

University of Groningen

## Vector Symbolic Finite State Machines in Attractor Neural Networks

Cotteret, Madison; Greatorex, Hugh; Ziegler, Martin; Chicca, Elisabetta

**IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.**

*Document Version*

Early version, also known as pre-print

*Publication date:*

2022

[Link to publication in University of Groningen/UMCG research database](#)

*Citation for published version (APA):*

Cotteret, M., Greatorex, H., Ziegler, M., & Chicca, E. (2022). *Vector Symbolic Finite State Machines in Attractor Neural Networks*. arXiv. <http://2212.01196v1>

**Copyright**

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).





The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

**Take-down policy**

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

*Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.*

# Vector Symbolic Finite State Machines in Attractor Neural Networks

Madison Cotteret <sup>1,2,3,†</sup>, Hugh Greatorex <sup>2,3</sup>, Martin Ziegler <sup>1</sup>, and Elisabetta Chicca <sup>2,3</sup>

<sup>1</sup>Micro- and Nanoelectronic Systems (MNES), Technische Universität Ilmenau, Germany

<sup>2</sup>Bio-Inspired Circuits and Systems (BICS) Lab, Zernike Institute for Advanced Materials, University of Groningen, Netherlands

<sup>3</sup>Groningen Cognitive Systems and Materials Center (CogniGron), University of Groningen, Netherlands

<sup>†</sup>Email: m.cotteret@rug.nl

**Abstract**—Hopfield attractor networks are robust distributed models of human memory. We propose construction rules such that an attractor network may implement an arbitrary finite state machine (FSM), where states and stimuli are represented by high-dimensional random bipolar vectors, and all state transitions are enacted by the attractor network’s dynamics. Numerical simulations show the capacity of the model, in terms of the maximum size of implementable FSM, to be linear in the size of the attractor network. We show that the model is robust to imprecise and noisy weights, and so a prime candidate for implementation with high-density but unreliable devices. By endowing attractor networks with the ability to emulate arbitrary FSMs, we propose a plausible path by which FSMs may exist as a distributed computational primitive in biological neural networks.

## I. INTRODUCTION

Hopfield attractor networks are robust models of human memory, as from a simple Hebbian learning rule they display emergent attractor dynamics which allow for reliable pattern recall, completion, and correction even in noisy situations [1–3]. Attractor models have since found widespread use in neuroscience as a functional and tractable model of human memory [4–7]. The assumption of these models is that the network represents different states by different, usually uncorrelated, global patterns of persistent activity. When the network is presented with an input that closely resembles one of the stored states, the network state switches to the corresponding fixed-point attractor.

This process of switching between discrete attractor states is thought to be fundamental both to describe biological neural activity, as well as to model higher cognitive decision making processes [8–12]. What attractor models currently lack, however, is the ability to perform state-dependent computation, a hallmark of human cognition [13–15]. That is, when the network is presented with an input, the attractor state to which the network switches ought to be dependent both upon the input stimulus as well as the state the network currently inhabits, rather than simply the input.

We thus seek to endow a classical neural attractor model, the Hopfield network, with the ability to perform state-dependent switching between attractor states, without resorting to the use of biologically implausible mechanisms, such as higher-order weight tensors or training via back-propagation algorithms. The resulting attractor networks will then be able to robustly emulate any arbitrary Finite State Machine (FSM), vastly improving their usefulness as a neural computational primitive.

We achieve this by leaning heavily on the framework of Vector Symbolic Architectures (VSAs). VSAs treat computation in an entirely distributed manner, by letting symbols be represented by high-dimensional random vectors: hypervectors [16–18]. When equipped with a few basic operators for binding and superimposing vectors together, corresponding often either to element-wise multiplication or addition respectively, these architectures are able to store primitives such as sets, sequences, graphs and arbitrary data bindings, as well as enabling more complex relations, such as analogical and figurative reasoning [19, 20]. Although different VSA implementations often have differing representations and binding operations [21], they all share the need for an auto-associative cleanup memory, which can recover a clean version of the most similar stored hypervector, given a noisy version of itself. We here use the recurrent dynamics of a Hopfield-like neural attractor network as a state-holding auto-associative memory [22].

Symbolic FSM states will thus be represented each by a hypervector and stored within the attractor network as a fixed-point attractor. Stimuli will also be represented by hypervectors, which, when input to the attractor network, will trigger the network dynamics to transition between the correct attractor states. We make use of common VSA techniques to construct a weights matrix to achieve these dynamics, where we use the Hadamard product between bipolar vectors  $\{-1, 1\}^N$  as our binding operation. We thus claim that attractor-based FSMs may be a plausible biological

computational primitive insofar as Hopfield networks are.

This represents a computational paradigm that is a departure from conventional von Neumann architectures, wherein the separation of memory and computation is a major limiting factor in current advances in conventional computational performance (the von Neumann bottleneck [23, 24]). Similarly, the high redundancy and lack of reliance on individual components makes this architecture fit for implementation with novel in-memory computing technologies such as resistive RAM (RRAM) or phase change memory (PCM) devices, which could perform the network’s matrix-vector-multiplication (MVM) in a single step [25–27].

## II. THEORY

Throughout this paper, symbols will be represented by high-dimensional randomly generated dense bipolar vectors

$$\mathbf{x} \in \{-1, 1\}^N \quad (1)$$

where the number of dimensions  $N > 10,000$ . Unless explicitly stated otherwise, any bold lowercase Latin letter may be assumed to be a new, independently generated hypervector, with the value  $X_i$  at any index  $i$  in  $\mathbf{x}$  generated according to

$$\mathbb{P}(X_i = 1) = \mathbb{P}(X_i = -1) = \frac{1}{2} \quad (2)$$

For any two arbitrary hypervectors  $\mathbf{a}$  and  $\mathbf{b}$ , we define the similarity between the two vectors by the normalised inner product

$$d(\mathbf{a}, \mathbf{b}) := \frac{1}{N} \mathbf{a}^\top \mathbf{b} = \frac{1}{N} \sum_{i=1}^N a_i b_i \quad (3)$$

where the similarity between a vector and itself  $d(\mathbf{a}, \mathbf{a}) = 1$ , and  $d(\mathbf{a}, -\mathbf{a}) = -1$ . Due to the high dimensionality of the vectors, the similarity between any two unrelated (and so independently generated) vectors is the mean of an unbiased random sequence of  $-1$  and  $1$ s

$$d(\mathbf{a}, \mathbf{b}) = 0 \pm \frac{1}{\sqrt{N}} \approx 0 \quad (4)$$

which tends to 0 for  $N \rightarrow \infty$ . It is from this result that we get the requirement of high dimensionality, as it ensures that the inner product between two random vectors is approximately 0. We can say that independently generated vectors are *pseudo-orthogonal* [20]. For a set of independently generated states  $\{\mathbf{x}^\mu\}$ , these results can be summarised by

$$d(\mathbf{x}^\mu, \pm \mathbf{x}^\nu) \stackrel{N \rightarrow \infty}{\approx} \pm \delta^{\mu\nu} \quad (5)$$

where  $\delta^{\mu\nu}$  is the Kronecker delta. Hypervectors may be combined in a so called *binding* operation to produce a new vector that is dissimilar to both its constituents. We here choose the Hadamard product, or element-wise multiplication, as our binding operation, denoted “ $\circ$ ”.

$$(\mathbf{a} \circ \mathbf{b})_i = a_i \cdot b_i \quad (6)$$

The statement that the product of two vectors is dissimilar to its constituents is written as

$$\begin{aligned} d(\mathbf{a} \circ \mathbf{b}, \mathbf{a}) &\approx 0 \\ d(\mathbf{a} \circ \mathbf{b}, \mathbf{b}) &\approx 0 \end{aligned} \quad (7)$$

where we implicitly assume that  $N$  is large enough that we can ignore the  $\mathcal{O}(\frac{1}{\sqrt{N}})$  noise terms.

If we wish to recover similarity between  $\mathbf{a} \circ \mathbf{b}$  and  $\mathbf{a}$ , we can mask the system using  $\mathbf{b}$ , such that only dimensions where  $b_i = 1$  are remaining. Then, we have

$$\begin{aligned} d(\mathbf{a} \circ \mathbf{b}, \mathbf{a} \circ H(\mathbf{b})) &= \frac{1}{N} \sum_{1 \leq i \leq N} a_i b_i a_i H(b_i) \\ &= \frac{1}{N} \left[ \sum_{\substack{1 \leq i \leq N \\ b_i = 1}} a_i^2 H(1) - \sum_{\substack{1 \leq i \leq N \\ b_i = -1}} a_i^2 H(-1) \right] \\ &= \frac{1}{N} \sum_{\substack{1 \leq i \leq N \\ b_i = 1}} 1 \\ &\approx \frac{1}{2} \end{aligned} \quad (8)$$

where we have used the Heaviside step function  $H(\cdot)$  defined by

$$(H(\mathbf{b}))_i = H(b_i) = \begin{cases} 1 & \text{if } b_i > 0 \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

to create a multiplicative mask  $H(\mathbf{b})$ , setting to 0 all dimensions where  $b_i = -1$ . In the second line, we have split the summation over all dimensions into summations over dimensions where  $b_i = 1$  and  $-1$  respectively. The final similarity of  $\frac{1}{2}$  is a consequence of approximately half of all values in a any vector being  $+1$  (Equation 2).

We choose this as a mechanism for recovering similarity, rather than simply applying another Hadamard multiplication  $\mathbf{b} \circ$ , as it is an operation that can easily and robustly be realised in a neural attractor network with asynchronous updates, as discussed later.

### A. Hopfield networks

A Hopfield network is a dynamical system defined by its internal state vector  $\mathbf{z}$  and fixed recurrent weights matrix  $\mathbf{W}$ , with a state update rule given by

$$\mathbf{z}_{t+1} = \text{sgn}(\mathbf{W}\mathbf{z}_t) \quad (10)$$

where  $\mathbf{z}_t$  is the network state at discrete time step  $t$ , and  $\text{sgn}(\cdot)$  is an element-wise sign function, with zeroes resolving<sup>1</sup> to +1.

From standard Hopfield theory, we know that if we want to store  $P$  uncorrelated patterns  $\{\mathbf{x}^\nu\}_{\nu=1}^P$  within a Hopfield network, we can construct the weights matrix  $\mathbf{W}$  according to

$$\mathbf{W} = \sum_{\text{patterns } \nu}^P \mathbf{x}^\nu \mathbf{x}^{\nu\top} \quad (11)$$

then as long as not too many patterns are stored ( $P < 0.14N$  [1]), the patterns will become fixed-point attractors of the network's dynamics, and the network can perform robust pattern completion and correction.

### B. Finite State Machines

A Finite State Machine (FSM)  $M$  is a discrete system with a finite state set  $Z_{\text{FSM}} = \{\zeta_1, \zeta_2, \dots, \zeta_{N_Z}\}$ , a finite input stimulus set  $S_{\text{FSM}} = \{\sigma_1, \sigma_2, \dots, \sigma_{N_S}\}$ , and finite output response set  $R_{\text{FSM}} = \{\rho_1, \rho_2, \dots, \rho_{N_R}\}$ .  $M$  is then fully defined with the addition of its two characterising functions  $F(\cdot)$  and  $G(\cdot)$

$$\begin{aligned} z_{t+1} &= F(z_t, s_t) \\ r_{t+1} &= G(z_t, s_t) \end{aligned} \quad (12)$$

where  $z_t \in Z_{\text{FSM}}$ ,  $r_t \in R_{\text{FSM}}$  and  $s_t \in S_{\text{FSM}}$  are the state, output and stimulus at time step  $t$  respectively.  $F(\cdot)$  thus provides the state update rule, while  $G(\cdot)$  provides the output for any state-stimulus pair.

