Constructing Memory: Consolidation as Teacher-Student Training of a Generative Model

Eleanor Spens
Institute of Cognitive Neuroscience
University College London
eleanor.spens.20@ucl.ac.uk

Neil Burgess
Institute of Cognitive Neuroscience
University College London
n.burgess@ucl.ac.uk

Abstract

Human episodic memories are (re)constructed, share neural substrates with imagination, and show systematic biases that increase as they are consolidated into semantic memory. Here we suggest that these main features of human memory are characteristic of 'teacher-student' training of a neocortical generative model by a one-shot memory system in the hippocampal formation (HF). As we simulate with image datasets, the 'students' (variational autoencoders) in association cortex develop a compressed 'latent variable' representation of experience by learning to reconstruct replayed samples from the 'teacher' (a modern Hopfield network). Recall and imagination require these representations to be decoded into sensory experience by the HF and return projections to sensory cortex, whereas semantic memory and inference rely directly on the latent variables without requiring the HF. We extend preceding models to explain the construction of episodic memory, and the interaction between semantic memory and episode generation more broadly.

1 Introduction

Episodic memory is thought to be constructive; recall is the (re)construction of a past experience, not the retrieval of a copy [1, 2]. Episodic and semantic memory are thought to complement each other, with the former rapidly capturing cross-modal experience via long term potentiation in the hippocampus, enabling the latter to learn statistical regularities over multiple experiences in the cortex [3, 4, 5, 6].

The standard model of systems consolidation is a simple transfer of information from the hippocampus to neocortex [7], whereas other views suggest that episodic and semantic information from the same events can exist in parallel [8]. However, consolidation does not just change which brain regions support memory traces; it also converts them into a more abstract representation, a process sometimes referred to as semanticisation [9, 10]. Whilst several models suggest how episodic memory shapes semantic memory [5], few account for the latter's influence on the former.

Here, we propose that consolidated memory takes the form of a generative model, trained to capture the statistical structure of stored events by learning to reproduce them. This builds on existing models of spatial cognition in which recall and imagination of scenes involve the same neural circuits [11, 12, 13]. The link between scene generation and recall follows from the reconstructive nature of memory; it is supported by evidence from neuropsychology that hippocampal damage leads to deficits in imagination [14], dreaming [15], and daydreaming [16], and by evidence from neuroimaging that recall and imagination involve similar neural processes [17, 18].

We model consolidation as the training of a generative neural network by an initial auto-associative encoding of memory, through the mechanism of teacher-student learning [19]. Recall after consolidation has occurred is a generative process mediated by schemas, as are other forms of scene construction.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

This adds to research on the relationship between generative models and consolidation [20, 21], and on the use of variational autoencoders to model the hippocampal formation (HF) [22, 23, 24]. Table 1 shows some of the key features required in a model of memory processing, and existing models of each; we suggest our model combines these features.

Figure 1: Key features required in a model of memory processing

No.	Feature	Evidence	Existing models
1	'One-shot' encoding followed by gradual consolidation	[3, 4, 7]	CLS [5] and most subsequent models
2	Semantic content of memory becomes HF-independent	Lesions to HF preserve remote memories in semantic form [25, 26]	CLS [5] and most subsequent models
3	But episodic recall stays HF-dependent	Vivid, detailed scene (re)construction requires HF [8]	Multiple trace theory [8]
4	Common mechanism for episode generation	Similar neural circuits involved in recall and imagination [11, 17, 18].	Bicanski and Burgess [13]

2 The model

Our model simulates how the initial representation of memories can be used to train a generative model. First the hippocampus rapidly encodes an event (modelled as one-shot memorisation in a modern Hopfield network). Then a generative network (modelled by a variational autoencoder) takes over, having been trained on replayed representations from the initial hippocampal network. This makes the memory more abstracted, more supportive of generalisation and relational inference, and also more prone to distortion. The generative network learns to encode experiences as latent variables (or 'schemas'), which can be used to reconstruct (for memory) or construct (for imagination) sensory experience via the HF and its return projections to neocortex, or to support semantic memory and inference. Semantic memory and inference depend directly on the latent variable representations and are therefore possible after lesions to the hippocampal formation. (See Figure 2.)

