stimuli that trigger concept recall is presented. The estimated fraction, γ, of MTL neurons involved in representing a given concept is approximately $\gamma = 0.23\%$ [4]. Assuming that each memory item is represented by the activation of a fixed, yet random, subset of active neurons, a single concept is expected to activate γN neurons, while two arbitrary concepts are expected to share $\gamma^2 N$ cells, where N represents the total number of neurons in the relevant brain regions. Suppose that the total number of MTL neurons is N= 200’000. Then the concept of “Sidney Opera” would be represented by an engram containing 460 neurons. Similarly, the “Tower of Pisa” would also be represented by a different engram of 460 neurons. Since the “Sidney Opera” has no relation to the “Tower of Pisa”, we expect from purely statistical arguments that at most one or two neurons are part of both engrams.

Association between concepts

It is possible to measure the fraction of neurons that respond to more than one concept. It was found that unrelated concepts share less than 1% of neurons, whereas assemblies representing previously associated concepts share approximately 4–5% of neurons [6] (Fig. 2B and 3). Moreover, during the process of building associations between pairs of concepts, individual neurons can become responsive to concepts to which they did not respond initially [5]. This suggests that an increased proportion of shared neurons facilitates the association between concepts [6–8].

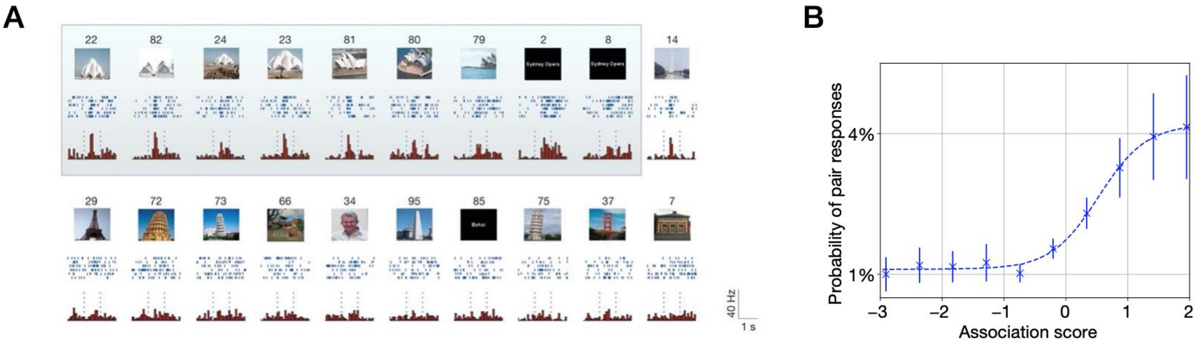


Figure 2 A) Recordings from a concept cell selective to the Sydney Opera House. Taken from [1]. B) Probability of pair response of concept cells. Adapted from [6].

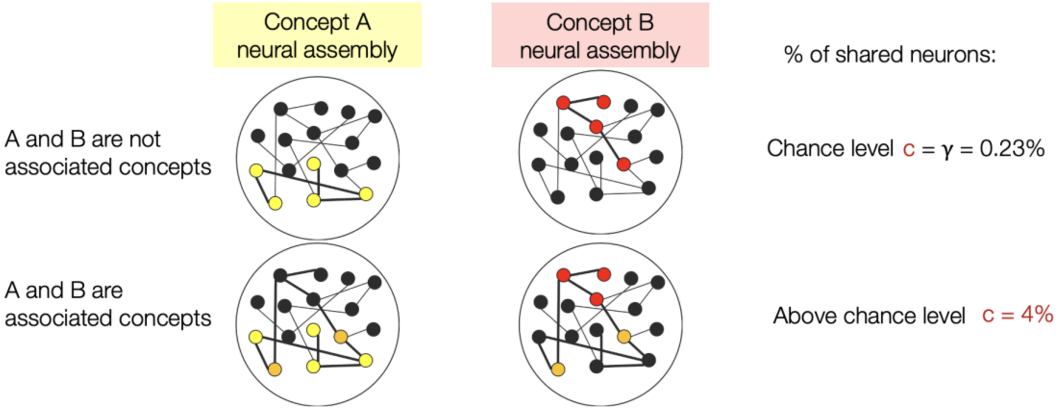


Figure 3 Schematic representation of the neural representation of two concepts in the human MTL.

Fundamental questions

In the presence of shared neurons, the activation of one engram (a place, in the initial example) may trigger the activation of another engram (such as a friend). This raises three fundamental issues which we address in this computational chapter:

First, for the brain to function effectively as a memory network, it must retain the ability to recall the two associated concepts separately (e.g., recalling the place without necessarily thinking of the friend). However, if the concepts share too many neurons, it becomes increasingly likely that the two concepts cannot be distinguished and instead merge into a broader engram counting a larger number of neurons. In the section "How can shared concept cells encode associations?" we address the following question: What is the maximum allowable fraction, c_{max}, of neurons shared between two assemblies before the possibility of separate memory recall breaks down?

Second, if mental concepts are related to engram cells, then associations between concepts should cause activation patterns of engrams cells, but how? The activation of a first concept can enhance the recall of a second, associated, concept that activates in parallel to the first one; or it may trigger a sequential activation of first, second, and potentially more concepts in the form of an association chain (as observed in free memory recall tasks [9–12]). If these transitions from one concept to the next are caused by neurons that are shared between engrams, we must pose a second question: is there a minimum fraction of shared neurons, c_{min}, required to facilitate a reliable activation of the associated concepts?

Neuronal assemblies as memory engrams

Memory engrams are mathematically linked to neural assemblies which play a central role in the network algorithm for memory retrieval that we will discuss in the section “Generalised Hopfield model”. Neuronal assemblies [45] are sub-networks of strongly connected neurons that represent an abstract concept. The assembly as a subgroup of strongly connected neurons has been an influential theoretical notion, introduced by Donald Hebb [45]. Moreover, Hebb suggested that the strong connections within the assembly could be the consequence of synaptic plasticity [45]. Finding real assemblies in the brain is technically challenging since neurons belonging to an assembly do not have to be neighbors but can be widely distributed across one, or even several, brain areas. If we define an assembly as a group of neurons that (i) respond simultaneously to a group of stimuli related to each other and (ii) have strong connectivity amongst themselves, then the second condition is hard to assess experimentally whereas the first condition has an obvious link to memory engrams.

