Place cells may simply be memory cells: Memory compression leads to spatial tuning and history dependence

Marcus K. Benna[a,b,c,1,2] and Stefano Fusi[a,b,d,1,2]

[a]Center for Theoretical Neuroscience, Columbia University, New York, NY 10027; [b]Mortimer B. Zuckerman Mind Brain Behavior Institute, Columbia University, New York, NY 10027; [c]Neurobiology Section, Division of Biological Sciences, University of California San Diego, La Jolla, CA 92093; and [d]Kavli Institute for Brain Sciences, Columbia University, New York, NY 10027

Edited by James McClelland, Center for Mind, Brain and Computation, Department of Psychology, Stanford University, Stanford, CA; received September 2, 2020; accepted November 2, 2021

The observation of place cells has suggested that the hippocampus plays a special role in encoding spatial information. However, place cell responses are modulated by several nonspatial variables and reported to be rather unstable. Here, we propose a memory model of the hippocampus that provides an interpretation of place cells consistent with these observations. We hypothesize that the hippocampus is a memory device that takes advantage of the correlations between sensory experiences to generate compressed representations of the episodes that are stored in memory. A simple neural network model that can efficiently compress information naturally produces place cells that are similar to those observed in experiments. It predicts that the activity of these cells is variable and that the fluctuations of the place fields encode information about the recent history of sensory experiences. Place cells may simply be a consequence of a memory compression process implemented in the hippocampus.

sparse autoencoders | place cells | hippocampus | memory | compression

Several studies show that neurons in the hippocampus encode the position of the animal in its environment, and as a consequence, they have been named "place cells" (e.g., ref. 1). Here, we propose an interpretation of the observation of place cells by suggesting that their response properties actually reflect a process of memory compression in which the hippocampus plays a fundamental role. The hypothesis that the hippocampus is involved in memory compression has already been proposed in several works, including the article by Gluck and Myers (2). These works mostly focused on how memories are recoded in order to be stored more efficiently (see also refs. 3–6 and the discussion below). Here, we show that a simple neural circuit implementing this memory compression process naturally leads to the formation of place cells. Hence, our interpretation of place cells is based on the idea that the hippocampus is essentially a memory device, and therefore, it contributes to the reconciliation between two dominant but apparently different points of view: one involving the hippocampus in spatial cognitive maps and navigation vs. another one that considers the hippocampus playing a broad role in episodic and declarative memory (e.g., refs. 7 and 8).

The first view is supported by the observation of place cells, some of which exhibit responses that are easily interpretable as the cells tend to fire only when the animal is in one particular location (single-field place cells). However, it is becoming clear that in many brain areas, including the hippocampus and entorhinal cortex (EC), neural responses are very diverse (9–12), highly variable in time (13–17), and modulated by multiple variables (12, 18–20). Place cells might respond at single or multiple locations, and multiple visitations of the same location typically elicit different responses. Part of this diversity can be explained by assuming that each neuron responds nonlinearly to a different combination of multiple external or internal variables (mixed selectivity) (e.g., ref. 10). The variability might be due to the fact that some of these variables are not being monitored in the experiment and hence, contribute to what appears to be noise. Some of the components of the variability probably depend on the variables that are represented at the current time, but some others might also depend on the recent history; in other words, they might be affected by the storage of recent memories.

The model of the hippocampus we propose is based on the idea that it is more efficient to compress memories before they are stored. Our model not only predicts that place cells should exhibit history effects but also predicts that their spatial tuning properties simply reflect an efficient strategy for storing correlated patterns. Much of the theoretical work on the memory capacity of neural networks is based on the assumption that the patterns representing memories are random and uncorrelated (e.g., refs. 21–25). This assumption is not unreasonable for long-term storage, despite the fact that most of our sensory experiences are highly correlated. Indeed, to efficiently store correlated episodes, it is desirable to preprocess or recode the new memories and compress them before they are placed in long-term memory, as already proposed in refs. 2–4, 26, and 27. Ideally, one would want to extract the uncorrelated (and hence,

Significance

Numerous studies on primates revealed the importance of the hippocampus in memory formation. The rodent literature instead focused on the spatial representations that are observed in navigation experiments. Here, we propose a simple model of the hippocampus that reconciles the main findings of the primate and rodent studies. The model assumes that the hippocampus is a memory system that generates compressed representations of sensory experiences using previously acquired knowledge about the statistics of the world. These experiences can then be memorized more efficiently. The sensory experiences during the exploration of an environment, when compressed by the hippocampus, lead naturally to spatial representations similar to those observed in rodent studies and to the emergence of place cells.

Author contributions: M.K.B. and S.F. designed research, performed research, analyzed simulation results, and wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

See online for related content such as Commentaries.

[1]M.K.B. and S.F. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: mbenna@ucsd.edu or sf2237@columbia.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2018422118/-/DCSupplemental.

Published December 16, 2021.

incompressible) components of the new input patterns and store only those. However, this form of preprocessing and compression has not been explicitly modeled in previous theoretical studies of memory capacity, like those discussed in ref. 22, and the performance improvement has not been estimated.

We hypothesize that this preprocessing is to some extent carried out in the hippocampus. The idea that the hippocampus is involved in memory storage and could implement some simple form of autoassociative memory is old, already proposed by David Marr in the 1970s (26). Autoassociative memories are naturally implemented by autoencoders whose function is basically to reconstruct the memory to be recalled. Autoencoders could be realized by a neural network with only two layers, an input layer and a reconstruction layer (6). These autoencoders would be equivalent to one step of the recurrent dynamics of a Hopfield network (21). However, it was clear from the early days that an intermediate layer that transforms the representations of the input can make the autoencoder significantly more efficient. This is why several mechanisms for recoding the inputs were proposed already in ref. 26 and then in refs. 2–4, 6, 27, and 28. The original idea behind many of these mechanisms is to orthogonalize the representations in the intermediate layer in order to make them more separable and facilitate the storage and reconstruction of memories. This is an approach that tries to mitigate the disruptive effects of correlations between sensory episodes that are similar to each other. However, it is possible to transform the representations in a way that not only avoids the problems of correlations but actually takes advantage of the similarities between memories to store a larger number of them. One possibility is to compress the information (2–5, 27, 29) as in a sparse autoencoder (30, 31). This is also a popular approach in the machine learning community and used not only to solve memory capacity problems. Indeed, the minimum description length (MDL) principle is often invoked to construct an efficient statistical model of the world. The MDL holds that the most compact (or compressed) description of the data is probably the best model of the observed data in terms of generalization. A neural network implementation of this principle also requires sparse compressed representations (e.g., ref. 32).

We constructed our model of the hippocampus following the sparse autoencoder approach; we assumed that the hippocampus is able to take advantage of the correlations between memories by building sparse compressed representations. We show that a neural system implementing this compression process is broadly consistent with the known neuroanatomy of the hippocampus; it allows efficient storage of a large number of correlated memories, and importantly, it naturally leads to neural representations that contain place cells.

Taking inspiration from the simplest classical architecture of autoencoders, our model consists of a three-layer network. The first layer represents the sensory inputs, and the feed-forward synapses that connect it to the second layer are continuously modified to create a compressed, sparse representation of the inputs. This layer implements a form of statistical learning, as the compressed representations are based on the statistics of recent sensory experiences. A third layer is used to store the memories (i.e., specific episodes). This architecture and the computational principles are similar to those proposed in refs. 2–5. We show that the plasticity of the feed-forward synapses improves the memory capacity compared with a network of the same architecture but with fixed, random feed-forward weights.

Furthermore, the model explains quite naturally the emergence of place cells in the second layer and the third layer. Compressing sensory inputs of an animal in a given environment automatically leads to the emergence of cells whose activity is strongly modulated by its position in the environment because many experiences of the animal depend on its position, and the latter is thus highly informative about the former. By the same token, processing sensory inputs correlated with other external variables would encourage cells to develop receptive fields in the space of those variables (33) since the computational principle of compressing inputs for efficient storage is agnostic to the nature of the variables that induce the correlations.

Such models with ongoing plasticity predict that the neural representation of a sensory episode will differ depending on the previous experiences of the animal: that is, it will be history dependent. In particular, synaptic weights are constantly modified (and correspondingly, the neural tuning properties change) in order to update the statistical model of the environment and to store new episodes in memory. The resulting place cell responses are modulated by any variable that describes relevant aspects of the sensory experiences. We will show in simulations that even in the absence of salient events, such as the delivery of a reward, the place fields constantly fluctuate to reflect the recent changes in the input statistics, although they remain sufficiently stable to decode position.