Before consolidation, a modern Hopfield network encodes the memory. Two properties of this network are particularly important for our simulation: firstly, memorisation can occur with only one exposure, and secondly, random inputs to the network retrieve attractor states, allowing sampling from the whole set of stored patterns. We consider the biological implementation of the modern Hopfield network as feature units and memory units suggested by Krotov and Hopfield [27]. For simplicity, the simulations here assume the number of feature units is equal to the dimensionality of the input data, however in reality they are likely to encode a compressed representation of the input (but one high-dimensional enough to capture the details of a memory).

After consolidation, a generative network also encodes the memory. Specifically, we use variational autoencoders (VAEs), so that the most compressed layer of the autoencoder represents a set of latent variables which can be sampled from [28]. In general, reliance on the generative model increases over time as it learns to remember a particular episode, and the hippocampal encoding decays. However, this is likely to be a gradual transition, and many memories involve both the initial hippocampal encoding and generative model in their recall. During perception, the generative model provides an ongoing estimate of novelty from its reconstruction error. Memories that are very consistent with previous experience do not need to be encoded in detail in the initial 'teacher' network (see [29, 30, 31, 32]).

The generative model trained through consolidation also supports imagination, in fitting with the evidence that episodic memory and imagination share neural substrates. Items can either be generated from scratch, or by transforming existing items. The former can be achieved by randomly sampling from the latent space. The latter can be achieved by interpolating between the latent representations of items, or by doing vector arithmetic in the latent space (i.e. inference). In addition, the system learns to support semantic memory, which relies on projections from the latent variable representation. The

latent variables encode the key facts about an episode, and semantic memory draws on these directly, rather than reconstructing them back into a sensory experience. We suggest that this corresponds to the persistence of semantic forms of autobiographical memory even when the hippocampus is damaged.

Teacher-student learning [19] has clear relevance to the transfer of memories from one neural network to another during consolidation [33]. Based on this concept, we use outputs from the initial network to train the generative network in our simulations. To do so, we give random noise as an input to the modern Hopfield network, then use just the outputs of the network to train the VAE. (These outputs represent the high-level sensory representations activated by hippocampal pattern completion, via return projections to sensory cortex.) The noise input to the initial network could potentially represent random activation during sleep [34, 35, 36]. In other words, random inputs to the hippocampus result in the reactivation of memories, and this reactivation results in consolidation.

Which brain regions do the components of this model represent? The initial auto-associative network involves the hippocampus binding together the constituents of a memory in the neocortex. The generative network involves these areas, and additionally the association cortex; in particular, the medial entorhinal cortex (mEC) and medial prefrontal cortex (mPFC) are both prime candidates for encoding a latent variable representation of experience. Firstly, the EC is the main route between the hippocampus and neocortex, and is also where grid cells are most often observed [37]. The mEC has been linked to structural inference, and prior models suggest it encodes latent structures underlying spatial and non-spatial tasks [22, 38]. Secondly, the mPFC is highly connected to the HF and plays a crucial role in episodic memory processing [39, 40, 41, 42, 43, 44]. It is thought to encode schemas [45], and has been implicated in transitive inference [46] and the integration of memories [47]. Furthermore, Mack et al. [48] suggest the mPFC performs dimensionality reduction on incoming data, compressing representations to remove irrelevant features as learning progresses. Therefore we suggest that these regions are involved in the extraction of underlying structure from experience in the generative model.

3 Results

In the following simulations, images represent events. Results are shown across four separate image datasets: MNIST [49], Fashion-MNIST [50], Shapes3D [51], and Symmetric Solids [52]. Each new event is initially encoded as an auto-associative trace in the hippocampus, which we simulate with a modern Hopfield network. Importantly, encoding is one-shot. We model recall as (re)constructing a memory from a partial input. In the case of image data, this can be conveniently modelled by presenting a network with a partial version of the image.

Firstly, we simulate recall in the initial network. The network memorises a set of images, representing events, as described above. When the network is given a partial input, it retrieves the closest attractor state, i.e. the closest memory. Activity propagates through the network, and the values of the nodes are adjusted to minimise the energy function. Even when the network is given random noise, it retrieves memories (see Figure 3).

Secondly, we simulate recall in the generative network. As shown in Figure 4, a generative network was trained on the reactivated memories from the initial network. (One variational autoencoder was trained per dataset from the corresponding modern Hopfield network's outputs.) When presented with a partial version of an item from the training data, the model is able to reconstruct the original. (See also Figure 5 for plots of reconstruction error over time.) Furthermore, it can 'imagine' new items, e.g. by interpolation as shown in Figure 6.

