Semantically-correlated memories in a dense associative model

Thomas F Burns [1 2 3]

Abstract

I introduce a novel associative memory model named *Correlated Dense Associative Memory* (CDAM), which integrates both auto- and hetero-association in a unified framework for continuous-valued memory patterns. Employing an arbitrary graph structure to semantically link memory patterns, CDAM is theoretically and numerically analysed, revealing four distinct dynamical modes: auto-association, narrow hetero-association, wide hetero-association, and neutral quiescence. Drawing inspiration from inhibitory modulation studies, I employ anti-Hebbian learning rules to control the range of hetero-association, extract multi-scale representations of community structures in graphs, and stabilise the recall of temporal sequences. Experimental demonstrations showcase CDAM's efficacy in handling real-world data, replicating a classical neuroscience experiment, performing image retrieval, and simulating arbitrary finite automata.

1. Introduction

1.1. Background

Mathematical models of ferromagnetism in statistical mechanics, as developed by Lenz, Ising, Schottky, and others (Brush, 1967; Folk & Holovatch, 2022), model the interactions between collections of discrete variables. When connected discrete variables disagree in their values, the energy of the system increases. The system trends toward low energy states via recurrent dynamics, but can be perturbed or biased by external input. Marr (1971) proposed a conceptual framework of associative memory in neurobiological systems using a similar principle but of interacting neurons, which was subsequently formalised in a similar way (Nakano, 1972; Amari, 1972; Little, 1974; Stanley, 1976; Hopfield, 1982)[1]. A key difference between these associative memory and ferromagnetism models is that the neurons are typically connected all-to-all with infinite-range interactions whereas in the ferromagnetism models variables were typically connected locally within a finite range.

The principle by which these associative memory networks store memories is by assigning recurrent connection weights and update rules such that the energy landscape of the network forms dynamic attractors (low energy states) around memory patterns (particular states of the neurons). In the case of pairwise connections, these weights translate to the synaptic strength between pairs of neurons in biological neural networks. The network therefore acts as a content addressable memory – given a partial or noise-corrupted memory, the network can update its states through recurrent dynamics to retrieve the full memory.

Of particular interest to the machine learning community is the recent development of dense associative memory networks (Krotov & Hopfield, 2016) (also referred to as modern Hopfield networks) and their close correspondence (Ramsauer et al., 2021) to the attention mechanism of Transformers (Vaswani et al., 2017). In particular, the dense associative memory networks introduced by Krotov & Hopfield (2016) (including with continuous variables) were generalised by using the SOFTMAX activation function, whereby Ramsauer et al. (2021) showed a connection to the attention mechanism of Transformers (Vaswani et al., 2017). Indeed, Krotov & Hopfield (2016) make a mathematical analogy between their energy-based update rule and setwise connections given their energy-based update rule can be interpreted as allowing individual pairs of pre- and post-synaptic neurons to make multiple synapses with each other – making pairwise connections mathematically as strong as equivalently-ordered setwise connections. Demircigil et al. (2017) later proved this analogy to be accurate in terms of theoretical memory capacity. As shown subsequently, by explicitly modelling higher-ordered connections in such networks, the energy landscape becomes sharper and memory capacity is increased (Burns & Fukai, 2023).

[1]Institute for Computational and Experimental Research in Mathematics, Brown University, USA [2]SciAI Center, Cornell University, USA [3]Neural Coding and Brain Computing Unit, OIST Graduate University, Japan. Correspondence to: Thomas F Burns <tfb43@cornell.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

[1]Simultaneously, work in spin glasses followed a similar mathematical trajectory in the works of Sherrington & Kirkpatrick (1975) and Pastur & Figotin (1977).

In the majority of the prior associative memory works discussed so far, memory recall is auto-associative, i.e., given some partial memory the dynamics of the network ideally lead to recalling the (same) full memory. However, hetero-association is just as valid dynamically (Amari, 1972; Gutfreund & Mezard, 1988; Griniasty et al., 1993; Gillett et al., 2020; Tyulmankov et al., 2021; Millidge et al., 2022; Karuvally et al., 2023; Chaudhry et al., 2023)[2]: instead of a partial memory directing the dynamics to recalling the same memory pattern, we can instead recall something else. Such hetero-associations are believed to naturally occur in the oscillatory dynamics of central pattern generators for locomotion (Stent et al., 1978), sequence memory storage in hippocampus (Treves & Amit, 1988), and visual working-memory in primate temporal cortex (Miyashita, 1988).

1.2. Motivations from neuroscience

A classical result in the hetero-association neuroscience literature is due to Miyashita (1988). This work demonstrated hetero-association of stimuli in monkey temporal cortex could arise semantically via repeated presentations of the same stimuli in the same order, not only spatially via similarities in the stimuli themselves. Miyashita (1988) showed neurons responsive to presentation of randomly-generated fractal patterns had a monotonically-decreasing auto-correlation between the firing rates due to the current pattern and the next expected patterns, up to a distance of 6 patterns into the past or future of the stimuli sequence.

Work on numerosity in birds, non-human primates, and humans (Nieder et al., 2002; Ditz & Nieder, 2015; Nieder, 2012; Kutter et al., 2018) have repeatedly provided evidence of neurons responding to specific numbers or quantities. In these experiments, the stimuli (numbers or quantities) can be both semantically and spatially correlated – i.e., they can have the known semantic ordering of '$1, 2, 3 \ldots$' or 'some, more, even more $\ldots$', as well as the spatial or statistical relationships between the stimuli, e.g., visually, the numerals '4' and '9'. Notably, even in abstract number experiments where spatial correlations are moot, semantic distances up to a range of ~ 5 numbers[3] (as measured by significant auto-correlations of the neural activity) are common.

This phenomenon extends beyond simple 1D, sequence relationships, however. Schapiro et al. (2013) presented human participants with a series of arbitrary visual stimuli which were ordered by a random walk on a graph with community structure (where each image was associated with a vertex in the graph). Functional magnetic resonance imaging analysis of the blood-oxygen-level-dependent response showed the representations of different stimuli were clustered by brain activity into the communities given by the underlying graph and unrelated to the actual stimuli features.

In all of these studies, both auto-association (for the present stimuli) and hetero-association (for the semantically-related stimuli) is present. And such mixtures, where they encode more general structures relevant for tasks, may be behaviourally useful. For instance, mice trained on goal-sequence tasks sharing a common semantic basis arising from a 2D lattice graph develop task-progress cells which generalise across tasks, physical distances, behavioural timescales, and stimuli modality (El-Gaby et al., 2023). Furthermore, similar dynamics may be modulated by inhibitory signals (King et al., 2013; Honey et al., 2017; Hertäg & Sprekeler, 2019; Haga & Fukai, 2019; 2021; Burns et al., 2022; Tobin et al., 2023) to shift the locus of attention, learning, or behaviour. Such function could account for the many instances of anti-Hebbian learning found throughout neural systems (Roberts & Leen, 2010; Shulz & Feldman, 2013), as well as their implications in the role of sleep for memory pruning (Crick & Mitchison, 1983; Hopfield et al., 1983; Diekelmann & Born, 2010; Poe, 2017; Zhou et al., 2020), motor control learning (Nashef et al., 2022), dendritic selectivity (Hayama et al., 2013; Paille et al., 2013) and input source separation (Brito & Gerstner, 2016).

1.3. Motivations from machine learning

Given the storied history of classical hetero-associative modelling work, extensions to dense associative memory are a natural next step. Some work in this direction has already begun. Millidge et al. (2022) present an elegant perspective which makes it straightforward to construct dense associative memory networks with hetero-association, and demonstrated recalling the opposite halves of MNIST or CIFAR10 images. Karuvally et al. (2023) construct an adiabatically-varying energy surface to entrain sequences in a series of meta-stable states, using temporal delays for memories to interact via a hidden layer. Application to a toy sequence episodic memory task showed how the delay signal can shift the attractive regime. And Chaudhry et al. (2023) studied a sequence-based extension of the dense associative memory model by adopting the polynomial or exponential update rule for binary-valued sequences of memories. This work also introduces a generalisation of the Kanter & Sompolinsky (1987) pseudoinverse rule to improve distinguishability between correlated memories. As Chaudhry et al. (2023) conclude, many potential research avenues remain, including extending these methods to continuous-valued patterns.

Chaudhry et al. (2023) also note the potential to study different network topologies. There are several distinct notions of

[2]An interesting alternative or supplementary technique is to use synaptic delays to generate such sequences (Tank & Hopfield, 1987; Kleinfeld & Sompolinsky, 1988; Karuvally et al., 2023), however here I will focus on non-delayed hetero-association where synapses all operate at the same timescale.

[3]Depending on the species, brain area, and stimuli modality.

network topology which we could study, including that of neuronal connections (as in Löwe & Vermet (2011); Burns & Fukai (2023)), spatial or statistical relationships between memory patterns (as in Löwe (1998); De Marzo & Iannelli (2023)), or semantic relationships between memory patterns (as in Amari (1972); Chaudhry et al. (2023)). A majority of classical work has focused on semantic correlation, likely due to its relevance to neuroscience (see Subsection 1.2). To extend the study of such semantic relationships to interesting topologies, it is necessary to introduce a basic topology, such as via embedding memories in graphs (as in Schapiro et al. (2013)). Being highly versatile mathematical structures, upon generalising semantic relationships with graphs, this additionally generates opportunities to study graph-based computations such as community detection and simulation of (finite) automata (Balle & Maillard, 2017; Ardakani et al., 2020; Liu et al., 2023).

In Appendix A.1, I summarise the technical connection between Transformers and associative memory, and how Transformers' attention mechanisms take the hetero-associative form mathematically. Functionally, however, the attention mechanism is not obliged to perform hetero-association, since its values and keys are created by their distinct weight matrices (see Vaswani et al. (2017)) and can in-principle align these functionally so as to perform auto-association, or otherwise some mixture of auto- and hetero-association. Taking this perspective seriously opens the way for analysing Transformers through the lens of potential mixtures of auto- and hetero-associative dynamics, à la the analysis of a large language model in Ramsauer et al. (2021) by considering the implied energy landscapes in each of its attention heads. For this to be possible, however, a first step is to rigorously develop and study a dense auto- and hetero-association model and its inherent computational capabilities. (In Appendix A.2, I provide suggestions for new interpretation approaches based on this paper's results.)

1.4. Contributions

With these joint motivations from neuroscience and machine learning in mind, I:

- Introduce a new dense associative memory model, *Correlated Dense Associative Memory (CDAM)*, which integrates a controllable mixture of auto- and hetero-association for dynamics on continuous-valued memory patterns, using an underlying (arbitrary) graph structure to semantically hetero-associate memories;
- Theoretically and numerically analyse CDAM's dynamics, demonstrating four distinct dynamical modes: auto-association, narrow hetero-association, wide hetero-association, and neutral quiescence;
- Taking inspiration from inhibitory modulation, I demonstrate how anti-Hebbian learning can be used to: (i) widen the range of hetero-association across memories; (ii) extract multi-scale representations of community structures in memory graph structures; (iii) stabilise recall of temporal sequences; and (iv) enhance performance in a non-traditional auto-association task; and
- Illustrate via experiments CDAM's capacity to work with real data, replicate a classical neuroscience result, and simulate arbitrary finite automata.

2. Correlated Dense Associative Memory

2.1. Model

To embed memories in the network, we first create p patterns as continuous-valued vectors of length n, the number of neurons in the network. These *memory patterns* can be random, partially-random, or themselves contain content we wish to store. In the random case, each component of a memory vector is independently sampled from the interval $[0, 1]$. In the partially-random case, we reserve some portion of the vector for structured memory and the rest is random in the same sense as before. We denote an individual memory pattern μ as the vector ξ^μ, where the ith component corresponds to neuron i. For convenience, we organise these vectors into a memory matrix $\Xi \in \mathbb{R}^{n\times p}$. We also define a *mean memory load* vector, $\tilde{\xi} := \frac{1}{p}\sum_{\mu=1}^{p}\xi^\mu$.

Next, we choose a finite graph $\mathcal{M} = (\mathcal{V}, \mathcal{E})$ with $|\mathcal{V}| = p$ vertices. We allow $\mathcal{E}$ to be a multiset in order to allow $\mathcal{M}$ to be a multigraph. We also allow elements of $\mathcal{E}$ to be weighted using a weight function, $\mathscr{W}$. The graph, which we refer to as the *memory graph*, forms the basis for the inter-pattern hetero-associations via its normalised adjacency matrix $M = D^{-1/2}AD^{-1/2}$, where D is the degree matrix of $\mathcal{M}$ and A is the adjacency matrix of $\mathcal{M}$.

We use discretised time and denote the network state at time t as $\sigma(t) \in \mathbb{R}^n$. To use the language of Millidge et al. (2022), we use SOFTMAX as our separation function, which is defined for each component in vector z as $\text{SOFTMAX}(z_i) := \frac{\exp(z_i)}{\sum_j \exp(z_j)}$. Starting at a chosen initial state $\sigma(0)$, subsequent states are given inductively by

$$\sigma(t+1) = \sigma(t) + \eta\left(\left[\text{SOFTMAX}(\beta\sigma(t)\Xi)Q - \frac{1}{n}\tilde{\xi}^T\right] - \sigma(t)\right), \quad Q := a\Xi + h\Xi M^T, \quad (1)$$

where $\eta \in \mathbb{R}^+$ is the magnitude of each update, $\beta \in \mathbb{R}^+$ is the inverse temperature (which can be thought of as controlling the level of mixing between memory patterns during retrieval), and $a, h \in \mathbb{R}$ are the strengths of auto- and hetero-association in the retrieval projection matrix Q, respectively.

2.2. Theoretical analysis

A typical analysis to perform on associative memory networks is to probe its memory storage capacity, i.e., how many memories can be stored and reliably retrieved given n neurons? In CDAM, when $a, h \neq 0$, the regular notions of 'capacity' seem inapplicable. This is because 'capacity' is normally measured in the pure auto-associative case by giving a noise-corrupted or partial memory pattern, and observing whether and how closely the model's dynamics converge to the uncorrupted or complete memory pattern (e.g., see Amit et al. (1985) for the classical model and Demircigil et al. (2017) for the dense model). In the pure hetero-associative case, 'capacity' has (to my knowledge) only ever been studied in the linear sequences case (e.g., see Löwe (1998) for the classical model and Chaudhry et al. (2023) for the dense model). However, in this model I study general mixtures of both auto- and hetero-association, as well as arbitrary memory graphs (not just linear cycles). It is therefore unclear what an appropriate notion of 'capacity' for this mixture would be and what it would measure[4].

One can, nevertheless, study the model in a similar spirit of analysis. To this end, I demonstrate the dynamics of the model in the thermodynamic limit. First, let us set aside the choice of η which controls the amplitude of each step's update. We define the *overlap* between a memory pattern and state as $m^\mu(t) = \frac{\sigma(t)^T \xi^\mu}{n}$. For an undirected memory graph $\mathcal{M}$ without loops, the energy function is

$$\begin{aligned}\mathfrak{E}(\sigma(t)) \propto &-\frac{1}{\beta}\, a \log \sum_{\mu=1}^{p} \exp[\beta m^\mu(t) m^\mu(t)] \\ &-\frac{1}{\beta}\, h\, 2 \log \sum_{\{\alpha,\kappa\}\in\mathcal{E}} \exp[\beta m^\alpha(t) m^\kappa(t)] \\ =&-\frac{1}{\beta}\, a \log \sum_{\mu=1}^{p} \exp[\beta m^\mu(t) m^\mu(t)] \\ &-\frac{1}{\beta}\, h \log \sum_{\alpha=1}^{p}\sum_{\kappa=1}^{p} M_{\alpha,\kappa} \exp[\beta m^\alpha(t) m^\kappa(t)]\ ,\end{aligned} \tag{2}$$

where $\mathcal{E}$ is the multiset of edges in $\mathcal{M}$.

Assume $\mathcal{M}$ is k–regular, meaning each vertex has degree k. This will cause there to be k non-zero values in the hetero-associative term. For a brief moment, let $M = A$. While in a Hebbian hetero-associative regime, i.e., $h > 0$, setting $a < -kh$ gives the trivial minimisation of letting all m^μ terms vanish, i.e., having a state which is far from any pattern. However, when $a > -kh$, minimisation of the energy demands maximising the auto-association under penalty of the consequent hetero-association. In the absence of the hetero-association penalty, we have the model of Equation 13 in Lucibello & Mézard (2023), where scaling comes from a; similarly, in the absence of auto-association, we have a model similar to Chaudhry et al. (2023) scaled by h and with arbitrary semantic correlations according to $\mathcal{M}$. In our case, however, where there is a mixture of auto- and hetero-associations (which act simultaneously), the hetero-associative component of Equation 2 causes a large number of pattern activations for negative values of a, i.e., while $0 > a > -kh$. When $a, h > 0$, hetero-association remains but across a narrower range. This gives us a neat analysis for k–regular graphs, which then generalises to arbitrary graphs when we use $M = D^{-1/2} A D^{-1/2}$, where we can let $k = 1$ in the above to reach the same behaviours.

In Appendix A.4, I show numerical simulations to demonstrate these four behavioural modes: auto-association, narrow hetero-association, wide hetero-association, and neutral quiescence. These simulations also demonstrate how we need $a + h = 1$ for the mean neural activity to converge to 0 in the limit of $t \to \infty$ and $n \to \infty$ when we have random memory patterns, i.e., to keep an unbiased excitatory–inhibitory (E–I) balance[5].