The concept cells described above are primary candidates of neurons forming memory assemblies. We hypothesize that the simultaneous activation of concept cells by stimuli related to the concept is the result of strong recurrent connectivity within the assembly. Moreover, the fact that some, but not all, of the neurons show responses that persist after the end of the stimulus presentation [1, 5-7, 28] is an indirect indication that information is held in some form of working memory and could be a sign of strong interconnectivity. In this chapter, we link the Hebbian notion of neural “assembly” with the notion of “engram”: We assume that the set of neurons participating in an engram form a Hebbian assembly and have relatively strong connectivity amongst themselves. Furthermore, we assume that the groups of concept cells participating in the same concept are the human equivalent of memory engrams in mice. However, it is worth emphasizing that, as of today, these are hypotheses and we cannot exclude other explanations.

Modeling memory networks

Attractor networks

“Attractor neural networks” have been widely used to model memory systems in recurrent neuronal networks such as the CA3 area of the hippocampus [13–17]. In attractor networks, each memory item is encoded as a memory engram [18, 19] consisting of a fixed random subset of neurons. The joint activity of a large fraction of neurons that are part of the same engram represents the memory item and indicates that the memorized ‘concept’ is recalled.

An attractor network is a recurrent dynamical network, that evolves toward a stable state, called a “fixed point” of the network dynamics. In computational neuroscience, attractor networks are built such that the memory engrams are the fixed points of the network dynamics. If a stimulus activates a subset of neurons within an engram, the interactions of neurons within the network activate other neurons of the same engram and suppress the activity of other neurons that are not part of the engram. These interactions assure that the memory network has the auto-associative property introduced at the beginning of this chapter: the full concept is recalled (a large fraction of engram neurons are active together) triggered by partial information (the initial cue that stimulated a subset of neurons). From a birds-eye perspective, the network activity is 'attracted' toward a state of memory recall.

Bio-plausibility of Attractor Memory Networks

Animal studies have provided evidence of attractor dynamics in the CA3 area of the hippocampus [20, 21]. Since concept cells have been found in the human hippocampus and its surroundings [1-3], attractor dynamics in the hippocampus is a likely candidate to describe the activity of concept cells in humans - or engram cells in mice.

Attractor memory networks exhibit two key functional properties: (i) the ability to retrieve memories when presented with partial cues and (ii) the capacity to sustain activity even after the stimulus is no longer present. Traditionally, the analysis of attractor networks using the replica [35] or cavity methods [36,37] has drawn criticism due to the unrealistic assumption of symmetric connections. However, the approach presented in this chapter is based on dynamic systems arguments [38] and enables a straightforward generalization to the case of asymmetric connectivity. Moreover, in the original model of Hopfield [25] each memory engram involved fifty percent of neurons whereas in the human hippocampus, the fraction g of neurons involved in a single memory engram is at most 1 percent ('sparse' engrams) [4,6]. This number is estimated indirectly from the probability that an experimentalist would find a neuron responding to a, say, famous person known to the patient if an electrode is placed randomly in a given brain area, knowing the number of presented stimuli and the number of neurons from which the experimentalist recorded [4, 6]. Modern attractor network theory does not rely on symmetric connectivity and is characterized by sparse [17] memory engrams and random [34] connectivity. Therefore, modern attractor networks have emerged as promising models for understanding biological memory.

Generalized Hopfield model

We work with an attractor neural network, made of *N* neurons, which has stored *P* memory engrams. Each engram μ (with μ =1, …, *P)*, is represented by a string of random binary variables $\xi_i^\mu \in \{0,1\}$.
For example, $\xi_i^3 = 1$ and $\xi_i^4 = 1$ indicate that neuron i is part of both memory engram 3 and 4; similarly, $\xi_j^3 = 0$ and $\xi_j^4 = 1$ indicates that neuron j is not part of the memory engram 3, but part of engram 4 (Figure 4B - redraw the figure with less math).

The state of each neuron i is characterized by its firing rate r_i (with *i* = 1, . . ., *N)*. The change dr_i/dt of the firing rate of neuron i is driven by the total input h_i

Equation 1

$$\tau \frac{dr_i}{dt} = -r_i + \phi(h_i),$$

where $\phi(h_i) = r_{max}/\{1 + exp[-b(h - h_0)]\}$ is frequency-current (f-I) curve, (also called transfer function), characterized by the firing threshold h_0, the maximal steepness b, and the maximal firing rate r_{max}.
The total input driving the neuron i is

Equation 2

$$h_i(t) = \sum_{j=1}^{N} w_{ij} r_j(t) + I_i(t).$$

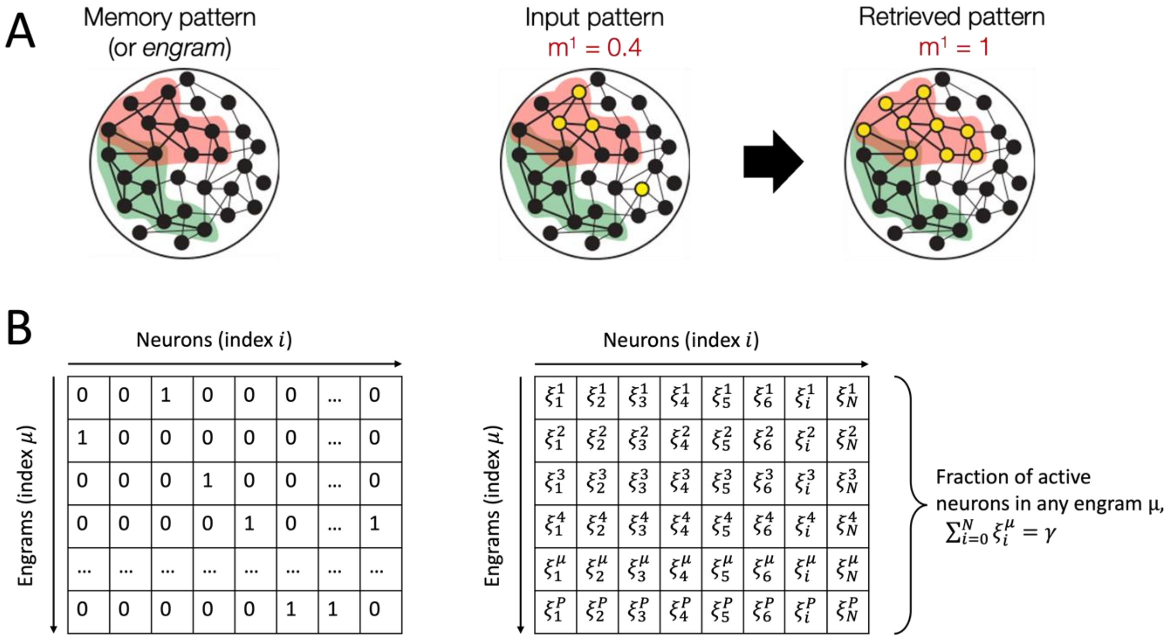
where the synaptic weights w_{ij} which can be interpreted as the strength of signal transmission from neuron j to neuron i is related to the amplitude of the (excitatory or inhibitory) postsynaptic current.