Results

Storing Correlated Patterns Efficiently. Most of the patterns we store in memory are likely to be highly correlated, as our experiences are often similar to each other. Storing correlated patterns in artificial neural networks typically requires a synaptic learning rule that is more complicated than simple Hebbian plasticity to avoid a bias toward the memory components shared by multiple memories. Simple extensions, such as the perceptron learning rule, can deal with many forms of correlations. However, storing correlated patterns in their original format is rarely the optimal strategy, and it is possible to greatly increase the memory capacity by constructing compressed neural representations that explicitly take into account correlations. This form of preprocessing or recoding can be illustrated with a simple example in which the patterns to be memorized are organized in an ultrametric tree (e.g., ref. 34) (Fig. 1*A*). To generate these correlated patterns, one starts from p uncorrelated random patterns (the ancestors

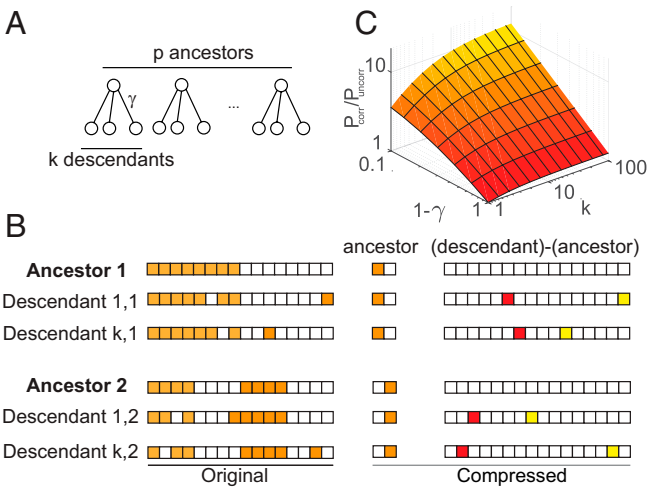

Fig. 1. Storing efficiently correlated patterns in memory. (*A*) Schematic of an ultrametric tree with p ancestors and k descendants per ancestor used to generate correlated patterns. (*B*) A possible scheme to take advantage of the correlations and generate compressed representations that are sparse and hence, more efficiently storable. (*C*) Total number P_{corr} of correlated patterns generated from a tree model with parameters p, k, and γ that can be stored using a simple compression strategy, divided by the number of patterns P_{uncorr} that could be stored (using approximately the same number of neurons and synapses) if the patterns were uncorrelated. The plot thus shows the relative advantage of using a compression strategy compared with storing incompressible patterns as a function of k and γ.

at the top of the tree). In these patterns, each neuron is either active or inactive with equal probability, as in the case of the Hopfield model (21). One can then generate k descendants for each ancestor by resampling randomly with a certain probability $1-\gamma$ the activation state of each of the neurons in the ancestors. k is called the branching ratio of the tree. The total number of descendants is $p\,k$, and these are the patterns that we intend to store in memory. The descendants that come from the same ancestor are correlated since they are all similar to their ancestor. This basic scheme for generating correlations between patterns has been studied to extend the Hopfield model to the case of correlated attractors (35–37).

In this section, we discuss a simple strategy to efficiently store these patterns in memory. This strategy is an intuitive way to recode correlated dense memories into sparse and approximately uncorrelated representations that contain the same information as the original memories. Sparseness makes these representations more suitable for storage in neural networks, as simple calculations show. In the next section, we will use a more general approach based on sparse autoencoders to generate the compressed representations.

The simple strategy is to store the ancestors in one network and the differences between the ancestors and their descendants in another (Fig. 1*B*). These differences are approximately uncorrelated, and they can be significantly sparser than the original patterns (for γ close to one) (also, ref. 38). Indeed, most of the neurons have the same activity in the ancestors and in its descendants, and hence, the difference is zero. In a sparse representation, only a relatively small fraction f of the neurons is active. Sparse random patterns contain less information than dense patterns (the information per neuron scales approximately as f, which is called the coding level) (e.g., ref. 39). However, the number of random sparse patterns that can be stored in memory is a factor $1/f$ larger than the number of storable dense patterns (i.e., patterns with $f=1/2$) (23–25, 39–42) because of the reduced interference between memories. In this scheme, it is possible to compress the information about the descendants simply by constructing sparse representations that take into account the already acquired information about the ancestors. Even though the amount of information per pattern is smaller for sparser representations, there is no loss of information because storing differences between ancestors and descendants requires fewer bits than storing the full representations of the descendants.

The relative advantage of this scheme as measured by the improvement factor of the total number of retrievable descendant patterns compared with storing uncorrelated patterns is estimated in *SI Appendix, Memorizing Ultrametric Patterns* and summarized in Fig. 1*C*. As k increases, more of the patterns to be stored become correlated, and their overall compressibility increases. The improvement factor also increases because the scheme that we discussed can take advantage of the increased compressibility. From the formula in *SI Appendix, Memorizing Ultrametric Patterns*, it is clear that the improvement increases approximately linearly with k, when γ is fixed close to one. The improvement also increases as γ tends to one (i.e., when the descendants become more similar to their common ancestor and hence, the patterns are more compressible).

Generating Compressed Representations with Sparse Autoencoders. The scheme we just discussed illustrates a simple strategy for taking advantage of the correlations between memories. It is unclear whether this strategy can actually be implemented in a neural network and whether it can be extended to more structured real-world memories. We now show that it is possible to construct a simple network that is able to generate compressed representations for arbitrary memories. We will first analyze the memory performance of this network in the case of ultrametric memories, and then, in the following sections, we will use the same model in a navigation task.

The network illustrated in Fig. 2*A* comprises two layers; the first (input) layer, which could be mapped onto EC, encodes sensory experiences, and the second layer (possibly the dentate gyrus [DG]) encodes their compressed representations. A third layer (perhaps cornu ammonis region 3 [CA3]) would store specific episodes (i.e., individual patterns that represent instantaneous sensory experiences), but in this first part of the article, we will not simulate it, as we will focus on the geometrical properties of the compressed representations. The compressed representations are not constructed by hand, as in Fig. 1, but by using the strategy of sparse autoencoders; the first two layers are complemented by a reconstruction layer, and the synaptic weights are modified to ensure that the input is faithfully reproduced in the reconstruction layer. We used an algorithm similar to the one introduced by Olshausen and Field (30, 31) to reproduce the neural representations of the visual cortex (*SI Appendix*, *SI Text*). Recent extensions of this algorithm apply to several important computational problems (43). The main idea is to modify the synaptic weights from the input to the second layer to build sparse representations of the inputs. The weights are chosen to minimize the reconstruction error (of the inputs) when one reads out these second-layer representations. The representations obtained using this approach are compressed because of the sparseness that is imposed on them by the algorithm. The reconstruction layer

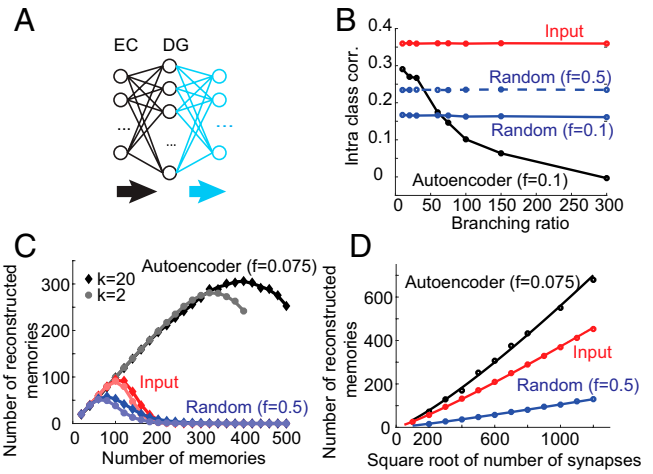

Fig. 2. (*A*) Scheme of the simulated autoencoder. The input layer (300 neurons; mappable to EC) projects to an intermediate layer (DG; 600 neurons). The weights to DG are chosen so that the output light blue neurons reproduce the input. (*B*) Geometry of the compressed representations: correlations between the representations of different descendants of the same ancestor for the inputs (red), the autoencoder (intermediate layer in *A*; black), and a random encoder (blue) as a function of the branching ratio when the total number of patterns is kept constant (and hence, the number of ancestors varies). As γ is fixed ($\gamma=0.6$), the correlations of the inputs and the random encoder are constant ($\gamma^2=0.36$ for the input). For the autoencoder, they decrease with the compressibility of the environment (i.e., when k increases). *SI Appendix*, Fig. S1*A* shows the average of the absolute value of the correlations between all descendants. (*C*) Memory performance of the autoencoder compared with a random encoder and a readout of the input; the number of reconstructed memories is plotted as a function of the total number of memory patterns (changing the number of ancestors). For the autoencoder, we show two curves that correspond to different branching ratios ($k=2, 20$) but the same $\gamma=0.6$ (different values of γ are shown in *SI Appendix*, Fig. S1*B*). As the number of ancestors increases, the quality of reconstruction decreases, and the number of reconstructed memories reaches a maximum. The autoencoder outperforms the input and the random encoder and performs better when the memories are more compressible. (*D*) Memory capacity as a function of the square root of the total number of synapses for autoencoder, random, and input representations. The autoencoder outperforms all the other models, even though it requires four times more synapses than the system that reads out inputs directly.

is used only to determine the weights between the input layer and the layer with the compressed representations. We will show later that the reconstruction layer is not needed when using other algorithmic approaches (e.g., refs. 44–46). Imposing sparseness is only one possible way of compressing information. Here, we assume that the intermediate layer contains more neurons than the input layer, and hence, we need to impose sparseness to limit the amount of information that is represented in the intermediate layer and encourage compression. An alternative possibility is to use dense representations with a smaller number of neurons, as suggested in ref. 2. This strategy leads to different response properties, and it is probably implemented in other parts of the brain (e.g., in the thalamus).

We now analyze the geometry of these representations by looking at the correlations between the compressed patterns. Given the simple ultrametric organization, the geometry of the input representations is completely characterized by some measure of the similarity between patterns that correspond to descendants of the same ancestor and between patterns of descendants of different ancestors. We will compare these correlations with those that characterize the compressed representations. We will also consider the representations obtained in a network in which the neurons in the second layer are randomly connected to the input neurons (random encoder) (e.g., refs. 26, 28, and 47–49). In this encoder, there is no learning except for the choice of the activation threshold, which is tuned to set the desired coding level. The random weights are chosen when the network is initialized, and then, they are frozen. Random encoders are universal encoders as they work for any statistics of the inputs. Moreover, they work surprisingly well considering that they do not require any training. For all these reasons, it is interesting to compare them with the trained sparse autoencoder and assess whether the plastic synapses confer on the autoencoder a significant computational advantage.

In Fig. 2*B*, we plotted the correlations between the descendants of the same ancestor as a function of the branching ratio when the total number of patterns is kept constant (by varying the number of ancestors). For the inputs and the random encoder, they are constant. For the random encoder, we show two curves: one for the same coding level $f = 0.1$ as the autoencoder and one for the coding level that maximizes the memory capacity of the random encoder ($f = 0.5$; see below). For the autoencoder, the correlations decrease with the branching ratio k. This is expected from the abstract scheme that we described above; as k increases, the number of ancestors decreases, the number of correlated inputs increases, and the full set of patterns to be memorized (the "environment") becomes more compressible. More memory resources (i.e., plastic synapses) are devoted to the differences between descendants and ancestors, which are approximately uncorrelated. Indeed, the correlations between the difference patterns are $(1-\gamma)/2$, and when the descendants are sufficiently similar to their ancestor ($\gamma \to 1$), these correlations are small.