Finally, to illustrate the idea that the latent variables capture the key 'facts' of a scene, we trained models to predict attributes of images from their latent vectors across the four datasets. Figure 5 shows that semantic 'decoding accuracy' increases as training progresses, thanks to structure developing in the latent space. (Decoding accuracy was measured by training a new support vector classifier on 200 examples at the end of each epoch, and measuring classification accuracy on a held-out test set.) In addition, as shown in Figure 6, when there is only a very small amount of training data, predicting the attributes of an image from its latent vector performs better than predicting the attributes from the original images from scratch. The parts of the model representing the HF are not required for this task, in agreement with the neuropsychology data [25, 26].

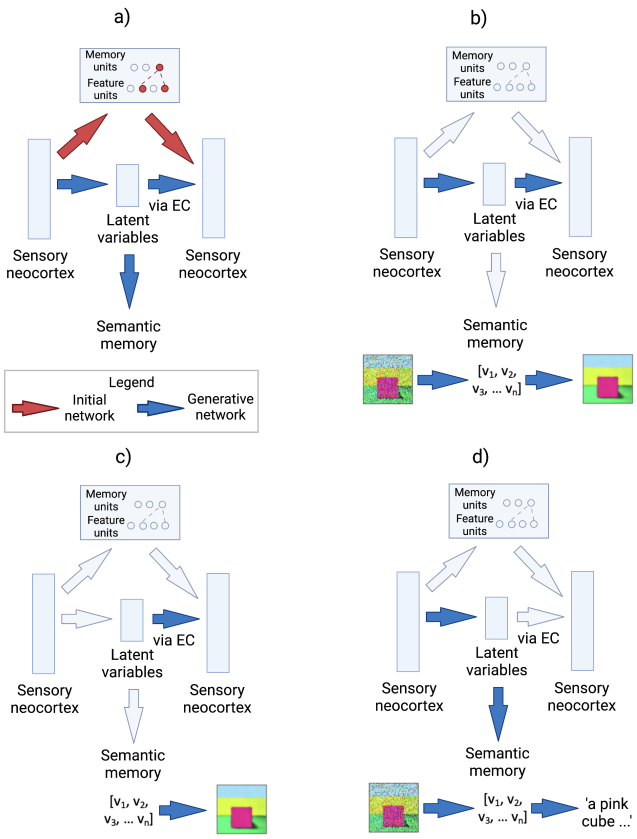

Figure 2: a) Basic architecture of the model. b) Episodic memory after consolidation: a partial input is mapped to latent variables. The HF and return projections to neocortex then decode these back into an experience. c) Imagination: latent variables are decoded into an experience by the HF and return projections to neocortex. d) Semantic memory: a partial input is mapped to latent variables, which capture the 'key facts' of the scene.

Figure 3: Random noise inputs to a modern Hopfield network reactivate its memories. Results are shown across four datasets: a) MNIST, b) Fashion-MNIST, c) Shapes3D (converted to black and white), and d) Symmetric Solids. One modern Hopfield network was used per dataset.

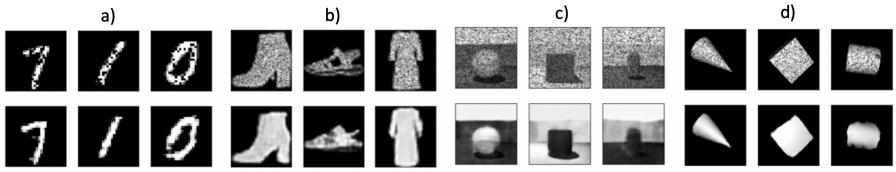

Figure 4: A generative model (in this case, a variational autoencoder) can recall images from a partial input, following training on reactivated memories from a modern Hopfield network. One generative model was trained per dataset.