We assume the learning process has already happened in the past so we consider the weights as fixed. The engrams ξ_i^μ are encoded in synaptic weights w_{ij} as follows: the basic idea is that two neurons i and j that participate in the same engram (e.g., engram 4) have strong excitatory connections, in both directions; moreover, if neuron i participates in engram 3, but neuron j does not, then the connections between the two neurons are weakly inhibitory. However, since the connection from neuron j to neuron i cannot be both excitatory and inhibitory at the same time, we take the average by summing over the contributions of all engrams

Equation 3

$$w_{ij} = A \sum_{\mu=1}^{P} (\xi_i^\mu - \gamma)(\xi_j^\mu - \gamma)$$

Here, the constant *A* can be interpreted as a normalization factor for averaging and γ=0.0023 is the fraction of neurons that participate in a given engram in MTL [4]. If the sum on the right-hand side of Eq. (3) is positive, then neuron j has an excitatory connection to neuron i. The set of weights defined in Equation 3, is called the Hopfield-Tsodyks connectivity [17, 24] for sparse engrams. By default, we assume that two unrelated concepts (e.g., Sidney Opera and Tower of Pisa) share only a small number of neurons that correspond to the statistical expectation, i.e., the memory engrams are statistically independent.



0	0	1	0	0	0	…	0
1	0	0	0	0	0	…	0
0	0	0	1	0	0	…	0
0	0	0	0	1	0	…	1
…	…	…	…	…	…	…	…
0	0	0	0	0	1	1	0

ξ_1^1	ξ_2^1	ξ_3^1	ξ_4^1	ξ_5^1	ξ_6^1	ξ_i^1	ξ_N^1
ξ_1^2	ξ_2^2	ξ_3^2	ξ_4^2	ξ_5^2	ξ_6^2	ξ_i^2	ξ_N^2
ξ_1^3	ξ_2^3	ξ_3^3	ξ_4^3	ξ_5^3	ξ_6^3	ξ_i^3	ξ_N^3
ξ_1^4	ξ_2^4	ξ_3^4	ξ_4^4	ξ_5^4	ξ_6^4	ξ_i^4	ξ_N^4
ξ_1^μ	ξ_2^μ	ξ_3^μ	ξ_4^μ	ξ_5^μ	ξ_6^μ	ξ_i^μ	ξ_N^μ
ξ_1^P	ξ_2^P	ξ_3^P	ξ_4^P	ξ_5^P	ξ_6^P	ξ_i^P	ξ_N^P

Figure 4 A) Memory recall in a generalized Hopfield network. B) Schematics of the memory engrams as represented into the generalized Hopfield model.

Mathematical analysis, as well as simulations of the model, show that with the connections defined in Equation (3) and the firing rate changes defined in Equation (1), the network has (at least) P stable stationary states where each state corresponds to a configuration where nearly all neurons that should be active in engram μ are indeed active and all others are inactive. Since these configurations are stable states, the network dynamics are 'attracted' towards these special configurations - and these states correspond exactly to the retrieval of a stored concept. In the stationary state where engram μ is retrieved, neuronal firing rates have a fixed value r_i which is high if $\xi_i^\mu = 1$ and low if $\xi_i^\mu = 0$. Attraction means that when the configurations of firing rates across the network are similar to one of the stored memories μ, then the attractor dynamics drive the network to recall memory μ, by showing persistent activity of all those neurons that belong to the assembly of concept μ. Importantly, it is possible to measure how close (or how far) is the network state is to retrieve one of the stored engrams, μ, thanks to the similarity measures m^μ:

Equation 4

$$m^\mu(t) = \frac{A}{r_{max}} \sum_{i=1}^{N} \left(\xi_i^\mu - \gamma\right) r_i(t).$$

The similarity measures the correlation between the firing rates {*rj*(*t*)}, with *j*=1,. . ., *N,* and the stored engrams ξ_i^μ such that if memory concept μ is retrieved, then m^μ ~ 1 (schematics in Fig 4A), and, if no memory is recalled (*resting state*), then m^μ ~ 0 for all μ. The similarity of the network activity with a stored memory changes over time as the network state, given by the firing rates $r_i(t)$ of the individual neurons, changes.

Theory, computer simulations, and low-dimensional factors

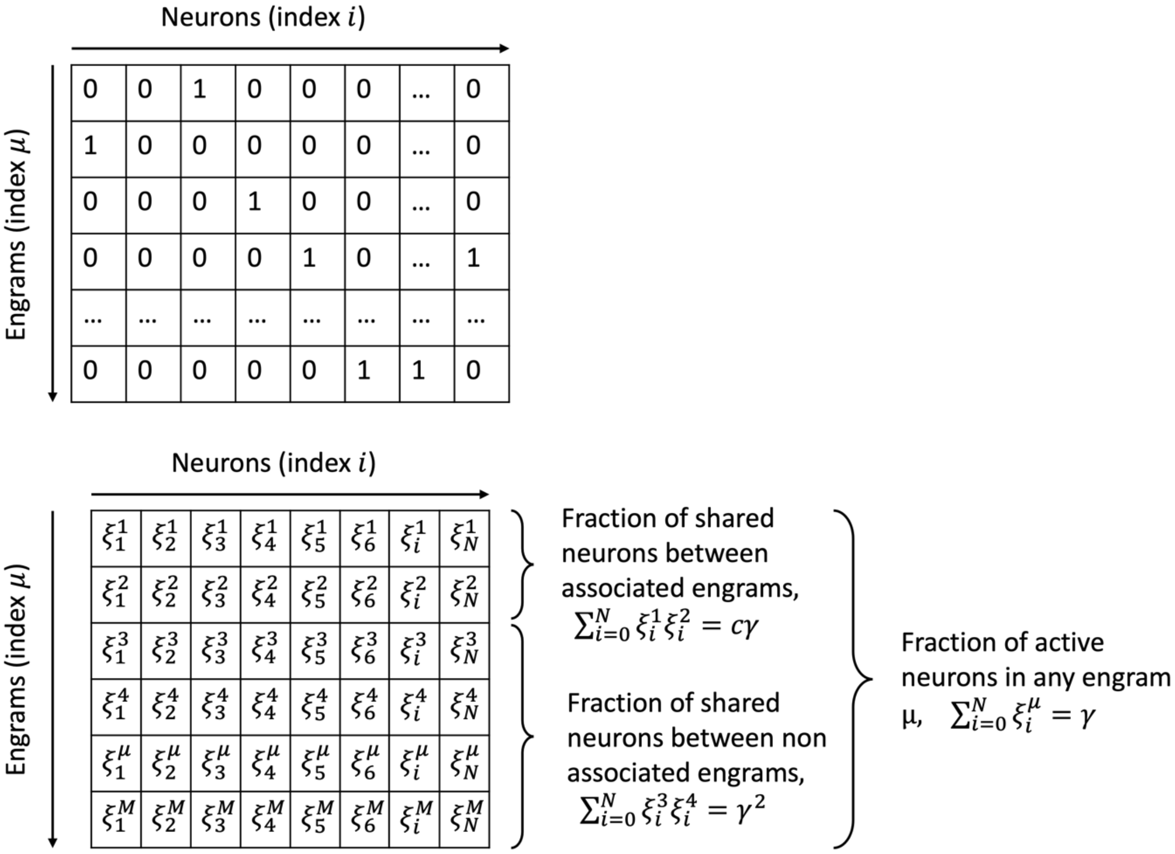
To answer the fundamental questions above, we consider a memory network with P memories stored and then focus on memory 1 and 2. By default their engram share neurons at chance level (memory 1 and 2 are not associated); alternatively, to implement associations between the two memories we also consider the case that the fraction of shared neurons is above chance. While the network dynamics can also be analyzed mathematically [43] we focus here on computer simulations, which can be done at two different levels of detail.