In *SI Appendix*, Fig. S1*A*, we show the average correlations between all patterns as a function of the branching ratio. These are the correlations that would be measured in an experiment when all the recorded patterns of activity are considered. We computed the average of the absolute value of the correlations because the average of positive and negative correlations could be close to zero, which might be misinterpreted to mean that the patterns were not correlated. The input representations are more correlated than the representations of the autoencoder and those of the random encoder. As the branching ratio increases, the representations become more correlated for the inputs and the random encoder, but this monotonic relationship does not hold for the trained autoencoder. In fact, for sufficiently large k (more compressible memories), the autoencoder representations become more decorrelated. If the autoencoder layer maps to the DG, this observation would be consistent with the notion that DG is involved in pattern separation (50–53). Interestingly, the model would predict that the representations in DG should become less correlated if the environment is more compressible (see the *Discussion*). Note that the random encoder also decorrelates the representations, as already shown in refs. 26, 28, and 54. However, the decorrelation does not depend on the compressibility of the environment but only on the coding level.

We then estimated the memory performance for the autoencoder, the random encoder, and for reconstruction directly from the input representations, without an intermediate layer. We kept the branching ratio fixed, and we increased the number of ancestors, so that the total number of memories increases. We estimated the number of memories that were correctly reconstructed. A memory was considered reconstructed when the pattern generated in the reconstruction layer (i.e., the light blue layer of Fig. 2*A*) had an overlap of at least 0.9 with the original memory. A noisy cue was imposed on the input layer to trigger memory reconstruction. The noise level of the cue was chosen to be large enough that the average overlap with the original memory was only 0.8. This noise level and the tight criterion for reconstruction guarantee that a memory is considered reconstructed only when the model learned the idiosyncratic features of each descendant, rather than merely the structure of the prototypes. Moreover, the network cannot simply replicate the specific memory cue used to trigger reconstruction because the noisy input is at a distance from the descendant to be retrieved that is larger than what is tolerated by the reconstruction criterion. To evaluate the memory performance without an intermediate layer, we connected the input layer directly to the reconstruction layer. This model, which we will call "input" model in what follows, would essentially correspond to one step of the dynamics of a Hopfield recurrent neural network (21) or to one of the autoassociative models proposed in ref. 6. As the total number of stored memories increases, the number of reconstructed memories also goes up and reaches a maximum, and then, it starts to smoothly decrease. This maximum defines the memory capacity. The autoencoder was trained with sigmoidal activation functions as described above, but we now use a binarized version of the compressed representations to make the comparison of the memory performance with the random encoder and the input as fair as possible. The memory performance is significantly higher when binarization is not imposed.

In Fig. 2*C*, we show the number of reconstructed memories for the autoencoder, the input, and the random encoder representations. The memory performance is significantly larger for the autoencoder, and it is higher for a larger branching ratio ($k = 20$), which corresponds to a more compressible environment. For each curve, we used the optimal coding level. The performance of the random encoder is the worst because of the elevated level of noise (*SI Appendix*, Fig. S11 demonstrates that the autoencoder representations are significantly more robust to noise). In *SI Appendix*, Fig. S1*B*, we show the number of reconstructed memories in the case in which the branching ratio is always the same but the similarity γ between the descendants and their ancestors varies. The best performance is achieved for the autoencoder in the case in which the descendants are most similar to their ancestor, which is again the case in which memories are most compressible.

In all these cases, as the number of stored memories increases, the fraction of correctly reconstructed memories decreases smoothly, unlike in other models where there is an abrupt transition [e.g., the Hopfield model (21, 22)]. Interestingly, the reconstructed patterns are progressively more similar to their ancestor than to the descendant as the number of memories increases (*SI Appendix*, Fig. S2). In the regime studied in Fig. 2*C*, the reconstructed pattern is always more similar to the descendant used to generate the noisy input (*SI Appendix*, Fig. S2*A*). However, if one increases the compression ratio by making the

representations sparser and by decreasing the number of neurons in the compressed layer, it is possible to observe situations in which the reconstructed pattern is more similar to the ancestor (*SI Appendix*, Fig. S2*B*).

Another interesting property of the autoencoder is its ability to learn more rapidly new descendants from a known ancestor (*SI Appendix*, Fig. S3), similarly to what has been observed in models of semantic cognition (55). Although the speed improvement is clear in some situations, the effect is relatively modest for at least two reasons. 1) The network is designed to store every single descendant, and the interference with the previously stored descendants from the same ancestor can slow down the process. A strategy like the one proposed in ref. 56 could alleviate the problem. 2) The information about the ancestors can be extracted rapidly if a sufficient number of descendants is stored. This rapidity reduces the advantage of already knowing the ancestor. This phenomenon is related to the one reported in ref. 57 in linear networks, in which the components with large input variances are learned faster than those with small variances. In the case of the ultrametric memories that we used in our simulations, this implies that the ancestors will be learned more quickly than the differences between descendants and their ancestors. Hence, the process of learning the ancestor is relatively fast, and this limits the advantage of already knowing an ancestor.

Finally, we estimated the memory capacity as a function of the total number of synapses in the system. We showed that the memory capacity is significantly better for the autoencoder than for the systems that read out the inputs directly. However, the autoencoder requires an intermediate layer and hence, a larger number of synapses. In Fig. 2*D*, we computed the memory capacity, as in Fig. 2*C*, for networks of different sizes. We progressively increased the size of the input (N neurons), and we scaled accordingly the number of neurons in the autoencoder intermediate layer ($2N$). The total number of synapses was then N^2 in the case in which we directly read out the input and $4N^2$ for the autoencoder. In Fig. 2, we show the memory capacity as a function of the square root of the number of synapses for the autoencoder, the random, and the input representations. The coding level was chosen to optimize the capacity in the case of $N = 300$. The autoencoder strategy outperforms all the others, and the improvement increases with the size of the network. Hence, the autoencoder representations can be legitimately called "compressed representations" because they allow for the same performance with a smaller number of synapses.

Compressing Sensory Inputs Experienced during Navigation. We now consider the specific case of navigation. In this case, the sensory experiences of an animal that visits the same location multiple times will be different but still correlated (Fig. 3*A*). Similarly to the case of ultrametric memory patterns described in the previous section, it is possible to take advantage of these correlations to compress the information that is stored in memory. We hypothesize that the hippocampus is involved in this process of compression, which leads to sparse, compressed representations of the sensory experiences of the animal during exploration. We now present a simple model to illustrate how the compressed representations can be generated, and then, we show that using these representations leads to a more efficient storage of correlated memories. As in the previous section, we construct the compressed representations using a sparse autoencoder. It is likely that in a more realistic situation in which the animal has to perform a task, the representations are not just shaped by the desire to reconstruct inputs but also affected by other factors.

We assume that the animal is exploring an environment enclosed by four walls, which has the shape of a square. The animal can use sensory cues (visual, tactile, olfactory) to determine its position when it is very close to the walls, and for simplicity, we assume that it explores the environment by walking along a straight line in a random direction until it reaches another wall. It then repeats this procedure by picking another direction and walking again in a straight line. We also hypothesize that the animal performs a simple form of path integration, and hence, it knows approximately the distance from the last wall it visited. Finally, we assume that the animal knows the direction of movement by using distal and other cues. These assumptions would be compatible with the observations that head direction is encoded in EC, one of the major cortical inputs to the hippocampus. Furthermore, we know that the estimate of position decoded from EC has an accuracy that decreases with the distance from the last visited wall (58), indicating that some form of path integration reset happens when the animal gets close to a wall.

We simulated the simple feed-forward network with three layers already described in the Introduction (Fig. 3*B*). In the specific case of navigation, the first layer (EC) represents the input and in our simple example, encodes 1) the direction of movement of the animal (59), 2) the distance from the last wall visited (60), and 3) the position along the last wall visited (61) (i.e., the initial position before the animal initiates its excursion to explore the environment). These variables are mixed nonlinearly through a random projection to obtain the putative EC representations we use as inputs to our network (*SI Appendix*, *SI Text* and Fig. S5).

The second layer (DG) contains the compressed representations of the sensory experiences. These representations are learned as described in the previous section by introducing an

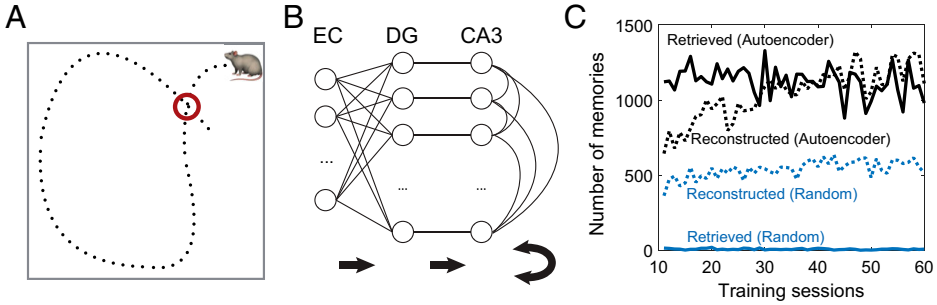

Fig. 3. (*A*) Schematic of a rodent exploring an open field arena. Whenever the animal returns to the same location, its sensory inputs will have some similarity with those experienced during previous visitations of that location. (*B*) Schematic of the architecture of the network with potential mapping of the layers onto EC and hippocampus. (*C*) The memory retrieval capacity (the number of patterns of 7,480 ± 150 stored inputs per session that can be recalled from noisy cues in the autoassociative network) as a function of the number of training sessions (exposures to the environment). This illustrates the computational advantage of using even a simple compression algorithm with one layer of learned weights as implemented in our network (black) compared with a network of the same architecture (and coding levels) but with fixed random feed-forward connections (blue). Note that the memory retrieval capacity is different from the reconstruction memory capacity studied in Fig. 2, which we also plot for comparison (dotted lines) and which is again larger for the autoencoder than for the random network.

artificial reconstruction layer (the light blue layer in Fig. 2*A*; not shown in Fig. 3*B*) and by imposing that its representations reproduce the inputs. Finally, the third layer (CA3) is an autoassociative memory system where memories are stored by modifying recurrent connections. The purpose of this layer in our simulations is to obtain a reasonable estimate of the memory capacity beyond simple feed-forward reconstruction.