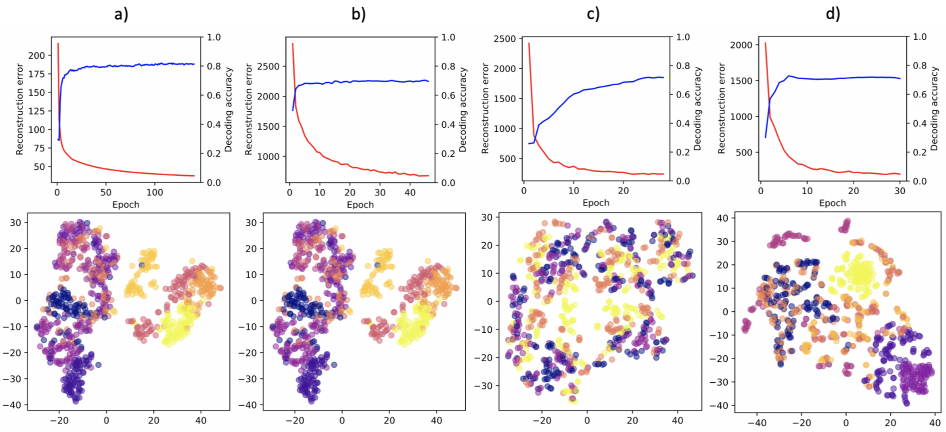

Figure 5: Upper row: reconstruction error (red) and decoding accuracy (blue) during training. The latter is measured by training a new support vector classifier on 200 examples at the end of each epoch. Lower row: the latent space projected into 2D with TSNE, colour coded by label. Results are shown across four datasets: a) MNIST, b) Fashion-MNIST, c) Shapes3D (converted to black and white), and d) Symmetric Solids.

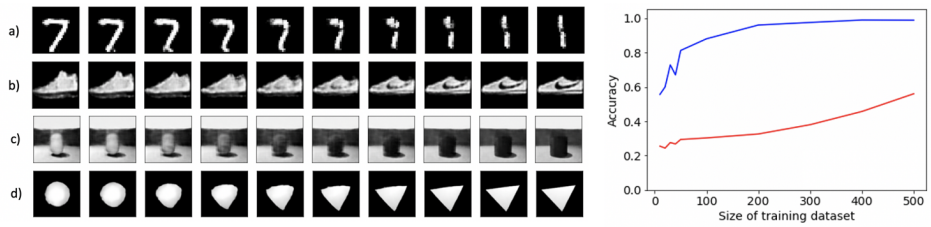

Figure 6: Left: Generating new items with interpolation in latent space. Each row shows points along a line in the latent space between two items from the training data. Right: Accuracy against training dataset size when learning to predict object shape from Shapes3D images in two ways: by training a perceptron on latent vectors (blue), and by training a convolutional neural network from scratch (red).

4 Discussion

We model systems consolidation as the training of a generative neural network through teacher-student learning. First the hippocampal 'teacher' rapidly encodes an event, modelled as one-shot memorisation in a modern Hopfield network. After exposure to replayed representations from the 'teacher', a generative 'student' model representing the association cortex and HF supports reconstruction of events. In contrast to the relatively veridical initial encoding, the generative model learns to capture the probability distributions underlying experiences, or 'schemas'. This enables not just efficient recall, in which the model reconstructs memories without the need to store them individually, but also imagination (by sampling from the distributions), inference (by using the learned statistics of experience to predict the values of unseen variables), and semantic memory.

Further research will extend the model to demonstrate how sensory and conceptual elements may be combined in memory, reducing the dimensionality of the initial network's input, and allowing very recent memory to exploit predictions from the generative model. (Each memory is represented as the sum of a predictable and an unpredictable component, where the predictable component is a schema, and the unpredictable component consists of parts of the stimuli that were poorly predicted by the existing generative model. The modern Hopfield network model of initial encoding in the HF can have both conceptual and sensory feature units, where conceptual feature units may correspond to concept cells [53].) Other areas for future work include extending these findings to sequences, and exploring how the system might achieve continual representation learning.

References

[1] Frederic Charles Bartlett. *Remembering: A study in experimental and social psychology*. Cambridge university press, 1932.

[2] Daniel L Schacter. Constructive memory: past and future. *Dialogues in clinical neuroscience*, 14(1):7, 2012.

[3] David Marr. A theory for cerebral neocortex. *Proceedings of the Royal society of London. Series B. Biological sciences*, 176(1043):161–234, 1970.

[4] David Marr. Simple memory: a theory for archicortex. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 262(841):23–81, 1971.

[5] James L McClelland, Bruce L McNaughton, and Randall 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. *Psychological review*, 102(3):419, 1995.

[6] Timothy J Teyler and Pascal DiScenna. The hippocampal memory indexing theory. *Behavioral neuroscience*, 100(2):147, 1986.

[7] Pablo Alvarez and Larry R Squire. Memory consolidation and the medial temporal lobe: a simple network model. *Proceedings of the national academy of sciences*, 91(15):7041–7045, 1994.