For the case of a directed memory graph $\overrightarrow{\mathcal{M}}$, the energy function is

$$\begin{aligned}\mathfrak{E}(\sigma(t)) \propto -\frac{1}{\beta} \log(a &\sum_{\mu=1}^{p} \exp[\beta m^\mu(t) m^\mu(t)] \\ &+ h \sum_{(\alpha,\kappa)\in\mathcal{E}} \exp[\beta m^\alpha(t) m^\kappa(t)]).\end{aligned} \tag{3}$$

As done for Equation 2, a similar analysis for Equation 3 is possible, but is complicated by the directed edges, e.g., consider the difference between an $\overrightarrow{\mathcal{M}}$ where all but one vertex μ point their edges to μ and an $\overrightarrow{\mathcal{M}}$ in which each vertex has equal in- and out-degree. Relatedly, I conjecture when $\overrightarrow{\mathcal{M}}$ is an Erdös-Renyi graph (a random graph constructed by allowing any edge with probability y), the critical value of a which marks the transition between neutral quiescence and wide hetero-association will be proportional to y when $y > \frac{(1-\varepsilon)\ln n}{n}$, i.e., when $\overrightarrow{\mathcal{M}}$ is asymptotically connected.

Of natural interest is when $h \neq 0$, which provides interactions between the patterns. What is interesting about Equation 1 is the possibility of both auto- and hetero-associative terms affecting the dynamics when both $a, h \neq 0$. As implied informally above, this means the model cannot per-

[4] From a practical standpoint, however, a non-traditional variant of auto-association wherein we care less about the absolute overlap between the pattern and the state and instead about which pattern has the highest overlap is possible, and something I explore in Appendix A.3.

[5] Note this does not imply there is no activity in the network, since we allow neurons to take negative values.

form *pure pattern retrieval*, i.e., retrieval of a single memory pattern ξ^μ without at least partial retrieval of other patterns. To show this, it is useful to refer to the overlap between a pattern ξ^μ and a state $\sigma(t)$. For this, we can use the Pearson product-moment correlation coefficient, which for pattern μ at time t I denote $r(\mu^{(t)})$.

Limited pure pattern retrieval. Hebbian auto- and hetero-associative mixtures cannot perform pure pattern retrieval for patterns not isolated in the memory graph. Suppose μ is not an isolated vertex in $\mathcal{M}$. Let $a, h > 0$. Then the model cannot perform pure pattern retrieval of ξ^μ. To see this, let $\{\xi^\mu, \xi^v\} \in \mathcal{E}$ if $\mathcal{M}$ is undirected, and let $(\xi^\mu, \xi^v) \in \mathcal{E}$ if $\overrightarrow{\mathcal{M}}$ is directed. Setting $\sigma(t) = \xi^\mu$ will cause the second term of Q to be non-negative because $h > 0$, and therefore $r(v^{(t+1)})$ will be proportionally large. Simultaneously, $r(\mu^{(t+1)})$ will be non-vanishing, since $a > 0$. Therefore, no non-isolated pattern can be purely retrieved.

Pure pattern retrieval. Pure pattern retrieval is possible for some memory graphs when the dynamics are Hebbian auto-associative or Hebbian hetero-associative, but not both. In particular, if:

- $a > 0$ and $h = 0$; or if
- $a = 0$, $h > 0$, the out-degree of all vertices in $\mathcal{M}$ is 1, and we have a sufficient β and η,

then the model can perform pure pattern retrieval of some memory patterns. The excitatory auto-associative result, where $a > 0$ and $h = 0$, is simply a weighted version of Theorems 1–3 from Ramsauer et al. (2021). The excitatory hetero-associative result, where $a = 0$ and $h > 0$, is indicated by the limited pure retrieval case above, with the added restriction that there exists only one memory pattern, ξ^v, projecting from ξ^μ in $\mathcal{M}$. This restriction is necessary because if the out-degree of ξ^μ was 0, then after setting $\sigma(t) = \xi^\mu$, the projection matrix Q would be filled with zeros since $a = 0$. Therefore, values of σ would converge to a value of $-\tilde{\xi}$. If the out-degree of ξ^μ was > 1, we would have a limited pure retrieval case, only with multiple memory patterns with large r values (with their strengths proportional to the weights of their respective in-edges from ξ^μ in $\mathcal{M}$). Finally, we need to achieve $\sigma(t+1) = \xi^v$ (or something arbitrarily close) to have pure pattern retrieval of ξ^v, since $a = 0$ means we will not have the luxury of additional time-steps to achieve convergence. Fortunately, by Theorem 4 of Ramsauer et al. (2021), we can get arbitrarily close by requiring sufficiently large values of β and η to update the state to ξ^v in a single step.

This naturally comports with Theorems 2.1 and 2.2 of Löwe (1998), wherein the classical associative memory model with binary-valued memories is studied when $\mathcal{M}$ is a 1D Markov chain[6]. There, Löwe (1998) showed that sequence capacity increases given large semantic correlations.

Retrieving connected components. Connected components in an undirected memory graph are retrieved in some Hebbian hetero-associative regimes. Let $\mathcal{Y} \subset \mathcal{M}$ be a connected component of $\mathcal{M}$. Set $h > a \geq 0$. Then setting $\sigma(t) = \xi^\mu$, where μ is a vertex in $\mathcal{Y}$, will cause $r(v^{(t+\lambda)})$ for all vertices v in $\mathcal{Y}$ to be non-vanishing, for some finite number of time-steps λ and thereafter for all time-steps. To see this, set $\sigma(t) = \xi^\mu$. If $a = 0$, then $r(v^{(t+1)})$ will be non-vanishing for all vertices v in $\mathcal{Y}$ which are adjacent to μ in $\mathcal{Y}$. Similarly, the vertices adjacent to v will have non-vanishing overlap at time $t+2$, and so on. If $h > a > 0$, the same argument applies.

Although I do not study this here, Appendix A.5 discusses biological mechanisms by which to learn M.

3. Numerical simulations

Now I investigate a wider collection of memory patterns and graphs, starting with a simple 1D cycle and gradually increasing complexity. Along the way, there are primarily two inter-weaving stories:

1. Anti-Hebbian auto-association increases the relative contribution of Hebbian hetero-association, which provides control over the range of hetero-association, extraction of multi-scale community structures in memory graphs, and stabilisation of temporal sequence recall; and

2. The flexibility of CDAM and its underlying graphical structure enables modelling a variety of phenomena, including graph community detection, sequence memory recall, and simulating finite automata.

Unless otherwise stated, in the following numerical analyses I used $n = 1,000$, $\beta = 1$, and $\eta = 0.1$. Simulations were run until convergence, at which point I measured the Pearson product-moment correlation coefficient between each memory μ and the state σ. To initialise the network state, I chose a memory pattern μ and set $\sigma(0) = \xi^\mu + c\zeta$, where ζ is a random vector with elements independently drawn from the interval $[-0.5, 0.5]$ and $c \in \mathbb{R}^+$ is the amplitude of the additive random noise, here $c = 1$.

[6] Löwe (1998) also studied the case of spatial correlations between neurons, as may arise in naturalistic data. This work has been recently continued for dense associative memory by De Marzo & Iannelli (2023).

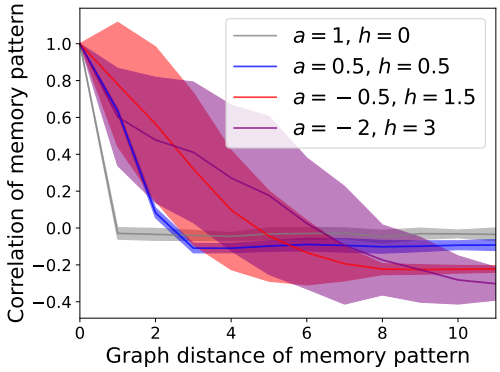

Figure 1. Mean correlations (± S.D.) of memory patterns within 10-hop neighbourhoods of the triggered memory pattern's vertex in $\mathcal{M} = \mathcal{C}_{30}$. The k-hop neighbourhood is the set of vertices within a distance of k edges from the triggered memory pattern. For each condition, all vertices ($n = 30$) are tested.

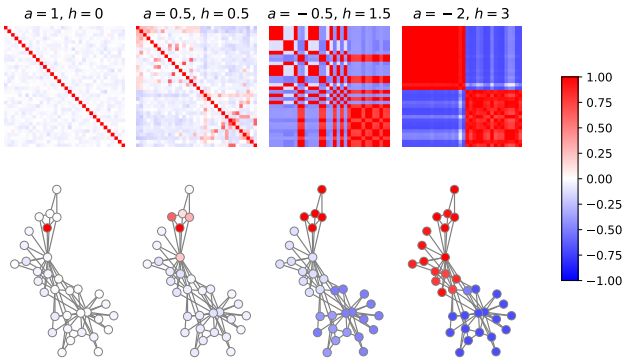

Figure 2. Correlations between the convergent (meta-)stable states ($\sigma(101)$ values from Figure 17) for all pairs of trigger stimuli (top row); and $\mathcal{M}$ drawn with vertices coloured by these (meta-)stable state correlations for a particular trigger stimulus (bottom row).

3.1. Controlling the range of recalled correlated memories

Modulating the balance of auto- and hetero-association using a and h allows us to control the range of memory retrieval in $\mathcal{M}$. To demonstrate this, I use an undirected cycle graph. A *cycle graph* $\mathcal{C}_n$ has n vertices connected by a single cycle of edges through all vertices. As described and illustrated in Appendix A.6, cycle graphs are the most commonly studied semantic hetero-associative memory structure previously studied, most likely due to it being a fitting representation of temporal sequences. In Appendix A.7, I show choices of a and h which achieve good fits ($R^2 = 0.997$) with the experimental data reported in Miyashita (1988).

Figure 1 measures the range of the spread across values of a and h, with significant differences observed between the tested conditions (one-way ANOVA, $F = 5.41$, *p-value* $= 0.001$); the range of recalled memories in terms of graph distance is controllable within the range of 0 to ~ 6. In Appendix A.8, I show the correlation matrices for all patterns.

3.2. Multi-scale representations of community structures in graphs

Now I will consider more interesting memory graph topologies. *Zachary's karate club graph* (Zachary, 1977) consists of 34 vertices, representing karate practitioners, where edges connect individuals who consistently interacted in extra-karate contexts. Notably, the club split into two halves. Setting Zachary's karate club graph as $\mathcal{M}$ and varying a and h, however, reveals that there were even finer social groupings than these, as Figure 2 reveals and as I discuss in Appendix A.9.

To more clearly illustrate the multi-scale representations of graph communities, I also test CDAM on the *barbell graph* (see Appendix A.10 for further details) and the *Tutte graph* (see Figure 3 and Appendix A.4).

3.3. Sparse temporal sequence recall of real video data

Hetero-association is naturally suited for encoding temporal sequences. Here I use a directed cycle graph $\overrightarrow{\mathcal{C}_{50}}$ where the patterns are sparsely sampled frames of videos (see Appendix A.11 for details).

Figure 4 shows activity over time in a network with $\mathcal{M} = \overrightarrow{\mathcal{C}_{50}}$. At each step t of the simulation, I calculate the correlation of $\sigma(t)$ with each pattern. I start the simulation by triggering the first pattern (frame) and thereafter leave the network to continue its dynamics according to Equation 1. Importantly, sufficient anti-Hebbian auto-association, i.e., $a < 0$, is required, in combination with relatively strong Hebbian hetero-associations, i.e., $h > 0$. Otherwise, the sequence recall can become stuck or lags due to auto-correlations between statistically-similar frames.

Notably, similar settings for anti-Hebbian auto-association and Hebbian hetero-association is required for a different sparsely sampled video, as shown in Figure 5. Only in the case of $a = -2, h = 3$ can we recall the sequence without skips or delays. Present in both Figures 4 and 5, we can see more global features in the video and sharp context switches. These structures can be also be seen in the correlations between the attractors (see Appendix A.11).

3.4. Simulation of arbitrary finite automata

CDAM is also capable of simulating arbitrary finite automata. To demonstrate this, I use the example of a family tree (as illustrated in Figure 6) composed of image and text data, which is capable of basic 'question-answering'.

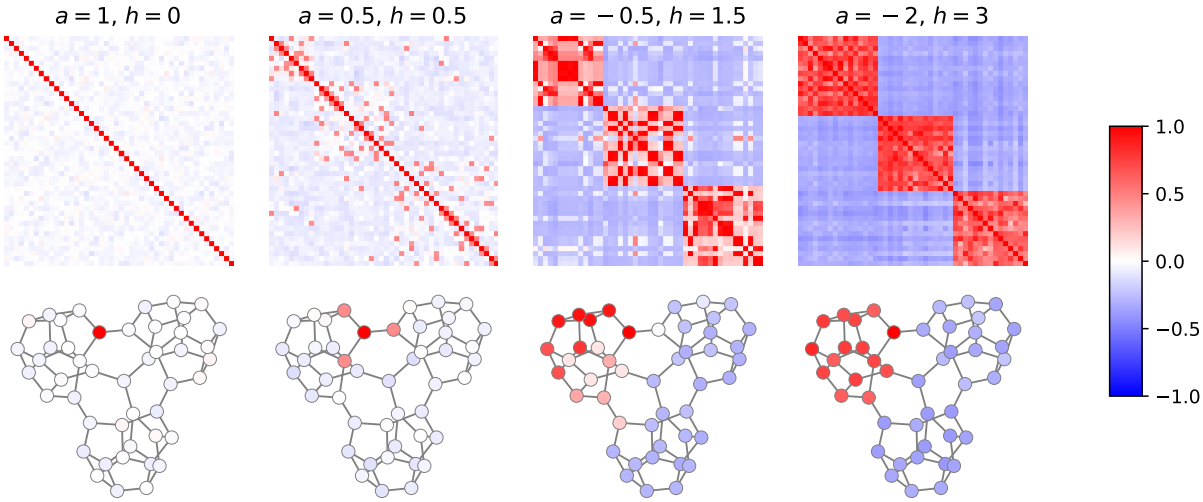

Figure 3. Correlations between the convergent (meta-)stable states ($\sigma(101)$ values from Figure 16) for all pairs of trigger stimuli (top row); and $\mathcal{M}$ drawn with vertices coloured by these (meta-)stable state correlations for a particular trigger stimulus (bottom row).

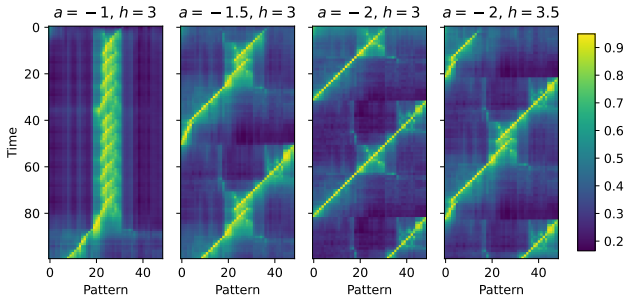

Figure 4. Correlations of memory patterns over time for each vertex in $\mathcal{M} = \overrightarrow{\mathcal{C}_{50}}$, where each memory pattern is a sparsely sampled video frame (see Appendix A.11 for details) from video 1.

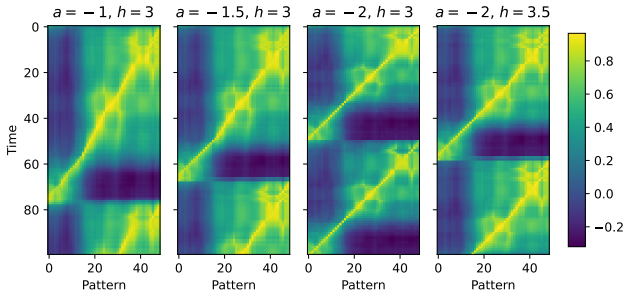

Figure 5. Correlations of memory patterns over time for each vertex in $\mathcal{M} = \overrightarrow{\mathcal{C}_{50}}$, where each memory pattern is a sparsely sampled video frame (see Appendix A.11 for details) from video 2.

Each individual (state) and each directed relation (transition) forms its own memory pattern. Importantly, the memory patterns are structured in the following way. In all memory patterns, 75% of the neurons are always reserved for representing individuals (using a fixed 75% random sampling of their image) and the remaining memory content is allocated to either the individuals (using the remaining 25% for their image) or a transition label (e.g., 'father', 'sister', etc.) consisting of text embeddings of these strings (Pennington et al., 2014). This generates memory patterns which consist, semantically, of 'Bart' and 'Bart+sister', but not 'Bart+brother', since there is no transition in Figure 6 stemming from the Bart character which has its transition labelled as 'brother'. The explicit structure of $\mathcal{M}$ is drawn in Figure 30 and further detailed in Appendix A.12.

As illustrated in Figure 7, using $\mathcal{M}$ we may perform a basic 'question-answering function' by starting at a memory pattern associated with an individual and then stimulating the neurons reserved for transitions. For example, if we set the state of the network to 'Marge' and wish to ask 'Who is Marge's husband?', we stimulate the transition neurons with the text embedding for 'husband'. This gives the state 'Marge+husband', which hetero-associates to 'Homer', our answer. If we then stimulate 'brother', however, the finite automaton will not transition. This is because no transition memory pattern of 'Homer+brother' exists, and since 75% of the neurons are reserved for representing the individual ('Homer', in this case), the network auto-associates back to 'Homer' due to the overall neural activity remaining close to this memory pattern. In Figure 31, I show simulations starting at all possible attractor states (individuals) and quasi-attractor states (directed relations), demonstrating precise recall of the entire finite automaton structure.