First, at the detailed level, we explicitly simulate a network of N neurons. This involves integrating numerically for each neuron the Equations (1) and (2) and (3). In other words, each neuron and each synaptic connection is modeled explicitly.

Second, at a coarser level, we use a common trick from the literature called the mean-field approach [13-17]. Such an approach assumes that the network is very large and that neurons belonging to the same memory engram behave similarly. It is then possible to fully describe the network dynamics using the similarity measures m^μ. Each similarity measure m^μ can be viewed as a single, network-wide variable that captures the global state of the network in terms of its similarity with memory μ. Since we are only interested in the retrieval of concepts μ = 1 and 2, we can assume the similarity of the present network state with other memories μ > 2 to be close to zero: we will refer to these non-activated memories as "background engrams". Under these assumptions, it is possible to derive dynamical mean-field equations that fully describe the network dynamics through the similarity variables m^1 and m^2. The details of such derivation are beyond the aim of this chapter and we refer to other texts to deepen the topic [34-39], however, we will illustrate the main results of this theoretical approach in what follows. The main take-home message is that the mean-field approach allows one to predict the collective behavior of a very large network using only a few equations and without the need to implement the dynamics of every single neuron in the network. The spirit of the mean-field approach is similar to that of identifying ‘factors’ [47] or low-dimensional neural manifolds [48] in experimental data.

How can shared concept cells encode associations?

As suggested in the subsection “Association between concepts'', the creation of episodic associations between different concepts (such as a person and a place) might be caused by common neurons shared across the corresponding memory engrams [5]. Drawing inspiration from these experiments, we artificially introduce shared neurons in the generalized Hopfield model to create pairwise associations between multiple concepts. We refer to "overlapping engrams'' when the number of shared concept cells exceeds the expected number of γ^2N cells shared by chance (see Fig. 5). We check by computer simulations whether increasing the overlap between two engrams causes a measurable increase in association performance.



Neurons (index i)
Engrams (index μ)

0	0	1	0	0	0	…	0
1	0	0	0	0	0	…	0
0	0	0	1	0	0	…	0
0	0	0	0	1	0	…	1
…	…	…	…	…	…	…	…
0	0	0	0	0	1	1	0

Neurons (index i)
Engrams (index μ)

ξ_1^1	ξ_2^1	ξ_3^1	ξ_4^1	ξ_5^1	ξ_6^1	ξ_i^1	ξ_N^1
ξ_1^2	ξ_2^2	ξ_3^2	ξ_4^2	ξ_5^2	ξ_6^2	ξ_i^2	ξ_N^2
ξ_1^3	ξ_2^3	ξ_3^3	ξ_4^3	ξ_5^3	ξ_6^3	ξ_i^3	ξ_N^3
ξ_1^4	ξ_2^4	ξ_3^4	ξ_4^4	ξ_5^4	ξ_6^4	ξ_i^4	ξ_N^4
ξ_1^μ	ξ_2^μ	ξ_3^μ	ξ_4^μ	ξ_5^μ	ξ_6^μ	ξ_i^μ	ξ_N^μ
ξ_1^M	ξ_2^M	ξ_3^M	ξ_4^M	ξ_5^M	ξ_6^M	ξ_i^M	ξ_N^M

Fraction of shared neurons between associated engrams, $\sum_{i=0}^{N} \xi_i^1 \xi_i^2 = c\gamma$
Fraction of shared neurons between non associated engrams, $\sum_{i=0}^{N} \xi_i^3 \xi_i^4 = \gamma^2$
Fraction of active neurons in any engram μ, $\sum_{i=0}^{N} \xi_i^\mu = \gamma$

Figure 5 Schematic representation of memory engrams

Let us imagine gradually increasing the fraction of shared neurons between the memory engrams 1 and 2 . At the lowest end, the two memories are not associated, since cell assemblies 1 and 2 share only a small fraction of neurons corresponding to chance level. It is well known, that in this case, each memory engram generates a separate attractive fixed point of the network dynamics [17], indicating that the two corresponding concepts can be retrieved separately. However, experimental data reports that, for associated concepts, the fraction of shared neurons *c* * 4–5% [6] is much larger than chance level *γ* ~0.23%. Let's now consider the case in which memory 1 and 2 share more neurons than by chance, i.e, the fraction c of shared neurons is larger than *γ*. In the upper limit case of a large fraction of shared neurons *c* close to 1, the two memory engrams share all neurons, and it is clearly impossible to retrieve one memory without the other. In other words, the two memories are indistinguishable so that there are no longer two memories, but only a single, larger one.