[8] Lynn Nadel and Morris Moscovitch. Memory consolidation, retrograde amnesia and the hippocampal complex. *Current opinion in neurobiology*, 7(2):217–227, 1997.

[9] Gordon Winocur and Morris Moscovitch. Memory transformation and systems consolidation. *Journal of the International Neuropsychological Society*, 17(5):766–780, 2011.

[10] Yitzhak Norman, Omri Raccah, Su Liu, Josef Parvizi, and Rafael Malach. Hippocampal ripples and their coordinated dialogue with the default mode network during recent and remote recollection. *Neuron*, 109(17):2767–2780, 2021.

[11] Suzanna Becker and Neil Burgess. Modelling spatial recall, mental imagery and neglect. *Advances in neural information processing systems*, 13, 2000.

[12] Patrick Byrne, Suzanna Becker, and Neil Burgess. Remembering the past and imagining the future: a neural model of spatial memory and imagery. *Psychological review*, 114(2):340, 2007.

[13] Andrej Bicanski and Neil Burgess. A neural-level model of spatial memory and imagery. *Elife*, 7:e33752, 2018.

[14] Demis Hassabis, Dharshan Kumaran, Seralynne D Vann, and Eleanor A Maguire. Patients with hippocampal amnesia cannot imagine new experiences. *Proceedings of the National Academy of Sciences*, 104(5):1726–1731, 2007.

[15] Goffredina Spanò, Gloria Pizzamiglio, Cornelia McCormick, Ian A Clark, Sara De Felice, Thomas D Miller, Jamie O Edgin, Clive R Rosenthal, and Eleanor A Maguire. Dreaming with hippocampal damage. *Elife*, 9:e56211, 2020.

[16] Cornelia McCormick, Clive R Rosenthal, Thomas D Miller, and Eleanor A Maguire. Mind-wandering in people with hippocampal damage. *Journal of Neuroscience*, 38(11):2745–2754, 2018.

[17] Donna Rose Addis, Alana T Wong, and Daniel L Schacter. Remembering the past and imagining the future: common and distinct neural substrates during event construction and elaboration. *Neuropsychologia*, 45(7):1363–1377, 2007.

[18] Demis Hassabis and Eleanor A Maguire. Deconstructing episodic memory with construction. *Trends in cognitive sciences*, 11(7):299–306, 2007.

[19] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.

[20] Szabolcs Káli and Peter Dayan. Hippocampally-dependent consolidation in a hierarchical model of neocortex. *Advances in Neural Information Processing Systems*, 13, 2000.

[21] Szabolcs Káli and Peter Dayan. Replay, repair and consolidation. *Advances in Neural Information Processing Systems*, 15, 2002.

[22] James CR Whittington, Timothy H Muller, Shirley Mark, Guifen Chen, Caswell Barry, Neil Burgess, and Timothy EJ Behrens. The tolman-eichenbaum machine: Unifying space and relational memory through generalization in the hippocampal formation. *Cell*, 183(5):1249–1263, 2020.

[23] David G Nagy, Balázs Török, and Gergő Orbán. Optimal forgetting: Semantic compression of episodic memories. *PLoS Computational Biology*, 16(10):e1008367, 2020.

[24] Gido M van de Ven, Hava T Siegelmann, and Andreas S Tolias. Brain-inspired replay for continual learning with artificial neural networks. *Nature communications*, 11(1):1–14, 2020.

[25] William Beecher Scoville and Brenda Milner. Loss of recent memory after bilateral hippocampal lesions. *Journal of neurology, neurosurgery, and psychiatry*, 20(1):11, 1957.

[26] Larry R Squire, Lisa Genzel, John T Wixted, and Richard G Morris. Memory consolidation. *Cold Spring Harbor perspectives in biology*, 7(8):a021766, 2015.

[27] Dmitry Krotov and John Hopfield. Large associative memory problem in neurobiology and machine learning. *arXiv preprint arXiv:2008.06996*, 2020.

[28] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[29] Daniel L Schacter, Donna Rose Addis, and Randy L Buckner. Remembering the past to imagine the future: the prospective brain. *Nature reviews neuroscience*, 8(9):657–661, 2007.

[30] Natalie Biderman, Akram Bakkour, and Daphna Shohamy. What are memories for? the hippocampus bridges past experience with future decisions. *Trends in Cognitive Sciences*, 24(7):542–556, 2020.