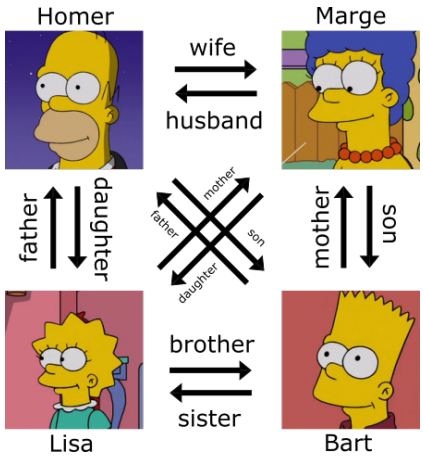

Figure 6. Family tree with labelled relationships.

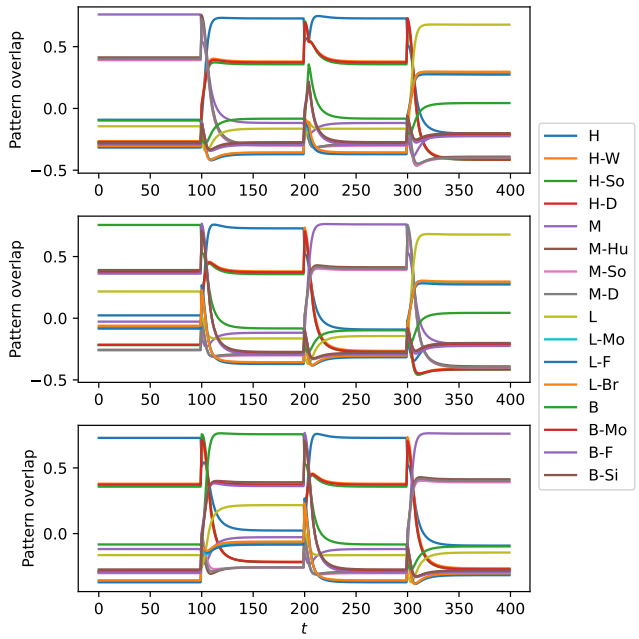

Figure 7. Pattern overlaps during simulations ($a = 0, h = 1$) of finite automaton based on Figure 6. First row inputs: 'M', 'Hu', 'Br', 'D'. Second row inputs: 'B', 'F', 'W', 'D'. Third row inputs: 'H', 'So', 'F', 'W'. Key: H='Homer', M='Marge', L='Lisa', B='Bart', W='Wife', Hu='Husband', So='Son', D='Daughter', Br='Brother', Si='Sister', Mo='Mother', F='Father'.

4. Discussion

In this paper I have introduced a new dense associative memory model, called *Correlated Dense Associative Memory (CDAM)*, which auto- and hetero-associates continuous-valued memory patterns using an underlying (arbitrary) graph structure. Using such memory graph structures, and especially by modulating recall using anti-Hebbian auto- or hetero-association, I demonstrated extraction of multiple scales of representation of the community structures present in the underlying graphs, as well as replication of a classic neuroscience experiment. I additionally tested CDAM with perhaps the most traditional and obvious application of hetero-associative memory networks – temporal sequence memory – with sparsely sampled real-world videos. Here, the benefits of anti-Hebbian modulations were highlighted once again, this time in its role as a stabiliser against internal correlations (natural distractors) within a sequence and of ordered recall generally. Finally, I demonstrate the ability of CDAM to simulate finite automata, hinting at its general computational power (beyond 'counting chimes' (Amit, 1988)). In an associationist framework, this also provides potential mechanistic insight for attention computations in Transformers, following the theme of Cabannes et al. (2024).

4.1. Future work

For neuroscience, this work highlights the highly non-trivial contributions of anti-Hebbian learning to proper functioning across a range of tasks, including controlling sequence recall or association ranges, multi-scale representation of correlated attractors, and temporal sequence retrieval. These findings invite experimentalists to further explore the contribution of inhibitory neurons in cognition.

For machine learning, perhaps one of the most impactful uses of this work will be in its application to improving the performance and/or understanding of Transformer models (Vaswani et al., 2017) through their connection to continuously-valued dense associative memory networks (Krotov & Hopfield, 2016; Ramsauer et al., 2021). Indeed, Ramsauer et al. (2021) used this connection to study the 'attractive schemas' of the implied energy landscape in a large language model. This generated hypotheses about the function of particular layers and attention heads in the model, and may potentially help us further elucidate the internal representational structure of similar models. As Millidge et al. (2022) note, Transformers' attention mechanism can be interpreted in its mathematical form as performing hetero-association between its keys and and values in the associative memory sense. Can we use these insights to identify the topology of the attractor or energy landscape? Do such models entrain particular structures such as memory graphs (or higher dimensional analogues) to reflect the topology of the underlying data structures and correlations within the training set? Could modulatory mechanisms such as anti-Hebbian learning help direct the 'flow' of temporally-evolving cognition, such as in-context or one-shot learning in large language models (Brown et al., 2020)? I expand on these and other questions in Appendix A.2, which I hope will lead to fruitful interpretability efforts and broader interaction between neuroscientists and machine learning researchers.

Impact statement

This paper presents work whose primary goals are to advance the fields of machine learning and theoretical neuroscience. There are many potential societal consequences of this work, particularly in its implications for the development and interpretability of Transformer architectures. Such architectures are widely used today by companies, academics, open source technical communities, and private individuals. I believe this work helps contribute to our theoretical understanding of such architectures, which has dual-uses – it may be used to make more powerful systems, and it may be used to better understand or control such systems.

Reproducibility statement

All numerical simulations were performed on a Lenovo x260 ThinkPad laptop computer using the Python 3 programming language. A copy of the code used is available at https://github.com/tfburns/CDAM.

Acknowledgements

Thank you to anonymous reviewers for their constructive feedback. Thank you to Dima Krotov, Mengsen Zhang, Horacio Rotstein, Vasiliki Liontou, Robert Tang, Dan Murfet, Adam Shai, Dia Taha, Christopher Earls, Tomoki Fukai, Tatsuya Haga, participants in the 'Math + Neuroscience' program at the Institute for Computational and Experimental Research in Mathematics, participants in the 'Physics of Life' symposium in November 2023 at the Center for the Physics of Biological Function, and many others for thoughtful comments and discussions.

This material is based on work supported by the National Science Foundation of the United States under Grant Number DMS-1929284 while in residence at the Institute for Computational and Experimental Research in Mathematics in Providence, Rhode Island, during the 'Math + Neuroscience: Strengthening the Interplay Between Theory and Mathematics' program.

References

Agrawal, P., Girshick, R., and Malik, J. Analyzing the performance of multilayer neural networks for object recognition. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII 13*, pp. 329–344. Springer, 2014.

Amari, S.-I. Learning patterns and pattern sequences by self-organizing nets of threshold elements. *IEEE Transactions on Computers*, C-21(11):1197–1206, 1972. doi: 10.1109/T-C.1972.223477.

Amit, D. J. Neural networks counting chimes. *Proceedings of the National Academy of Sciences*, 85(7):2141–2145, 1988.

Amit, D. J., Gutfreund, H., and Sompolinsky, H. Storing infinite numbers of patterns in a spin-glass model of neural networks. *Phys. Rev. Lett.*, 55: 1530–1533, Sep 1985. doi: 10.1103/PhysRevLett.55.1530. URL https://link.aps.org/doi/10.1103/PhysRevLett.55.1530.

Ardakani, A., Ardakani, A., and Gross, W. Training linear finite-state machines. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 7173–7183. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4fc28b7093b135c21c7183ac07e928a6-Paper.pdf.

Balle, B. and Maillard, O.-A. Spectral learning from a single trajectory under finite-state policies. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 361–370. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/balle17a.html.

Berkeley, I. Density plots of hidden value unit activations reveal interpretable bands. *Connection Science*, 7(2):167–187, 1995. doi: 10.1080/09540099550039336. URL https://doi.org/10.1080/09540099550039336.

Brito, C. S. and Gerstner, W. Nonlinear hebbian learning as a unifying principle in receptive field formation. *PLoS computational biology*, 12(9):e1005070, 2016.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Brush, S. G. History of the lenz-ising model. *Rev. Mod. Phys.*, 39:883–893, Oct 1967. doi: 10.1103/RevModPhys.39.883. URL https://link.aps.org/doi/10.1103/RevModPhys.39.883.

Burns, T. F. and Fukai, T. Simplicial hopfield networks. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=_QLsH8gatwx.

Burns, T. F., Haga, T., and Fukai, T. Multiscale and extended retrieval of associative memory structures in a cortical model of local-global inhibition balance. *eNeuro*, 9(3), 2022. doi: 10.1523/ENEURO.0023-22.2022. URL https://www.eneuro.org/content/9/3/ENEURO.0023-22.2022.

Buzsáki, G. Neural syntax: Cell assemblies, synapsembles, and readers. *Neuron*, 68(3):362–385, 2010. ISSN 0896-6273. doi: https://doi.org/10.1016/j.neuron.2010.09.023. URL https://www.sciencedirect.com/science/article/pii/S0896627310007658.

Cabannes, V., Dohmatob, E., and Bietti, A. Scaling laws for associative memories. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Tzh6xAJSll.

Chang, L. and Tsao, D. Y. The code for facial identity in the primate brain. *Cell*, 169(6):1013–1028.e14, 2017. ISSN 0092-8674. doi: https://doi.org/10.1016/j.cell.2017.05.011. URL https://www.sciencedirect.com/science/article/pii/S009286741730538X.

Chaudhry, H. T., Zavatone-Veth, J. A., Krotov, D., and Pehlevan, C. Long sequence hopfield memory, 2023.

Crick, F. and Mitchison, G. The function of dream sleep. *Nature*, 304(5922):111–114, 1983.

De Marzo, G. and Iannelli, G. Effect of spatial correlations on hopfield neural network and dense associative memories. *Physica A: Statistical Mechanics and its Applications*, 612:128487, 2023. ISSN 0378-4371. doi: https://doi.org/10.1016/j.physa.2023.128487. URL https://www.sciencedirect.com/science/article/pii/S0378437123000420.

de Ruyter van Steveninck, R. and Bialek, W. Reliability and statistical efficiency of a blowfly movement-sensitive neuron. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 348(1325):321–340, 1995.

Demircigil, M., Heusel, J., Löwe, M., Upgang, S., and Vermet, F. On a model of associative memory with huge storage capacity. *Journal of Statistical Physics*, 168(2):288–299, Jul 2017. ISSN 1572-9613. doi: 10.1007/s10955-017-1806-y. URL https://doi.org/10.1007/s10955-017-1806-y.

Diekelmann, S. and Born, J. The memory function of sleep. *Nature Reviews Neuroscience*, 11(2):114–126, 2010.

Ditz, H. M. and Nieder, A. Neurons selective to the number of visual items in the corvid songbird endbrain. *Proceedings of the National Academy of Sciences*, 112(25):7827–7832, 2015. doi: 10.1073/pnas.1504245112. URL https://www.pnas.org/doi/abs/10.1073/pnas.1504245112.

Donnelly, J. and Roegiest, A. On interpretability and feature representations: an analysis of the sentiment neuron. In *Advances in Information Retrieval: 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14–18, 2019, Proceedings, Part I 41*, pp. 795–802. Springer, 2019.

El-Gaby, M., Harris, A. L., Whittington, J. C. R., Dorrell, W., Bhomick, A., Walton, M. E., Akam, T., and Behrens, T. E. J. A cellular basis for mapping behavioural structure. *bioRxiv*, 2023. doi: 10.1101/2023.11.04.565609. URL https://www.biorxiv.org/content/early/2023/11/05/2023.11.04.565609.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

Folk, R. and Holovatch, Y. Schottky's forgotten step to the ising model. *The European Physical Journal H*, 47(1):9, Sep 2022. ISSN 2102-6467. doi: 10.1140/epjh/s13129-022-00041-0. URL https://doi.org/10.1140/epjh/s13129-022-00041-0.

Gillett, M., Pereira, U., and Brunel, N. Characteristics of sequential activity in networks with temporally asymmetric hebbian learning. *Proceedings of the National Academy of Sciences*, 117(47):29948–29958, 2020. doi: 10.1073/pnas.1918674117. URL https://www.pnas.org/doi/abs/10.1073/pnas.1918674117.

Griniasty, M., Tsodyks, M. V., and Amit, D. J. Conversion of Temporal Correlations Between Stimuli to Spatial Correlations Between Attractors. *Neural Computation*, 5(1):1–17, 01 1993. ISSN 0899-7667. doi: 10.1162/neco.1993.5.1.1. URL https://doi.org/10.1162/neco.1993.5.1.1.

Gross, C. G. Genealogy of the "grandmother cell". *The Neuroscientist*, 8(5):512–518, 2002.

Gutfreund, H. and Mezard, M. Processing of temporal sequences in neural networks. *Phys. Rev. Lett.*, 61:235–238, Jul 1988. doi: 10.1103/PhysRevLett.61.235. URL https://link.aps.org/doi/10.1103/PhysRevLett.61.235.

Haga, T. and Fukai, T. Extended temporal association memory by modulations of inhibitory circuits. *Physical review letters*, 123(7):078101, 2019.

Haga, T. and Fukai, T. Multiscale representations of community structures in attractor neural networks. *PLOS Computational Biology*, 17(8):1–26, 08 2021. doi:

10.1371/journal.pcbi.1009296. URL https://doi.org/10.1371/journal.pcbi.1009296.

Hayama, T., Noguchi, J., Watanabe, S., Takahashi, N., Hayashi-Takagi, A., Ellis-Davies, G. C., Matsuzaki, M., and Kasai, H. Gaba promotes the competitive selection of dendritic spines by controlling local ca2+ signaling. *Nature neuroscience*, 16(10):1409–1416, 2013.

Hertäg, L. and Sprekeler, H. Amplifying the redistribution of somato-dendritic inhibition by the interplay of three interneuron types. *PLOS Computational Biology*, 15(5):1–29, 05 2019. doi: 10.1371/journal.pcbi.1006999. URL https://doi.org/10.1371/journal.pcbi.1006999.

Honey, C. J., Newman, E. L., and Schapiro, A. C. Switching between internal and external modes: A multiscale learning principle. *Network Neuroscience*, 1(4):339–356, 12 2017. ISSN 2472-1751. doi: 10.1162/NETN_a_00024. URL https://doi.org/10.1162/NETN_a_00024.

Hopfield, J. J. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982. doi: 10.1073/pnas.79.8.2554. URL https://www.pnas.org/doi/abs/10.1073/pnas.79.8.2554.

Hopfield, J. J., Feinstein, D. I., and Palmer, R. G. 'unlearning'has a stabilizing effect in collective memories. *Nature*, 304(5922):158–159, 1983.

Kanter, I. and Sompolinsky, H. Associative recall of memory without errors. *Phys. Rev. A*, 35:380–392, Jan 1987. doi: 10.1103/PhysRevA.35.380. URL https://link.aps.org/doi/10.1103/PhysRevA.35.380.

Karuvally, A., Sejnowski, T., and Siegelmann, H. T. General sequential episodic memory model. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 15900–15910. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/karuvally23a.html.

King, P. D., Zylberberg, J., and DeWeese, M. R. Inhibitory interneurons decorrelate excitatory cells to drive sparse code formation in a spiking model of v1. *Journal of Neuroscience*, 33(13):5475–5485, 2013. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.4188-12.2013. URL https://www.jneurosci.org/content/33/13/5475.

Kleinfeld, D. and Sompolinsky, H. Associative neural network model for the generation of temporal patterns. theory and application to central pattern generators. *Biophys J*, 54(6):1039–1051, December 1988.

Krotov, D. and Hopfield, J. J. Dense associative memory for pattern recognition. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 1180–1188, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

Kutter, E. F., Bostroem, J., Elger, C. E., Mormann, F., and Nieder, A. Single neurons in the human brain encode numbers. *Neuron*, 100(3):753–761, 2018.

Le, Q. V., Ranzato, M., Monga, R., Devin, M., Chen, K., Corrado, G. S., Dean, J., and Ng, A. Y. Building high-level features using large scale unsupervised learning. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, pp. 507–514, Madison, WI, USA, 2012. Omnipress. ISBN 9781450312851.

Lin, L., Chen, G., Kuang, H., Wang, D., and Tsien, J. Z. Neural encoding of the concept of nest in the mouse brain. *Proceedings of the National Academy of Sciences*, 104 (14):6066–6071, 2007.

Little, W. The existence of persistent states in the brain. *Mathematical Biosciences*, 19 (1):101–120, 1974. ISSN 0025-5564. doi: https://doi.org/10.1016/0025-5564(74)90031-5. URL https://www.sciencedirect.com/science/article/pii/0025556474900315.

Liu, B., Ash, J. T., Goel, S., Krishnamurthy, A., and Zhang, C. Transformers learn shortcuts to automata. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=De4FYqjFueZ.

Löwe, M. and Vermet, F. The Hopfield model on a sparse Erdös-Renyi graph. *Journal of Statistical Physics*, 143 (1):205–214, Apr 2011. ISSN 1572-9613. doi: 10.1007/s10955-011-0167-1. URL https://doi.org/10.1007/s10955-011-0167-1.

Lucibello, C. and Mézard, M. The exponential capacity of dense associative memories, 2023.

Löwe, M. On the storage capacity of Hopfield models with correlated patterns. *The Annals of Applied Probability*, 8(4):1216 – 1250, 1998. doi: 10.1214/aoap/1028903378. URL https://doi.org/10.1214/aoap/1028903378.

Marr, D. Simple memory: a theory for archicortex. *Philos Trans R Soc Lond B Biol Sci*, 262(841):23–81, July 1971.

Millidge, B., Salvatori, T., Song, Y., Lukasiewicz, T., and Bogacz, R. Universal hopfield networks: A general framework for single-shot associative memory models. In *International Conference on Machine Learning*, pp. 15561–15583. PMLR, 2022.