$M$  can thus be represented by a directed graph, where each node represents a different state  $\zeta$ , and every edge has a stimulus  $\sigma$  and optional output  $\rho$  associated with it.

## III. ATTRACTOR NETWORK CONSTRUCTION

We now show how a Hopfield-like attractor network may be constructed to emulate an arbitrary FSM, where the states within the FSM are now attractors within the attractor network, and the stimuli for transitions between node states in the FSM trigger all corresponding transitions between attractors. More specifically, for every node  $\zeta \in Z_{\text{FSM}}$ , an

<sup>1</sup>Though this arbitrary choice may seem to incur a bias to a particular state, in practise the postsynaptic sum very rarely equals 0.

associated hypervector  $\mathbf{x}$  is randomly generated and stored as an attractor within the attractor network. We use  $Z_{\text{AN}}$  to denote the set of nodal hypervectors stored as attractors within the attractor network. Every unique stimulus  $\sigma \in S_{\text{FSM}}$  in the FSM is also now associated with a randomly generated hypervector  $\mathbf{s} \in S_{\text{AN}}$ , where  $S_{\text{AN}}$  is the set of hypervectors associated with a unique stimulus. For the FSM edge outputs, a corresponding set of output hypervectors is similarly generated. These correspondences are summarised in Table I.

FSM (Symbols)		Attractor Net. (Vectors)	
Nodes	$\zeta \in Z_{\text{FSM}}$	Attractors	$\mathbf{x} \in Z_{\text{AN}}$
Input stimuli	$\sigma \in S_{\text{FSM}}$	Input stimuli	$\mathbf{s} \in S_{\text{AN}}$
Outputs	$\rho \in R_{\text{FSM}}$	Output vectors	$\mathbf{r} \in R_{\text{AN}}$

TABLE I: A comparison of the notation used to represent states, inputs and outputs in the FSM picture, and the corresponding hypervectors used to represent the FSM within the attractor network.

### A. Constructing transitions

We consider the general situation that we want to initiate a transition from *source* attractor state  $\mathbf{x} \in Z_{\text{AN}}$  to *target* attractor state  $\mathbf{y} \in Z_{\text{AN}}$ , by imposing some stimulus state  $\mathbf{s} \in S_{\text{AN}}$  as input onto the network.

$$\mathbf{x} \xrightarrow{\mathbf{s}} \mathbf{y} \quad (13)$$

Crucial to the functioning of the network transitions is how we model input to the network. We choose to model input to the network as a masking of the network state, such that all dimensions where the stimulus  $\mathbf{s}$  is -1 are forced to be 0. This may be likened to saying we are considering input to the network that selectively silences half of all neurons according to the stimulus vector. While a stimulus vector  $\mathbf{s}$  is being imposed upon the network, the modified state update rule is thus

$$\mathbf{z}_{t+1} = \text{sgn}(\mathbf{W}(\mathbf{z}_t \circ H(\mathbf{s}))) \quad (14)$$

where the Hadamard multiplication of the network state with  $H(\mathbf{s})$  enacts the masking operation, and the weights matrix  $\mathbf{W}$  is constructed such that  $\mathbf{z}_{t+1}$  will resemble the desired target state.

For every edge in the FSM, we generate an "edge state"  $\mathbf{e}$ , which is also stored as an attractor within the network. Each edge will use this  $\mathbf{e}$  state as a "halfway-house", en route to  $\mathbf{y}$ . Additionally, each unique edge label will now have *two* stimulus hypervectors associated with it,  $\mathbf{s}_a$  and  $\mathbf{s}_b$  which trigger transitions from source state  $\mathbf{x}$  to edge state  $\mathbf{e}$  and edge state  $\mathbf{e}$  to target state  $\mathbf{y}$  respectively. A general transition now looks like

$$\mathbf{x} \xrightarrow{\mathbf{s}_a} \mathbf{e} \xrightarrow{\mathbf{s}_b} \mathbf{y} \quad (15)$$

where  $\mathbf{x}, \mathbf{y} \in Z_{AN}$  correspond to nodal states in the FSM but  $\mathbf{e}$  exists purely to facilitate the transition. The weights matrix is constructed as

$$\mathbf{W} = \underbrace{\frac{1}{N} \sum_{\text{nodes } \nu} \mathbf{x}^\nu \mathbf{x}^{\nu\top}}_{\text{Hopfield attractor terms}} + \underbrace{\frac{1}{N} \sum_{\text{edges } \eta} \mathbf{E}^\eta}_{\text{Asymmetric transition terms}} \quad (16)$$

where  $\mathbf{x}^\nu \in Z_{AN}$  is the state corresponding to the  $\nu$ 'th node in the graph to be implemented,  $N_Z$  and  $N_E$  are the number of nodes and edges respectively, and  $\mathbf{E}^\eta$  is the addition to the weights matrix required to implement an individual edge, given by

$$\begin{aligned} \mathbf{E}^{(\eta)} &= \mathbf{e}\mathbf{e}^\top \\ &+ H(\mathbf{s}_a) \circ (\mathbf{e} - \mathbf{x})(\mathbf{x} \circ \mathbf{s}_a)^\top \\ &+ H(\mathbf{s}_b) \circ (\mathbf{y} - \mathbf{e})(\mathbf{e} \circ \mathbf{s}_b)^\top \end{aligned} \quad (17)$$

where  $\mathbf{x}$ ,  $\mathbf{e}$  and  $\mathbf{y}$  are the source, edge, and target states of the edge  $\eta$  respectively, and  $\mathbf{s}_a$  and  $\mathbf{s}_b$  are the input stimulus vectors associated with this edge's label. The edge index  $\eta$  has been dropped for brevity. The  $\mathbf{e}\mathbf{e}^\top$  term is the attractor we have introduced as a halfway-house for the transition. The second set of terms enacts the  $\mathbf{x} \xrightarrow{\mathbf{s}_a} \mathbf{e}$  transition, by giving a nonzero inner product with the network state only when the network is in state  $\mathbf{x}$ , and the network is being masked by the input  $\mathbf{s}_a$ . This allows terms to be stored in  $\mathbf{W}$  which are effectively obfuscated, not affecting network dynamics considerably, until a specific stimulus is applied as a mask to the network. Likewise, the third set of terms enacts the  $\mathbf{e} \xrightarrow{\mathbf{s}_b} \mathbf{y}$  transition.

In the absence of input, the network functions like a standard Hopfield attractor network,

$$\mathbf{W}\mathbf{x} \approx \mathbf{x} \pm \sigma \mathbf{n} \quad \forall \quad \mathbf{x} \in Z_{AN} \quad (18)$$

where  $\mathbf{n} \in \mathbb{R}^N$  is a standard normally distributed random vector, and

$$\sigma = \sqrt{\frac{N_Z + 3N_E}{N}} \quad (19)$$

is the magnitude of noise due to the undesired finite inner product with other stored terms (see Appendix). Thus as long as the magnitude of the noise is not too large,  $\mathbf{x}$  will be a solution of  $\mathbf{z} = \text{sgn}(\mathbf{W}\mathbf{z})$  and so a fixed-point attractor of the dynamics.

When a valid stimulus is presented as input to the network however, masking the network state, the previously obfuscated asymmetric transition terms become significant and dominate the dynamics. Assuming there is a stored transition term  $\mathbf{E}$  corresponding to a valid edge with vectors

$\mathbf{x}, \mathbf{e}, \mathbf{y}, \mathbf{s}_a, \mathbf{s}_b$  having the same meaning as in Equation 17, we have

$$\mathbf{W}(\mathbf{x} \circ H(\mathbf{s}_a)) \approx H(\mathbf{s}_a) \circ \mathbf{e} + H(-\mathbf{s}_a) \circ \mathbf{x} \pm \sqrt{2}\sigma \mathbf{n} \quad (20)$$

where  $\approx$  implies approximate proportionality. The second set of terms can be ignored, as they project only to neurons which are currently being masked. Thus the only significant term is that containing the edge state  $\mathbf{e}$ , which consequently drives the network to the  $\mathbf{e}$  state, enacting the  $\mathbf{x} \xrightarrow{\mathbf{s}_a} \mathbf{e}$  transition. Since the state  $\mathbf{e}$  is also stored as an attractor within the network, we have

$$\mathbf{W}(\mathbf{e} \circ H(\mathbf{s}_a)) \approx \mathbf{e} \pm \sqrt{2}\sigma \mathbf{n} \quad (21)$$

and

$$\mathbf{W}\mathbf{e} \approx \mathbf{e} \pm \sigma \mathbf{n} \quad (22)$$

thus the edge states  $\mathbf{e}$  are also fixed-point attractors of the network dynamics. To complete the transition from state  $\mathbf{x}$  to  $\mathbf{y}$ , the second stimulus  $\mathbf{s}_b$  is applied, giving

$$\mathbf{W}(\mathbf{e} \circ H(\mathbf{s}_b)) \approx H(\mathbf{s}_b) \circ \mathbf{y} + H(-\mathbf{s}_b) \circ \mathbf{e} \pm \sqrt{2}\sigma \mathbf{n} \quad (23)$$

which drives the network state towards  $\mathbf{y} \in Z_{AN}$ , the desired target attractor state. By consecutive application of the inputs  $\mathbf{s}_a$  and  $\mathbf{s}_b$ , the transition terms  $\mathbf{E}^\eta$  stored in  $\mathbf{W}$  have thus caused the network to controllably transition from the source to target attractor states. Transition terms  $\mathbf{E}^\eta$  may be iteratively added to  $\mathbf{W}$  to achieve any arbitrary transition between attractor states, and so any arbitrary FSM may be implemented within a large enough attractor network.

Note that we have here ignored that the diagonal of  $\mathbf{W}$  is set to 0 (no self connections), but this does not significantly affect these results.

## B. Edge outputs

Until now we have not mentioned the other critical component of FSMs: the output associated with every edge. We have separated the construction of transitions and edge outputs for clarity, since the two may be effectively decoupled. Much like for the nodes and edges in the FSM to be implemented, for every unique FSM output  $\rho \in R_{\text{FSM}}$ , we generate a corresponding hypervector  $\mathbf{r} \in R_{AN}$ , where  $R_{AN}$  is the set of all output vectors. Different however, is that we let these be sparse ternary vectors  $\mathbf{r} \in \{-1, 0, 1\}^N$  with coding level  $f_r := \frac{1}{N} \sum_i |r_i|$ , the fraction of nonzero elements. These output states are then embedded in the edge state attractors, altering the  $\mathbf{e}\mathbf{e}^\top$  terms in each  $\mathbf{E}$  term according to

$$\mathbf{e}\mathbf{e}^\top \rightarrow \mathbf{e}_r \mathbf{e}_r^\top := \left[ \mathbf{e} \circ (\mathbf{1} - H(\mathbf{r} \circ \mathbf{r})) + \mathbf{r} \right] \mathbf{e}^\top \quad (24)$$

where  $\mathbf{e}_r$  is here defined and  $\mathbf{1}$  is a vector of all ones. As a result of this modification, the edge states  $\mathbf{e}$  themselves will no longer be exact attractors of the space. The composite state  $\mathbf{e}_r$  will however be stable, in which the presence of  $\mathbf{r}$  can be easily detected ( $\mathbf{e}_r \cdot \mathbf{r} = Nf_r$ ). This has been achieved without incurring any similarity and thus interference between attractors, which would otherwise alter the dynamics of the previously described transitions. A full transition term  $\mathbf{E}$ , including its output, is thus given by

$$\begin{aligned} \mathbf{E}^{(n)} = & \left[ \mathbf{e} \circ (\mathbf{1} - H(\mathbf{r} \circ \mathbf{r})) + \mathbf{r} \right] \mathbf{e}^\top \\ & + H(\mathbf{s}_a) \circ (\mathbf{e} - \mathbf{x})(\mathbf{x} \circ \mathbf{s}_a)^\top \\ & + H(\mathbf{s}_b) \circ (\mathbf{y} - \mathbf{e})(\mathbf{e} \circ \mathbf{s}_b)^\top \end{aligned} \quad (25)$$

which combined with the network state masking operation is solely responsible for storing the FSM connectivity and enabling the desired inter-attractor transition dynamics.