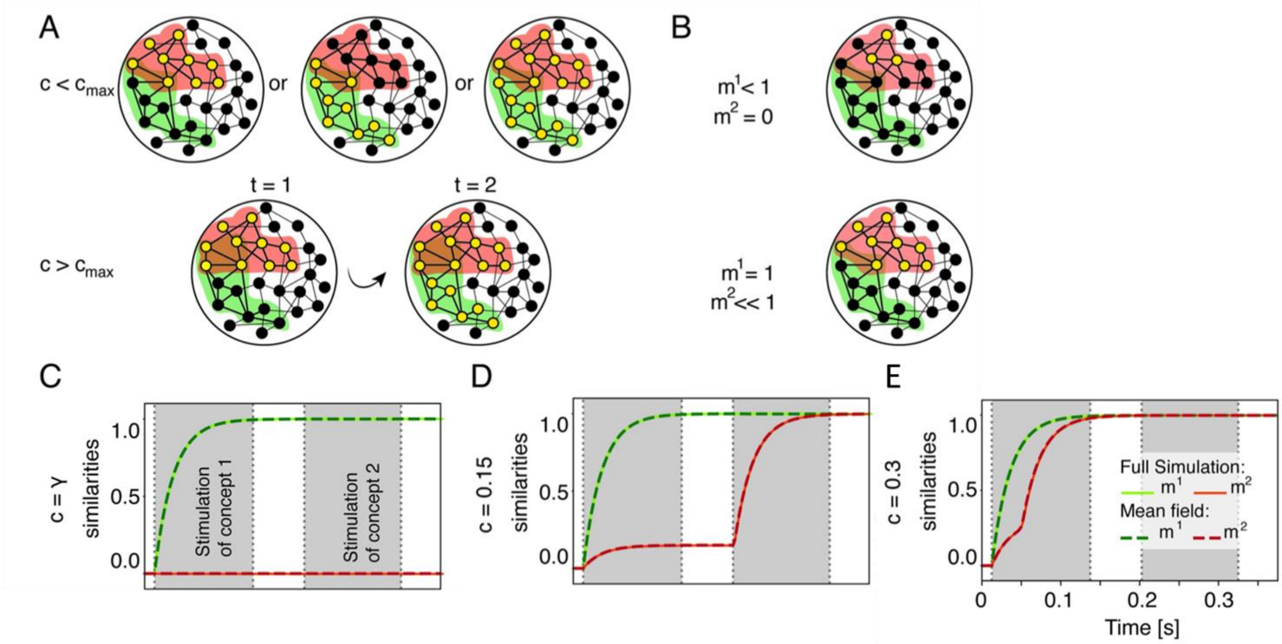


Figure 6 Overlapping concepts can be retrieved separately and jointly (adapted from [43]).

Computer simulations of a network of *N* = 10000 interacting neurons indicate that, if one of two engrams that share concept cells is stimulated for 120ms, then the similarity of the network activity with this engram increases to a value close to one, indicating that the memory has been recalled (Fig 6C middle) while the second memory is only weakly activated quantified by a small, but non-zero similarity. However, if the fraction of shared neurons is above a maximally allowed fraction c_{max}, then the second memory always gets activated even before it is stimulated (Fig 6C bottom) indicating that associations are so strong that the two concepts have been merged. Hence, the fraction of neurons shared between two engrams should be above chance but remain relatively low so as to guarantee that concepts (e.g. your friend and 'Sidney Opera') can remain separated if needed, but can also be recalled jointly if so desired.

Association chains

The notion of common neurons shared among memory engrams has also been proposed as the foundation for recalling a list of memorized words. In earlier research [9–12, 26], Romani and Tsodyks analyzed human behavioral experiments, during which the subjects had some time to memorize a list of words and then to freely recall as many words as possible (Fig.7). They noticed that the subjects doing better at the task were those who associated the words in the list in small groups and they seemed to recall the words following those personal associations. The same scientists also proposed a computational explanation [9-12] of the experimental results, using a generalized Hopfield model similar to that introduced above, but with two additional components that we now add to our model. Firstly, we introduce global inhibitory feedback that is periodically modulated in strength mimicking hippocampal oscillatory activity. These oscillations serve as a clock signal, triggering transitions between overlapping concepts. Secondly, we introduce an adaptation current, $\theta_i(t)$, to each neuron i, preventing the network state from immediately reverting to the previous concept. With this expanded model, the network state transitions from one concept to the next (Fig 8A). These transitions are repeated, but eventually, the network state returns to one of the previously retrieved memories, resulting in a cyclic pattern [9] (Fig 8A).

In network simulations where concepts are represented by sparse memory engrams (γ = 0.2%), we allow a subgroup of p = 2, 4, or 16 memory engrams to share a fraction c = 20% of neurons. As the number of shared concept cells is identical across all concept pairs within the same subgroup, the order of recalled concepts is dependent on the initial condition. When the subgroup of overlapping engrams is small (p = 2 or 4), all memory items are successfully retrieved. However, in the case of a larger group of overlapping engrams (p = 16), the cycle closes once a subgroup of the overlapping memory engrams has been recalled (Fig 8B).

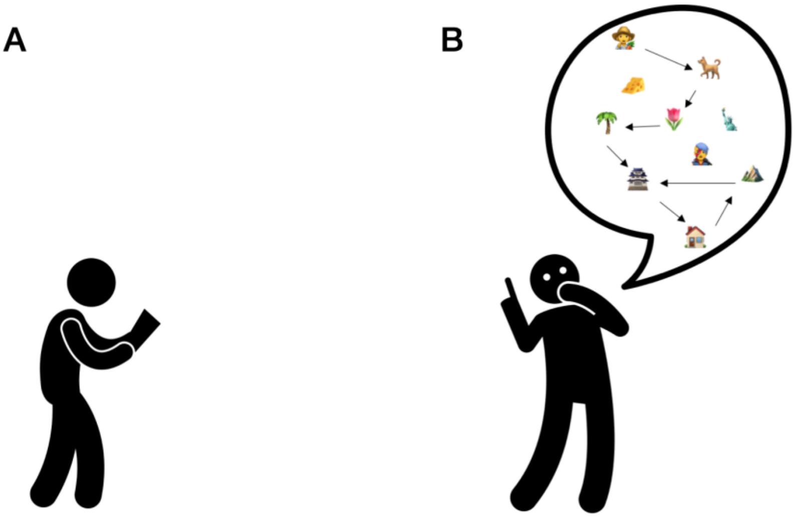


Figure 7 Memory recall task.

The previous studies [9–12, 26] have assumed that memory engrams have a substantial fraction (γ = 10%) of active neurons, allowing transitions to rely on the chance-based number of shared units. However, considering the much lower sparsity value in the MTL (γ = 0.23%), it is natural to question whether the number of neurons shared by chance ($c = \gamma$) is sufficient to induce a sequence of memory retrievals. The simulations from [43] demonstrate that this is not the case (Fig 8C). In a memory network with a realistic level of sparsity ($\gamma \sim 0.2\%$), associations between memory engrams require a fraction of shared neurons above chance level for the successful retrieval of concept chains.

The same conclusions are confirmed and reinforced by the mean-field approach. With the theoretical approach, it is possible to determine the lower bound of the fraction of shared neurons, denoted as c_{0min}, corresponding to the minimum overlap between two engrams that is required for a reliable transition. Importantly, with suitable choices of neuronal and network parameters, association chains can be achieved for the values of γ and c as observed in the human MTL. This suggests that, in principle, associations can be implemented as sequences of transitions if the number of shared neurons exceeds c_{min}.