Miyashita, Y. Neuronal correlate of visual associative long-term memory in the primate temporal cortex. *Nature*, 335 (6193):817–820, 1988.

Müller, M. G., Papadimitriou, C. H., Maass, W., and Legenstein, R. A model for structured information representation in neural networks of the brain. *eNeuro*, 7(3): ENEURO.0533–19.2020, May 2020.

Nakano, K. Associatron-a model of associative memory. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-2(3):380–388, 1972. doi: 10.1109/TSMC.1972.4309133.

Nashef, A., Cohen, O., Perlmutter, S. I., and Prut, Y. A cerebellar origin of feedforward inhibition to the motor cortex in non-human primates. *Cell Reports*, 39(6), 2022.

Nieder, A. Supramodal numerosity selectivity of neurons in primate prefrontal and posterior parietal cortices. *Proceedings of the National Academy of Sciences*, 109(29):11860–11865, 2012. doi: 10.1073/pnas.1204580109. URL https://www.pnas.org/doi/abs/10.1073/pnas.1204580109.

Nieder, A., Freedman, D. J., and Miller, E. K. Representation of the quantity of visual items in the primate prefrontal cortex. *Science*, 297(5587):1708–1711, 2002.

Paille, V., Fino, E., Du, K., Morera-Herreras, T., Perez, S., Kotaleski, J. H., and Venance, L. Gabaergic circuits control spike-timing-dependent plasticity. *Journal of Neuroscience*, 33(22):9353–9363, 2013.

Papadimitriou, C. H., Vempala, S. S., Mitropolsky, D., Collins, M., and Maass, W. Brain computation by assemblies of neurons. *Proceedings of the National Academy of Sciences*, 117(25):14464–14472, 2020. doi: 10.1073/pnas.2001893117. URL https://www.pnas.org/doi/abs/10.1073/pnas.2001893117.

Pastur, L. and Figotin, A. Rigorously solvable model of spin glass. *JETP Lett*, 25(8), 1977.

Pennington, J., Socher, R., and Manning, C. D. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014. URL http://www.aclweb.org/anthology/D14-1162.

Poe, G. R. Sleep is for forgetting. *Journal of Neuroscience*, 37(3):464–473, 2017.

Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., and Fried, I. Invariant visual representation by single neurons in the human brain. *Nature*, 435(7045):1102–1107, 2005.

Quiroga, R. Q., Kreiman, G., Koch, C., and Fried, I. Sparse but not 'grandmother-cell' coding in the medial temporal lobe. *Trends in Cognitive Sciences*, 12(3):87–91, 2008. ISSN 1364-6613. doi: https://doi.org/10.1016/j.tics.2007.12.003. URL https://www.sciencedirect.com/science/article/pii/S1364661308000235.

Ramsauer, H., Schäfl, B., Lehner, J., Seidl, P., Widrich, M., Gruber, L., Holzleitner, M., Adler, T., Kreil, D., Kopp, M. K., Klambauer, G., Brandstetter, J., and Hochreiter, S. Hopfield networks is all you need. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=tL89RnzIiCd.

Roberts, P. and Leen, T. Anti-hebbian spike-timing-dependent plasticity and adaptive sensory processing. *Frontiers in Computational Neuroscience*, 4, 2010. ISSN 1662-5188. doi: 10.3389/fncom.2010.00156. URL https://www.frontiersin.org/articles/10.3389/fncom.2010.00156.

Schapiro, A. C., Rogers, T. T., Cordova, N. I., Turk-Browne, N. B., and Botvinick, M. M. Neural representations of events arise from temporal community structure. *Nature Neuroscience*, 16(4):486–492, Apr 2013. ISSN 1546-1726. doi: 10.1038/nn.3331. URL https://doi.org/10.1038/nn.3331.

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

Sherrington, D. and Kirkpatrick, S. Solvable model of a spin-glass. *Phys. Rev. Lett.*, 35:1792–1796, Dec 1975. doi: 10.1103/PhysRevLett.35.1792. URL https://link.aps.org/doi/10.1103/PhysRevLett.35.1792.

Shulz, D. and Feldman, D. Spike timing-dependent plasticity. *Rubenstein JLR, Rakic P. Comprehensive developmental neuroscience: neural circuit development and function in the healthy and diseased brain. San Diego, CA: Elsevier*, pp. 155–81, 2013.

Stanley, J. C. Simulation studies of a temporal sequence memory model. *Biological Cybernetics*, 24(3):121–137, Sep 1976. ISSN 1432-0770. doi: 10.1007/BF00364115. URL https://doi.org/10.1007/BF00364115.

Stent, G. S., Kristan, W. B., Friesen, W. O., Ort, C. A., Poon, M., and Calabrese, R. L. Neuronal generation of the leech swimming movement. *Science*, 200(4348):1348–1357, 1978. doi: 10.1126/science.663615. URL https://www.science.org/doi/abs/10.1126/science.663615.

Tank, D. W. and Hopfield, J. J. Neural computation by concentrating information in time. *Proceedings of the National Academy of Sciences*, 84(7):1896–1900, 1987. doi: 10.1073/pnas.84.7.1896. URL https://www.pnas.org/doi/abs/10.1073/pnas.84.7.1896.

Tobin, M., Sheth, J., Wood, K. C., and Geffen, M. N. Localist versus distributed representation of sounds in the auditory cortex controlled by distinct inhibitory neuronal subtypes. *bioRxiv*, 2023. doi: 10.1101/2023.02.01.526470. URL https://www.biorxiv.org/content/early/2023/06/22/2023.02.01.526470.

Treves, A. and Amit, D. J. Metastable states in asymmetrically diluted Hopfield networks. *Journal of Physics A: Mathematical and General*, 21(14):3155–3169, jul 1988. doi: 10.1088/0305-4470/21/14/016. URL https://doi.org/10.1088/0305-4470/21/14/016.

Tutte, W. T. On hamiltonian circuits. *Journal of the London Mathematical Society*, s1-21(2):98–101, 1946. doi: https://doi.org/10.1112/jlms/s1-21.2.98. URL https://londmathsoc.onlinelibrary.wiley.com/doi/abs/10.1112/jlms/s1-21.2.98.

Tyulmankov, D., Fang, C., Vadaparty, A., and Yang, G. R. Biological learning in key-value memory networks. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 22247–22258. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/bacadc62d6e67d7897cef027fa2d416c-Paper.pdf.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Whittington, J. C., Muller, T. H., Mark, S., Chen, G., Barry, C., Burgess, N., and Behrens, T. E. The tolman-eichenbaum machine: unifying space and relational memory through generalization in the hippocampal formation. *Cell*, 183(5):1249–1263, 2020.

Whittington, J. C. R., Warren, J., and Behrens, T. E. Relating transformers to models and neural representations of the hippocampal formation. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=B8DVo9B1YE0.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

Zachary, W. W. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33(4):452–473, 1977. doi: 10.1086/jar.33.4.3629752. URL https://doi.org/10.1086/jar.33.4.3629752.

Zeiler, M. D. and Fergus, R. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pp. 818–833. Springer, 2014.

Zhou, Y., Lai, C. S. W., Bai, Y., Li, W., Zhao, R., Yang, G., Frank, M. G., and Gan, W.-B. Rem sleep promotes experience-dependent dendritic spine elimination in the mouse cortex. *Nature communications*, 11(1):4819, 2020.

A. Appendix

A.1. Connection between associative memory and Transformers

A network of n neurons is modelled by n spins. In the binary neuron case, the spin of neuron j at time-step t is $\sigma_j^{(t)} = \pm 1$. The configuration of spins across all neurons correspond to patterns of neural firing, which are determined dynamically by neurons becoming active in response to signals received from other neurons they are connected to. We can use an abstract simplicial complex to model these connections. An abstract simplicial complex Δ is a collection of finite sets closed under taking subsets, i.e.:

Definition A.1. Let Δ be a subset of $2^{[n]}$. The subset Δ is an abstract simplicial complex if for any $\delta \in \Delta$, the condition $\rho \subseteq \delta$ gives $\rho \in \Delta$, for any $\rho \subseteq \delta$.

A member of Δ is called a *simplex* δ. We call a simplicial complex Δ a k*–skeleton* when all possible faces of dimension k exist and $\dim(\Delta) = k$. Let Δ be a simplicial complex on n vertices. Given a set of neurons δ (which is a unique $(|\delta| - 1)$–simplex in Δ), $w(\delta)$ is the associated simplicial weight and σ_δ the product of their spins. The traditional associative memory model is defined by the energy function

$$\mathfrak{E} = -\sum_{\delta \in \Delta} w(\delta)\sigma_\delta, \qquad \text{where} \qquad w(\delta) = \frac{1}{n}\sum_{\mu=1}^{p} \xi_\delta^\mu, \quad (4)$$

with ξ_i^μ $(= \pm 1)$ static variables being the p binary memory pattern vectors stored, and where ξ_δ^μ is the product of the δ-indexed components in ξ^μ.

Traditional associative memory models pairwise interactions between neurons, i.e., Δ is a 1-skeleton on n neurons. Let $i \in \delta$. The difference equation which governs a spin update for neuron i is

$$\sigma_i^{(t+1)} = \sum_{\delta \in \Delta} w(\delta)\sigma_{\delta \setminus i}^{(t)}$$

which we could also write as

$$\sum_{j=1}^{n} w_{ij}\sigma_j^{(t)}, \quad \text{where} \quad w_{ij} = \frac{1}{n}\sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu.$$

If we set $\sigma = \xi^\mu$, the dynamics will be stationary so long as p is finite and the other memories aren't distributed by a demon. And, under certain conditions, if we set σ as a noise-corrupted version of ξ^μ, we may recover the uncorrupted ξ^μ by applying the above dynamics (Hopfield, 1982). In fact, in the thermodynamic limit of $n \to \infty$, this behaviour is guaranteed with up to $\sim 0.138n$ independent memory pattern vectors (Amit et al., 1985). The model property which induces this capacity limitation is the order of spin interactions. Indeed, higher-order interactions increases the memory capacity of the network, however it is linear in the number of weighted simplices for k–skeleton models (Burns & Fukai, 2023).

Relatedly, we may re-write the energy function (Krotov & Hopfield, 2016) and consider an exponential order of multispin interactions (the 'limit' case) (Demircigil et al., 2017; Ramsauer et al., 2021), where the energy and update functions are

$$\mathfrak{E} = -\sum_{\mu=1}^{p} F(\xi_\delta^\mu \sigma_\delta),$$

$$\sigma_i^{(t+1)} = \text{sgn}\left[\sum_{\mu=1}^{p}\left(F(1 \cdot \xi_i^\mu + \sum_{j=1}^{n} \xi_j^\mu \sigma_j^{(t)}) - F(-1 \cdot \xi_i^\mu + \sum_{j=1}^{n} \xi_j^\mu \sigma_j^{(t)})\right)\right]$$

where the function F can be chosen, for example, to be of a polynomial $F(x) = x^n$ or exponential $F(x) = e^x$ form, which improves the memory capacity to n^{d-1} (where $d \geq 2$) and $2^{n/2}$, respectively (Demircigil et al., 2017).

This model can be readily extended to use continuous spin and pattern values, i.e., $\sigma_j, \xi_j^\mu \in \mathbb{R}$. It is then convenient to arrange the pattern vectors into a matrix $\Xi = (\xi^1, ..., \xi^p)$ and define the *log-sum-exp function* (lse) for $\beta > 0$ as

$$\text{lse}(\beta, x) = \frac{1}{\beta}\left(\sum_{\mu=1}^{p} \exp(\beta x_\mu)\right).$$

The energy function of this model can then be succinctly written as

$$\mathfrak{E} = -\text{lse}(\beta, \Xi^T \sigma) + \frac{1}{2}\sigma^T \sigma.$$

Using vector notation in the case of a 1-skeleton model, the update rule is

$$\sigma^{(t+1)} = \text{SOFTMAX}\left(\beta \Xi^T \sigma^{(t)}\right)\Xi. \quad (5)$$

Those familiar with Transformers (Vaswani et al., 2017) from the machine learning literature will recognise that Equation 5 is very close that of the attention mechanism. A difference between Equation 5 and the attention mechanism is that one of the Ξ variables should be replaced by a unique matrix. This connection was noticed by Ramsauer et al. (2021), and as Section 3.5 of Millidge et al. (2022) notes, associative memory models of this kind perform auto-association (σ dynamically moves towards the memory patterns stored in Ξ) whereas the attention mechanism of Transformers performs hetero-association (σ dynamically moves towards columns of a matrix Ξ' based on the corresponding closeness to columns in a separate matrix Ξ). However, as I note in the current work, the Transformer

creates Ξ and Ξ' from distinct weight matrices and there is nothing preventing those weight matrices from being partially or even entirely equivalent. The consequence of this is that Transformers, while their mathematical form is of a hetero-associative nature, can in fact implement any mixture of auto- and hetero-association.

A.2. Analysing Transformers from an associative memory perspective

The primary goal of this work is to develop an understanding of some types of computations which are possible with general mixtures of auto- and hetero-association in associative memory networks. One motivation behind this goal is, through the connection to Transformers (see Appendix A.1), to help interpret the inner-workings of Transformers. In this section, I suggest some analyses motivated by this work and offer a hypothesis related to the 'superposition' phenomenon demonstrated in Elhage et al. (2022).

In the case illustrated in Figure 7, there exists fixed stable points (the states), and we may transition between these stable points by systematically perturbing the state towards a nearby 'quasi-attractor' which is set up to direct that 'edge transition state' towards the target state in the finite automaton. Combining auto- and hetero-associative elements in this way therefore further enriches the network's capabilities, allowing both content-addressable retrieval and directed input-output mapping. However, as seen in Figure 30, this relies on a particular structure in $\mathcal{M}$. I hypothesise that general structures such as these, if efficiently learnt in Transformers, will cause the network to exhibit 'superposition' and that this is influenced by symmetries within the function maps, à la Liu et al. (2023).

Suppose an individual neuron in the brain of a person responds exactly and only to the sensory stimuli which identify their grandmother. We may say such a 'grandmother cell' (Gross, 2002) is *monosemantic*. Although there exist biological (de Ruyter van Steveninck & Bialek, 1995; Quiroga et al., 2005; Lin et al., 2007) and artificial (Berkeley, 1995; Le et al., 2012; Zeiler & Fergus, 2014) neurons with very high stimulus specificity, empirically it appears many neurons in commonly-studied neural networks are *polysemantic* (their activities correspond with multiple variables) (Quiroga et al., 2008; Agrawal et al., 2014; Chang & Tsao, 2017; Donnelly & Roegiest, 2019).

Elhage et al. (2022) studied small rectified linear unit (ReLU) auto-encoder networks tasked with faithfully encoding m independent features using n neurons, where $m > n$. They found neurons become systematically responsive to multiple features, depending on the relative importance and sparsity of the features, i.e., they become systematically polysemantic. What is particularly interesting in the Elhage et al. (2022) study is that these polysemantic neurons also arranged themselves in such a way that their shared feature axes reliably form specific geometric patterns given sparsity in the input data. Elhage et al. (2022) call this phenomenon *superposition* and note its close relation to ideas from neural coding in neuroscience.

In neuroscience, a notable biological detail of the neural activity corresponding to semantically-related objects (memories, sensations, etc.) is that they typically share many of the same physical substrates. For instance, two engrams of related memories might share some of the same neurons (the extent to which they share neurons in their distinct engrams might even correspond to the extent they are semantically related). Because of this, it is natural from a neuroscientific perspective to believe that superposition and polysemanticity occurs whenever there is a structure required of or naturally occurring in the individual cognitive items (memories, sensations, etc.). This is perhaps necessary given energetic and resource restrains in nature.

In the context of such shared physical substrates, the phenomenon of superposition becomes a likely candidate for being related to 'context switching' and 'data dependent geometry', i.e., different partitions of external inputs possessing different loss functions – and therefore different geometry, in the computational and dynamical sense – which is switched between depending on contextual cues. Supposing there is a relationship between geometry in the above sense and internal computational structures, this suggests these structures are far more data-dependent than there merely being a 'circuit' with simple IF statements. Instead, it suggests the notion of there being a single master circuit neatly routing to discrete computational pieces for different inputs is fundamentally mistaken, and rather that the picture is a more integrated and messy one (and naturally so), albeit with an underlying structure related to these tasks (Liu et al., 2023).

Such a picture may seem daunting to study. However, there are some very computationally inexpensive approaches which now bear trying, such as in the vein of Ramsauer et al. (2021), who interpret attention heads in a large language model. Ramsauer et al. (2021), taking the correspondence between associative memory energy and the attention mechanism of Transformers seriously, studied the implied energy landscapes and dynamic regimes of different attention heads using the following heuristic: geometrically, we can crudely classify the implied energy landscape of an attention head $\mathcal{H}$ by how many memories it typically appears to interpolate between or be influenced by – if the number of influencing 'memories' is high, we can crudely say it is producing 'meta-stable-like states' whereas if the number is low, we can crudely say it is operating in 'point-like attractor states'.

Technically, this is done by presenting the attention head with inputs, one-by-one, and observing its distribution of

outputs. Let s be the number of input tokens required to sum the resulting SOFTMAX value of the attention head to 0.9 for a given input. After recording these s values, we will generate a distribution S. Let k be the median of S. If k is small, we say $\mathcal{H}$ has a *point-like attraction schema*; if k is relatively large, we say $\mathcal{H}$ has a *metastable-like attraction schema*. In other words, k is an approximate measure of how many 'memories' contribute to the dynamics at a given point in the energy landscape. It is important to emphasise that in the Transformers context, referring to input tokens as 'memories' is misleading – we should probably not think of such tokens as memories in the traditional sense. It might instead be more appropriate to consider them as 'dynamical attractors', which (as I show in the current study) should not only be studied geometrically, but also topologically. I therefore suggest that S distributions be studied with the CDAM perspective, for example: How might S distributions imply $\mathcal{M}$-like structures? Are these $\mathcal{M}$-like structures necessary for an attention head's capacities to perform CDAM-like functions, such as widening the range of hetero-association across input tokens, extracting multi-scale representations of community structures in the $\mathcal{M}$-like structures, stabilising recall of temporal sequences, or simulating finite automata? How do attention heads and their $\mathcal{M}$-like structures interact across layers? These, and many other questions, are now available, and have theoretical connections and possible clues in the associative memory and neuroscience literatures.