#### IV. RESULTS

##### A. FSM Emulation

To show the generality of FSM construction, we chose to implement a directed graph representing the relationships between gods in ancient Greek mythology, due to the graph's dense connectivity. The graph and thus FSM to be implemented is shown in Figure 1. From the graph it is clear that a state machine representing the graph must explicitly be capable of state-dependent transitions, e.g. the input "overthrown\_by" must result in a transition to state "Kronos" when in state "Uranus", but to state "Zeus" when in state "Kronos". To construct  $\mathbf{W}$ , the necessary random hypervectors are first generated. For every node state  $\zeta \in Z_{\text{FSM}}$  within the FSM (e.g. "Zeus", "Kronos") a random bipolar vector  $\mathbf{x}$  is generated. For every unique edge label  $\sigma \in S_{\text{FSM}}$  (e.g. "overthrown\_by", "father\_is") a pair of random stimulus hypervectors  $\mathbf{s}_a$  and  $\mathbf{s}_b$  is generated. Similarly, sparse ternary output vectors  $\mathbf{r}$  are also generated. The weights matrix  $\mathbf{W}$  is then iteratively constructed as per Equations 16 and 25, with a new hypervector  $\mathbf{e}$  also being generated for every edge. The matrix generated from this procedure we denote  $\mathbf{W}^{\text{ideal}}$ . For all of the following results, the attractor network is first initialised to be in a certain attractor state, in this case, "Hades". The network is then allowed to freely evolve for 10 time steps (chosen arbitrarily) as per Equation 10, with every neuron being updated simultaneously on every time step. During this period, it is desired that the network state  $\mathbf{z}_t$  remains in the attractor state in which it was initialised. An input stimulus  $\mathbf{s}_a$  is then presented to the network for 10 time steps, during which time the network state is masked by the stimulus vector, and the network evolves synchronously according to Equation 14. If the stimulus corresponds to a valid

edge in the FSM, the network state  $\mathbf{z}_t$  should then be driven towards the correct edge state attractor  $\mathbf{e}$ . After these 10 time steps, the second stimulus vector  $\mathbf{s}_b$  for a particular input is presented for 10 time steps. Again, the network evolves according to Equation 14, and the network should be driven towards the target state attractor  $\mathbf{y}$ , completing the transition. This process is repeated every 30 time steps, causing the network state  $\mathbf{z}_t$  to travel between attractor states  $\mathbf{x} \in Z_{\text{AN}}$ , corresponding to a valid walk between states  $\zeta \in Z_{\text{FSM}}$  in the represented FSM. To view the resulting network dynamics, the similarity between the network state  $\mathbf{z}_t$  and the edge and node states is calculated as per Equation 3, such that a similarity of 1 between  $\mathbf{z}_t$  and some attractor state  $\mathbf{x}'$  implies  $\mathbf{z}_t = \mathbf{x}'$  and thus that the network is inhabiting that attractor. The similarity between the network state  $\mathbf{z}_t$  and the outputs states  $\mathbf{r} \in R_{\text{AN}}$  is also calculated, but due to the output vectors being sparse, the maximum value that the similarity can take is  $d(\mathbf{z}, \mathbf{r}) = f_r$ , which would be interpreted as that output symbol being present.

An attractor network performing a walk is shown in Figure 2, with parameters  $N = 10,000$ ,  $Nf_r = 200$ ,  $N_Z = 8$ , and  $N_E = 16$ . This corresponds to the network having a per-neuron noise (the finite size effect resulting from random hypervectors having a nonzero similarity to each-other) of  $\sigma \approx 0.07$ , calculated via Equation 19. The magnitude of the noise is thus small compared with the desired signal of magnitude 1 (Equation 18), and so we are far away from reaching the memory capacity of the network. The network performs the walk as intended, transitioning between the correct nodal attractor states and corresponding edge states with their associated outputs. The specific sequence of inputs was chosen to show the generality of implementable state transitions. First, there is the explicit state dependence in the repeated input of "father\_is, father\_is". Second, it contains an input stimulus that does not correspond to a valid edge for the currently inhabited state ("Zeus overthrown\_by"), which should not cause a transition. Third, it contains bidirectional edges ("consort\_is"), whose repeated application causes the network to flip between two states (between "Kronos" and "Rhea"). And fourthly self-connections, whose target states and source states are identical. Since the network traverses all these edges as expected, we do not expect the precise structure of an FSM's graph to limit whether or not it can be emulated by the attractor network.

##### B. Network robustness

One of the advantages of attractor neural networks that make them suitable as plausible biological models is their robustness to imperfect weights [2]. That is, individual synapses may have very few bits of precision or become damaged, yet the relevant brain region must still be able to carry out its

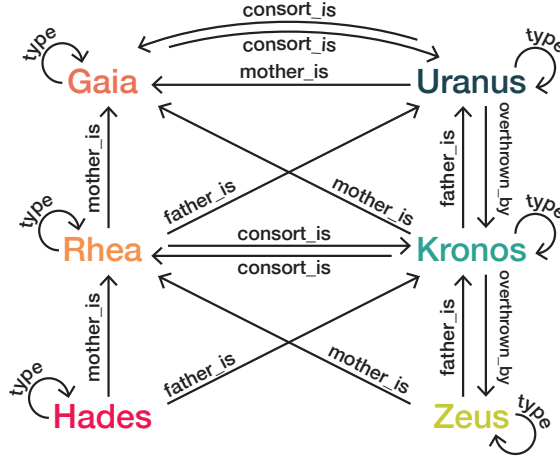


Fig. 1: An example FSM which we implement within the attractor network. Each node within the graph (e.g. "Zeus") is represented by a new hypervector  $\mathbf{x}^\mu$  and stored as an attractor within the network. Every edge is labelled by its stimulus (e.g. "father\_is"), for which corresponding hypervectors  $\mathbf{s}_a$  and  $\mathbf{s}_b$  are also generated. When a stimulus' hypervector is input to the network, it should allow all corresponding attractor transitions to take place. Each edge may also have an associated output symbol, where we here choose the edges labelled "type" to output the generation of the god {"Primordial", "Titans", "Olympians"}. This graph was chosen as it displays the generality of the embedding: it contains cycles, loops, bidirectional edges and state-dependent transitions.

functional task. To this end, we subject the network presented here to similar non-idealities, to check that the network retains the feature of global stability and robustness despite being implemented with low-precision and noisy weights. In the first of these tests, the ideal weights matrix  $\mathbf{W}^{\text{ideal}}$  was binarised and then additive noise was applied, via

$$W_{ij}^{\text{noisy}} = \text{sgn}(W_{ij}^{\text{ideal}}) + \sigma_{\text{noise}} \cdot \chi_{ij} \quad (26)$$

where  $\chi_{ij} \in \mathbb{R}$  are independently sampled standard Gaussian variables, sampled once during matrix construction, and  $\sigma_{\text{noise}} \in \mathbb{R}$  is a scaling factor on the strength of noise being imposed. The  $\text{sgn}(\cdot)$  function forces the weights to be bipolar, emulating that the synapses may have only 1 bit of precision, while the  $\chi_{ij}$  random variables act as a smearing on the weight state, emulating that the two weight states have a finite width. A  $\sigma_{\text{noise}}$  value of 2 thus corresponds to the magnitude of the noise being equal to that of the signal (whether  $W_{ij}^{\text{ideal}} \geq 0$ ), and so, for example, for a damaged weight value of  $W_{ij}^{\text{noisy}} = +1$  there is a 38% chance that the pre-damaged weight  $W_{ij}^{\text{ideal}} = -1$ . This level of degradation is far worse than is expected even from novel binary memory devices [25], and presumably also for biology. We used the same set of hypervectors and sequence of inputs as in Figure 2, but this time using the degraded weights matrix  $\mathbf{W}^{\text{noisy}}$ , to test the network's robustness. The results are shown in Figure 3 for weight degradation values of  $\sigma_{\text{noise}} = 2$

and  $\sigma_{\text{noise}} = 5$ . We see that for  $\sigma_{\text{noise}} = 2$  the attractor network performs the walk just as well as in Figure 2, which used the ideal weights matrix, despite the fact that here the binary weight distributions overlap each-other considerably. Furthermore, we have that  $d(\mathbf{z}_t, \mathbf{x}^\nu) \approx 1$  where  $\mathbf{x}^\nu$  is the attractor that the network should be inhabiting at any time, indicating that the attractor stability and recall accuracy is unaffected by the non-idealities. For  $\sigma_{\text{noise}} = 5$ , a scenario where the realised weight carries very little information about the ideal weight's value, we see that the network nonetheless continues to function, performing the correct walk between attractor states. However, there is a degradation in the recall of stored states, with the network state no longer converging to a similarity of 1 with the stored attractor states. For greater values of  $\sigma_{\text{noise}}$ , the network ceases to perform the correct walk, and indeed does not converge on any stored attractor state (not shown).