The recurrent synaptic weights are modified according to a simple covariance rule, similar to the Hopfield rule (21), to create stable fixed points corresponding to the patterns of activity imposed by the inputs from the second layer (also, refs. 4, 6, and 27). For simplicity, we assumed that the number of neurons in the third layer is equal to the number of neurons in the second one and that the third layer basically just copies the compressed sparse representations prepared in the previous layer. This is a reasonable idealization since CA3 pyramidal cells receive only a small number of strong synapses from the DG (62). While the second-layer neurons are continuous valued (and typically exhibit a bimodal activity distribution in our simulations), we threshold their activity to obtain binary neural representations suitable for storage in an autoassociative network with binary neurons.

We simulated an animal exploring an unfamiliar environment by discretizing time and computing at every time step the inputs for the current position and direction of motion of the animal by a random projection of the variables described above (the representation is constructed by computing the weighted sum of the inputs, with random weights, and then we pass it through a nonlinearity). These input patterns, which are higher dimensional than the original inputs, represent the memories that we intend to store.

We consider a situation in which the animal explores for several hundred (straight-line) trajectories crossing the environment, which is similar to the typical situations studied in experiments on rodent navigation. We compared the performance of the proposed memory system with the performance of an analogous network with the same architecture and coding level in which the input layer is connected to the second layer with fixed random connections. To quantify this memory performance, we estimated the number of memories that can be correctly reconstructed, as in Fig. 2. We also computed the number of patterns that can be retrieved from the recurrent network (presumably CA3) with sufficiently high fidelity, namely with a pattern overlap of at least 0.8, from noisy cues with an initial overlap of at most 0.7 (with the binary pattern that was stored in the autoassociative network). These memory patterns correspond to the sensory inputs experienced by the animal during the preceding exploration of the environment. The results are reported in Fig. 3*C*, which shows that both the reconstruction and the retrieval memory performance of the proposed network are substantially better than those of the network with random connections (*SI Appendix*, *Memory Capacity and Decoding Analyses* has details, and *SI Appendix*, Fig. S10 shows a summary of the pattern statistics). The coding level in the random network is kept equal to the coding level in the autoencoder to make the performance comparison as fair as possible.

Single-Neuron Properties: The Emergence of Place Cells. Now that we have established that the compressed neural representations allow for a better memory performance, it is interesting to inspect the neural representations obtained in the second layer of the network. These are the representations that we expect to observe in the hippocampus. One common way to represent the responses of recorded individual neurons is to plot their place fields. In Fig. 4*B*, we show the place fields (averaged over training epochs) of 36 randomly selected cells of the second layer of the network. Their fields have been measured during the simulated exploration along trajectories sampled from a distribution illustrated in Fig. 4*A*. We observe a number of cells with localized

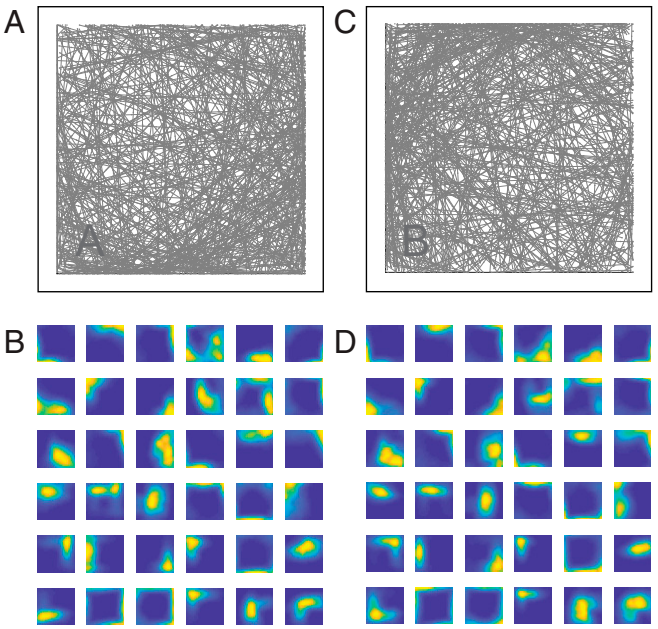

Fig. 4. (*A*) Trajectories of a simulated animal in an open arena (exploration statistics A) and (*B*) the spatial tuning profiles emerging from training the autoencoder network on an artificial sensory input corresponding to these trajectories for 36 neurons randomly selected from the second (DG-like) layer of the model. We find a very heterogeneous set of spatial tuning profiles: some consistent with simple place cells, some exhibiting multiple place fields, and some that look more like boundary cells. The statistics of this diverse set of responses appear to be consistent with calcium imaging data from the dentate (12). (*C* and *D*) Same as *A* and *B* but for a set of trajectories with a slightly different exploration bias (exploration statistics B). Half of the trajectories on both sides have the same statistics and are drawn from an isotropic distribution of initial positions. The other half of the trajectories are drawn from different distributions with initial positions biased toward the lower right corner in *A* and *B* and the upper left corner in *C* and *D*. As a result, the two sets of place fields that correspond to exploration statistics A and B are slightly different.

place fields, which are similar to those of classical place cells. Interestingly, the responses are highly diverse, which is typically what is observed in the hippocampus and in particular, in DG (12, 50). The place fields generated by a random encoder are shown in *SI Appendix*, Fig. S6; the fields are substantially different and more noisy than in the case in which the weights are learned.

The Instability of Place Fields Reflects Their History Dependence. Neurons with spatial tuning properties can be obtained in many ways, and while the fact that we found cells with spatial tuning in our model is reassuring, it was not unexpected given that the inputs contain information about the animal's position. Less obvious is the fact that the spatial tuning properties of these units are consistent with those of place cells exhibiting few well-localized fields, even though the inputs consist of highly mixed representations of the spatial variables. If the inputs were provided by cells which themselves had smoothly localized spatial tuning profiles, random connectivity in combination with sparsification of the neural activity would be sufficient to achieve this (e.g., ref. 63). For highly mixed inputs, obtaining spatial tuning properties resembling those of place cells requires some learning of the weights in addition to a penalty enforcing sparse coding.

Another interesting aspect of the cell responses is their dependence on recent experiences. According to our model, the neural representations are continuously updated during exploration through ongoing synaptic plasticity. This means that the neural fields can be rather unstable. We illustrate this in Fig. 4 *C* and *D*, where we show the place fields of the same 36 cells resulting from

two different exploration statistics of the same environment; in Fig. 4*A*, the simulated animal tended to visit the bottom right corner more often (exploration statistics A), whereas in Fig. 4*C*, the animal prefers the top left corner (exploration statistics B). Many of the neural fields are similar in the two cases, but there are clear differences reflecting the exploration bias. Due to the ongoing plasticity, these differences remain even after many training sessions with both types of exploration statistics (i.e., they do not stem from the initial formation of the place fields during early exposures to the environment, and in fact, we excluded these early sessions when computing the place field maps).

In Fig. 5*A*, we show the changes in the fields and quantify them in Fig. 5*B*. We compute the average normalized overlap between place fields in a simulation during which the animal experiences explorations statistics A and B repeatedly in a random order (while training the autoencoder). The average overlap between fields estimated in different sessions with the same exploration statistics is less than 0.7 and even smaller when sessions with different statistics are considered (comparing A and B sessions). This indicates that the response properties of individual neurons are continuously modified and that they reflect the recent exploration statistics. The relatively small overlap in the case of the same exploration statistics is partly due to the stochasticity of the algorithm used to determine the synaptic weights (which includes sampling a new set of random trajectories for every session) and partly due to the dependence of the fields on the exploration statistics of the previous session (see below). Note that the overlaps of the place fields of different sessions only decay very slowly as a function of the time interval elapsed between them. This would not be the case during the initial phase of rapid learning (not shown) that occurs during the first few sessions.

Despite the continual modifications of the fields, it is still possible to decode the position of the simulated animal (Fig. 5*C*). We trained a linear regression decoder to predict the x and y coordinates of the animal from the second-layer representations. The decoder is trained on the data of one session and tested on the data of a different session, as in ref. 13. The median error is plotted as a function of the interval between these two sessions (expressed in number of sessions). The decoder is more accurate when the exploration statistics are the same, but it is still significantly better than chance (dashed lines) if they are not. Despite the instability of the fields, it is still possible to decode the animal's position. This is similar to what has been observed in ref. 13, although the statistics of the field modifications are probably different in the experiment (in which some cells respond with significant spatial tuning only in a subset of sessions) (also, ref. 64). These differences might be due to the simplicity of our model, the fact that we are considering a two-dimensional (2D) arena rather than a one-dimensional (1D) track, and potentially, the way the activity is recorded in the experiment. However, the model captures the basic observation that it is possible to decode position despite the relative instability of the fields.

Note that the between-session variability we are quantifying here depends on the parameters of the algorithm used to train the model. In our simulations, the length of an individual training session with a given exploration statistic determines the level of stability of the place fields (the learning rate would play a similar role). Furthermore, in a real experiment, the within-session fluctuations of the place cell responses may be larger than the variability due to synaptic plasticity because of noise or additional variables encoded in the neural activity that we have not modeled here.