A.3. Non-traditional auto-association performance

As discussed in Section 2.2, analysing the 'capacity' of CDAM is conceptually dubious due to its auto- and hetero-associative mixture. Nevertheless, here I present a non-traditional auto-association task wherein we load CDAM with different numbers of memories and attempt a kind of auto-associative recall of individual patterns.

First, it bears repeating that if we set $a = 1$ and $h = 0$, CDAM becomes the model of Equation 13 in Lucibello & Mézard (2023), where scaling comes from a. Analysing the performance of auto-associative recall is then done in the usual way of setting σ as a noisy version of a memory pattern, then running the dynamics forward until convergence, and measuring the final overlap between the original (denoised) memory pattern and the final, convergent state. From a practical standpoint, however, it is possible to study the following non-traditional variation: we no longer care about the absolute overlap between the pattern and the state, instead we simply care about which pattern has the highest overlap.

Questions which immediately arise are how to structure $\mathcal{M}$ and choose a and h. For this, I take inspiration from the work of Sharma et al. (2022), who introduce a model called 'Memory Scaffold with Hetero-association' (MESH). MESH can be said to perform auto-encoding using three layers: a features/input layer (where noisy input patterns are given), a smaller hidden layer that 'cleans-up' the input and hetero-associates it into the 'memory scaffold', whose dynamical core exists in an even smaller labels layer and which contains a set of fixed, pre-defined, well-separated attractors. Whichever fixed point attractor the hidden layer's encoding of the noisy input pattern leads to is then decoded by the hidden layer's projection back to the features layer. From the CDAM perspective, we could say MESH arranges a set of well-behaved 'auto-associative cores' onto which we may hook or link items through hetero-association, for improvement of those linked patterns' recall performance.

As in MESH, I test CDAM on the FashionMNIST dataset (Xiao et al., 2017), which consists of 28×28 grayscale images associated with a label from 10 classes of clothing (shirts, pants, shoes, etc.). I arrange the maximum-normalised pixel values of these images into vectors and use them as the memory states, i.e., $n = 784 = 28 \times 28$. As in previous simulations, I use $\beta = 1$, and $\eta = 0.1$. To initialise the network state, I choose a memory pattern μ (an image from the FashionMNIST dataset) and set $\sigma(0) = \xi^\mu + c\zeta$, where ζ is a random vector with elements independently drawn from the interval $[-0.5, 0.5]$ and $c \in \mathbb{R}^+$ is the amplitude of the additive random noise. Here I use $c = 1$. I run the simulation until convergence and measure the overlap of the final state with all memory vectors. Whichever memory vector has the highest overlap is considered to be 'predicted' (see Figure 8).

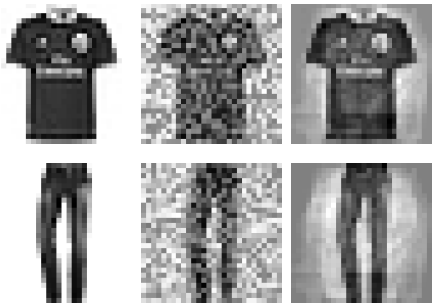

Figure 8. Examples of two FashionMNIST memory items in original (left), noisy (center), and 'predicted' (right) forms.

To construct $\mathcal{M}$, we start with an empty graph $\overline{\mathcal{K}_p}$, i.e., a graph with p memory vertices and no edges between them. Then, from each memory vertex, we create an undirected edge to another memory vertex corresponding to its closest memory vector as measured by Euclidean distance between all memory vectors. Upon doing so, we introduce a basic amount of memory scaffolding in the form of topological support between similar memory patterns.

In Figure 9, I test a range of positive values for a and h, keep-

ing the model in E–I balance by ensuring $a+h=1$. We see that setting a close to but above 0 and h close to but below 1, we achieve a more graceful trade-off between memory storage capacity and memory pattern precision, reminiscent of Sharma et al. (2022). Notably, if $a=h$, CDAM performs very poorly, and using a purely auto-associative structure ($a=1, h=0$) demonstrates catastrophic forgetting, seen as sudden drops in performance.

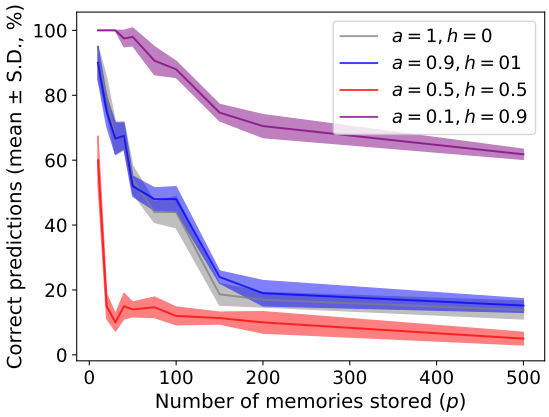

Figure 9. Performance of CDAM on auto-association task using FashionMNIST data over 5 trials, using various settings of a and h. (Tested levels of $p=10, 20, 30, 40, 50, 75, 100, 150, 200, 500$.)

A.4. Numerical simulations of the four dynamical modes

In the numerical simulations below, I use $n=1,000$, $\beta=1$, and $\eta=0.1$. I choose fixed values of a and h to demonstrate the four dynamical modes described in Subsection 2.2 across the tested graphs: pure auto-association ($a=1, h=0$), narrow ($a=0.5, h=0.5$) and wide ($a=-0.5, h=1.5$) hetero-association, and neutral quiescence ($a=-2.5, h=1$). Simulations are terminated at $t=101$, which in all cases is a fixed point or limit cycle. The memory patterns stored are random vectors as described in Subsection 2.1.

To initialise the network state in each simulation, I choose a memory pattern $\mu \in \mathcal{M}$ and set $\sigma(0)=\xi^\mu + c\zeta$, where ζ is a random vector with elements independently drawn from the interval $[-0.5, 0.5]$ and $c \in \mathbb{R}^+$ is the amplitude of the additive random noise. Here I use $c=1$.

Figure 10 shows results for a 2–regular graph (the only type of which are unions of 1D cycles). We can notice some apparent 'clusters' of vertices which appear to commonly become co-active. This represents a common meta-stable state shared by the surrounding trigger stimuli. The reason for these particular meta-stable groups is due to the random biases present in the random patterns, which likely become amplified by a finite field effect. Notably, in the wide hetero-association condition ($a=-0.5, h=1.5$), the meta-stable groups are fewer in number and larger in size than the narrow hetero-association condition ($a=0.5, h=0.5$).

Figure 11 shows results for the Tutte graph (Tutte, 1946), which is 3–regular. Unlike the 1D cycle graph shown in Figure 10, the Tutte graph has a more interesting topology, in the form of 3 clusters of highly-connected vertices. These clusters become noticeable by looking at the emergent structure of the correlations for the wide hetero-association case ($a=-0.5, h=1.5$). However, it becomes even more noticeable by looking at the correlations between the convergent meta-stable states, which I show in Figure 3.

Finally, in Figure 12 I analyse a random 3–regular graph with $p=46$ vertices (the same number as in the Tutte graph – making their size and degree distributions equal). As in the previous memory graphs, we can see the network converges to an unbiased E–I balance for the first three settings (bottom row of Figure 12). We can also see that as a decreases in value, the spread of hetero-association becomes gradually wider. However, unlike in the Tutte graph, there are no natural clusters of vertices. Therefore, the resulting correlation matrices reflect the random topology insofar as having no discernible regularity, besides the approximately uniform distribution of noise (which is uniform due to the regular nature of the graph, causing each trigger stimulus to activate a similar number of other memory patterns).

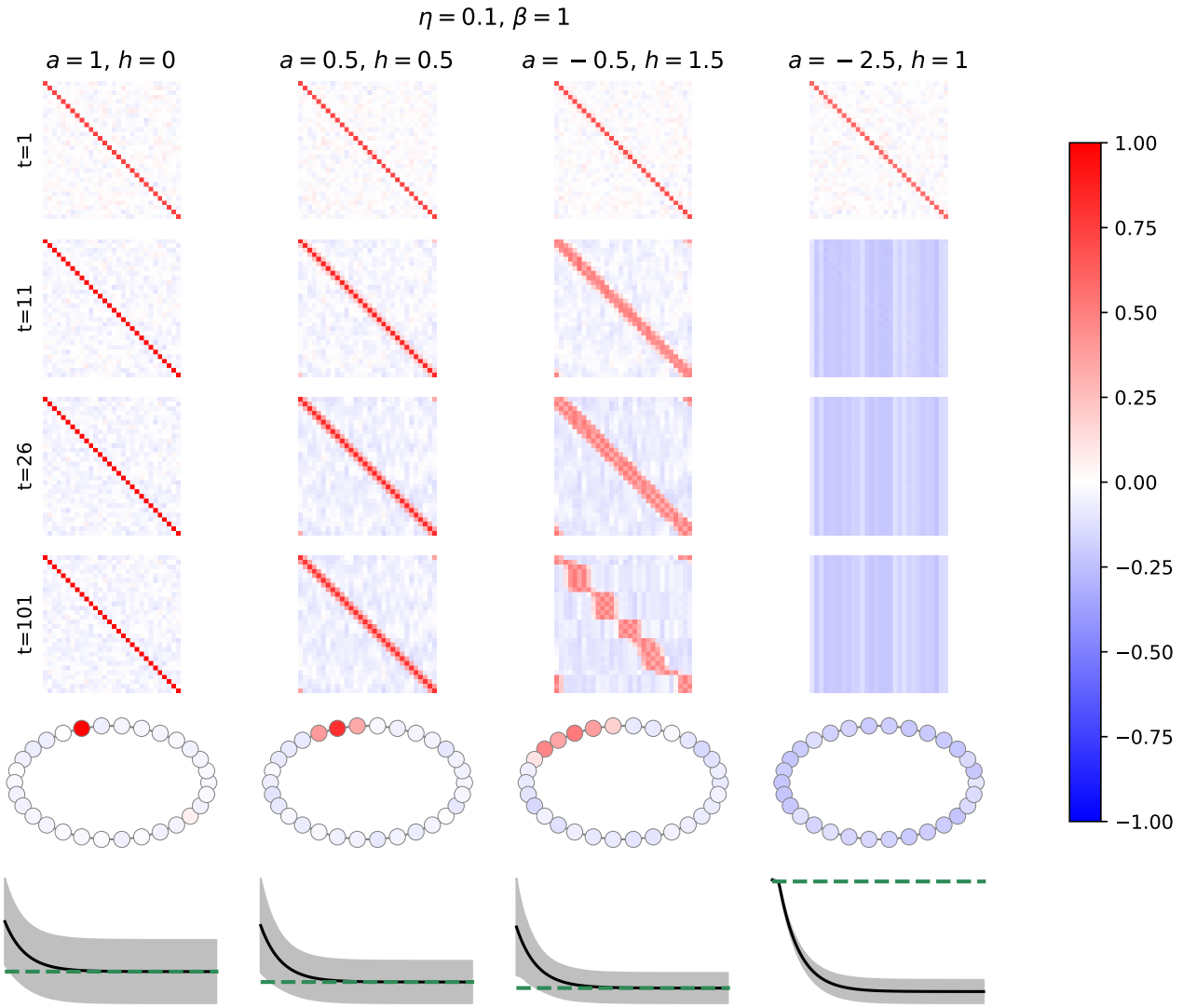

Figure 10. Setting $\mathcal{M} = \mathcal{C}_{30}$, I demonstrate the four dynamical modes of the network (from left to right): auto-association, narrow hetero-association, wide hetero-association, and neutral quiescence. At $t = 1, 11, 26, 101$, I plot the correlation between each memory pattern and the current state, $r(\mu^{(t)})$. In the penultimate row, I draw $\mathcal{M}$ with vertices coloured by $r(\mu^{(t)})$ for one initial trigger stimulus. And in the final row, I plot the mean $\pm$ standard deviation of the neural activities over time, with the dotted green line at 0.

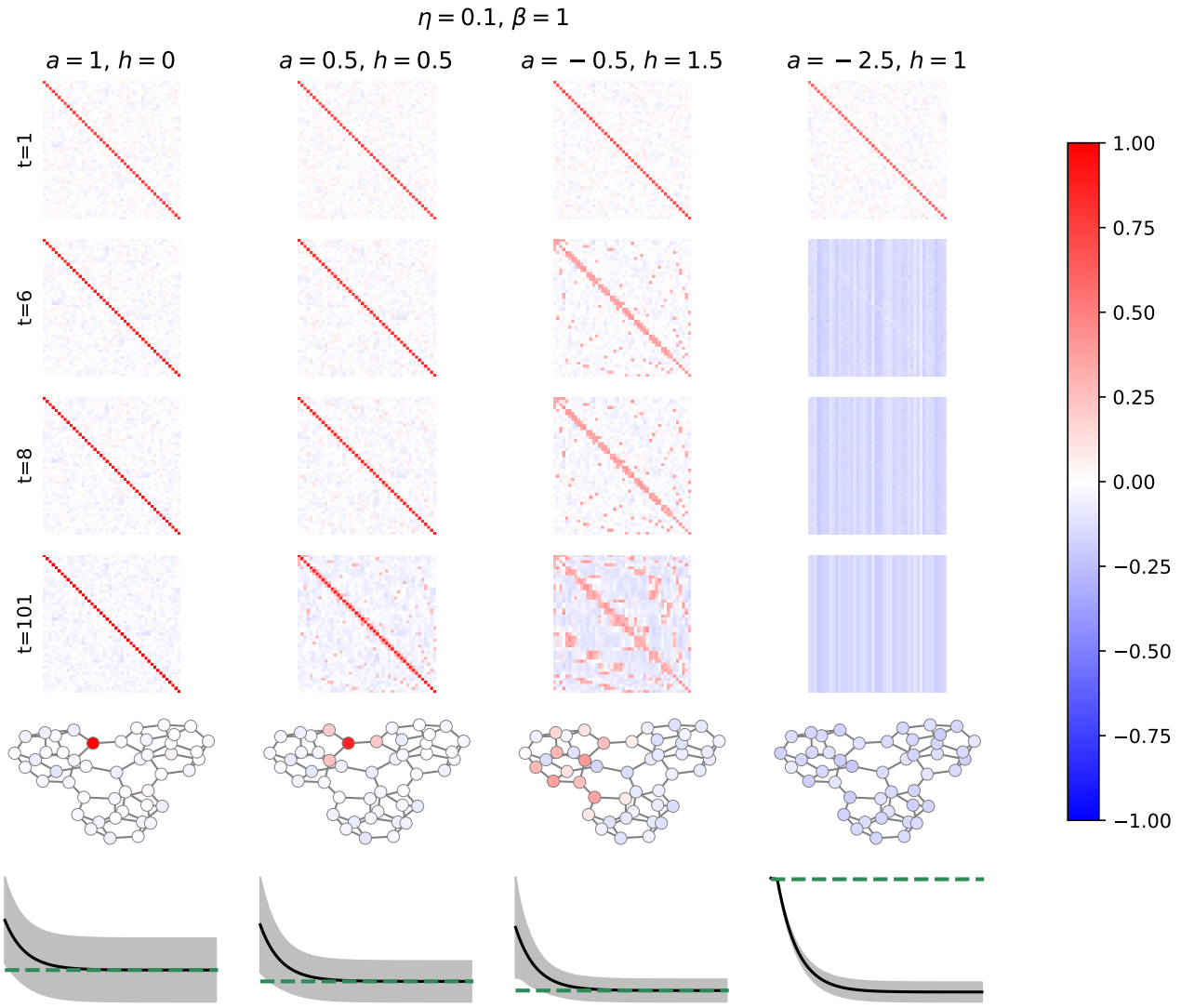

Figure 11. Setting $\mathcal{M}$ as the Tutte graph (Tutte, 1946) ($p = 46$), I demonstrate the four dynamical modes of the network (from left to right): auto-association, narrow hetero-association, wide hetero-association, and neutral quiescence. At $t = 1, 11, 26, 101$, I plot the correlation between each memory pattern and the current state, $r(\mu^{(t)})$. In the penultimate row, I draw $\mathcal{M}$ with vertices coloured by $r(\mu^{(t)})$ for one initial trigger stimulus. And in the final row, I plot the mean $\pm$ standard deviation of the neural activities over time, with the dotted green line at 0.