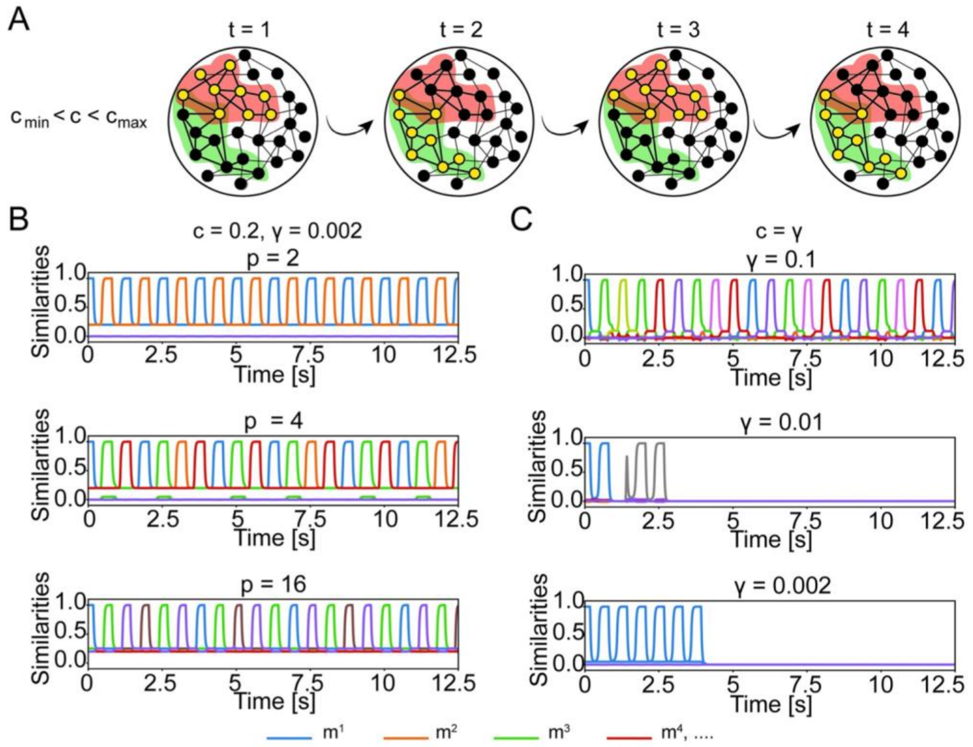


Figure 8 Chain of associations requires shared concepts cells Adapted from [43].

In summary, overlaps between memory engrams are necessary to encode associations. More precisely, the fraction of concept cells shared across concepts must significantly exceed chance level, to account for the phenomenon of free memory recall as a chain of associations in recurrent networks like the CA3 region of the human brain. Notably, in these networks, each memory engram is represented by only a small fraction of neurons.

Conclusions

This chapter bridges experimental observations and theories from four distinct fields: (1) experimental investigations of concept cells in the human MTL, (2) empirical studies on memory engrams in mice, (3) the theory of association chains for free memory recall, and (4) the classic theory of attractor neural networks. Our theory assumes that concepts are represented by the activation of sparse subgroups of neurons participating in the same engram. If engrams represent the memorized concepts, then associations between concepts can be implemented by an increased fraction of concept cells shared across engrams. This increased fraction must surpass the chance level but remain below a certain maximal threshold to enable reliable encoding of associations. Experimental evidence indicates a 4-5% overlap between memory engrams in the human MTL [6] - and these numbers lie in the range of values that are supported by simulations of association chains [43].

Association chains could form the basis of a “stream of thought” where the direction of transitions from one concept to the next is based on learned associations. In large networks with sparse coding levels (γ ~ 0.23%), neurons shared by chance are not enough to reliably induce the retrieval of a chain of concepts. Sequential memory retrieval is possible only for overlaps larger than chance, potentially representing associations learned during real-life episodes. Potentially, the fraction of shared concept cells could increase due to Hebbian learning if two concepts repeatedly occur simultanesoulsy. However, the existence of a maximal fraction of shared neurons highlights the need for Hebbian learning to operate in conjunction with an intrinsic control mechanism to prevent the undesired merging of distinct concepts.

Overall, our computational approach shows that the creation and retrieval of memory engrams is amenable to mathematical analysis. Here we focused on the challenges and advantages of overlapping memory engrams - and for this we had to rely on data from human MTL, but future experiments in mice might bring further evidence on how overlapping engrams can build or reinforce associations between memory concepts.

Other applications of Hopfield and generalized Hopfield networks

Other areas: ITC

The area CA3 of the hippocampus is not the only one that can be modeled with attractor neural networks. Indeed, this model was first introduced with the intention of modeling cortical areas. The Inferior Temporal Cortex (ITC) has often been taken as an example area to be modeled with generalized Hopfield networks. [15-21, 25, 46]

The Inferotemporal Cortex (IT) is a region of the brain that plays a crucial role in processing visual stimuli related to objects within our field of vision. It is responsible for extracting complex visual features and attributes, enabling us to recognize and identify objects.

One of the key functions of the IT cortex is its involvement in memory and memory recall processes. Once we have seen and processed an object, the IT cortex stores representations of its visual features and characteristics as memory engrams. These memory engrams are subsequently utilized during memory recall to recognize and identify the object when encountered again.

We can represent the objects memorized in the ITC as attractor states within the network. Each attractor state corresponds to a particular object or category, and when the network is presented with a visual input, it undergoes dynamics that converge to the nearest attractor state, thus recognizing the object in the input.

Non-binary engrams

Above we have considered the so-called “binary” memory engrams, where a neuron either participates in the memory engram or it does not. This is a modeling simplification, but experimental results do not exclude other types of memory engrams. Indeed, in the concept cells experiments a neuron [1, 5-7, 28] is considered to be a concept cell if its firing rate is deviating of at least 3 or 5 (depending on the experiment) standard deviations from the firing rate distribution of all other recorded neurons during the stimulus presentation. This means that the criterion of assigning concept cells is binary but not necessary the concept cells responds, which might follow a different distribution.

It is possible to define the generalized Hopfield model for non-binary firing rate engrams, for example in [24] they proposed a model for Gaussian firing rate engrams. In this case retrieving a memory means that the neurons' firing rate follows the same Gaussian distribution of the equivalent memory engram.

Spiking Hebbian networks

So far, we have considered the firing rate of neurons as the only relevant parameter of the neuron. This is an approximation, but ideally, we would like to achieve the same type of memory neural network with more realistic spiking neurons [41, 42, 49].

Moreover, in the generalized Hopfield model, synaptic weights can be both positive and negative, allowing for bidirectional signaling. However, experimental observations have revealed a phenomenon known as Dale's law. According to Dale's law, all connections originating from the same presynaptic neuron have the same sign, either excitatory or inhibitory. This empirical finding has led to the primary classification of neurons into two categories: excitatory neurons, which promote the firing of postsynaptic neurons, and inhibitory neurons, which suppress or inhibit the firing of postsynaptic neurons.