History Effects and the Ability to Decode the Recent Past. The instability of the fields is compatible with several experimental observations (e.g., refs. 13 and 15). Here, we propose an interpretation of these fluctuations that can be tested in experiments; they reflect the recent history of experiences of the animal, and therefore, any bias in the exploration statistics or any other events that are represented in the input to the hippocampus should affect the neuronal responses. This means that by studying the fluctuations of the neural responses, we should be able to decode

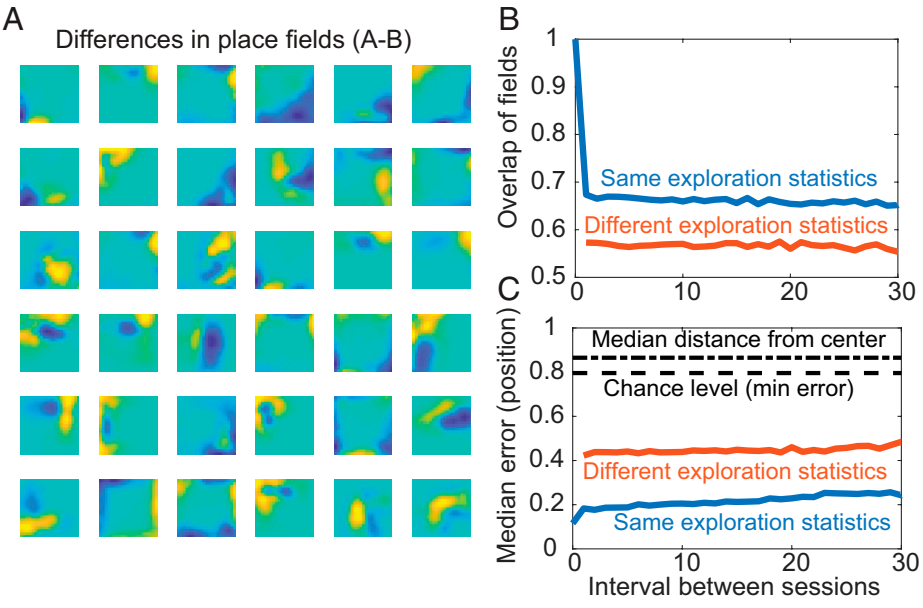

Fig. 5. (*A*) Maps of differences of average place fields between A and B sessions in a simulated experiment in which the animal experiences a random sequence of the two types of sessions with different exploration statistics (as in Fig. 4). (*B*) Normalized overlap between the place fields of two sessions with the same (blue) or with different (red) statistics as a function of the time interval between sessions. The overlap is larger in the former case and stays rather high even for long time intervals between sessions, indicating relative long-term stability despite short-term fluctuations. (*C*) Median decoding error for position from simple regression predictors for the x and y coordinates of the animal. Position can be predicted more accurately if the decoder was trained on the same type of exploration statistics as in the session used for testing, but even for different statistics, this works significantly better than chance level. The decoding error grows only slowly with the interval between training and test sessions.

at least some information about the recent history of the animal's sensory experiences. This is a prediction that can be tested if sufficient numbers of neurons are simultaneously recorded for a long-enough period.

In simulations, it is indeed possible to decode the recent history by reading out the fluctuations of the firing fields, but taking this approach literally, one first has to construct place field maps, which requires knowledge of the position of the animal. Here, we show that one can also decode some information about the previous exposure to the environment that the animal experienced directly from the neural activity patterns elicited in the current session (without requiring additional spatial information, even though much of it is of course contained in the neural representations).

We consider a random sequence of biased exploration sessions like those shown in Fig. 4 *A* and *C*. Discriminating A and B is challenging because the environment and therefore, the sensory input the animal receives given its position and head direction are the same in A and B sessions. At the end of each session, we estimate the place fields of the neurons in the second layer, and as shown in Figs. 4 *B* and *D* and 5*A*, the resulting fields depend on the exploration bias. Interestingly, they also depend on the bias in the previous session; if an A session is preceded by another A, the fields (evaluated during the latter session) are different from the case in which A is preceded by B. We plot these differences between the fields in Fig. 6*A*. Similarly, we show the differences in place fields (in a B session) between the case in which B is preceded by A and the case in which B is preceded by another B in Fig. 6*B*. These differences are relatively small but consistent enough that it is possible to train a decoder to read them out and report successfully whether the current session was preceded by A or B. In Fig. 6*C*, we show that even without first computing place fields, we can train linear classifiers to decode not just whether the current session but also, whether the previous session was of type A or B. While the performance of these classifiers is far from perfect when predictions are made based on a single activity vector, they can achieve very high accuracy when combining the predictions from many neural representations (corresponding to different time points) using a majority vote. These simulations illustrate one possible experiment that can be performed to test some of the central ideas of our theory.

Sparse Compression Using Local Learning Rules without a Reconstruction Layer. While it is possible for the brain to implement an autoencoder and learn its synaptic efficacies using a biologically plausible approximation to backpropagation, such an implementation with DG as the hidden layer may be difficult to reconcile with the known neuroanatomy of the hippocampus. In this sense, the reconstruction layer of our autoencoder model (light blue in Fig. 2*A*) could be considered biologically implausible. The reconstruction layer is useful to learn the encoding weights onto DG but no longer required for the system (in Fig. 3*B*) to perform its memory function after the statistics of the environment relevant for compression have been learned. However, the hippocampus does not actually need to explicitly reconstruct its inputs to learn compressed representations, and we can construct an alternative model that achieves sparse compression without the reconstruction layer of the autoencoder. In this more biologically plausible model, the synaptic weight updates follow entirely local (anti-)Hebbian learning rules, and no backpropagation is required. This model is based on the similarity matching–inspired network of ref. 65, which contains only an input layer and a compressed layer that extracts the leading eigenmodes of the input data and rescales them to put them on an equal footing. This can be understood as a form of compression related to linear dimensionality reduction. We modified this network to make it more similar to the autoencoder used above by introducing a sigmoidal nonlinearity in the layer learning the compressed representations and adding a sparsity-inducing penalty (*SI Appendix*, *Input Compression Using Local Learning Rules in a Network without Reconstruction Layer* has details). As in the autoencoder network, sparse compression in this model creates place field–like spatial response profiles of the units learning the compressed representations (*SI Appendix*, Fig. S8), enhances the

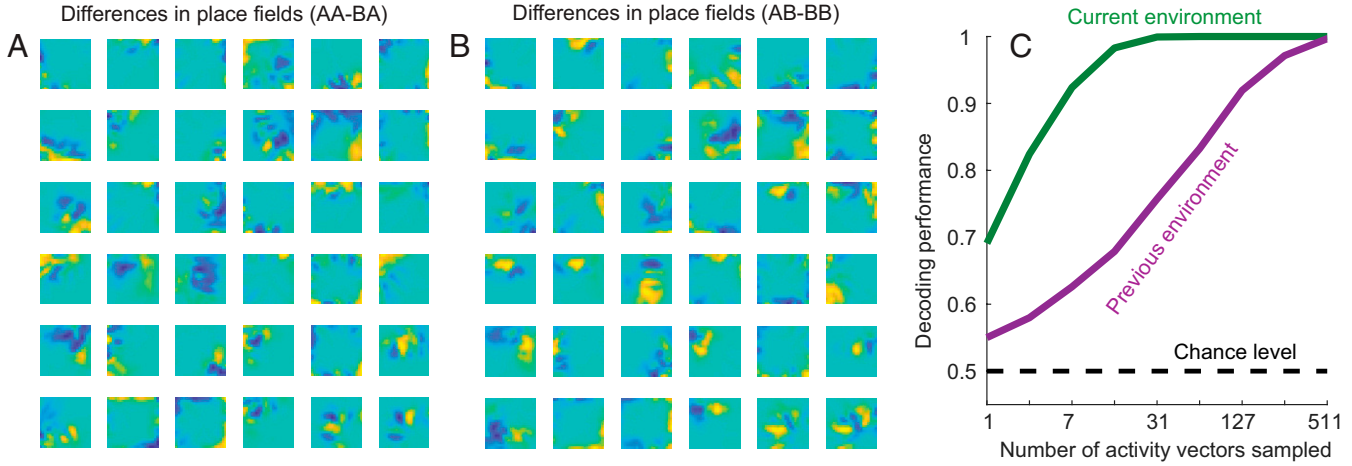

Fig. 6. (*A*) Difference maps of average place fields in A sessions between the cases when the previous session was A vs. B (i.e., sequences AA–BA). (*B*) Similar difference maps for B sessions (corresponding to sequences AB–BB). Note that these differences are more subtle than those between A and B shown in Fig. 5*A*. (*C*) To demonstrate that the fluctuations of the previous two panels are not just noise but reliably capture history-dependent information, we show that one can decode from the neural (DG) representations of the simulated animal exploring an environment not just the statistics of the current session (i.e., A vs. B; green) but also, the statistics of the previous session it experienced (purple). We decode using simple maximum margin linear classifiers in combination with a form of boosting (by combining the predictions made from several neural representations experienced at different points in time) and report the resulting performance as a function of the number of neural representations (snapshots of the second-layer activity in the current session) used for decoding. While the performance is only slightly above chance level when decoding from a single snapshot of the neural activity, a linear classifier can almost perfectly discriminate A and B sessions when combining together the predictions of the trained classifier for many such activity patterns by taking a simple majority vote of the predicted labels. Crucially, the decoder for the statistics of the previous session only uses activity patterns from the current session.

memory retrieval capacity of an autoassociative memory network storing these representations, and allows us to decode from them the position and direction of motion of the simulated animal (*SI Appendix*, Fig. S9).

The absence of a reconstruction layer whose mismatch error could be backpropagated necessitates a mechanism to prevent different units in the compression network from learning redundant representations. This can be achieved with mutual inhibition between the units carrying the compressed representations (65), which is mediated in the model by a population of interneurons, consistent with the neuroanatomy of the DG. In addition to the Hebbian learning of the feed-forward weights between EC and DG, the weights between the principal DG units and these interneurons are also learned. In particular, the weights from the DG units to the interneurons follow a Hebbian rule, and those from the interneurons to the DG units follow an anti-Hebbian rule. This has the important functional consequence of causing units that tend to be excited at the same time to inhibit each other.