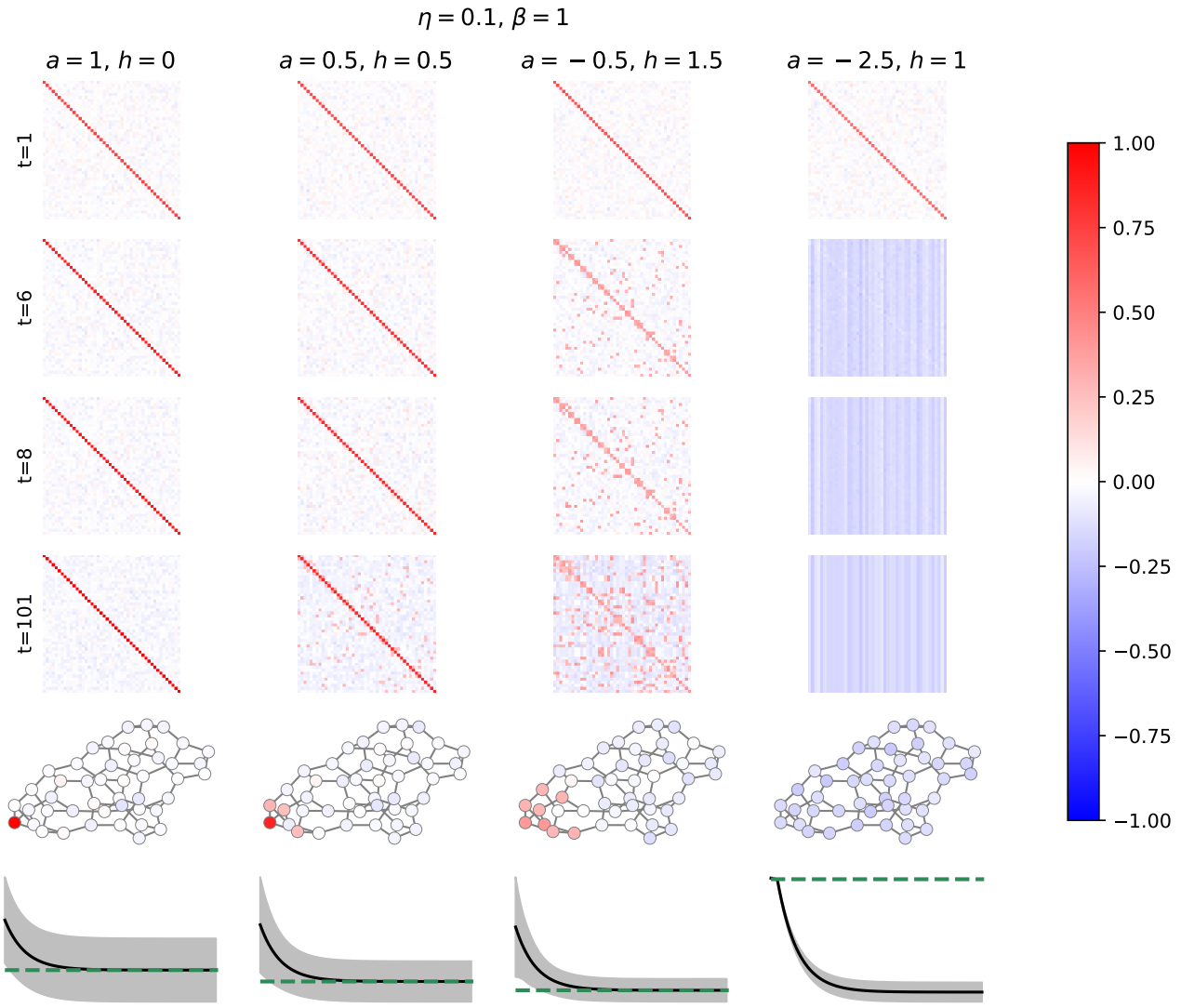

Figure 12. Setting $\mathcal{M}$ as a random 3–regular graph with $p = 46$, I demonstrate the four dynamical modes of the network (from left to right): auto-association, narrow hetero-association, wide hetero-association, and neutral quiescence. At $t = 1, 11, 26, 101$, I plot the correlation between each memory pattern and the current state, $r(\mu^{(t)})$. In the penultimate row, I draw $\mathcal{M}$ with vertices coloured by $r(\mu^{(t)})$ for one initial trigger stimulus. And in the final row, I plot the mean $\pm$ standard deviation of the neural activities over time, with the dotted green line at 0.

A.5. Biological mechanisms by which to learn M

As discussed in Section 1.2, there are clearly many examples in humans and non-human animals of hetero-associative learning, and this ought to have a physical basis in the brain. Classically, most associative memory work assumes a prior process of Hebbian learning, whereby 'cells that fire together, wire together'. In doing so, neurons form clusters or 'assemblies' which together correspond to internal or external cues. The biological mechanisms for these processes are well-studied (Buzsáki, 2010) and have led theoretical and computational neuroscientists to propose mechanisms by which such assemblies might perform cognitive functions (Papadimitriou et al., 2020; Müller et al., 2020).

In this context, it does not seem far-fetched to imagine a biological organism explicitly learning arbitrary hetero-associative structures of the form M, in CDAM, or some approximation thereof. This would most easily be achieved through repeated exposure to sequences of memory items as a random walk on $\mathcal{M}$, as was studied in human participants by Schapiro et al. (2013), who indeed found the participants learnt the community structure of the $\mathcal{M}$ equivalent in that study. However, it is quite likely animals encounter stimuli which have similar structure, and thereby use past experience as a cognitive bias to more efficiently store related structures and learn new instances of similar structures. A computational example of this idea comes from Whittington et al. (2020), who introduced the 'Tolman-Eichenbaum Machine' (TEM), a joint where-what model of hippocampus and entorhinal cortex which demonstrates close correspondence to biological data as well as having a mathematical relationship to Transformers (Whittington et al., 2022). This has been followed by alluring biological studies (El-Gaby et al., 2023), further solidifying the potential biological bases of learning and leveraging such structures and biases.

A.6. 1D cycle memory graphs

Practically all of the past semantic hetero-associative literature (Amari, 1972; Tank & Hopfield, 1987; Kleinfeld & Sompolinsky, 1988; Gutfreund & Mezard, 1988; Griniasty et al., 1993; Gillett et al., 2020; Tyulmankov et al., 2021; Karuvally et al., 2023; Chaudhry et al., 2023; Karuvally et al., 2023) has studied the case of $\mathcal{M}$ being a 1D cycle. This is because such models typically consider p memory patterns, and construct weights between neurons i and j in a form such as

$$J_{ij} = \frac{1}{n}\sum_{\mu}^{p}(\xi_i^{\mu+1}\xi_j^{\mu} + \xi_i^{\mu}\xi_j^{\mu+1}), \tag{6}$$

where, crucially, the memory patterns are semantically correlated along a single line. This would make $\mathcal{M}$ a line graph, where it not for the fact that most studies let $\xi^{p+1} = \xi^1$, which connects the two ends of the line to form a circle, as shown in Figure 13.

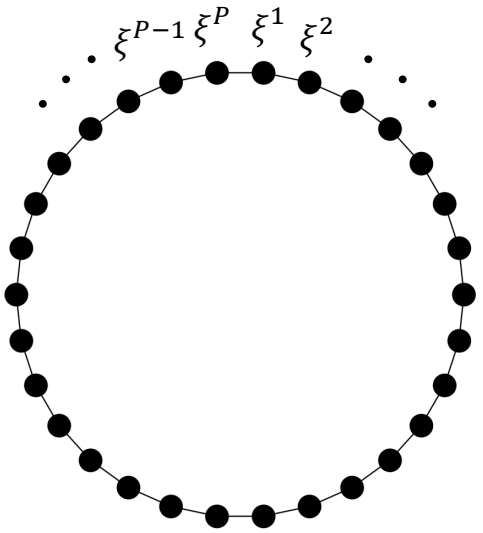

Figure 13. Illustration of the 1D cycle memory graph, with the vertices labelled by the memory index. Note that this would be a line if we do not identify $\xi^{p+1} = \xi^1$.

In other appendices and the main text, I denote a 1D cycle graph with n vertices as $\mathcal{C}_n$.

A.7. Replication of Miyashita (1988)

As described in Subsection 1.2, Miyashita (1988) is a a classical study in the semantic hetero-association neuroscience literature, which showed neurons from monkey temporal cortex were responsive to the presentation of stimuli according to the order in which they were presented. These semantic links were developed without general regard to any spatial or statistical similarities shared between the stimuli. In neurons which were significantly responsive to the stimuli, their activity was significantly auto-correlated with the activity elicited by the stimuli up to a distance of 6 patterns into the past or future of the stimuli sequence.

I manually transcribed data from Figure 3C of Miyashita (1988) by printing the enlarged figure and carefully using a pencil and ruler to measure data for the 28 cell group (illustrated as square symbols in the original figure), which showed the largest hetero-associations. The mean and standard error of the mean (SEM) which I measured and used in the subsequent analysis are shown in Table 1.

Here I model the results of Miyashita (1988) by setting $\mathcal{M} = \mathcal{C}_{30}$ (see Appendix A.6) and choosing a and h to

Table 1. Auto-correlations between neural activities responsive to visual stimuli in the monkey temporal cortex. The data are transcribed from the 28 cell group (square symbols) of Figure 3C in Miyashita (1988). Distance refers to the temporal distance between the stimuli.

Distance	0	1	2	3	4	5	6
Mean	1	0.33810	0.19700	0.11940	0.08806	0.07015	0.06493
SEM	0	0.03731	0.03582	0.02985	0.02388	0.02015	0.02239

match the data. As shown in Figure 14, I find that an anti-Hebbian auto-association and Hebbian hetero-association ($a = -2.45, h = 3.45$) correlated well with the experimental results.

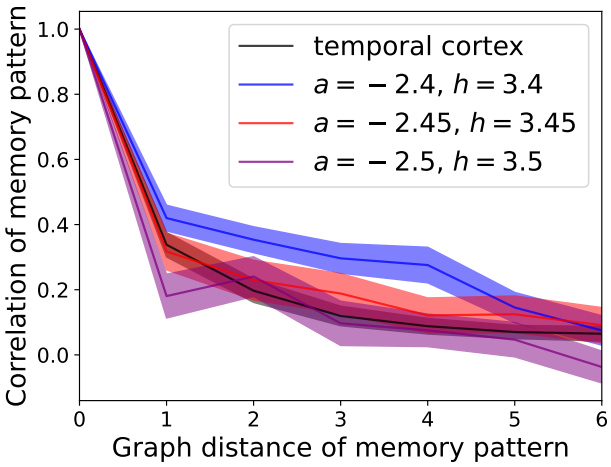

Figure 14. Mean $\pm$ standard error of the means for $\mathcal{M} = \mathcal{C}_{30}$ with different values of a and h, alongside the transcribed Miyashita (1988) data shown Table 1. The closest matching model tested was $a = -2.45, h = 3.45$, the means of which had a high correlation ($R^2 = 0.997$) with the means reported in Miyashita (1988).

A.8. Controlling the range of attractors

Figure 15 shows the case of $\mathcal{M} = \mathcal{C}_{30}$ with varying levels of a and h and random memory patterns (and draws its data from the same simulations as for Figure 1). At $a = 1$ and $h = 0$, we have the expected auto-associative behaviour. However, as we decrease a and set $h = 1 - a$ (to maintain unbiased E–I balance), we see an increase in hetero-association and a gradually increasing spread of excitation through the graph, with excitation emanating from the triggered memory pattern that was set at $\sigma(0)$.

Figure 16 shows the case of $\mathcal{M}$ as the Tutte graph (Tutte, 1946), using the same tested values for a and h as in Figure 15. Again, we see an increase in hetero-association and a gradually increasing spread of excitation through the graph.

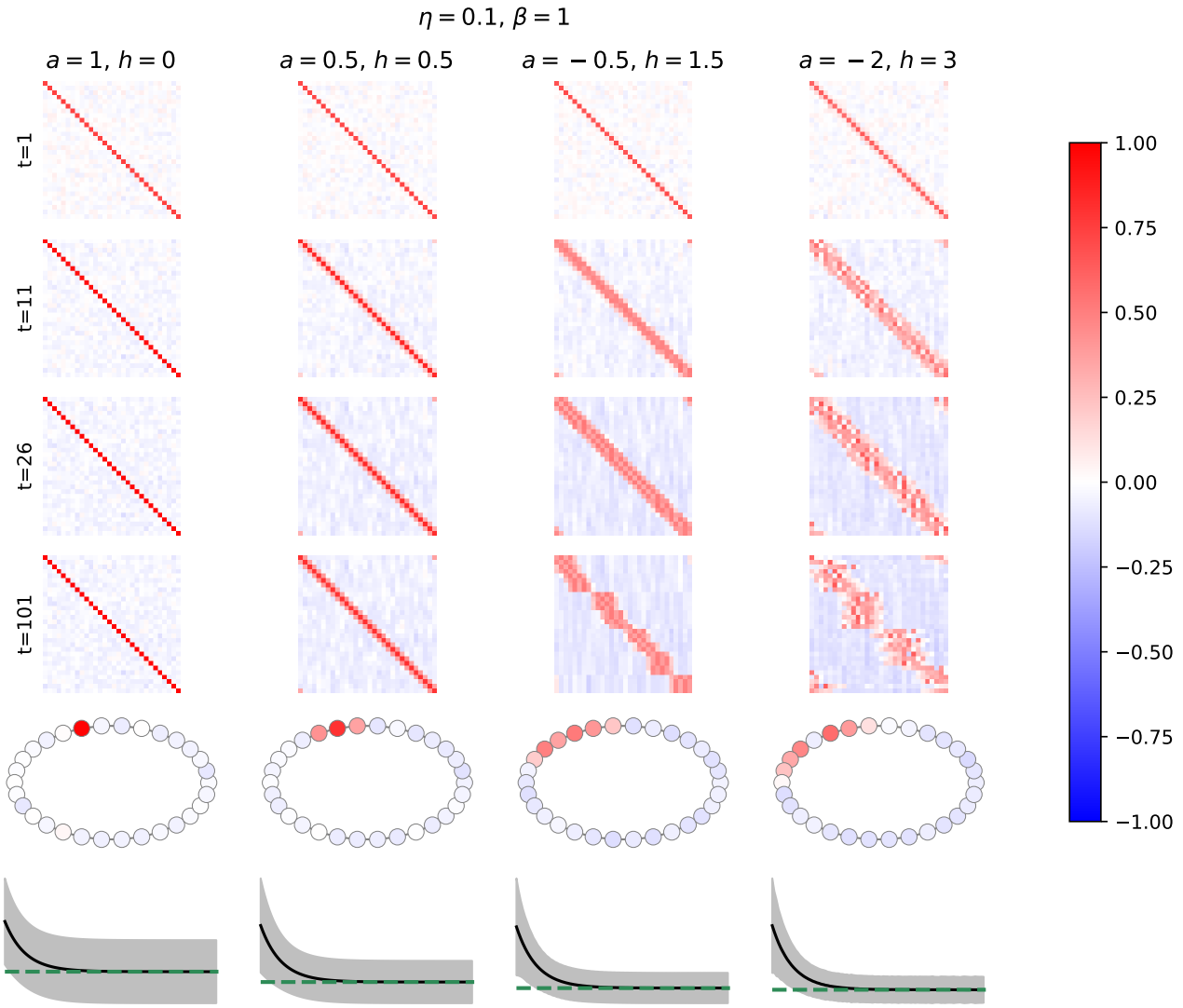

Figure 15. Memory pattern correlations for each vertex in $\mathcal{M} = \mathcal{C}_{30}$ with increasing range of hetero-association (left column to right column). The first four rows show the correlations as a heatmap of dimensions 30×30, where each cell is coloured by its correlation coefficient, 1 (red) to -1 (blue). The penultimate row draws $\mathcal{C}_{30}$ with vertices coloured by the correlations at the end of the simulation for the same trigger stimulus (the vertex coloured red in the left-most column) for each tested set of parameters.

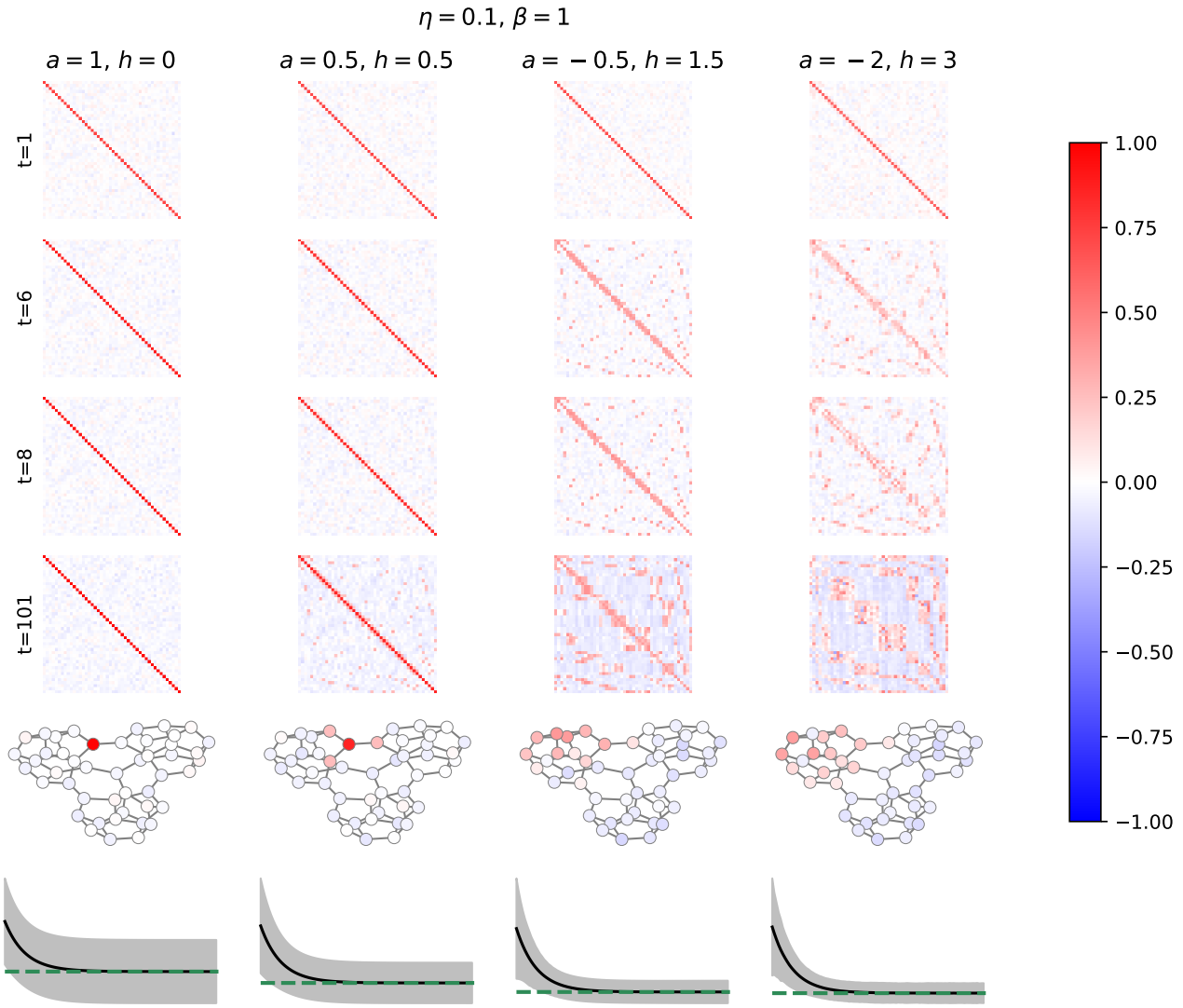

Figure 16. Memory pattern correlations for each vertex in $\mathcal{M}$ as the Tutte graph (Tutte, 1946), with increasing range of hetero-association (left column to right column). The first four rows show the correlations as a heatmap of dimensions 46×46, where each cell is coloured by its correlation coefficient, 1 (red) to -1 (blue). The penultimate row draws $\mathcal{M}$ with vertices coloured by the correlations at the end of the simulation for the same trigger stimulus (the vertex coloured red in the left-most column) for each tested set of parameters.