A further test of robustness was to restrict the weights matrix to be sparse, as a dense all-to-all connectivity may not be feasible in biology, where synaptic connections are spatially constrained and have an associated chemical cost. Similar to the previous test, the sparse weights matrix was generated via

$$W_{ij}^{\text{sparse}} = \text{sgn}(W_{ij}^{\text{ideal}}) \cdot H(|W_{ij}| - \theta) \quad (27)$$

where  $\theta$  is a threshold set such that  $\mathbf{W}^{\text{sparse}} \in$

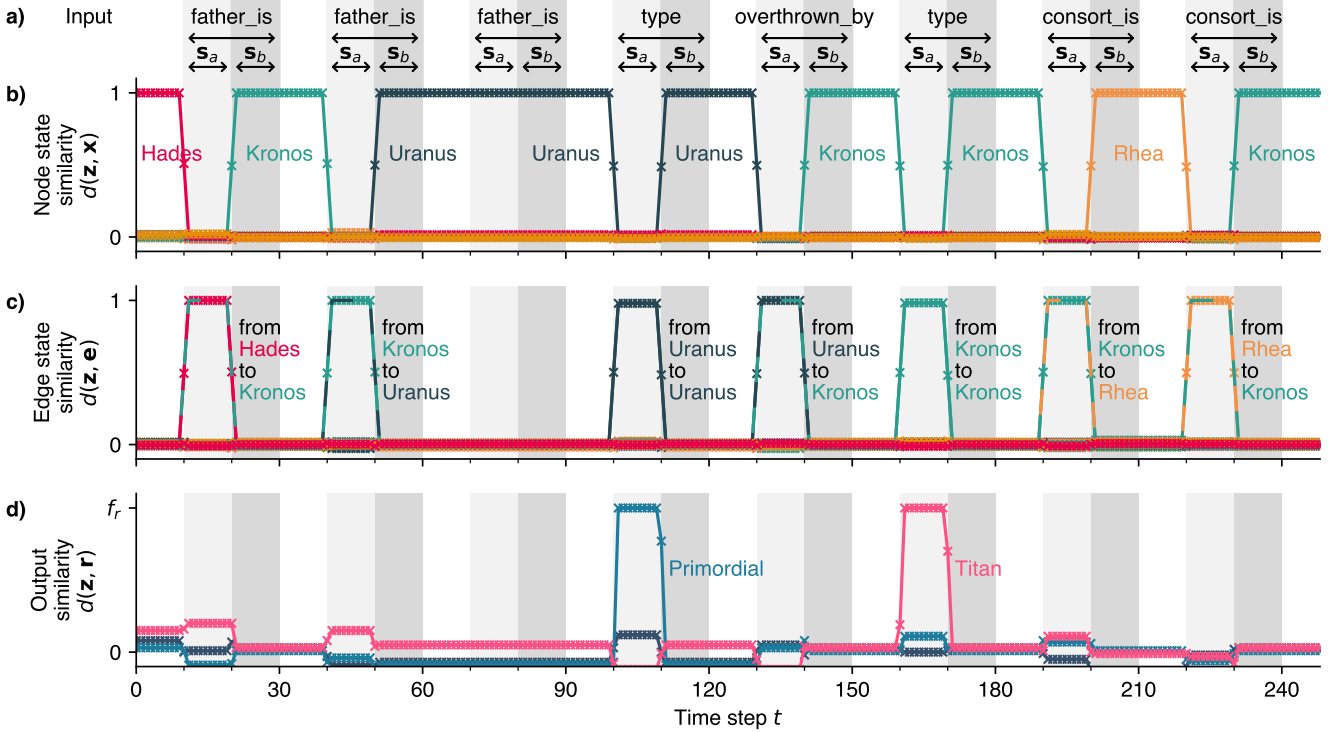


Fig. 2: An attractor network transitioning through attractor states in a state-dependent manner, as a sequence of input stimuli is presented to the network. **a)** The input stimuli to the network, where for each unique edge label (e.g. "father\_is") in the FSM to be implemented (Figure 1) a pair of hypervectors  $s_a$  and  $s_b$  have been generated. No stimulus, a stimulus  $s_a$ , then a stimulus  $s_b$  are input for 10 time steps each in sequence. **b)** & **c)** The similarity of the network state  $\mathbf{z}_t$  to stored node states  $\mathbf{x} \in Z_{AN}$  and stored edge states  $\mathbf{e}$  respectively, computed via the inner product (Equation 3). **d)** The similarity of the network state  $\mathbf{z}_t$  to the sparse output states  $\mathbf{r} \in R_{AN}$ . All similarities have been labelled with the state they represent and the colours are purely illustrative. The state transitions shown here are explicitly state dependent, as can be seen from the repeated input of "father\_is", which results in a transition to state "Kronos" when in "Hades", but to "Uranus" when in "Kronos". Additionally, the network is unaffected by nonsense input that does not correspond to a stored edge, as the network remains in the attractor "Uranus" when presented with the input "father\_is".

$\{-1, 0, 1\}^{N \times N}$  has the desired sparsity. Through this procedure, only the most extreme weight values are allowed to be nonzero. Since the terms inside  $\mathbf{W}^{\text{ideal}}$  are symmetrically distributed around 0, there are approximately as many +1 entries in  $\mathbf{W}^{\text{sparse}}$  as -1s. Using the same hypervectors and sequence of inputs as before, an attractor network performing a walk using the sparse weights matrix  $\mathbf{W}^{\text{sparse}}$  is shown in Figure 4, with sparsities of 98% and 99%. We see that for the 98% sparse case, there is again very little difference with the ideal case shown in Figure 2, with the network still having a similarity of  $d(\mathbf{z}_t, \mathbf{x}) \approx 1$  with stored attractor states, and performing the correct walk. When the sparsity is pushed further to 99% however, we see that despite the network performing the correct walk, the attractor states are again

slightly degraded, with the network converging on states with  $d(\mathbf{z}_t, \mathbf{x}^\nu) < 1$  with stored attractor states  $\mathbf{x}^\nu$ . For greater sparsities, the network ceases to perform the correct walk, and again does not converge on any stored attractor state (not shown).

These two tests thus highlight the extreme robustness of the model to imprecise and unreliable weights. The network may be implemented with 1 bit precision weights, whose weight distributions are entirely overlapping, or set 98% of the weights to 0, and still continue to function without any discernible loss in performance. The extent to which the weights matrix may be degraded and the network still remain stable is of course a function not only of the level of degradation, but also of the size of the network  $N$ , as



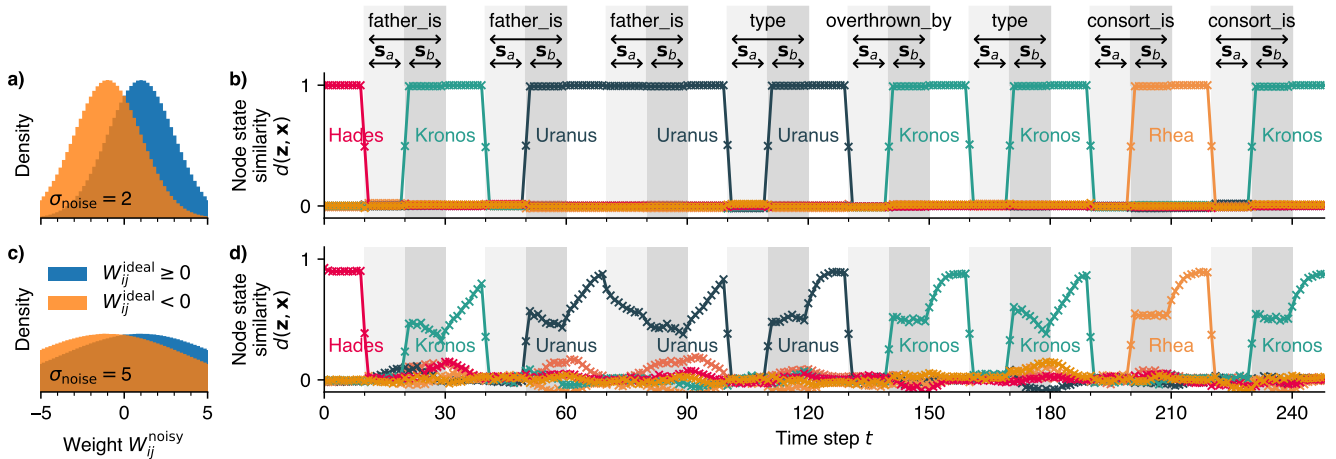


Fig. 3: The attractor network performing a walk as in Figure 2, but using the damaged weights matrix  $\mathbf{W}^{\text{noisy}}$ , whose entries have been binarised and then independent additive noise has been applied, as per Equation 16. **a)** The distribution of weights after they have been thusly damaged with noise of magnitude  $\sigma_{\text{noise}} = 2$ . Weights whose ideal values were positive or negative have been plotted separately. **b)** The similarity of the network state  $\mathbf{z}_t$  to stored node states, with the network using the weights from a). Shown above is the sequence of inputs given to the network, identical to in Figure 2. **c)** The distribution of weights damaged with  $\sigma_{\text{noise}} = 5$ . **d)** The similarity of the network state to stored node states, but with the network using the damaged weights from c). The network transitions are thus highly robust to unreliable weights, and show a gradual degradation in performance, even when the network’s weights are majorly imprecise and noisy. For both b) and d) the edge state and output similarity plots have been omitted for visual clarity.

well as the the number of states  $N_Z$  and edges  $N_E$  stored within the network. For conventional Hopfield models with Hebbian learning, these two factors are normally theoretically treated alike, as contributing an effective noise to the postsynaptic sum as in Equation 19, and so the magnitude of withstandable synaptic noise increases with increasing  $N$  [2, 28]. Although a thorough mathematical investigation into the scaling of weight degradation limits is justified, as a first result we have here given numerical data showing stability even in the most extreme cases of non-ideal weights, and expect that any implementation of the network with novel devices would be far away from such extremities.

### C. Asynchronous updates

Another useful property of Hopfield networks is the ability to robustly function even with asynchronously updating neurons, wherein not every neuron experiences a simultaneous state update. This property is especially important for any architecture claiming to be biologically plausible, as biological neurons update asynchronously and largely independent of each-other, without the the need for global clock signals. To this end, we ran a similar experiment to that in Figure 2, using the undamaged weights matrix  $\mathbf{W}^{\text{ideal}}$ , but with an asynchronous neuron update rule, wherein on each time step every neuron has only a 10% chance of updating its state.

The remaining 90% of the time, the neuron retains its state from the previous time step, regardless of its postsynaptic sum. There is thus no fixed order of neuron updates, and indeed it is not even a certainty that a neuron will update in any finite time. To account for the slower dynamics of the network state, the time for which inputs were presented to the network, as well as the periods without any input, was increased from 10 to 40 time steps. To be able to easily view the gradual state transition, three of the attractor states were chosen to be columns of the  $N$ -dimensional Hadamard matrix, rather than being randomly generated. The results are shown in Figure 5, for a shorter sequence of stimulus inputs. We see that the network functions as intended, but with the network now converging on the correct attractors in a finite number of updates rather than in just one. The model proposed here is thus not reliant on synchronous dynamics, which is important not only for biological plausibility, but also when considering possible implementations on asynchronous neuromorphic hardware.

### D. Storage capacity

It is well known that the storage capacity of a Hopfield network, the number of patterns  $P$  that can be stored and reliably retrieved, is proportional to the size of the network, via  $P < 0.14N$  [1, 2]. When one tries to store more than  $P$