Discussion

Preprocessing or recoding correlated patterns can greatly increase memory capacity. This is an old idea that has been proposed and discussed in multiple works (e.g., refs. 2–5 and 26). One efficient way of preprocessing memory representations is to extract the uncorrelated components or more generally, the components that are truly independent from previously stored inputs. This preprocessing would enable the memory system to store only the information that is not already in memory. Any similarity with previous experiences can be exploited to reduce the amount of information for each input that actually needs to be stored (because it is truly novel), which decreases the amount of synaptic resources required. Following refs. 2–5, we proposed that the hippocampus plays an important role in this process of compressing memories, and we presented a simple neural network model that illustrates the main ideas of memory compression. We showed that a memory system that incorporates a sparse autoencoder has a significantly larger memory capacity than one that stores directly unprocessed representations. Not only is the overall memory capacity larger but also the memory capacity per synapse, which is a nontrivial result given that the autoencoder network requires an additional layer of neurons. The representations constructed by the autoencoder are sparse and hence, may contain less information than those of the inputs, which are dense. However, we showed that there are several regimes in which these sparse representations contain all the information that is needed to reconstruct the input or to retrieve a memory. These sparse representations are not only more efficient for memory storage, but they also allow for fast storage of new episodes. The process of generating these efficient representations is of course slower as it has to involve some statistical learning of the features of the environment. Whether the full process of compression based on sparse autoencoders accounts for the ability of humans to learn a new association from just a small number of arbitrary pairings is something that will be investigated in the future in significantly more complex tasks.

Random Recoding Schemes. Some of the original recoding schemes (3, 26, 28) were based on random connections between the neurons representing the inputs and the neurons encoding the compressed representation. These schemes lead to better pattern separation, possibly increasing the memory capacity. Although the trained autoencoder outperforms these random encoders (Fig. 2), it is interesting to consider them because they are universal encoders; they work for any input statistics, and they do not require any training. In *SI Appendix*, Fig. S4, we tested a few random schemes, in particular those proposed in ref. 28, in which the representations are constructed by projecting the inputs onto random weights and then suppressing the activation of all neurons except the k most active (k winner take all). Although these schemes can perform better than the one we extensively analyzed in Fig. 2, their memory performance is still much lower than the one of the sparse autoencoder, indicating that there is a significant advantage in training the encoder.

Finally, it is important to remember that random projections lead to place fields that are qualitatively different from those obtained with a trained autoencoder. While they still exhibit some residual spatial selectivity, the activity maps are not as coherent and much less reminiscent of experimentally observed place fields (*SI Appendix*, Fig. S6).

Memory Compression and Navigation. Our model provides an interpretation of the observations of experiments on navigation conducted on rodents. These experiments show that the recorded neural activity in the hippocampus encodes the position of the animal, suggesting that the hippocampus plays an important role in navigation. However, there is an ongoing debate on whether the hippocampus is actually needed to navigate in familiar environments (e.g., refs. 66–69). Several studies indicate that the hippocampus is important primarily in situations of navigation that require the formation of new memories (70). Similarly to what has been suggested by several memory researchers (2–5) but also by investigators who focused on spatial encoding (9, 70), we propose that the hippocampus is a general memory device used for compressing correlated inputs into efficiently storable episodic memories. It is only due to the nature of the navigation experiments that many investigators were led to put so much emphasis on the role of the rodent hippocampus in encoding the position of the animal. Because the sensory experiences during navigation are highly correlated (for similar locations of the animal), a simple network compressing such inputs can reproduce the response properties of typical hippocampal neurons.

Mapping the Hippocampus to Our Model. The hippocampus is highly structured, and while we believe in the general principle that it implements some form of compression, we are less clear about the mapping between its different parts and the layers of our model. Our input layer maps naturally to EC, the compressed layer could be DG, and the recurrent network for the episodic memories could be CA3. While this mapping is broadly compatible with the known hippocampal anatomy, it is not the only possible one. The reasons why we proposed this mapping are essentially two; the first one is that the realistic model of the encoder does not require a reconstruction layer but only lateral inhibition. This is compatible with the architecture of DG. Moreover, DG is also known to exhibit sparse activity, which is another requirement. The second reason is that the recurrent network that stores episodic memories needs to be downstream of the encoder and requires recurrent connections that are also excitatory. This maps nicely to CA3. It is still unclear why we need CA1 in our model. However, it is important to remember that we are modeling the hippocampus in isolation, and hence, CA1 could be important to mediate the interactions with the cortex and be involved in the process of decoding. Another possible function of CA1 is to compare CA3 output with self-organized feed-forward EC input (4). If the comparison shows a match between CA3 output and EC input, then CA1 inhibits the medial septal acetylcholine input to push the network into retrieval mode; if there is a mismatch, acetylcholine levels remain high, and the network remains in encoding mode. Alternatively, it is possible that the encoder is actually implemented by CA1, which also exhibits sparse activity and has direct bidirectional connections to EC. Area CA3 also projects to CA1 (Schaffer collaterals), so a memory of an EC pattern retrieved from CA3 can reinstate that pattern in EC. In this scheme, proposed in ref. 3, DG represents a parallel encoding pathway, which is based on random connectivity to achieve pattern separation. Other works proposed that the EC–CA1–EC pathway supports

statistical learning, while the EC–DG–CA3–CA1 pathway learns individual episodes (5). Future extensions of our model, which incorporate the interactions with the cortex, might indicate what critical experiments should be done to map the components of the model to the different parts of the hippocampus.

Response Variability and Remapping. In our model, the neural representations reflect not just current inputs but also recent memories since the synaptic weights are continuously updated, which would explain some of the observed high variability of the neuronal responses (e.g., ref. 15). Moreover, our model is in line with the elevated sensitivity of the neuronal responses to any change in the environment [e.g., the delivery of reward (19, 71) or the manipulation of landmarks (72, 73)]. This phenomenon is often described as "remapping" (74) and is widely observed. In ref. 75, the authors suggest that it is important for memory. According to our interpretation, the hippocampus would encode any memory, not just those that are related to navigation in physical space. The presence of an item or the delivery of reward at a particular location, which often constitute salient episodes, would certainly alter the neuronal responses, as observed in experiments (19, 71, 73, 76). We show such a modification of spatial response profiles explicitly in *SI Appendix*, Fig. S7, where we use our model to compress inputs created by nonlinearly mixing the spatial variables used in Figs. 3 and 4 with an additional localized input that represents a reward signal. Moreover, any structured sensory experience involving correlated inputs parameterized by some external variable would also be reflected in the hippocampal representations, as in the case of auditory stimuli (33).

The History Effect and Other Predictions. One of the predictions of our model is that the neuronal responses should be affected by the recent history to the point that the fluctuations of the firing fields should contain decodable information about the recent exploration statistics. However, history effects are not a unique feature of our model for at least two reasons. The first one is that during learning, it is likely that any model would exhibit history effects. However, if the hippocampus was designed to encode only the position of the animal, it appears unlikely that in a stationary environment, the representations would keep changing substantially after the environment is familiar and the position of the animal can be decoded from the neural activity. Our model predicts that even in situations in which position is strongly encoded, we should observe continual modifications that reflect the recent history of sensory experiences. The second reason is that there are many real-world tasks where it is important to encode temporal correlations. As soon as these correlations are encoded, we have a different source of history dependence; if the input to our model already encodes the recent history, then of course the activity of the neurons in the compressed representations will also be history dependent. This dependence is conceptually different from the one we discussed because ours is due to the continual update of the compressed representations. One of the reasons we decided to study only problems with spatial and not temporal correlations is to highlight the history dependence that is characteristic of our model: the one due to the ever-changing synaptic weights of the encoder. This history dependence is likely to happen on a different timescale, comparable with the timescale of synaptic modifications (minutes, not seconds). However, this clearly poses a problem in terms of predictions. If we observe a clear history effect on multiple timescales, as the authors of ref. 17, is it due to the explicit encoding of temporal correlations as in ref. 77 or due to the continual update of the representations, as we propose? In both cases, the changes in the response properties of the neurons would not be a random drift, but they would depend on the specific history of recent events. If the timescales that are encoded in the input and those that characterize the learning process are not well separated, we cannot really discriminate between these two scenarios, and hence, we cannot consider the history effect a unique signature of our compression model.

To confirm the role of the hippocampus in memory compression suggested by the model, we would need additional experiments in which the sensory experiences are structured and for example, organized as in the ultrametric case. Comparing the neural representations before and after learning should reveal whether the changes are compatible with a compression process. For example, one could look at the geometry of the representations, as in Fig. 2, and its dependence on the compressibility of the environment. A possible experiment could be to train a rodent to run in a virtual 1D environment and control the statistics of the encountered landmarks to change compressibility. For example, if one walks on a desert road, the sensory experiences are very similar to each other, and the environment is highly compressible. A pedestrian area in a city center would be significantly less compressible.

Compressing Temporal Sequences. There are other recent theoretical works that emphasize the general role of the hippocampus in learning and memory (e.g., refs. 78–85) at the expense of the specific role it plays in navigation. These works mostly focus on the encoding of temporal sequences with one notable exception (85); some assume that the hippocampus tries to store only the information that is relevant for predicting the next state of the environment, while others postulate that the goal of the hippocampus is to represent the probability distribution of future locations (conditioned on the current position of the animal). These predictions are then used to drive reinforcement learning. In our case, we considered for simplicity only the compression of the instantaneous sensory experiences. We focused on the lower-level questions of how spatially modulated cells may arise mechanistically and which of their features may be explained without postulating any higher cognitive goals other than simply remembering the past. Our model can easily be extended to deal with simple temporal correlations: for example, by replacing the input layer with a recurrent network. Even in the case of random connectivity, the network activity would then contain information about the temporal sequence of recent sensory experiences (86). For longer timescales and more complex temporal correlations, a different mechanism would be required (e.g., refs. 77 and 87), which might also involve synaptic plasticity on multiple timescales (24, 88). In all these cases, the instantaneous input of the autoencoder contains information about a recent temporal sequence. The downstream autoencoder can then be modeled exactly in the same way as we modeled it here.