A.9. Zachary's karate club graph

An interesting and naturally-constructed graph is *Zachary's karate club graph* (Zachary, 1977). It consists of 34 vertices, representing karate practitioners, where edges connect individuals who consistently interacted in extra-karate contexts. Notably, the club split into two halves. Setting Zachary's karate club graph as $\mathcal{M}$ and varying a and h, however, reveals that there were even finer social groupings than these, as seen in Figures 2 and 17. Smaller groups are particularly noticeable in some of the individual pattern trigger stimuli for $a = 0.4, h = 0.05$ (Figure 19) and $a = 0.3, h = 0.1$ (Figure 20). Contrastingly, $a = 1, h = 0$ selects for individuals (Figure 18) and $a = -0.1, h = 0.1$ selects for the two major groups post-split (Figure 21).

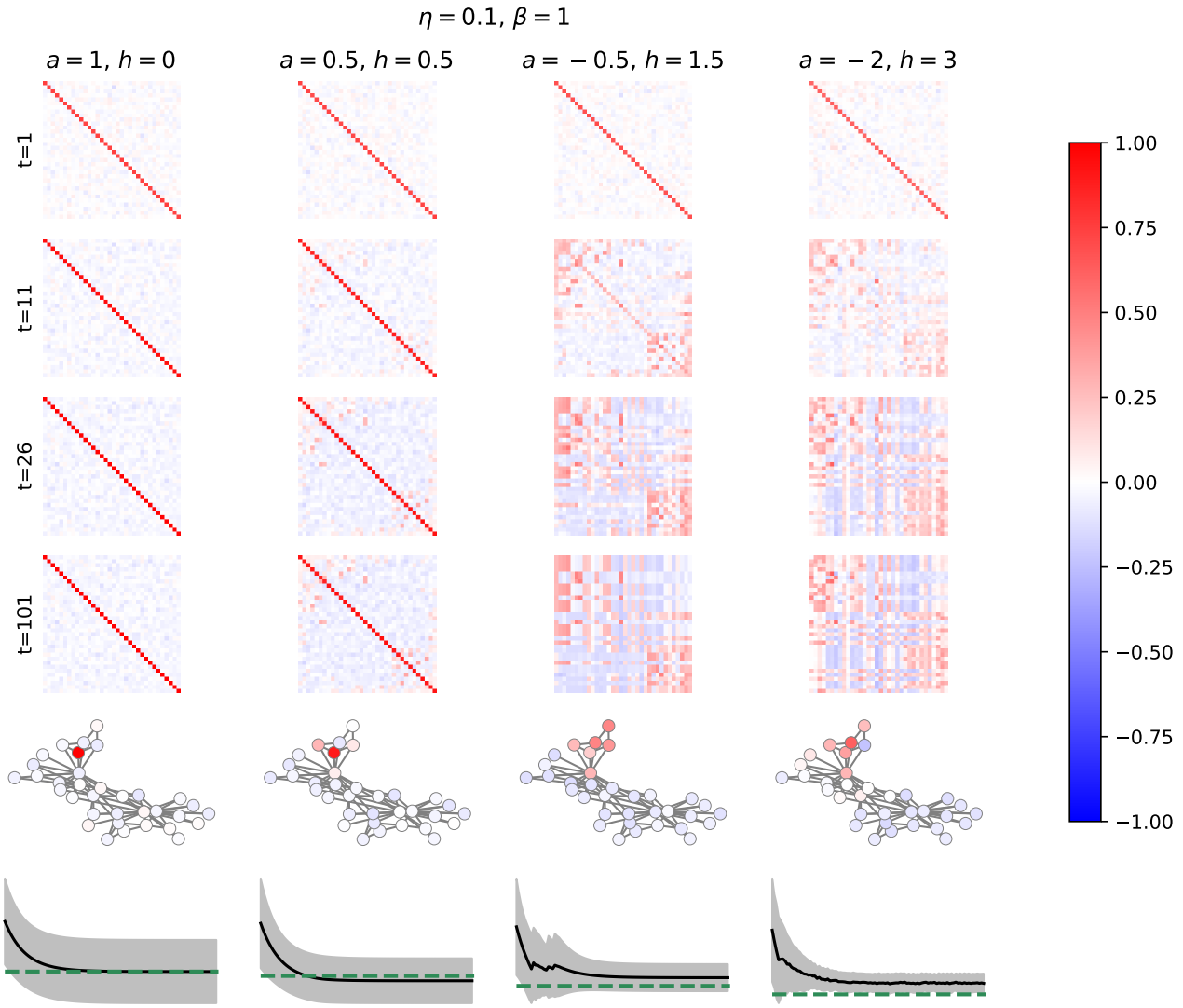

Figure 17. Setting $\mathcal{M}$ as Zachary's karate club graph (Zachary, 1977), I demonstrate multi-scale graph segmentation. At $t = 1, 11, 26, 101$, I plot the correlation between each memory pattern and the current state, $r(\mu^{(t)})$. In the penultimate row, I draw $\mathcal{M}$ with vertices coloured by $r(\mu^{(t)})$ for one initial trigger stimulus. And in the final row, I plot the mean $\pm$ standard deviation of the neural activities over time, with the dotted green line at 0, demonstrating dynamic stability.

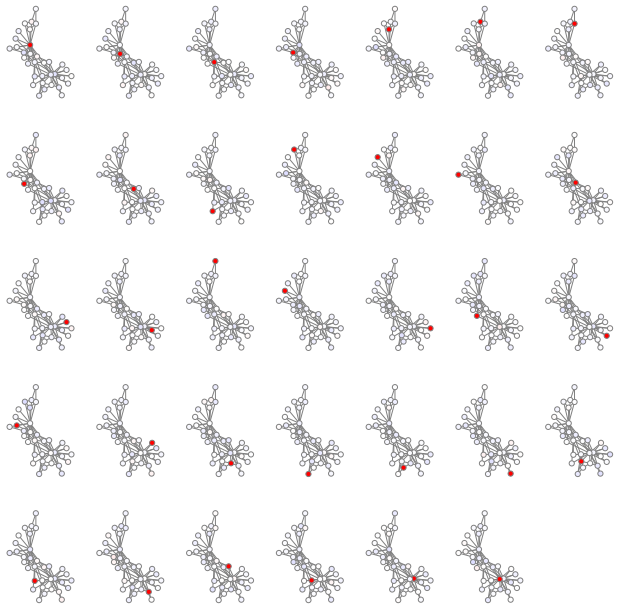

Figure 18. Correlations, $r(\mu^{(1)})$, for every possible trigger stimulus in Zachary's karate club graph with $a = 1$ and $h = 0$.

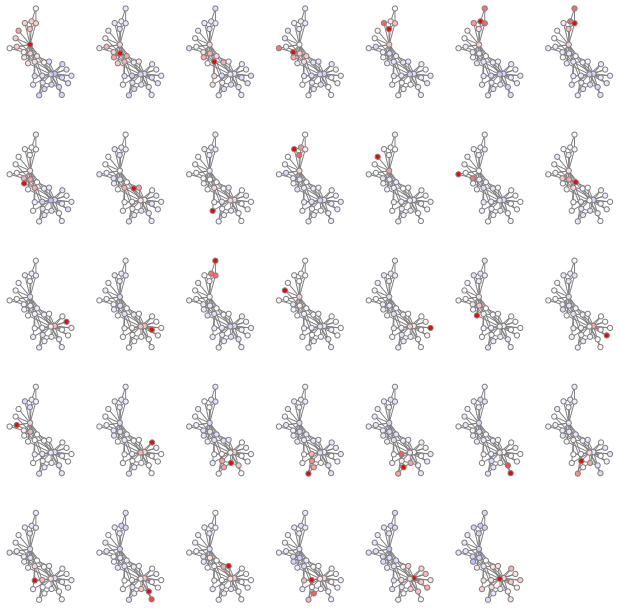

Figure 19. Correlations, $r(\mu^{(11)})$, for every possible trigger stimulus in Zachary's karate club graph with $a = 0.5$ and $h = 0.05$.

Figure 20. Correlations, $r(\mu^{(26)})$, for every possible trigger stimulus in Zachary's karate club graph with $a = -0.5$ and $h = 1.5$.

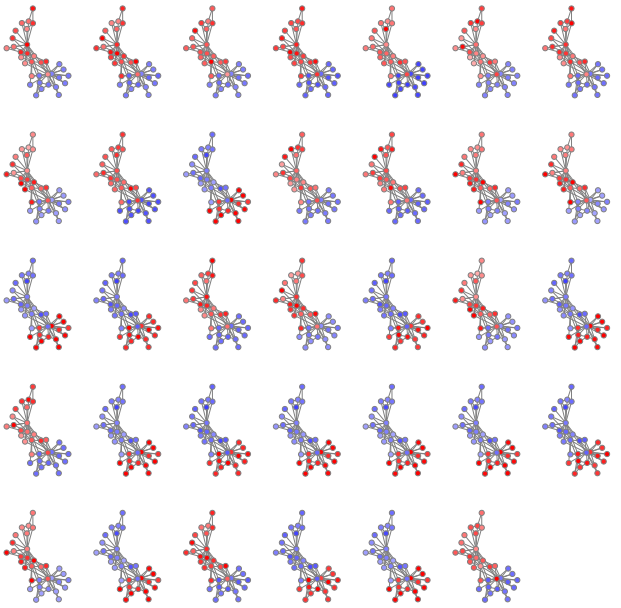

Figure 21. Correlations, $r(\mu^{(101)})$, for every possible trigger stimulus in Zachary's karate club graph with $a = -2$ and $h = 3$.

A.10. Barbell memory graph

A *barbell graph* $\mathcal{B}_{n,m}$ is the union of two copies of the complete (fully-connected) graph $\mathcal{K}_n$ on n vertices, connected by a single path vertices of size $\mathcal{M}$. Here I choose $n = m = 10$. This simple example helps to demonstrate the two extremes of the attractive regime scales – where one scale maintains individual pattern activities and the other identifies the local pattern cliques in $\mathcal{M}$.

Figure 22 shows correlations for random memory patterns embedded in $\mathcal{M} = \mathcal{B}_{10,10}$ with varying levels of a and h. At $a = 1$ and $h = 0$, there is no hetero-associative activity, only auto-association. However, as we decrease a with $h = 1 - a$ (for E–I balance) the two $\mathcal{K}_n$ cliques quickly show correlated group activity. Along the path connecting the two complete graphs, we also see a lengthening in the spread of activity along the path (like in the cycle graph). This can be further verified by inspection of the individual attractors, where Figures 23–26 show terminal Correlations, $r(\mu^{(101)})$, for each trigger stimulus pattern in the graph across tested values of a and h.

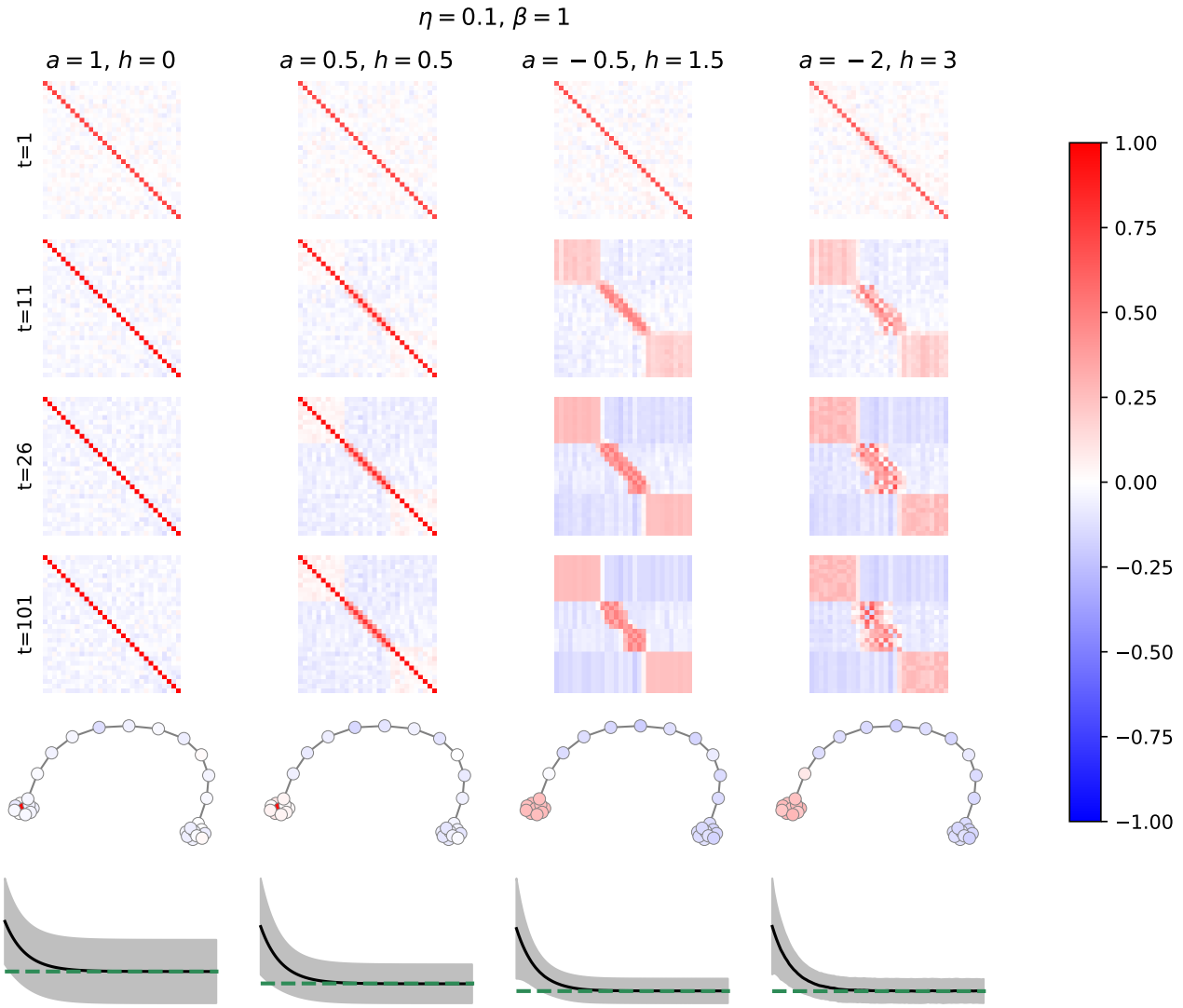

Figure 22. Memory pattern correlations for each vertex in $\mathcal{M} = \mathcal{B}_{10,10}$ with decreasing a (left column to right column). The top row shows the correlations as a heatmap of dimensions 30×30, where each cell is coloured by its correlation coefficient, 1 (red) to -1 (blue). The bottom row shows an example of the mean terminal activity states given the same pattern trigger stimulus (the vertex coloured red in the left column) for each tested pair of a and h values.

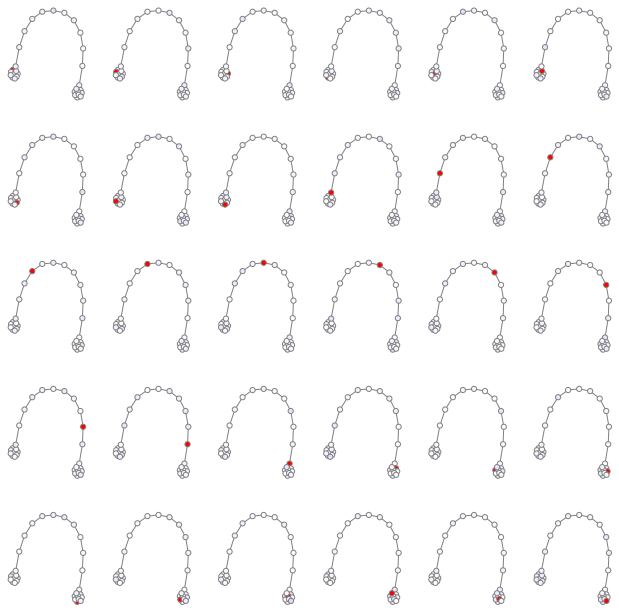

Figure 23. Correlations, $r(\mu^{(1)})$, for every possible trigger stimulus in $\mathcal{B}_{10,10}$ with $a = 1$ and $h = 0$.