[31] Oded Bein, Natalie A Plotkin, and Lila Davachi. Mnemonic prediction errors promote detailed memories. *Learning & Memory*, 28(11):422–434, 2021.

[32] Brynn E Sherman, Kathryn N Graves, David M Huberdeau, Imran H Quraishi, Eyiyemisi C Damisah, and Nicholas B Turk-Browne. Temporal dynamics of competition between statistical learning and episodic memory in intracranial recordings of human visual cortex. *bioRxiv*, 2022.

[33] Weinan Sun, Madhu Advani, Nelson Spruston, Andrew Saxe, and James E Fitzgerald. Organizing memories for generalization in complementary learning systems. *BioRxiv*, 2021.

[34] Federico Stella, Peter Baracskay, Joseph O'Neill, and Jozsef Csicsvari. Hippocampal reactivation of random trajectories resembling brownian diffusion. *Neuron*, 102(2):450–461, 2019.

[35] Oscar C González, Yury Sokolov, Giri P Krishnan, Jean Erik Delanois, and Maxim Bazhenov. Can sleep protect memories from catastrophic forgetting? *Elife*, 9:e51005, 2020.

[36] Giovanni Pezzulo, Marco Zorzi, and Maurizio Corbetta. The secret life of predictive brains: what's spontaneous activity for? *Trends in Cognitive Sciences*, 25(9):730–743, 2021.

[37] Edvard I Moser, Emilio Kropff, and May-Britt Moser. Place cells, grid cells, and the brain's spatial representation system. *Annu. Rev. Neurosci.*, 31:69–89, 2008.

[38] Alexandra O Constantinescu, Jill X O'Reilly, and Timothy EJ Behrens. Organizing conceptual knowledge in humans with a gridlike code. *Science*, 352(6292):1464–1468, 2016.

[39] Asaf Gilboa and Hannah Marlatte. Neurobiology of schemas and schema-mediated memory. *Trends in cognitive sciences*, 21(8):618–631, 2017.

[40] Atsuko Takashima, Karl Magnus Petersson, F Rutters, I Tendolkar, O Jensen, MJ Zwarts, BL McNaughton, and G Fernández. Declarative memory consolidation in humans: a prospective functional magnetic resonance imaging study. *Proceedings of the National Academy of Sciences*, 103(3):756–761, 2006.

[41] Steffen Gais, Geneviève Albouy, Mélanie Boly, Thien Thanh Dang-Vu, Annabelle Darsaud, Martin Desseilles, Géraldine Rauchs, Manuel Schabus, Virginie Sterpenich, Gilles Vandewalle, et al. Sleep transforms the cerebral trace of declarative memories. *Proceedings of the National Academy of Sciences*, 104(47):18778–18783, 2007.

[42] Paul W Frankland and Bruno Bontempi. The organization of recent and remote memories. *Nature reviews neuroscience*, 6(2):119–130, 2005.

[43] Marlieke TR Van Kesteren, Guillén Fernández, David G Norris, and Erno J Hermans. Persistent schema-dependent hippocampal-neocortical connectivity during memory encoding and postencoding rest in humans. *Proceedings of the National Academy of Sciences*, 107(16):7550–7555, 2010.

[44] Karim Benchenane, Adrien Peyrache, Mehdi Khamassi, Patrick L Tierney, Yves Gioanni, Francesco P Battaglia, and Sidney I Wiener. Coherent theta oscillations and reorganization of spike timing in the hippocampal-prefrontal network upon learning. *Neuron*, 66(6):921–936, 2010.

[45] Vanessa E Ghosh and Asaf Gilboa. What is a memory schema? a historical perspective on current neuroscience literature. *Neuropsychologia*, 53:104–114, 2014.

[46] Timothy R Koscik and Daniel Tranel. The human ventromedial prefrontal cortex is critical for transitive inference. *Journal of cognitive neuroscience*, 24(5):1191–1204, 2012.

[47] Kelsey N Spalding, Margaret L Schlichting, Dagmar Zeithamova, Alison R Preston, Daniel Tranel, Melissa C Duff, and David E Warren. Ventromedial prefrontal cortex is necessary for normal associative inference and memory integration. *Journal of Neuroscience*, 38(15):3767–3775, 2018.

[48] Michael L Mack, Alison R Preston, and Bradley C Love. Ventromedial prefrontal cortex compression during concept learning. *Nature communications*, 11(1):1–11, 2020.