How to Get Place Cells with Autoencoders. The place cells of Fig. 4 are obtained under the assumption that the inputs contain implicit information about the position of the animal: the direction of movement, the distance from the last wall visited, and the position of the animal when it was at the wall. These are not the only possible inputs to get place cells. Indeed, the same model would produce rather realistic place fields for a broad class of different inputs that contain implicitly the information about position. The only requirement is that nearby locations should lead to similar inputs. In other words, there should be a topology in the environment that is preserved in the input space. If the environment is 2D, it is not necessary that the inputs that correspond to all possible locations also lie on a 2D linear subspace of the input space. In fact, in our case, the inputs that are fed to the autoencoder come from a relatively low-dimensional latent space, but they are actually high-dimensional (we pass the inputs through a nonlinear random projection). All we require is that the intrinsic dimensionality of the manifold is 2D as the original environment (or three-dimensional if we take into account heading direction in addition to position). If temporal correlations are considered, as described in the previous

section, it is likely that one can get place fields even when the representations of different locations are completely random, as in refs. 83 and 84.

The Missing Loop with the Cortex. In our work, we focused on the role of the hippocampus in compressing correlated sensory experiences represented in the input. However, it will be important in the future to also consider the correlations between sensory inputs and long-term memory, which probably resides in the cortex (e.g., refs. 54 and 89). The hippocampus is anatomically part of a loop that involves the cortex. This loop complicates the validation of our model because it is difficult to isolate the components of the input in EC that are present from the very beginning and those that result from the projections that come back from the hippocampus. The loop is important for at least two reasons. The first one is that the hippocampus can contribute to organizing long-term memories stored in the cortex, and it might be important for the process of abstraction that underlies the creation of schemas (90). The second one is that the hippocampus should also take into account the similarities between the current episode and all the memories already stored in the cortex. Even more importantly, the hippocampus should be able to use the abstract information that might be stored in the long-term memory: for example, the information stored in a schema (90). We focused only on the learning dynamics of the hippocampus in the early stages, when there is still no information about the new memories in the cortex. A recent model (84) addressed these issues by introducing a mechanism that maps specific episodes to low-dimensional structures that were previously learned and then encoded in the cortex. The model also assumes that the hippocampus plays an important role in the formation of these cortical representations. Such a model leads to the formation of place cells in the hippocampus and grid cells in EC. Our model focused more on the role of the hippocampus in episodic memory, and it considers for simplicity only spatial correlations. In ref. 84, the authors instead assume that the instantaneous sensory experiences are completely random and uncorrelated, and the only correlations that are considered are temporal. Incorporating into our model a loop with the cortex in a way that is similar to what has been proposed in ref. 84 would probably lead to even more efficient ways of storing episodic memories and will be considered in future studies.

Materials and Methods

Methods Summary. The detailed description of the simulations with ultrametric patterns is reported in *SI Appendix*, *Memorizing Ultrametric Patterns*. Fig. 2 results are obtained using the simulations described in *SI Appendix, Simulations of the Compression of Ultrametric Patterns*. Details of the simulated mouse that explores an environment are in *SI Appendix, Simulations of Input Compression in Navigational Experiments* for the autoencoder model and in *SI Appendix, Input Compression Using Local Learning Rules in a Network without Reconstruction Layer* for the biologically plausible model. *SI Appendix, Memory Capacity and Decoding Analyses* describes methods for Figs. 3*C* and 6*C*.

Data Availability. There are no data underlying this work.

ACKNOWLEDGMENTS. We thank A. Losonczy, J. Priestley, and L. Posani for useful discussions and D. Aronov, R. Gulli, A. Losonczy, E. Mackevicius, and J. Priestley for comments on the article. This work was supported by Defense Advanced Research Projects Agency Lifelong Learning Machine (L2M), NSF NeuroNex Grant DBI-1707398, the Gatsby Charitable Foundation, the Simons Foundation, the Kavli Foundation (Kavli Institute for Brain and Mind), and the Swartz Foundation.

1. J. O'Keefe, J. Dostrovsky, The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Res.* **34**, 171–175 (1971).
2. M. A. Gluck, C. E. Myers, Hippocampal mediation of stimulus representation: A computational theory. *Hippocampus* **3**, 491–516 (1993).
3. J. L. McClelland, N. H. Goddard, Considerations arising from a complementary learning systems perspective on hippocampus and neocortex. *Hippocampus* **6**, 654–665 (1996).
4. M. E. Hasselmo, B. P. Wyble, Free recall and recognition in a network model of the hippocampus: Simulating effects of scopolamine on human memory function. *Behav. Brain Res.* **89**, 1–34 (1997).
5. A. C. Schapiro, N. B. Turk-Browne, M. M. Botvinick, K. A. Norman, Complementary learning systems within the hippocampus: A neural network modelling approach to reconciling episodic memory with statistical learning. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **372**, 20160049 (2017).
6. B. L. McNaughton, R. G. Morris, Hippocampal synaptic enhancement and information storage within a distributed memory system. *Trends Neurosci.* **10**, 408–415 (1987).
7. D. Schiller *et al.*, Memory and space: Towards an understanding of the cognitive map. *J. Neurosci.* **35**, 13904–13911 (2015).
8. J. Lisman *et al.*, Viewpoints: How the hippocampus contributes to memory, navigation and cognition. *Nat. Neurosci.* **20**, 1434–1447 (2017).
9. H. Eichenbaum, Barlow versus Hebb: When is it time to abandon the notion of feature detectors and adopt the cell assembly as the unit of cognition? *Neurosci. Lett.* **680**, 88–93 (2017).
10. S. Fusi, E. K. Miller, M. Rigotti, Why neurons mix: High dimensionality for higher cognition. *Curr. Opin. Neurobiol.* **37**, 66–74 (2016).
11. K. Hardcastle, N. Maheswaranathan, S. Ganguli, L. M. Giocomo, A multiplexed, heterogeneous, and adaptive code for navigation in medial entorhinal cortex. *Neuron* **94**, 375–387.e7 (2017).
12. F. Stefanini *et al.*, A distributed neural code in the dentate gyrus and in ca1. *Neuron* **107**, 703–716.e4 (2020).
13. Y. Ziv *et al.*, Long-term dynamics of CA1 hippocampal place codes. *Nat. Neurosci.* **16**, 264–266 (2013).
14. B. J. Kraus, R. J. Robinson II, J. A. White, H. Eichenbaum, M. E. Hasselmo, Hippocampal "time cells": Time versus path integration. *Neuron* **78**, 1090–1101 (2013).
15. A. A. Fenton, R. U. Muller, Place cell discharge is extremely variable during individual passes of the rat through the firing field. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 3182–3187 (1998).
16. N. R. Kinsky *et al.*, Trajectory-modulated hippocampal neurons persist throughout memory-guided navigation. *Nat. Commun.* **11**, 2443 (2020).
17. Y. Liu *et al.*, Consistent population activity on the scale of minutes in the mouse hippocampus. bioRxiv [Preprint] (2021). https://doi.org/10.1101/2021.02.07.430172 (Accessed 1 May 2021).
18. P. E. Jercog *et al.*, Heading direction with respect to a reference point modulates place-cell activity. *Nat. Commun.* **10**, 2333 (2019).
19. N. B. Danielson *et al.*, Sublayer-specific coding dynamics during spatial navigation and learning in hippocampal area ca1. *Neuron* **91**, 652–665 (2016).
20. G. W. Diehl, O. J. Hon, S. Leutgeb, J. K. Leutgeb, Grid and nongrid cells in medial EC represent spatial location and environmental features with complementary coding schemes. *Neuron* **94**, 83–92.e6 (2017).
21. J. J. Hopfield, Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. U.S.A.* **79**, 2554–2558 (1982).
22. D. Amit, *Modeling Brain Function* (Cambridge University Press, 1989).
23. D. J. Amit, S. Fusi, Learning in neural networks with material synapses. *Neural Comput.* **6**, 957–982 (1994).
24. M. K. Benna, S. Fusi, Computational principles of synaptic memory consolidation. *Nat. Neurosci.* **19**, 1697–1706 (2016).
25. S. Fusi, Memory capacity of neural network models. arXiv [Preprint] (2021). https://arxiv.org/abs/2108.07839 (Accessed 17 August 2021).
26. D. Marr, Simple memory: A theory for archicortex. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **262**, 23–81 (1971).
27. A. Treves, E. T. Rolls, Computational analysis of the role of the hippocampus in memory. *Hippocampus* **4**, 374–391 (1994).
28. R. C. O'Reilly, J. L. McClelland, Hippocampal conjunctive encoding, storage, and recall: Avoiding a trade-off. *Hippocampus* **4**, 661–682 (1994).
29. Y. Lian, A. N. Burkitt, Learning an efficient place cell map from entorhinal inputs using non-negative sparse coding. *eNeuro* **8**, ENEURO.0557-20.2021 (2021).
30. B. A. Olshausen, D. J. Field, Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **381**, 607–609 (1996).
31. B. A. Olshausen, D. J. Field, Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Res.* **37**, 3311–3325 (1997).
32. R. S. Zemel, G. E. Hinton, Learning population codes by minimizing description length. *Neural Comput.* **7**, 549–564 (1995).
33. D. Aronov, R. Nevers, D. W. Tank, Mapping of a non-spatial dimension by the hippocampal-entorhinal circuit. *Nature* **543**, 719–722 (2017).
34. R. Rammal, G. Toulouse, M. A. Virasoro, Ultrametricity for physicists. *Rev. Mod. Phys.* **58**, 765–788 (1986).
35. M. Feigelman, L. Ioffe, *The Augmented Models of Associative Memory Asymmetric Interaction and Hierarchy of Patterns in 30 Years of the Landau Institute Selected Papers* (World Scientific, 1996), pp. 270–287.
36. H. Gutfreund, Neural networks with hierarchically correlated patterns. *Phys. Rev. A Gen. Phys.* **37**, 570–577 (1988).
37. S. Fusi, *Prototype Extraction in Material Attractor Neural Networks with Stochastic Dynamic Learning in Applications and Science of Artificial Neural Networks* (International Society for Optics and Photonics, 1995), **vol. 2492**, pp. 1027–1039.
38. L. Fontolan, "*Learning hierarchical memories with binary synapses*," Master's thesis, University la Sapienza, Rome, Italy (2010).
39. N. Brunel, J. P. Nadal, G. Toulouse, Information capacity of a perceptron. *J. Phys. A Math. Gen.* **25**, 5017 (1992).
40. M. V. Tsodyks, M. V. Feigel'man, The enhanced storage capacity in neural networks with low activity level. *Europhys. Lett.* **6**, 101–105 (1988).