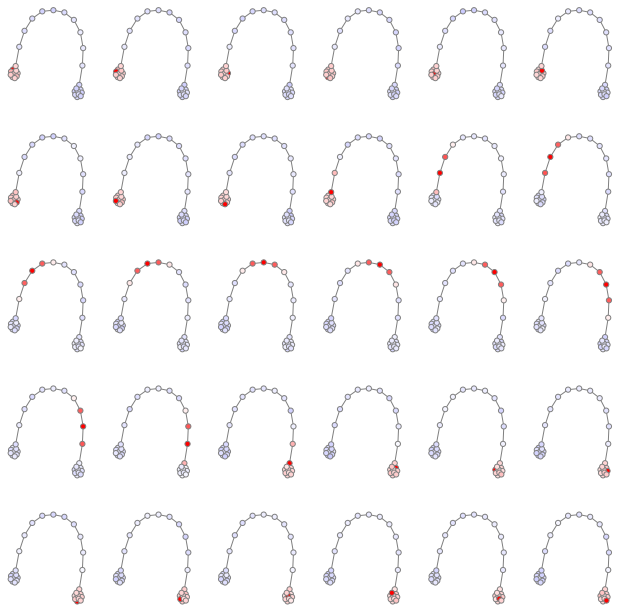

Figure 24. Correlations, $r(\mu^{(11)})$, for every possible trigger stimulus in $\mathcal{B}_{10,10}$ with $a = 0.5$ and $h = 0.5$.

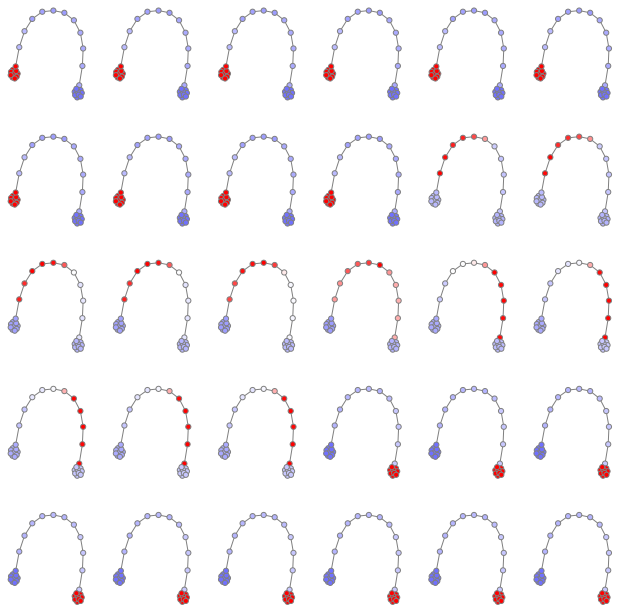

Figure 25. Correlations, $r(\mu^{(26)})$, for every possible trigger stimulus in $\mathcal{B}_{10,10}$ with $a = -0.5$ and $h = 1.5$.

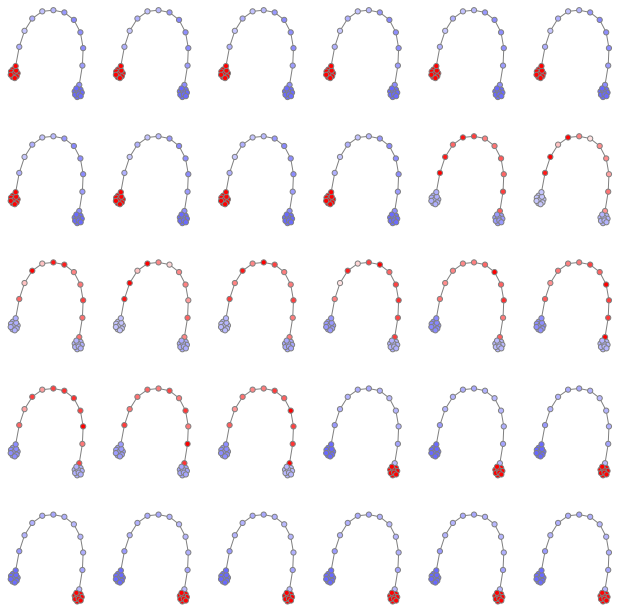

Figure 26. Correlations, $r(\mu^{(101)})$, for every possible trigger stimulus in $\mathcal{B}_{10,10}$ with $a = -2$ and $h = 3$.

A.11. Video sequence recall

The two videos used were sourced from Wikimedia Commons and were uploaded by User:Raul654 on 24 January 2006. They can found at the below URLs:

https://commons.wikimedia.org/wiki/File:Gorilla_gorilla_gorilla1.ogv

https://commons.wikimedia.org/wiki/File:Gorilla_gorilla_gorilla4.ogv

I used the first 50 frames of each video. The videos are size 320px $\times$ 240px with bit depth of 24. The pixel and colour information was flattened into a single vector of length $230,400$, with the values normalised by the maximum value, 240. I then randomly sampled $n = 2,000$ values from this flattened vector and treat these as our neural patterns and states. For illustration, the first frames of each video are shown in Figure 27.

Figure 27. First frames of the two videos used in the video recall experiment: video 1 (top) and video 2 (bottom).

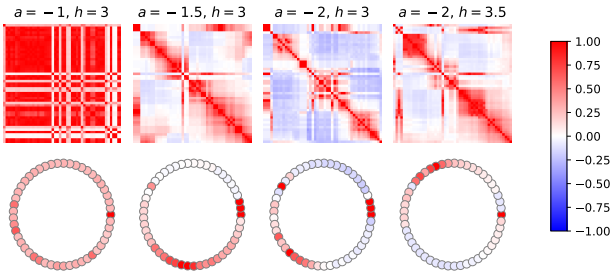

Figure 28. Correlations between attractors for each target stimuli of video 1.

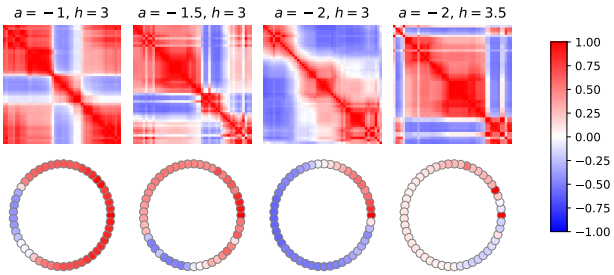

Figure 29. Correlations between attractors for each target stimuli of video 2.

A.12. Finite automaton simulation

Table 2 shows the adjacency matrix and Figure 30 draws the vertices and edges of $\mathcal{M}$ for simulating a finite automaton with the information from Figure 6. The individuals' memory patterns are set to be auto-associative (there is a self-edge in $\mathcal{M}$ for that memory pattern) and transitions' memory patterns are set to be hetero-associative (there is a directed edge leading from the transition memory pattern to the relevant individual).

Figure 31 shows simulations starting at each possible vertex of $\mathcal{M}$, and 32 shows the same with using random data instead of the image and text embeddings data. Compared to the random data memory patterns, we can see that there are some additional correlations between memory patterns in Figure 31 due to the semantic correlations, particularly those associated with the text data.

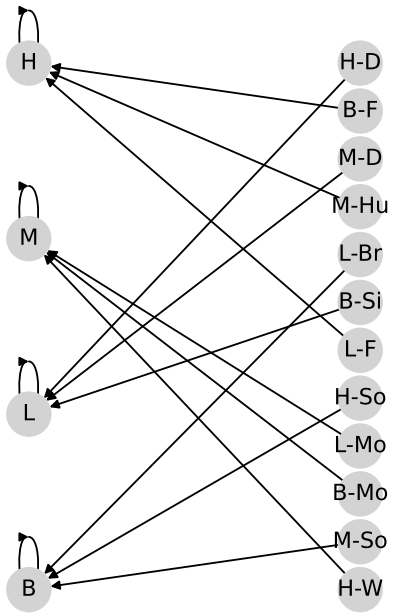

Figure 30. Drawing of $\mathcal{M}$ for simulating a finite automaton with the information from Figure 6. Key: H='Homer', M='Marge', L='Lisa', B='Bart', W='Wife', Hu='Husband', So='Son', D='Daughter', Br='Brother', Si='Sister'.

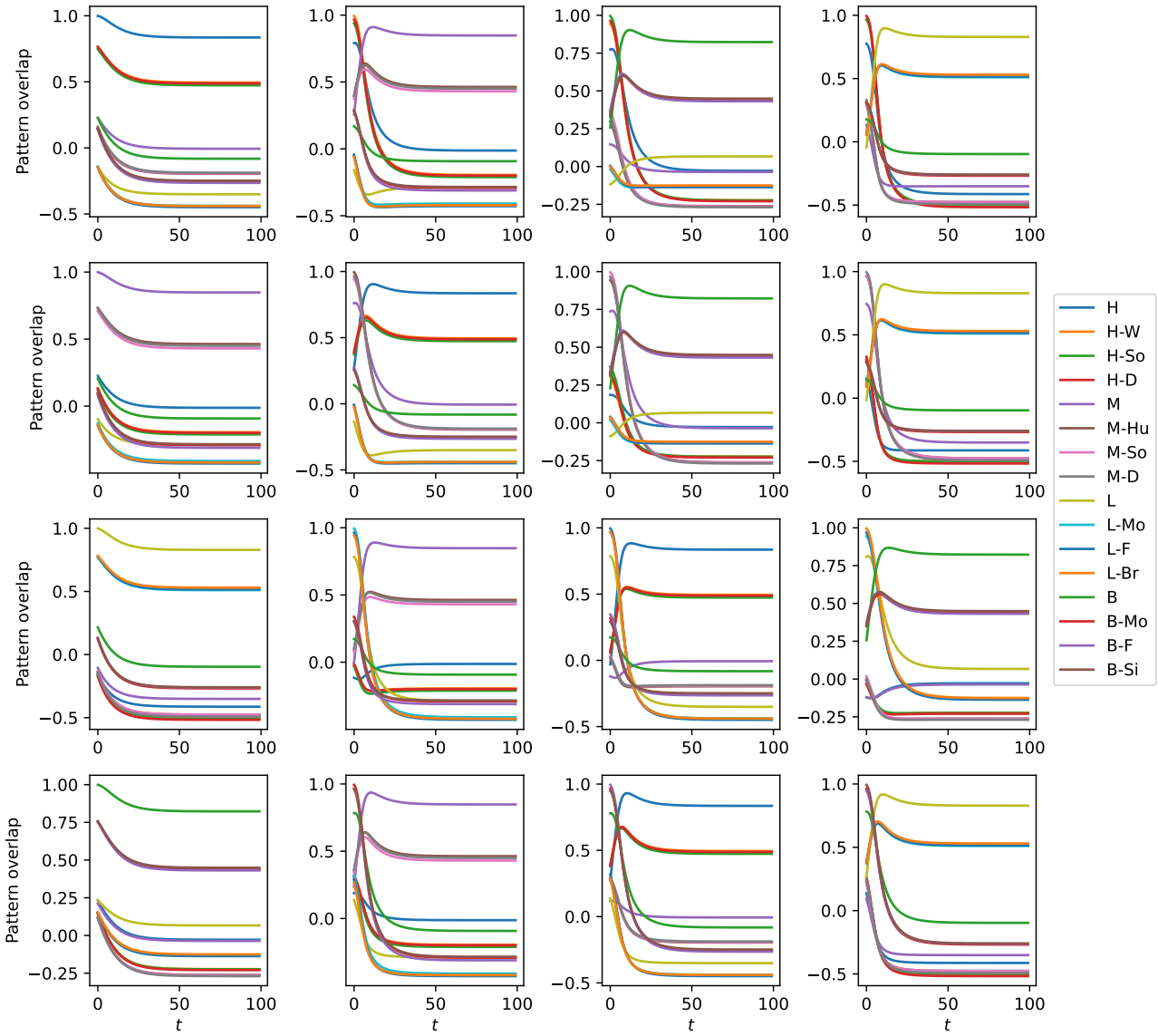

Figure 31. Simulations beginning at all possible starting vertices in $\mathcal{M}$ of Figure 30, using $a = 0, h = 1$. Key: H='Homer', M='Marge', L='Lisa', B='Bart', W='Wife', Hu='Husband', So='Son', D='Daughter', Br='Brother', Si='Sister'.

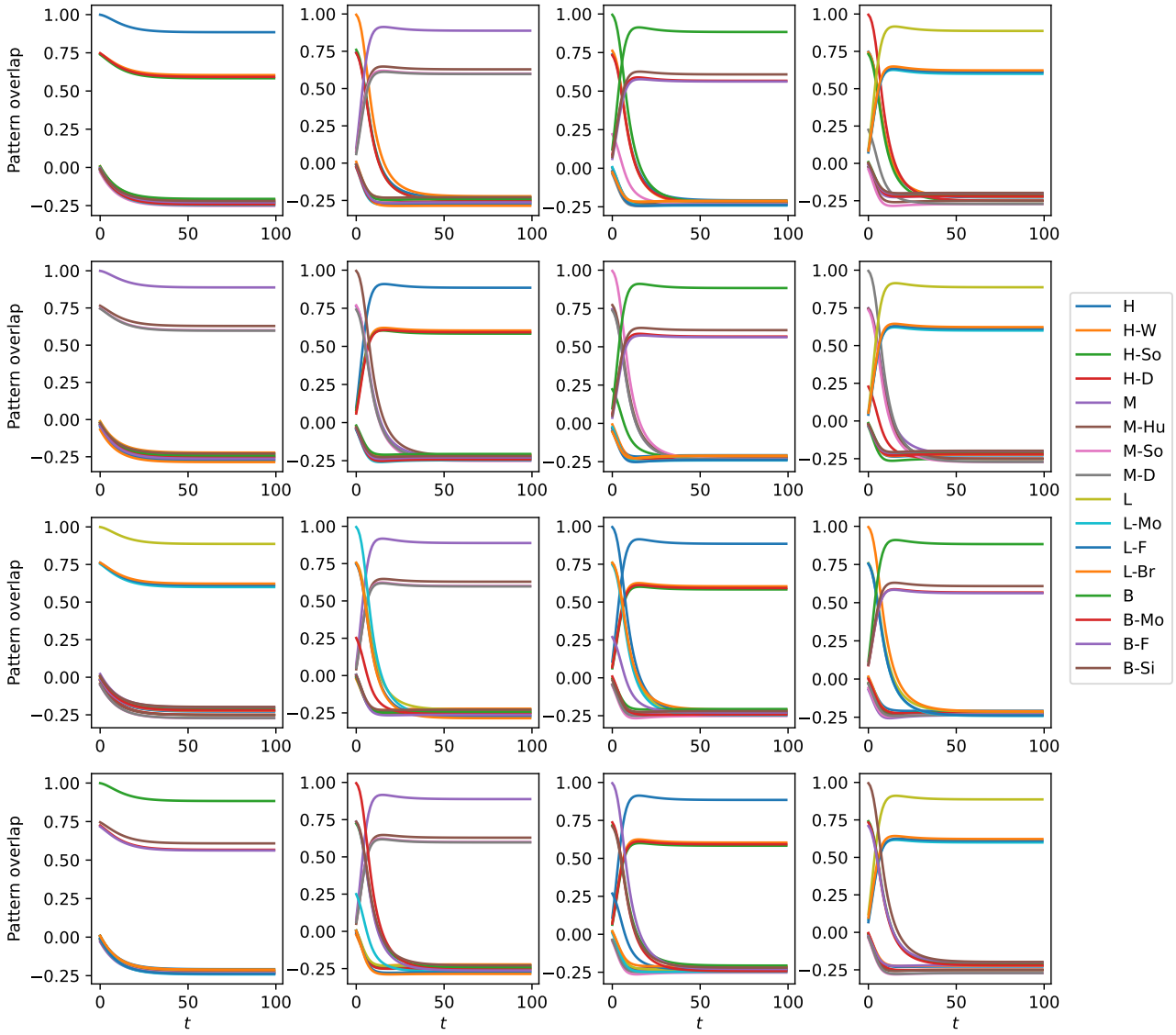

Figure 32. Simulations beginning at all possible starting vertices in $\mathcal{M}$ of Figure 30, using $a = 0, h = 1$, and random data for the attractors (instead of image and text embeddings data). Key: H='Homer', M='Marge', L='Lisa', B='Bart', W='Wife', Hu='Husband', So='Son', D='Daughter', Br='Brother', Si='Sister'.

Table 2. Adjacency matrix of $\mathcal{M}$ in the example finite automaton simulation for the family tree shown in Figure 6. Cells with values of 0 entries have been omitted for visual clarity. Key: H=‘Homer’, M=‘Marge’, L=‘Lisa’, B=‘Bart’, W=‘Wife’, Hu=‘Husband’, So=‘Son’, D=‘Daughter’, Br=‘Brother’, Si=‘Sister’.

	H	H-W	H-So	H-D	M	M-Hu	M-So	M-D	L	L-Mo	L-F	L-Br	B	B-Mo	B-F	B-Si
H	1					1					1				1	
H-W																
H-So																
H-D																
M		1			1					1				1		
M-Hu																
M-So																
M-D																
L				1				1	1							1
L-Mo																
L-F																
L-Br																
B			1				1					1	1			
B-Mo																
B-F																
B-Si																