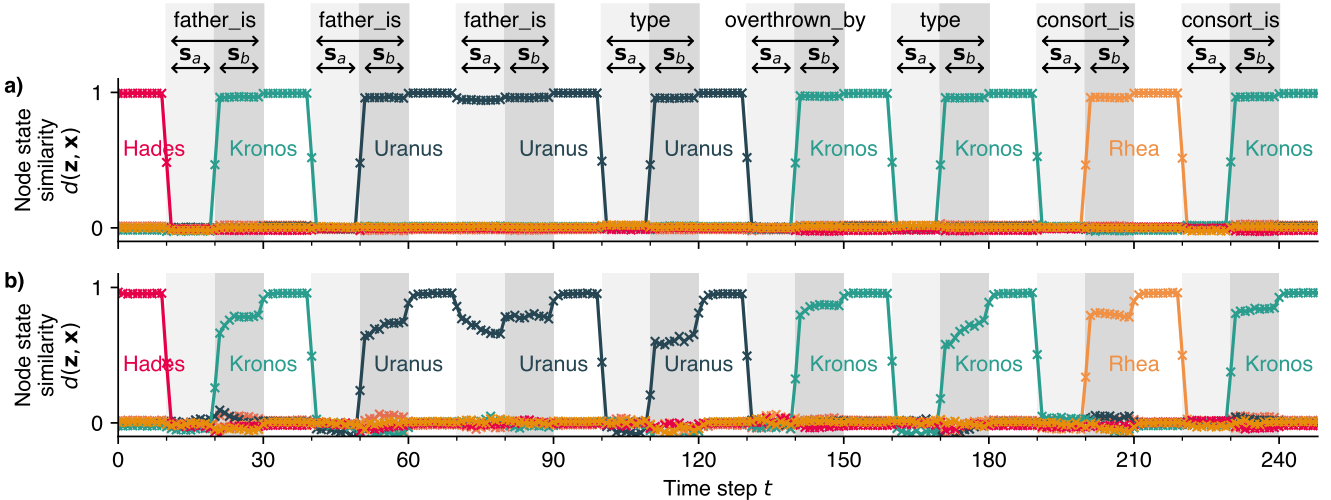


Fig. 4: The attractor network performing a walk as in Figure 2, but using a sparse ternary weights matrix  $\mathbf{W}^{\text{sparse}} \in \{-1, 0, 1\}^{N \times N}$ , generated via Equation 27. The weights matrices for **a)** and **b)** are 98% and 99% sparse respectively. Shown are the similarities of the network state  $\mathbf{z}_t$  with stored node states  $\mathbf{x}^\nu \in Z_{\text{AN}}$ , with the stimulus input hypervector to the network at any time shown above. We see that even when 98% of the entries in  $\mathbf{W}$  are zeroes, the network continues to function with negligible loss in stability, as the correct walk between attractor states is performed, and the network converges on stored attractors with similarity  $d(\mathbf{z}_t, \mathbf{x}) \approx 1$ . At 99% sparsity there is a degradation in the accuracy of stored attractors, with the network converging on states with  $d(\mathbf{z}_t, \mathbf{x}) < 1$ , but with the correct walk still being performed. Beyond 99% sparsity the attractor dynamics break down (not shown). Thus although requiring a large number of neurons  $N$  to enforce state pseudo-orthogonality, the network requires far fewer than  $N^2$  nonzero weights to function robustly.

attractors within the network, the so-called memory blackout occurs, after which no pattern can be retrieved. We thus perform numerical simulations for a large range of attractor network and FSM sizes, to see if an analogous relationship exists. Said otherwise, for an attractor network of finite size  $N$ , what sizes of FSM can the network successfully emulate?

For a given  $N$ , number of FSM nodes  $N_Z$  and edges  $N_E$ , a random FSM was generated and an attractor network constructed to represent it as described in Section III. To ensure a reasonable FSM was generated, the FSM's graph was first generated to have all nodes connected in a sequential ring structure, i.e. every state  $\zeta^\nu \in Z_{\text{FSM}}$  connects to  $\zeta^{\nu+1 \bmod N_Z}$ . The remaining edges between nodes were selected at random, until the desired number of edges  $N_E$  was reached. For each edge an associated stimulus is then required. Although one option would be to allocate as few unique stimuli as possible, so that the state transitions are maximally state dependent, this results in some advantageous cancellation effects between the  $\mathbf{E}^n$  transition terms and the stored attractors  $\mathbf{x}^\nu \mathbf{x}^{\nu T}$ . To instead probe a worst-case scenario, each edge was assigned a unique stimulus.

With the FSM now generated, an attractor network with  $N$  neurons was constructed as previously described. An initial

attractor state was chosen at random, and then a random valid walk between states was chosen to be performed (chosen arbitrarily to be of length 6, corresponding to each run taking 180 time steps). The corresponding sequence of stimuli was input to the attractor network via the same procedure as in Figure 2, each masking the network state. Each run was then evaluated to have either passed or failed, with a pass meaning that the network state inhabited the correct attractor state with overlap  $d(\mathbf{z}_t, \mathbf{x}^\nu) > 0.75$  in the middle of all intervals when it should be in a certain node attractor state. A pass thus corresponds to the network performing the correct walk between attractor states. The results are shown in Figure 6. We see that for a given  $N$ , there is a linear relationship between the the number of nodes  $N_Z$  and number of edges  $N_E$  in the FSM that can be implemented before failure. That this trade-off exists is not surprising, since both contribute additively to the SNR within the attractor network (Equation 19). For each  $N$ , a linear Support Vector Machine (SVM) was fitted to the data, to find the separating boundary at which failure and success of the walk are approximately equiprobable. The boundary is given by  $N_Z + \beta N_E = c(N)$ , where  $\beta$  represents the relative cost of adding nodes and edges, and  $c(N)$  is an offset. For all of the fitted boundaries,

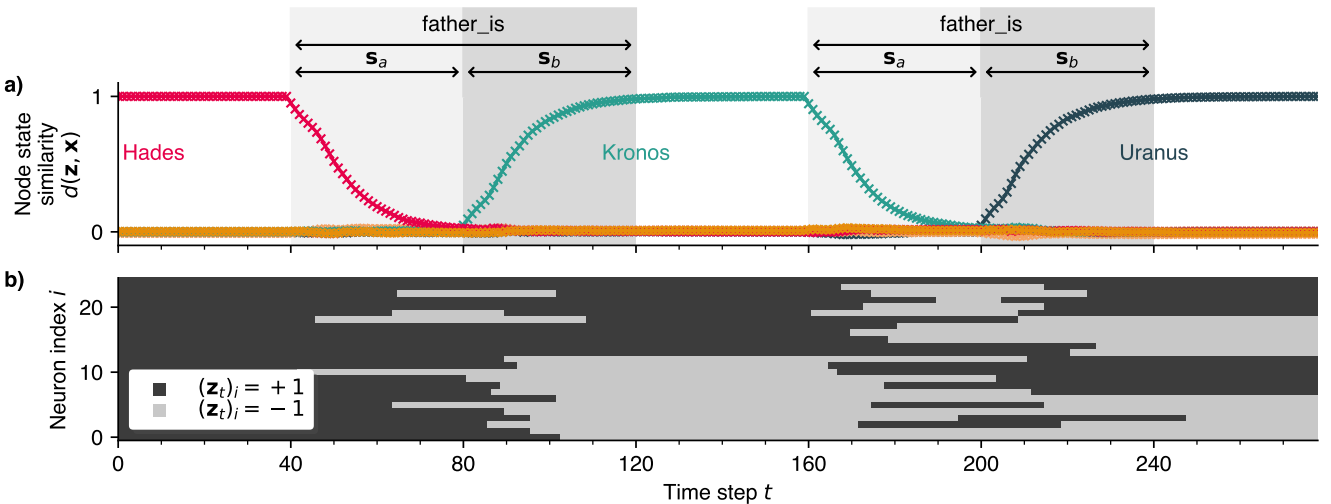


Fig. 5: An attractor network performing a shorter walk than in Figure 2, but where neurons are updated asynchronously, with each neuron having a 10% chance of updating on any time step. **a)** The similarity of the network state  $\mathbf{z}_t$  to stored attractor states, with the input stimuli to the network shown above. **b)** The evolution of a subset of neurons within the attractor network, where for visual clarity, three nodal states shown have hypervectors taken from columns of the  $N$ -dimensional Hadamard matrix, rather than being randomly generated. The network functions largely the same as in the synchronous case, but with transitions between states now taking a finite number of time steps to complete. The model is thus not dependent on the precise timing of neuron updates, and should function robustly in asynchronous systems where timing is unreliable.

the value of  $\beta$  was found to be approximately constant, with  $\beta = 2.4 \pm 0.1$ , and so is assumed to be independent of  $N$ . For every value of  $N$ , we define the capacity  $C$  to be the maximum size of FSM which can be implemented before failure, for which  $N_E = N_Z$ . The capacity  $C$  is then given by  $C(N) = \frac{c(N)}{1+\beta}$ , and is also plotted in Figure 6. A linear fit reveals an approximate proportionality relationship of  $C(N) \approx 0.025N$ . In all, the boundary which limits the size of FSM which can be emulated is then given by

$$N_Z + 2.4N_E < 0.085N \quad (28)$$

It is expected that additional edges consume more of the network's storage capacity than additional nodes, since for every edge, 5 additional terms are added to  $\mathbf{W}$  (Equation 25), contributing  $3\times$  as much cross-talk noise as adding a node would (Equation 19). We can compare this storage capacity relation with that of the standard Hopfield model, by considering the case  $N_E = 0$ , i.e. there are no transition terms in the network, and so the network is identical to a standard Hopfield network. In this case, our failure boundary would become  $N_Z < 0.085N$ , in comparison to Hopfield's  $P < 0.14N$ . Although this may seem like a drastic reduction in memory capacity, we must remember that as a result of input stimuli applying a masking operation to our network,

the actual size of the network during these periods is actually  $N' := \frac{1}{2}N$ . In this case, the failure boundary can be rewritten as  $N_Z < 0.17N'$ , which is more in keeping with the Hopfield estimate<sup>2</sup>.

## V. DISCUSSION

### A. Hardware implementability

Since the network described in this paper differs very little from the conventional Hopfield description, having changed only the prescription for the generation of the weights matrix and the specific way that input is modelled, we can lean heavily on previous works concerning VLSI implementations of Hopfield networks [29–31].

The main difficulty associated with an implementation would be the size requirements, requiring a high dimensionality  $N$  to ensure pseudo-orthogonality of randomly generated hypervectors. Although we may need  $N > 10^4$  neurons, we have shown this does not necessarily mean we need  $10^8$  synapses, as the network still functions when the weights matrix is sparse (Figure 4). Despite having shown the robust functioning of the network only in the two cases of sparse

<sup>2</sup>Although it might seem we are claiming that the network implemented here has a greater storage capacity than standard Hopfield networks, our boundary at  $0.17N$  is for equiprobable success and failure, while the  $0.14N$  figure is given for overwhelmingly likely success.

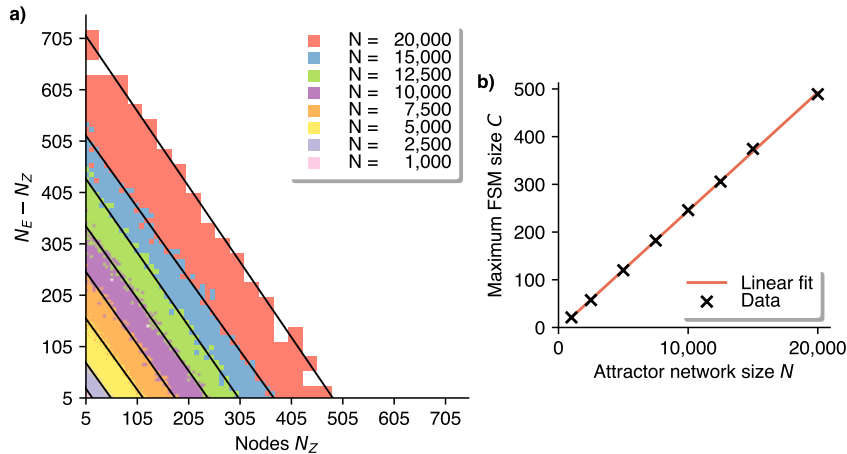


Fig. 6: The capacity of the attractor network for varying size  $N$ , in terms of the size of FSM that can be emulated before failure. For a given  $N$ , a random FSM was generated with number of nodes  $N_Z$  and number of edges  $N_E$ . An attractor network was then constructed as described in Section III, and a sequence of stimuli input to the network that should trigger a specific walk between attractor states. **a)** Every coloured square is a successful walk, with no unique  $(N_Z, N_E, N)$  triplet being sampled more than once, and lower- $N$  squares occlude higher- $N$  squares. Since only graphs with at least as many edges as nodes were sampled,  $N_E - N_Z$  is given on the  $y$ -axis rather than  $N_E$ . The overlain black lines are the SVM-fitted decision boundaries, distinguishing between values that succeeded and values that failed. **b)** The capacity  $C$  for varying Hopfield network sizes  $N$ , where  $C$  is defined to be the maximum size of of FSM which can be implemented before failure, for which  $N_E = N_Z$ . A linear fit is overlain, and shows a linear relationship in the capacity  $C$  in terms of  $N$  over the range explored. Assuming that the gradients of the linear fit in a) are equal, the boundary at which failure and success are equiprobable is given by  $N_Z + 2.4N_E = 0.085N$ .

ternary weights and dense noisy weights, we expect there to be a trade-off between the amount of each non-ideality that the network can withstand before failure. That is, an attractor network with dense noisy weights may withstand a greater degree of noise before failure than a network with sparse noisy weights.