It is possible to create a more detailed and realistic computational spiking neuron model in which Dale's law is respected, yet it preserves the same dynamical behavior of a generalized Hopfield network. An example of such a spiking neural network is presented in Fig. 10, where excitatory and inhibitory neurons have been separated. In a neural network scenario, there exists a population of excitatory neurons that interact with two distinct populations of inhibitory neurons. The encoding of memory engrams occurs through the formation of Hebbian assemblies within the excitatory population. All neurons in the network follow an integrate-and-fire behavior.

According to theoretical predictions, the first inhibitory population is expected to be activated to a degree where the gain function (as shown in the left inset) exhibits an approximately linear response. This linear activation range ensures effective regulation of excitatory neuron activity and helps maintain a stable network state.

On the other hand, the activation of the second inhibitory population occurs when the total input to the network surpasses a specific threshold value (as depicted in the right inset). This threshold-based activation mechanism serves as a control mechanism to prevent excessive excitation or maintain the stability of the network by suppressing neuronal firing when necessary.

Overall, this architecture and activation scheme involving excitatory and inhibitory populations contribute to the dynamic regulation of neural activity and play a crucial role in shaping the network's behavior and information processing capabilities.

Bibliography

1. Quiroga R Quian, Reddy Leila, Kreiman Gabriel, Koch Christof, and Fried Itzhak. Invariant visual representation by single neurons in the human brain. *Nature*, 435(7045):1102, 2005. https://doi.org/10.1038/ nature03687 PMID: 15973409
2. Ison Matias J and Quiroga Rodrigo Quian. Selectivity and invariance for visual object perception. *Front Biosci*, 13:4889–4903, 2008. https://doi.org/10.2741/3048 PMID: 18508554
3. Quiroga Rodrigo Quian. Neural representations across species. *Science*, 363(6434):1388–1389, 2019. https://doi.org/10.1126/science.aaw8829 PMID: 30923208
4. Waydo Stephen, Kraskov Alexander, Quiroga Rodrigo Quian, Fried Itzhak, and Koch Christof. Sparse representation in the human medial temporal lobe. *Journal of Neuroscience*, 26(40):10232–10234, 2006. https://doi.org/10.1523/JNEUROSCI.2101-06.2006 PMID: 17021178
5. Ison Matias J, Quiroga Rodrigo Quian, and Fried Itzhak. Rapid encoding of new memories by individual neurons in the human brain. *Neuron*, 87(1):220–230, 2015. https://doi.org/10.1016/j.neuron.2015.06. 016 PMID: 26139375
6. De Falco Emanuela, Ison Matias J, Fried Itzhak, and Quiroga Rodrigo Quian. Long-term coding of personal and universal associations underlying the memory web in the human brain. *Nature communica- tions*, 7:13408, 2016. https://doi.org/10.1038/ncomms13408 PMID: 27845773
7. Rey Hernan G, De Falco Emanuela, Ison Matias J, Valentin Antonio, Alarcon Gonzalo, Selway Richard, Richardson Mark P, and Quiroga Rodrigo Quian. Encoding of long-term associations through neural unitization in the human medial temporal lobe. *Nature Communications*, 9(1):1–13, 2018. https://doi. org/10.1038/s41467-018-06870-2 PMID: 30348996
8. Rey Hernan G., Gori Belen, Chaure Fernando J., Collavini Santiago, Blenkmann Alejandro O., Seoane Pablo, Seoane Eduardo, Kochen Silvia, and Quiroga Rodrigo Quian. Single neuron coding of identity in the human hippocampal formation. *Current Biology*, 30(6):1152–1159.e3, 2020. https://doi.org/10. 1016/j.cub.2020.01.035 PMID: 32142694
9. Romani Sandro, Pinkoviezky Itai, Rubin Alon, and Tsodyks Misha. Scaling laws of associative memory retrieval. *Neural Computation*, 25(10):2523–2544, 2013. https://doi.org/10.1162/NECO_a_00499 PMID: 23777521
10. Recanatesi Stefano, Katkov Mikhail, Romani Sandro, and Tsodyks Misha. Neural network model of memory retrieval. *Frontiers in computational neuroscience*, 9:149, 2015. https://doi.org/10.3389/ fncom.2015.00149 PMID: 26732491
11. Recanatesi Stefano, Katkov Mikhail, and Tsodyks Misha. Memory states and transitions between them in attractor neural networks. *Neural computation*, 29(10):2684–2711, 2017. https://doi.org/10.1162/ neco_a_00998 PMID: 28777725
12. Naim Michelangelo, Katkov Mikhail, Romani Sandro, and Tsodyks Misha. Fundamental law of memory recall. *Physical Review Letters*, 124(1):018101, 2020. https://doi.org/10.1103/PhysRevLett.124. 018101 PMID: 31976719
13. Hopfield John J. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982. https://doi.org/10.1073/ pnas.79.8.2554 PMID: 6953413
14. Weisbuch Gerard and Fogelman-Soulie´ Françoise. Scaling laws for the attractors of Hopfield networks. *Journal de Physique Lettres*, 46(14):623–630, 1985. https://doi.org/10.1051/ jphyslet:019850046014062300
15. Amit Daniel J and Amit Daniel J. *Modeling brain function: The world of attractor neural networks*. Cambridge University Press, 1992.
16. Kanter I and Sompolinsky Haim. Associative recall of memory without errors. *Physical Review A*, 35 (1):380, 1987. https://doi.org/10.1103/PhysRevA.35.380 PMID: 9897963
17. Tsodyks Mikhail V and Feigel'man Mikhail V. The enhanced storage capacity in neural networks with low activity level. *EPL (Europhysics Letters)*, 6(2):101, 1988. https://doi.org/10.1209/0295-5075/6/2/ 002
18. Tonegawa Susumu, Pignatelli Michele, Roy Dheeraj S, and Ryan Toma´s J. Memory engram storage and retrieval. *Current opinion in neurobiology*, 35:101–109, 2015. https://doi.org/10.1016/j.conb.2015. 07.009 PMID: 26280931