41. E. T. Rolls, A. Treves, The relative advantages of sparse versus distributed encoding for associative neuronal networks in the brain. *Network* **1**, 407–421 (1990).
42. A. Treves, E. T. Rolls, What determines the capacity of autoassociative memories in the brain? *Network* **2**, 371–397 (1991).
43. Y. Chen, D. Paiton, B. Olshausen, The sparse manifold transform. *Adv. Neural Inf. Process. Syst.* **31**, 10532–10543 (2018).
44. C. J. Rozell, D. H. Johnson, R. G. Baraniuk, B. A. Olshausen, Sparse coding via thresholding and local competition in neural circuits. *Neural Comput.* **20**, 2526–2563 (2008).
45. C. Pehlevan, A. M. Sengupta, D. B. Chklovskii, Why do similarity matching objectives lead to Hebbian/anti-Hebbian networks? *Neural Comput.* **30**, 84–124 (2018).
46. A. Sengupta *et al.*, Manifold-tiling localized receptive fields are optimal in similarity-preserving neural networks. *Adv. Neural Inf. Process. Syst.* **31**, 7080–7090 (2018).
47. O. Barak, M. Rigotti, S. Fusi, The sparseness of mixed selectivity neurons controls the generalization-discrimination trade-off. *J. Neurosci.* **33**, 3844–3856 (2013).
48. B. Babadi, H. Sompolinsky, Sparseness and expansion in sensory representations. *Neuron* **83**, 1213–1226 (2014).
49. A. Litwin-Kumar, K. D. Harris, R. Axel, H. Sompolinsky, L. F. Abbott, Optimal degrees of synaptic connectivity. *Neuron* **93**, 1153–1164.e7 (2017).
50. J. K. Leutgeb, S. Leutgeb, M. B. Moser, E. I. Moser, Pattern separation in the dentate gyrus and CA3 of the hippocampus. *Science* **315**, 961–966 (2007).
51. A. Bakker, C. B. Kirwan, M. Miller, C. E. Stark, Pattern separation in the human hippocampal CA3 and dentate gyrus. *Science* **319**, 1640–1642 (2008).
52. M. A. Yassa, C. E. Stark, Pattern separation in the hippocampus. *Trends Neurosci.* **34**, 515–525 (2011).
53. N. A. Cayco-Gajic, R. A. Silver, Re-evaluating circuit mechanisms underlying pattern separation. *Neuron* **101**, 584–602 (2019).
54. J. L. McClelland, B. L. McNaughton, R. C. O'Reilly, Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychol. Rev.* **102**, 419–457 (1995).
55. J. L. McClelland, T. T. Rogers, The parallel distributed processing approach to semantic cognition. *Nat. Rev. Neurosci.* **4**, 310–322 (2003).
56. K. McRae, P. A. Hetherington, "Catastrophic interference is eliminated in pretrained networks" in *Proceedings of the 15h Annual Conference of the Cognitive Science Society* (Cognitive Science Society, 1993), pp. 723–728.
57. A. M. Saxe, J. L. McClelland, S. Ganguli, Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. arXiv [Preprint] (2014). https://arxiv.org/abs/1312.6120 (Accessed 1 May 2021).
58. K. Hardcastle, S. Ganguli, L. M. Giocomo, Environmental boundaries as an error correction mechanism for grid cells. *Neuron* **86**, 827–839 (2015).
59. F. Sargolini *et al.*, Conjunctive representation of position, direction, and velocity in entorhinal cortex. *Science* **312**, 758–762 (2006).
60. C. Lever, S. Burton, A. Jeewajee, J. O'Keefe, N. Burgess, Boundary vector cells in the subiculum of the hippocampal formation. *J. Neurosci.* **29**, 9771–9777 (2009).
61. T. Solstad, C. N. Boccara, E. Kropff, M. B. Moser, E. I. Moser, Representation of geometric borders in the entorhinal cortex. *Science* **322**, 1865–1868 (2008).
62. D. A. Henze, L. Wittner, G. Buzsáki, Single granule cells reliably discharge targets in the hippocampal CA3 network in vivo. *Nat. Neurosci.* **5**, 790–795 (2002).
63. J. D. Monaco, L. F. Abbott, Modular realignment of entorhinal grid cell activity as a basis for hippocampal remapping. *J. Neurosci.* **31**, 9414–9425 (2011).
64. J. S. Lee, J. Briguglio, S. Romani, A. K. Lee, The statistical structure of the hippocampal code for space as a function of time, context, and value. *Cell* **183**, 620–635.e22 (2019).
65. C. Pehlevan, D. B. Chklovskii, "A normative theory of adaptive dimensionality reduction in neural networks" in *Proceedings of the 28th International Conference on Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, R. Garnett, Eds. (MIT Press, Cambridge, MA, 2015), vol. 2, pp. 2269–2277.
66. A. D. Redish, *Beyond the Cognitive Map: From Place Cells to Episodic Memory* (MIT Press, 1999).
67. R. S. Rosenbaum, G. Winocur, C. L. Grady, M. Ziegler, M. Moscovitch, Memory for familiar environments learned in the remote past: fMRI studies of healthy people and an amnesic person with extensive bilateral hippocampal lesions. *Hippocampus* **17**, 1241–1251 (2007).
68. R. E. Clark, *Current Topics Regarding the Function of the Medial Temporal Lobe Memory System* (Springer, 2018).
69. Z. J. Urgolites, S. Kim, R. O. Hopkins, L. R. Squire, Map reading, navigating from maps, and the medial temporal lobe. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 14289–14293 (2016).
70. H. Eichenbaum, The role of the hippocampus in navigation is memory. *J. Neurophysiol.* **117**, 1785–1796 (2017).
71. S. A. Hollup, S. Molden, J. G. Donnett, M. B. Moser, E. I. Moser, Accumulation of hippocampal place fields at the goal location in an annular watermaze task. *J. Neurosci.* **21**, 1635–1644 (2001).
72. K. M. Scaplen, A. A. Gulati, V. L. Heimer-McGinn, R. D. Burwell, Objects and landmarks: Hippocampal place cells respond differently to manipulations of visual cues depending on size, perspective, and experience. *Hippocampus* **24**, 1287–1299 (2014).
73. S. McKenzie *et al.*, Hippocampal representation of related and opposing memories develop within distinct, hierarchically organized neural schemas. *Neuron* **83**, 202–215 (2014).
74. R. U. Muller, J. L. Kubie, The effects of changes in the environment on the spatial firing of hippocampal complex-spike cells. *J. Neurosci.* **7**, 1951–1968 (1987).
75. L. L. Colgin, E. I. Moser, M. B. Moser, Understanding memory through hippocampal remapping. *Trends Neurosci.* **31**, 469–477 (2008).
76. A. M. Wikenheiser, A. D. Redish, Changes in reward contingency modulate the trial-to-trial variability of hippocampal place cells. *J. Neurophysiol.* **106**, 589–598 (2011).
77. Y. Liu, Z. Tiganj, M. E. Hasselmo, M. W. Howard, A neural microcircuit model for a scalable scale-invariant representation of time. *Hippocampus* **29**, 260–274 (2019).
78. P. Dayan, Improving generalization for temporal difference learning: The successor representation. *Neural Comput.* **5**, 613–624 (1993).
79. K. L. Stachenfeld, M. Botvinick, S. J. Gershman, "Design principles of the hippocampal cognitive map" in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, K. Q. Weinberger, Eds. (MIT Press, 2014), pp. 2528–2536.
80. K. L. Stachenfeld, M. M. Botvinick, S. J. Gershman, The hippocampus as a predictive map. *Nat. Neurosci.* **20**, 1643–1653 (2017).
81. S. J. Gershman, C. D. Moore, M. T. Todd, K. A. Norman, P. B. Sederberg, The successor representation and temporal context. *Neural Comput.* **24**, 1553–1568 (2012).
82. K. Hardcastle, S. Ganguli, L. M. Giocomo, Cell types for our sense of location: Where we are and where we are going. *Nat. Neurosci.* **20**, 1474–1482 (2017).
83. S. Recanatesi *et al.*, Predictive learning as a network mechanism for extracting low-dimensional latent space representations. *Nat. Commun.* **12**, 1417 (2021).
84. J. C. R. Whittington *et al.*, The Tolman-Eichenbaum machine: Unifying space and relational memory through generalisation in the hippocampal formation. *Cell* **183**, 1249–1263.e23 (2020).
85. M. Harsh, J. Tubiana, S. Cocco, R. Monasson, 'Place-cell' emergence and learning of invariant data with restricted Boltzmann machines. *J. Phys. A* **53**, 174002 (2020).
86. D. V. Buonomano, W. Maass, State-dependent computations: Spatiotemporal processing in cortical networks. *Nat. Rev. Neurosci.* **10**, 113–125 (2009).
87. I. Momennejad, M. W. Howard, Predicting the future with multi-scale successor representations. bioRxiv [Preprint] (2018). https://doi.org/10.1101/449470 (Accessed 1 May 2021).
88. S. Fusi, P. J. Drew, L. F. Abbott, Cascade models of synaptically stored memories. *Neuron* **45**, 599–611 (2005).
89. R. C. O'Reilly, K. A. Norman, Hippocampal and neocortical contributions to memory: Advances in the complementary learning systems framework. *Trends Cogn. Sci.* **6**, 505–510 (2002).
90. D. Tse *et al.*, Schemas and memory consolidation. *Science* **316**, 76–82 (2007).