An efficient and high-density implementation of a dense weights matrix may however be enabled by novel memristive crossbar technologies, which execute the dense matrix-vector-multiplication (MVM) step required in the state update rule in one operation [25–27]. Such devices are already of great interest due to their immediate application in image processing and deep learning acceleration, and hybrid CMOS-memristor hardware implementations have thus already been pursued for their usefulness as an associative memory [32–34], as well as for more direct FSM implementations [35, 36] using memristive ternary content addressable memory (TCAM) cells. Since we have shown that individual weights may be bistable, incredibly unreliable and noisy without incurring a significant loss to performance, then a large enough crossbar would be a highly suitable platform on which to implement the network presented here.

Since the additions to the weights matrix for each state

and transition are composed purely of outer products, fully parallel one-shot in-memory online learning of the weights matrix may be achievable. As long as the updates in the memristors’ conductances are sufficiently linear and symmetric, then attractors and transitions may be learned in one-shot by specifying the two vectors at the crossbar’s inputs and outputs [37, 38].

### B. Relation to other architectures

While there is a large body of work concerning the equivalence between RNNs and FSMs, their implementations broadly fall into a few categories. There are those that require iterative gradient descent methods to mimic an FSM [39–41], which makes them difficult to train for large FSMs, and improbable for use in biology. There are those that require creating a new FSM with an explicitly expanded state set,  $Z' := Z \times S$ , such that there is a new state for every old state-stimulus pair [42, 43], which is unfavourable due to the the explosion of (usually one-hot) states needing to be represented, as well as the difficulty of adding new states or stimuli iteratively. There are those that require higher-order weight tensors in order to explicitly provide a weight entry for every unique state-stimulus pair [44–46]. As well

as being non-distributed, the weight tensors require synapses to connect between more than two neurons, which is difficult to implement and non-biological [47]. In [48] transitions are triggered by adiabatically modulating a global inhibition parameter, such that the network may transition between similar stored patterns. Lacking however is a method to construct a network to perform arbitrary, controllable transitions between states. In [49] an in-depth analysis of small populations of rate-based neurons is conducted, wherein synapses with short-term synaptic depression enable a rich behaviour of itinerancy between attractor states, but does not scale to large systems and arbitrary stored memories.

Most closely resembling our approach, however, are earlier works concerned with the related task of creating a sequence of transitions between attractor states in Hopfield-like neural networks. The majority of these efforts rely upon the use of synaptic delays, such that the postsynaptic sum on a time step  $t$  depends, for example, also on the network state at time  $t - 10$ , rather than just  $t - 1$ . These delay synapses thus allow attractor cross-terms of the form  $\mathbf{x}^{\nu+1}\mathbf{x}^{\nu\top}$  to become influential only after the network has inhabited an attractor state for a certain amount of time, triggering a walk between attractor states [50, 51]. This then also allowed for the construction of networks with state-dependent input-triggered transitions [52–54]. Similar networks were shown to function without the need for synaptic delays, but require fine tuning of network parameters and suffer from extremely low storage capacity [2, 55]. In any case, the need for synaptic delay elements represents a large requirement on any substrate which might implement such a network, and indeed are problematic to implement in neuromorphic systems [56].

State-dependent computation in spiking neural networks was realised in [57] and [58], where they used population attractor dynamics to achieve robust state representations via sustained spiking activity. However, these approaches differ from this work in that the state representations are still fundamentally population-based rather than distributed, and so pose difficulties such as the requirement of finding a new population of neurons to represent any new state.

This work also differs from conventional methods to implement graphs in VSAs [20, 59], in that the network state does not need to be read by an outsider in order to implement state-dependent switching. That is, where in previous works a graph is encoded by a hypervector such that it may be reliably decoded by external circuitry, we instead encode the graph’s connectivity in the attractor network’s weights matrix, such that its recurrent dynamics realise the desired state machine. Our implementation could however have been brought closer to previous works, and indeed made simpler, if there were a way to reliably achieve a flipping of neuron states when input is received. Then, the transition

dynamics could be achieved just by storing edge terms like  $\mathbf{E} = \mathbf{y}(\mathbf{x} \circ \mathbf{s})^\top$ . Although achieving a state flip may be easy in a digital synchronous system, it would be very difficult to robustly achieve in an analogue asynchronous system. To avoid flickering, the flip would need to be reliably initiated by a single event. These events would also need to arrive at all neurons simultaneously, lest attractor dynamics take over and correct these apparent errors. Both of these factors prohibit such an operation from existing in biological systems.

The lack of a flipping mechanism is also discussed in [60], wherein they show the necessity of a population of neurons with mixed selectivity, connected to both the input and attractor neurons, in order to achieve the flipping-like behaviour necessary for complex state switching. This requirement arose by demanding that the network state switch to resembling the target state immediately upon receiving a stimulus. We instead show that similar results can be achieved without this extra population, if we relax to instead demanding only that the network soon evolve to the target state.

### C. Biological plausibility

Transitions between discrete neural attractor states are thought to be a crucial mechanism for performing context-dependent decision making in biological neural systems [8–11]. Attractor dynamics enable a temporary retention of received information, and ensure that irrelevant inputs do not produce stable deviations in the neural state. As such, in this work we provide a description of an attractor-based network that can perform controllable context-dependent transitions in a simple manner, while abiding by the principles of distributed representation.

The procedure for generating the weights matrix  $\mathbf{W}$ , as a result of this simplicity, makes the proposed network more plausible than other more complex approaches, e.g. those utilising gradient descent methods. It can be learned in one-shot in a fully online fashion, since adding a new node or edge involves only an additive contribution to the weights matrix, which does not require knowledge of irrelevant edges, nodes, their hypervectors, or the weight values themselves. Furthermore, as a result of the entirely distributed representation of states and transitions, new behaviours may be added to the weights matrix at a later date, both without having to allocate new hardware, and without having to recalculate  $\mathbf{W}$  with all previous data. Both of these factors are critical for continual online learning.

Evaluating the local learnability of  $\mathbf{W}$  to implement transitions is also necessary to evaluate the biological plausibility of the model. In the original Hopfield paper the weights could

be learned using the simple Hebbian rule

$$\delta w_{ij} = x'_i x'_j \quad (29)$$

where  $x'_i$  and  $x'_j$  are the activities of the post- and presynaptic neurons respectively, and  $\delta w_{ij}$  the online synaptic efficacy update [1, 61]. While the attractor terms within the network can be learned in this manner, the transition cross-terms that we have introduced require an altered version of the learning rule. If we simplify our network construction by removing the edge state attractors, then the local weight update required to learn a transition between states is given by

$$\delta w_{ij} = H(s_i) y_i x_j s_j \quad (30)$$

where  $\mathbf{y}$ ,  $\mathbf{x}$  and  $\mathbf{s}$  are as previously defined. In removing the edge states, we disallow FSMs with consecutive edges with the same stimulus (e.g. "father\_is, father\_is"), but this is not a problem if completely general FSM construction is not the goal per se. This state-transition learning rule is just as local as the original Hopfield learning rule, as the weight update from presynaptic neuron  $j$  to postsynaptic neuron  $i$  is dependent only upon information that may be made directly accessible in the pre- and postsynaptic neurons, and does not depend on information in other neurons to which the synapse is not connected [62, 63].

The robust functioning of the network despite noisy and unreliable weights is another prerequisite for the model to plausibly be able to exist in biological systems. The network weights may be considerably degraded without affecting the behaviour of the network, and indeed beyond this the network exhibits a so-called graceful degradation in performance. Furthermore, biological synapses are expected to have only a few bits of precision [64–66], and the network has been shown to function even in the worst case of binary weights. These properties stem from the massive redundancy arising from storing the attractor states across the entire synaptic matrix in a distributed manner, a technique that the brain is expected to utilise [67, 68]. Since the network is still an attractor network, it retains all of the properties that make attractor networks suitable for modelling cognitive function, such as that the network can perform robust pattern completion and correction, i.e. the recovery of a stored prototypical memory given a damaged, incomplete or noisy version, and thereafter function as a stable working memory [1, 2].

## VI. CONCLUSION

Attractor neural networks are robust abstract models of human memory, but previous attempts to endow them with complex and controllable attractor-switching capabilities have suffered mostly from being either non-distributed, not

scalable, or not robust. We have here introduced a simple procedure by which any arbitrary FSM may be embedded into a large enough Hopfield-like attractor network, where states and stimuli are replaced by high-dimensional random vectors, and all information pertaining to FSM transitions is stored in the network's weights matrix in a fully distributed manner. Our method of modelling input to the network as a masking of the network state allows cross-terms between attractors to be stored in the weights matrix in a way that they are effectively obfuscated until the correct state-stimulus pair is present, much in a manner similar to the standard binding-unbinding operation in more conventional VSAs.

We showed that the network retains many of the features of attractor networks which make them suitable for biology, namely that the network is robust to unreliable and imprecise weights, thus also making them highly suitable for implementation with high-density but noisy devices. We presented numerical results showing that the network capacity in terms of implementable FSM size scales linearly with the size of the attractor network, and also that the network continues to function when the synchronous neuron update rule is replaced with a stochastic, asynchronous variant.

In summary, we introduced an attractor-based neural state machine which overcomes many of the shortcomings that made previous models unsuitable for use in biology, and propose that attractor-based FSMs may thus represent a plausible path by which FSMs may exist as a distributed computational primitive in biological neural networks.

## ACKNOWLEDGEMENTS

We thank Dr. Federico Corradi, Dr. Nicoletta Risi and Dr. Matthew Cook for their invaluable input and suggestions, as well as their help with proofreading this document.

Funded by the Deutsche Forschungsgemeinschaft (DFG German Research Foundation) - Projects NMVAC (432009531) and MemTDE (441959088).

The authors would like to acknowledge the financial support of the CogniGron research center and the Ubbo Emmius Funds (Univ. of Groningen).

## REFERENCES

1. Hopfield, J. J. Neural networks and physical systems with emergent collective computational abilities. en. *Proceedings of the National Academy of Sciences* **79**. Publisher: National Academy of Sciences Section: Research Article, 2554–2558 (Apr. 1982).
2. Amit, D. *Modeling Brain Function: The World of Attractor Neural Networks, 1st Edition* in (1989).
3. Eliasmith, C. A unified approach to building and controlling spiking attractor networks. eng. *Neural Computation* **17**, 1276–1314 (June 2005).