19. Josselyn Sheena A and Tonegawa Susumu. Memory engrams: Recalling the past and imagining the future. *Science*, 367 (6473), 2020. https://doi.org/10.1126/science.aaw4325 PMID: 31896692
20. Renno´ -Costa Ce´ sar, Lisman John E, and Verschure Paul FMJ. A signature of attractor dynamics in the ca3 region of the hippocampus. *PLoS Comput Biol*, 10(5):e1003641, 2014. https://doi.org/10.1371/ journal.pcbi.1003641 PMID: 24854425
21. Wills Tom J, Lever Colin, Cacucci Francesca, Burgess Neil, and O'Keefe John. Attractor dynamics in the hippocampal representation of the local environment. *Science*, 308(5723):873–876, 2005. https:// doi.org/10.1126/science.1108905 PMID: 15879220
22. ~~Bo¨ s Siegfried, Ku¨ hn R, and van Hemmen JL. Martingale approach to neural networks with hierarchically structured information. *Zeitschrift fu¨ r Physik B Condensed Matter*, 71(2):261–271, 1988. https://doi. org/10.1007/BF01312798~~
23. ~~Boboeva Vezha, Brasselet Romain, and Treves Alessandro. The capacity for correlated semantic memories in the cortex. *Entropy*, 20(11):824, 2018. https://doi.org/10.3390/e20110824 PMID: 33266548~~
24. Pereira Ulises and Brunel Nicolas. Attractor dynamics in networks with learning rules inferred from in vivo data. *Neuron*, 99(1):227–238.e4, 2018. https://doi.org/10.1016/j.neuron.2018.05.038 PMID: 29909997
25. Hopfield J. J. Neurons with graded response have collective computational properties like those of two- state neurons. *Proceedings of the National Academy of Sciences*, 81(10):3088–3092, 1984. https:// doi.org/10.1073/pnas.81.10.3088 PMID: 6587342
26. Katkov Mikhail, Romani Sandro, and Tsodyks Misha. Effects of long-term representations on free recall of unrelated words. *Learning & Memory*, 22(2):101–108, 2015. https://doi.org/10.1101/lm.035238.114 PMID: 25593296
27. Guzman Segundo Jose, Schlo¨ gl Alois, Frotscher Michael, and Jonas Peter. Synaptic mechanisms of pattern completion in the hippocampal ca3 network. *Science*, 353(6304):1117–1123, 2016. https://doi. org/10.1126/science.aaf1836 PMID: 27609885
28. Quiroga Rodrigo Quian. Concept cells: the building blocks of declarative memory functions. *Nature Reviews Neuroscience*, 13(8):587–597, 2012. https://doi.org/10.1038/nrn3251 PMID: 22760181
29. Quiroga Rodrigo Quian. Plugging in to human memory: advantages, challenges, and insights from human single-neuron recordings. *Cell*, 179(5):1015–1032, 2019. https://doi.org/10.1016/j.cell.2019.10. 016
30. Podlaski William F, Agnes Everton J, and Vogels Tim P. Context-modular memory networks support high-capacity, flexible, and robust associative memories. *BioRxiv*, 2020.
31. Russo Eleonora, Namboodiri Vijay MK, Treves Alessandro, and Kropff Emilio. Free association transitions in models of cortical latching dynamics. *New Journal of Physics*, 10(1):015008, 2008. https://doi. org/10.1088/1367-2630/10/1/015008
32. Russo Eleonora and Treves Alessandro. Cortical free-association dynamics: Distinct phases of a latching network. *Physical Review E*, 85(5):051920, 2012. https://doi.org/10.1103/PhysRevE.85.051920 PMID: 23004800
33. Akrami Athena, Russo Eleonora, and Treves Alessandro. Lateral thinking, from the Hopfield model to cortical dynamics. *Brain research*, 1434:4–16, 2012. https://doi.org/10.1016/j.brainres.2011.07.030 PMID: 21839426
34. Amit Daniel J and Brunel Nicolas. Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex. *Cerebral cortex (New York, NY: 1991)*, 7(3):237–252, 1997. PMID: 9143444
35. Amit Daniel J, Gutfreund Hanoch, and Sompolinsky Haim. Information storage in neural networks with low levels of activity. *Physical Review A*, 35(5):2293, 1987. https://doi.org/10.1103/PhysRevA.35.2293 PMID: 9898407
36. Mezard Marc, Parisi Giorgio, and Virasoro Miguel. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific Publishing Company, 1987.
37. Shamir Maoz and Sompolinsky Haim. Thouless-Anderson-Palmer equations for neural networks. *Physical Review E*, 61(2):1839, 2000. https://doi.org/10.1103/PhysRevE.61.1839 PMID: 11046469
38. Shiino Masatoshi and Fukai Tomoki. Self-consistent signal-to-noise analysis and its application to analogue neural networks with asymmetric connections. *Journal of Physics A: Mathematical and General*, 25(7):L375, 1992. https://doi.org/10.1088/0305-4470/25/7/017

39. Amit Daniel J, Gutfreund Hanoch, and Sompolinsky Haim. Storing infinite numbers of patterns in a spin- glass model of neural networks. *Physical Review Letters*, 55(14):1530, 1985. https://doi.org/10.1103/ PhysRevLett.55.1530 PMID: 10031847
40. Andersen Per, Morris Richard, Amaral David, Bliss Tim, and O'Keefe John. *The hippocampus book*. Oxford University Press, 2006.
41. Wilson Hugh R and Cowan Jack D. Excitatory and inhibitory interactions in localized populations of model neurons. *Biophysical journal*, 12(1):1–24, 1972. https://doi.org/10.1016/S0006-3495(72)86068- 5 PMID: 4332108
42. Muscinelli S. P., Gerstner W., and Schwalger T. How single neuron properties shape chaotic dynamics and signal transmission in random neural networks. *PLOS Comput. Biol.*, 15(6):e1007122, 06 2019. https://doi.org/10.1371/journal.pcbi.1007122 PMID: 31181063
43. Gastaldi, Chiara, Tilo Schwalger, Emanuela De Falco, Rodrigo Quian Quiroga, and Wulfram Gerstner. "When shared concept cells support associations: Theory of overlapping memory engrams." *PLOS Computational Biology* 17, no. 12 (2021): e1009691.
44. ~~Gerstner, Wulfram, Werner M. Kistler, Richard Naud, and Liam Paninski. *Neuronal dynamics: From single neurons to networks and models of cognition*. Cambridge University Press, 2014.~~
45. D. O. Hebb (1949) *The Organization of Behavior*. Wiley, New York.
46. Khona, M., Fiete, I.R. Attractor and integrator networks in the brain. *Nat Rev Neurosci* **23**, 744–766 (2022). https://doi.org/10.1038/s41583-022-00642-0
47. DePasquale, Brian, Mark M. Churchland, and L. F. Abbott. "Using firing-rate dynamics to train recurrent networks of spiking model neurons." *arXiv preprint arXiv:1601.07620*(2016).
48. Gallego, Juan A., Matthew G. Perich, Lee E. Miller, and Sara A. Solla. "Neural manifolds for the control of movement." *Neuron* 94, no. 5 (2017): 978-984.
49. Zenke, Friedemann, Everton J. Agnes, and Wulfram Gerstner. "Diverse synaptic plasticity mechanisms orchestrated to form and retrieve memories in spiking neural networks." *Nature communications* 6, no. 1 (2015): 6922.