[49] Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. at&t labs, 2010.

[50] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[51] Chris Burgess and Hyunjik Kim. 3d shapes dataset, 2018.

[52] Kieran Murphy, Carlos Esteves, Varun Jampani, Srikumar Ramalingam, and Ameesh Makadia. Implicit-pdf: Non-parametric representation of probability distributions on the rotation manifold. *arXiv preprint arXiv:2106.05965*, 2021.

[53] Rodrigo Quian Quiroga. Concept cells: the building blocks of declarative memory functions. *Nature Reviews Neuroscience*, 13(8):587–597, 2012.

[54] Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. Deep learning for classical japanese literature. *arXiv preprint arXiv:1812.01718*, 2018.

[55] Diederik P Kingma and Max Welling. An introduction to variational autoencoders. *arXiv preprint arXiv:1906.02691*, 2019.

[56] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

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

[58] Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, et al. Hopfield networks is all you need. *arXiv preprint arXiv:2008.02217*, 2020.

[59] Dmitry Krotov and John J Hopfield. Dense associative memory for pattern recognition. *Advances in neural information processing systems*, 29, 2016.

[60] Mete Demircigil, Judith Heusel, Matthias Löwe, Sven Upgang, and Franck Vermet. On a model of associative memory with huge storage capacity. *Journal of Statistical Physics*, 168(2):288–299, 2017.

5 Appendix

Code for all simulations can be found in the supplementary materials.

Diagrams were created using BioRender.com.

5.1 Datasets

The following datasets (all covered by the Creative Commons Attribution 4.0 License) were used in the simulations:

Dataset	Origin
MNIST [49]	https://www.tensorflow.org/datasets/catalog/mnist
Fashion-MNIST [50]	https://www.tensorflow.org/datasets/catalog/fashion_mnist
Shapes3D [51]	https://www.tensorflow.org/datasets/catalog/Shapes3D
Symmetric Solids [52]	https://www.tensorflow.org/datasets/catalog/symmetric_solids
KMNIST [54]	https://www.tensorflow.org/datasets/catalog/kmnist

5.2 Model details

5.2.1 Variational autoencoders

An autoencoder is a neural network which encodes an input into a shorter vector, and then decodes this compressed representation back to the original. It learns by minimising the difference between the inputs and outputs. There is no guarantee that decoding an arbitrary compressed representation produces a sensible output, so standard autoencoders do not perform well as generative models. In other words, there are many regions in the vector space of the compressed representations which do not correspond to anything meaningful. However, one can train an autoencoder with special properties, such that each latent variable is normally distributed for a given input, which allow one to sample realistic items. The result is called a variational autoencoder [28, 55]. (Latent variables can be thought of as hidden factors behind the observed data.)

The VAEs in these simulations use convolutional layers to better encode and decode image features. Convolutional layers learn sliding windows that scan the image for a relevant feature, outputting a stack of feature maps [56]. Applying such a layer to the output of a preceding convolutional layer has the effect of finding higher-level features in the stacked feature maps, i.e. if the first convolutional layer learns to identify simple features such as lines at different orientations, the second convolutional layer might learn features consisting of combinations of lines.

In the encoder used in most simulations, five convolutional layers, plus a final pooling layer, gradually decrease the width and height of the representation and increase the depth (as is standard when using convolutional neural networks to encode images). In the decoder, five convolutional layers alternate with up-sampling layers to increase the width and height of the representation and decrease the depth. The encoder has filters (i.e. convolution windows, or feature detectors) of 4x4 pixels per layer. The convolutional layers one to five have 32, 64, 128, 256 and 512 filters respectively. The decoder has filters of 3x3 pixels per layer. The convolutional layers one to five have 128, 64, 32, 16 and 3 filters respectively (i.e. the final three filters are the three channels of an RGB image).

5.2.2 Modern Hopfield networks

A Hopfield network uses a simple Hebbian learning rule to memorise patterns after a single exposure [57]. However one issue is their limited capacity; a Hopfield network can only recall approximately 0.14d states, where d is the dimension of the input data [58]. It therefore seems unlikely that classical Hopfield networks are a good model of hippocampal memory encoding – even if we assume that only a temporary store is required until consolidation occurs. In addition, they frequently recall incorrect memories, as the energy function can get 'stuck' in a local minimum.