4. Little, W. A. The existence of persistent states in the brain. en. *Mathematical Biosciences* **19**, 101–120 (Feb. 1974).
5. Schneidman, E., Berry, M. J., Segev, R. & Bialek, W. Weak pairwise correlations imply strongly correlated network states in a neural population. en. *Nature* **440**, 1007–1012 (Apr. 2006).
6. Chaudhuri, R. & Fiete, I. Computational principles of memory. en. *Nature Neuroscience* **19**. Number: 3 Publisher: Nature Publishing Group, 394–403 (Mar. 2016).
7. Khona, M. & Fiete, I. R. Attractor and integrator networks in the brain. en. *Nature Reviews Neuroscience* **23**. Number: 12 Publisher: Nature Publishing Group, 744–766 (Dec. 2022).
8. Daelli, V. & Treves, A. Neural attractor dynamics in object recognition. eng. *Experimental Brain Research* **203**, 241–248 (June 2010).
9. Mante, V., Sussillo, D., Shenoy, K. V. & Newsome, W. T. Context-dependent computation by recurrent dynamics in prefrontal cortex. en. *Nature* **503**. Number: 7474 Publisher: Nature Publishing Group, 78–84 (Nov. 2013).
10. Miller, P. Itinerancy between attractor states in neural systems. *Current opinion in neurobiology* **40**, 14–22 (Oct. 2016).
11. Tajima, S. *et al.* Task-dependent recurrent dynamics in visual cortex. *eLife* **6** (ed Latham, P.) Publisher: eLife Sciences Publications, Ltd, e26868 (July 2017).
12. Brinkman, B. a. W. *et al.* Metastable dynamics of neural circuits and networks. *Applied Physics Reviews* **9**. Publisher: American Institute of Physics, 011313 (Mar. 2022).
13. Dayan, P. Simple substrates for complex cognition. *Frontiers in Neuroscience* **2**, 31 (2008).
14. Buonomano, D. V. & Maass, W. State-dependent computations: spatiotemporal processing in cortical networks. en. *Nature Reviews Neuroscience* **10**. Number: 2 Publisher: Nature Publishing Group, 113–125 (Feb. 2009).
15. Granger, R. Toward the quantification of cognition. *arXiv:2008.05580 [cs, q-bio]*. arXiv: 2008.05580 (Aug. 2020).
16. Kanerva, P. Fully Distributed Representation. *Proc. 1997 Real World Computing Symposium (RWC97, Tokyo)* (Nov. 2002).
17. Plate, T. A. *Holographic Reduced Representation: Distributed Representation for Cognitive Structures* en (Center for the Study of Language and Information, Apr. 2003).
18. Gayler, R. W. Vector Symbolic Architectures answer Jackendoff’s challenges for cognitive neuroscience. *arXiv:cs/0412059*. arXiv: cs/0412059 (Dec. 2004).
19. Kanerva, P. Hyperdimensional Computing: An Introduction to Computing in Distributed Representation with High-Dimensional Random Vectors. en. *Cognitive Computation* **1**, 139–159 (June 2009).
20. Kleyko, D. *et al.* Vector Symbolic Architectures as a Computing Framework for Nanoscale Hardware. en. *arXiv:2106.05268 [cs]*. arXiv: 2106.05268 (June 2021).
21. Kleyko, D., Rachkovskij, D. A., Osipov, E. & Rahimi, A. A Survey on Hyperdimensional Computing aka Vector Symbolic Architectures, Part I: Models and Data Transformations. *ACM Computing Surveys*. Just Accepted (May 2022).
22. Gritsenko, V. I. *et al.* Neural Distributed Autoassociative Memories: A Survey. *Kibernetika i vychislitel’naâ tehnika* **2017**. arXiv:1709.00848 [cs], 5–35 (June 2017).
23. Backus, J. Can programming be liberated from the von Neumann style? A functional style and its algebra of programs. *Communications of the ACM* **21**, 613–641 (Aug. 1978).
24. Indiveri, G. & Liu, S.-C. Memory and Information Processing in Neuromorphic Systems. *Proceedings of the IEEE* **103**. Conference Name: Proceedings of the IEEE, 1379–1397 (Aug. 2015).
25. Xia, Q. & Yang, J. J. Memristive crossbar arrays for brain-inspired computing. en. *Nature Materials* **18**. Number: 4 Publisher: Nature Publishing Group, 309–323 (Apr. 2019).
26. Ielmini, D. & Wong, H.-S. P. In-memory computing with resistive switching devices. en. *Nature Electronics* **1**. Number: 6 Publisher: Nature Publishing Group, 333–343 (June 2018).
27. Zidan, M. A. & Lu, W. D. en. in *Memristive Devices for Brain-Inspired Computing* (eds Spiga, S., Sebastian, A., Querlioz, D. & Rajendran, B.) 221–254 (Woodhead Publishing, Jan. 2020).
28. Sompolinsky, H. *The theory of neural networks: The Hebb rule and beyond* en. in *Heidelberg Colloquium on Glassy Dynamics* (eds van Hemmen, J. L. & Morgenstern, I.) (Springer, Berlin, Heidelberg, 1987), 485–527.
29. Howard, R. *et al.* An associative memory based on an electronic neural network architecture. *IEEE Transactions on Electron Devices* **34**. Conference Name: IEEE Transactions on Electron Devices, 1553–1556 (July 1987).
30. Verleysen, M. & Jespers, P. An analog VLSI implementation of Hopfield’s neural network. *IEEE Micro* **9**. Conference Name: IEEE Micro, 46–55 (Dec. 1989).

31. Weinfeld, M. en. in *VLSI for Artificial Intelligence* (eds Delgado-Frias, J. G. & Moore, W. R.) 169–178 (Springer US, Boston, MA, 1989).
32. Guo, X. *et al.* Modeling and Experimental Demonstration of a Hopfield Network Analog-to-Digital Converter with Hybrid CMOS/Memristor Circuits. *Frontiers in Neuroscience* **9**, 488 (2015).
33. Yang, J., Wang, L., Wang, Y. & Guo, T. A novel memristive Hopfield neural network with application in associative memory. en. *Neurocomputing. Dynamical Behaviors of Coupled Neural Networks with Reaction-Diffusion Terms: Analysis, Control and Applications* **227**, 142–148 (Mar. 2017).
34. Molahasani Majdabadi, M., Shamsi, J. & Baradaran Shokouhi, S. Hybrid CMOS/memristor crossbar structure for implementing hopfield neural network. en. *Analog Integrated Circuits and Signal Processing* **107**, 249–261 (May 2021).
35. Graves, C. E. *et al.* In-Memory Computing with Memristor Content Addressable Memories for Pattern Matching. en. *Advanced Materials* **32**, 2003437 (Sept. 2020).
36. De Lima, J. P. C., Brandalero, M., Hübner, M. & Carro, L. STAP: An Architecture and Design Tool for Automata Processing on Memristor TCAMs. *ACM Journal on Emerging Technologies in Computing Systems* **18**, 39:1–39:22 (Dec. 2022).
37. Alibart, F., Zamanidoost, E. & Strukov, D. B. Pattern classification by memristive crossbar circuits using ex situ and in situ training. en. *Nature Communications* **4**. Number: 1 Publisher: Nature Publishing Group, 2072 (June 2013).
38. Li, Y. *et al.* In situ Parallel Training of Analog Neural Network Using Electrochemical Random-Access Memory. *Frontiers in Neuroscience* **15**, 636127 (Apr. 2021).
39. Zeng, Z., Goodman, R. M. & Smyth, P. Learning Finite State Machines With Self-Clustering Recurrent Networks. *Neural Computation* **5**, 976–990 (Nov. 1993).
40. Lee Giles, C., Horne, B. G. & Lin, T. Learning a class of large finite state machines with a recurrent neural network. en. *Neural Networks* **8**, 1359–1365 (Jan. 1995).
41. Das, S. & Mozer, M. C. A Unified Gradient-Descent/Clustering Architecture for Finite State Machine Induction in *Advances in Neural Information Processing Systems* **6** (Morgan-Kaufmann, 1994).
42. Minsky, M. L. *Computation: finite and infinite machines* (Prentice-Hall, Inc., USA, 1967).
43. Sanfeliu, A. An Algebraic Framework to Represent Finite State Machines in Single-Layer Recurrent Neural Networks. *Neural Computation* **7** (Oct. 1999).
44. Omlin, C., Thornber, K. & Giles, C. Fuzzy finite-state automata can be deterministically encoded into recurrent neural networks. *IEEE Transactions on Fuzzy Systems* **6**. Conference Name: IEEE Transactions on Fuzzy Systems, 76–89 (Feb. 1998).
45. Forcada, M. & Carrasco, R. C. *Finite-State Computation in Analog Neural Networks: Steps towards Biologically Plausible Models?* in *Emergent Neural Computational Architectures Based on Neuroscience* (2001).
46. Mali, A. A., Ororbia II, A. G. & Giles, C. L. A Neural State Pushdown Automata. *IEEE Transactions on Artificial Intelligence* **1**. Conference Name: IEEE Transactions on Artificial Intelligence, 193–205 (Dec. 2020).
47. Krotov, D. & Hopfield, J. Large Associative Memory Problem in Neurobiology and Machine Learning. *arXiv:2008.06996 [cond-mat, q-bio, stat]*. arXiv: 2008.06996 (Apr. 2021).
48. Recanatesi, S., Katkov, M. & Tsodyks, M. Memory States and Transitions between Them in Attractor Neural Networks. *Neural Computation* **29**, 2684–2711 (Oct. 2017).
49. Chen, B. & Miller, P. Attractor-state itinerancy in neural circuits with synaptic depression. *The Journal of Mathematical Neuroscience* **10**, 15 (Sept. 2020).
50. Sompolinsky, H. & Kanter, I. Temporal Association in Asymmetric Neural Networks. *Physical Review Letters* **57**. Publisher: American Physical Society, 2861–2864 (Dec. 1986).
51. Kleinfeld, D. Sequential state generation by model neural networks. *Proceedings of the National Academy of Sciences of the United States of America* **83**, 9469–9473 (Dec. 1986).
52. Gutfreund, H. & Mezard, M. Processing of Temporal Sequences in Neural Networks. *Physical Review Letters* **61**. Publisher: American Physical Society, 235–238 (July 1988).
53. Amit, D. J. Neural Networks Counting Chimes. *Proceedings of the National Academy of Sciences of the United States of America* **85**. Publisher: National Academy of Sciences, 2141–2145 (1988).
54. Drossaers, M. F. J. *Hopfield models as nondeterministic finite-state machines* in *Proceedings of the 14th conference on Computational linguistics - Volume 1* (Association for Computational Linguistics, USA, Aug. 1992), 113–119.
55. Buhmann, J. & Schulten, K. Noise-Driven Temporal Association in Neural Networks. en. *Europhysics Letters (EPL)* **4**. Publisher: IOP Publishing, 1205–1209 (Nov. 1987).