However, recent research has shown that the storage capacity of a Hopfield network can be increased in several ways. Krotov and Hopfield [59] devise a new energy function involving a polynomial

function, and a corresponding update rule to minimise this; the activation of a node flips from -1 to 1 or vice versa if the energy is lower in the flipped state. Demircigil et al. [60] develop this idea further, increasing the capacity from approximately 0.14d to $2^{d/2}$ with the use of an exponential energy function. Ramsauer et al. [58] extend this to memories involving continuous variables and further amend the energy function, enabling the recall of much more complex data. (For example, whilst classical Hopfield networks can only recall black and white images, the modern variant can recall greyscale ones.)

However, understanding these new variants of Hopfield networks in terms of neural networks is less straightforward. To recap, the equations below from Krotov and Hopfield [59] give the energy of a standard Hopfield Network. During recall, a node's value is updated to the sign of the weighted sum of its inputs; in other words, a node's value is flipped if it decreases the energy. The matrix T gives the weights of the network, and the calculation of T is simply Hebbian learning.

$$E = -\frac{1}{2}\sum_{i,j=1}^{N} \sigma_i T_{ij} \sigma_j, \qquad T_{ij} = \sum_{\mu=1}^{K} \xi_i^\mu \xi_j^\mu$$

The equation below from Krotov and Hopfield [27] gives the energy of a dense Hopfield network. In this example F(x) is x^3, but it can be any polynomial function. As above, at recall time a node's value flips if it decreases the energy. When F(x) is x^2, the equation reduces to the one above for a standard Hopfield network. In any other case, the tensor T has more than two indices, and can no longer be thought of a matrix produced by Hebbian learning. This means the energy is no longer a function of weights and activations in a neural network. Modern Hopfield networks [27] suffer from the same problem.

$$E = -\sum_{\mu} F(\sum_{i} \xi_{\mu i} \sigma_i) = -\sum_{i,j,k} T_{ijk} \sigma_i \sigma_j \sigma_k$$

Krotov and Hopfield [27] suggest a way to overcome this problem by using hidden units (which they call 'memory units') in addition to the 'feature units' which represent the input. As a result, a modern Hopfield network can be understood as a neural network, like its predecessor. The authors provide two equations for the evolution of the feature neurons and hidden neurons over time. Rather than using discrete time steps as in a classical Hopfield network, time is modelled as continuous. They therefore give a pair of differential equations, in which change to each set of currents is driven by the weighted sum of currents in the other layer. They then define an energy function, chosen 'so that the energy function decreases on the dynamical trajectory'. The energy function has three terms: energy in the feature neurons, energy in the hidden neurons, and energy from the interaction between the two groups. Importantly, the interaction term can be described in terms of two-body synapses, so once again the energy is a function of weights and activations in a neural network.

The authors state that 'the memory patterns ... can be interpreted as the strengths of the synapses connecting feature and memory neurons'. To understand the intuition behind this, suppose we set the weights connecting a particular hidden node with the feature neurons to the values of the pattern to be memorised. Then activating the hidden node results in the pattern being reinstated in the feature neurons. In other words, each hidden node represents a memory, and each memory could be encoded using Hebbian learning. The key point is that the energy does not require a matrix of stored patterns, unlike in earlier formulations of modern Hopfield networks – the patterns are encoded in the weights, and the energy is a function of weights and activations as explained above.

Krotov and Hopfield [27] show that under different circumstances, their formulation can be simplified to dense associative memory [59], or modern Hopfield networks [58]. Having established that modern Hopfield networks increase memory performance and are biologically plausible (in the sense that they involve only 'two-body synapses', and that memories can be stored as weights), we use them to model the initial learning in the hippocampus; Ramsauer et al. [58] provide a Python implementation that our code is based on.

An important question is how the memories get encoded as the weights of a bipartite graph in the Krotov and Hopfield [27] formulation of a modern Hopfield network. Each memory is bound together by a single node, which connects the features that comprise that memory. The weights between a given memory node and the feature nodes are simply the values of the features for that memory;

these weights can be learned by Hebbian learning. Therefore encoding in a modern Hopfield network is very similar to previous models of the hippocampus as 'indexing', or binding together, a set of memory components [6]. The innovative aspect of modern Hopfield networks is the update rule, which is cleverly designed to guarantee the desired properties.

It should be noted that the initial model could be swapped out for other computational models of associative memory, providing they i) are high capacity, ii) can retrieve memories from noise, and iii) are capable of one-shot memorisation.