56. Nielsen, C., Qiao, N. & Indiveri, G. *A compact ultra low-power pulse delay and extension circuit for neuromorphic processors* in *2017 IEEE Biomedical Circuits and Systems Conference (BioCAS)* (Oct. 2017), 1–4.
57. Neftci, E. *et al.* Synthesizing cognition in neuromorphic electronic systems. en. *Proceedings of the National Academy of Sciences* **110**, E3468–E3476 (Sept. 2013).
58. Liang, D. & Indiveri, G. A Neuromorphic Computational Primitive for Robust Context-Dependent Decision Making and Context-Dependent Stochastic Computation. *IEEE Transactions on Circuits and Systems II: Express Briefs* **66**. Conference Name: IEEE Transactions on Circuits and Systems II: Express Briefs, 843–847 (May 2019).
59. Poduval, P. *et al.* GrapHD: Graph-Based Hyperdimensional Memorization for Brain-Like Cognitive Learning. *Frontiers in Neuroscience* **16** (2022).
60. Rigotti, M., Ben Dayan Rubin, D., Wang, X.-J. & Fusi, S. Internal Representation of Task Rules by Recurrent Dynamics: The Importance of the Diversity of Neural Responses. *Frontiers in Computational Neuroscience* **4** (2010).
61. Hebb, D. O. *The organization of behavior; a neuropsychological theory* Pages: xix, 335 (Wiley, Oxford, England, 1949).
62. Zenke, F. & Neftci, E. O. Brain-Inspired Learning on Neuromorphic Substrates. *Proceedings of the IEEE* **109**, 935–950 (May 2021).
63. Khacef, L. *et al.* *Spike-based local synaptic plasticity: A survey of computational models and neuromorphic circuits* arXiv:2209.15536 [cs]. Nov. 2022.
64. O’Connor, D. H., Wittenberg, G. M. & Wang, S. S.-H. Graded bidirectional synaptic plasticity is composed of switch-like unitary events. eng. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 9679–9684 (July 2005).
65. Bartol, T. M. *et al.* *Hippocampal Spine Head Sizes Are Highly Precise* en. Tech. rep. Section: New Results Type: article (bioRxiv, Mar. 2015), 016329.
66. Baldassi, C., Gerace, F., Lucibello, C., Saglietti, L. & Zecchina, R. Learning may need only a few bits of synaptic precision. *Physical Review E* **93** (Feb. 2016).
67. Rumelhart, D. E. & McClelland, J. L. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations — MIT Press eBooks — IEEE Xplore* 1987.
68. Crawford, E., Gingerich, M. & Eliasmith, C. Biologically Plausible, Human-Scale Knowledge Representation. eng. *Cognitive Science* **40**, 782–821 (May 2016).

## APPENDIX

### A. Dynamics without masking

For the following calculations we assume that the coding level of the output states  $f_r$  is low enough that their effect can be ignored. With this in mind, if we ignore the semantic differences between attractors for node states and attractors for edge states, the two summations over states can be absorbed into one summation over both types of attractor, here both denoted  $\mathbf{x}^\nu$ . Similarly there is then no difference between the two transition cross-terms within each  $\mathbf{E}$  term, and they too can be absorbed into one summation. Our simplified expression for  $\mathbf{W}$  is now given by

$$\mathbf{W} = \frac{1}{N} \sum_{\text{attr's } \nu}^{N_Z+N_E} \mathbf{x}^\nu \mathbf{x}^{\nu \top} + \frac{1}{N} \sum_{\text{tran's } \lambda}^{2N_E} H(\mathbf{s}^{\pi(\lambda)}) \circ (\mathbf{x}^{v(\lambda)} - \mathbf{x}^{\chi(\lambda)}) (\mathbf{x}^{\chi(\lambda)} \circ \mathbf{s}^{\pi(\lambda)})^\top \quad (31)$$

where  $\chi(\lambda)$  and  $v(\lambda)$  are functions  $\{1, \dots, 2N_E\} \rightarrow \{1, \dots, N_Z + N_E\}$  determining the indices of the source and target states for transition  $\lambda$ , and  $\pi(\lambda) : \{1, \dots, 2N_E\} \rightarrow \{1, \dots, N_{\text{stimuli}}\}$  determines the index of the associated stimulus. We then wish to calculate the statistics of the postsynaptic sum  $\mathbf{W}\mathbf{z}$  while the attractor network is currently in an attractor state. When in an attractor state  $\mathbf{x}^\mu$ , the postsynaptic sum is given by

$$\begin{aligned} [\mathbf{W}\mathbf{x}^\mu]_i &= \frac{1}{N} \sum_{\text{attr's } \nu}^{N_Z+N_E} x_i^\nu \underbrace{[\mathbf{x}^\nu \cdot \mathbf{x}^\mu]}_{\substack{N \text{ if } \mu=\nu \\ \text{else } \mathcal{N}(0,N)}} + \frac{1}{N} \sum_{\text{tran's } \lambda}^{2N_E} H(s_i^{\pi(\lambda)}) \circ (x_i^{v(\lambda)} - x_i^{\chi(\lambda)}) \underbrace{[(\mathbf{x}^{\chi(\lambda)} \circ \mathbf{s}^{\pi(\lambda)}) \cdot \mathbf{x}^\mu]}_{\mathcal{N}(0,N)} \\ &= x_i^\mu + \sum_{\substack{\text{attr's} \\ \nu \neq \mu}}^{N_Z+N_E} \underbrace{x_i^\nu}_{\text{Var.}=1} \left[ \mathcal{N}^\nu \left( 0, \frac{1}{N} \right) \right] + \sum_{\text{tran's } \lambda}^{2N_E} \underbrace{H(s_i^{\pi(\lambda)}) \circ (x_i^{v(\lambda)} - x_i^{\chi(\lambda)})}_{\text{Var.}=1} \left[ \mathcal{N}^\lambda \left( 0, \frac{1}{N} \right) \right] \\ &\approx x_i^\mu + \mathcal{N} \left( 0, \frac{N_Z + N_E - 1}{N} \right) + \mathcal{N} \left( 0, \frac{2N_E}{N} \right) \\ &\approx x_i^\mu + \mathcal{N} \left( 0, \frac{N_Z + 3N_E}{N} \right) \end{aligned} \quad (32)$$

where we have used the notation  $\mathcal{N}(\mu, \sigma^2)$  to denote a normally-distributed random variable (RV) with mean  $\mu$  and variance  $\sigma^2$ . In the third line we have made the approximation in the transition summation that the linear sum of attractor hypervectors, each multiplied by a Gaussian RV, is itself a separate Gaussian RV in each dimension. This holds as long as there are "many" attractor terms appearing on the LHS of the transition summation. Said otherwise, if the summation over transition terms has only very few unique attractor terms on the LHS ( $N_E \gg N_Z$ ), then the noise will be a random linear sum of the same few (masked) hypervectors, each with approximate magnitude  $\frac{1}{\sqrt{N}}$ , and so will be highly correlated between dimensions. Nonetheless we assume we are far away from this regime, and let the effect of the sum of these unwanted terms be approximated by a normally-distributed random vector, and so we have

$$\mathbf{W}\mathbf{x}^\mu \approx \mathbf{x}^\mu + \sigma \mathbf{n} \quad (33)$$

where  $\sigma = \sqrt{\frac{N_Z + 3N_E}{N}}$  is the strength of cross-talk noise, and  $\mathbf{n}$  a vector composed of IID standard normally-distributed terms. This procedure of quantifying the signal-to-noise ratio (SNR) is adapted from that in the original Hopfield paper [1, 2].

### B. Dynamics with masking

We can similarly calculate the postsynaptic sum when in an attractor state  $\mathbf{x}^\mu$ , while the network is being masked by a stimulus  $\mathbf{s}^\kappa$ , with this (state, stimulus) tuple corresponding to a certain valid transition  $\lambda'$ , with source, target, and stimulus vectors  $\mathbf{x}^\mu$ ,  $\mathbf{x}^\phi$ , and  $\mathbf{s}^\kappa$  respectively:

$$\begin{aligned}
\left[ \mathbf{W}(\mathbf{x}^\mu \circ H(\mathbf{s}^\kappa)) \right]_i &= \frac{1}{N} \sum_{\text{attr's } \nu}^{N_Z+N_E} x_i^\nu \underbrace{\left[ \mathbf{x}^\nu \cdot (\mathbf{x}^\mu \circ H(\mathbf{s}^\kappa)) \right]}_{\substack{\frac{1}{2}N \text{ if } \mu=\nu \\ \text{else } \mathcal{N}(0, \frac{1}{2}N)}} \\
&+ \frac{1}{N} \sum_{\text{tran's } \lambda}^{2N_E} H(s_i^{\pi(\lambda)}) (x_i^{v(\lambda)} - x_i^{\chi(\lambda)}) \underbrace{\left[ (\mathbf{x}^{\chi(\lambda)} \circ \mathbf{s}^{\pi(\lambda)}) \cdot (\mathbf{x}^\mu \circ H(\mathbf{s}^\kappa)) \right]}_{\substack{\frac{1}{2}N \text{ if } \chi(\lambda)=\mu \text{ and } \pi(\lambda)=\kappa \\ \text{else } \mathcal{N}(0, \frac{1}{2}N)}} \\
&= \frac{1}{2} x_i^\mu + \frac{1}{2} H(s_i^\kappa) (x_i^\phi - x_i^\mu) + \sum_{\substack{\text{attr's} \\ \nu \neq \mu}}^{N_Z+N_E} \underbrace{x_i^\nu}_{\text{Var.}=1} \left[ \mathcal{N}^\nu(0, \frac{1}{2N}) \right] \\
&+ \sum_{\substack{\text{tran's} \\ \lambda \neq \lambda'}}^{2N_E} \underbrace{H(s_i^{\pi(\lambda)}) (x_i^{v(\lambda)} - x_i^{\chi(\lambda)})}_{\text{Var.}=1} \left( \mathcal{N}^\lambda(0, \frac{1}{2N}) \right) \\
&\approx \frac{1}{2} [H(s_i^\kappa) + H(-s_i^\kappa)] x_i^\mu + \frac{1}{2} H(s_i^\kappa) (x_i^\phi - x_i^\mu) + \mathcal{N}\left(0, \frac{N_Z + N_E - 1}{2N}\right) + \mathcal{N}\left(0, \frac{2N_E - 1}{2N}\right) \\
&= \frac{1}{2} H(s_i^\kappa) x_i^\phi + \frac{1}{2} H(-s_i^\kappa) x_i^\mu + \mathcal{N}\left(0, \frac{N_Z + 3N_E - 2}{2N}\right) \\
&\approx \frac{1}{2} \left[ H(s_i^\kappa) x_i^\phi + H(-s_i^\kappa) x_i^\mu + \sqrt{2} \cdot \mathcal{N}\left(0, \frac{N_Z + 3N_E}{N}\right) \right]
\end{aligned} \tag{34}$$

where in the third line we have made the same approximations as previously discussed. The postsynaptic sum is thus approximately  $\mathbf{x}^\phi$  in all indices that are not currently being masked, which drives the network towards that (target) attractor. In vector form, the above is written as

$$\mathbf{W}(\mathbf{x}^\mu \circ H(\mathbf{s}^\kappa)) \approx H(\mathbf{s}^\kappa) \circ \mathbf{x}^\phi + H(-\mathbf{s}^\kappa) \circ \mathbf{x}^\mu + \sqrt{2} \sigma \mathbf{n} \tag{35}$$

where it is assumed that there exists a stored transition from state  $\mathbf{x}^\mu$  to  $\mathbf{x}^\phi$  with stimulus  $\mathbf{s}^\kappa$ , and  $\approx$  denotes approximate proportionality. A similar calculation can be performed in the case that a stimulus is imposed which does not correspond to a valid transition for the current state. In this case, no terms of significant magnitude emerge from the transition summation, and we are left with

$$\mathbf{W}(\mathbf{x}^\mu \circ H(\mathbf{s}^{\text{invalid}})) \approx \mathbf{x}^\mu + \sqrt{2} \sigma \mathbf{n} \tag{36}$$

i.e. the attractor dynamics are largely unaffected. Since we have not distinguished between our above attractor terms being node attractors or edge attractors, or our stimuli from being  $\mathbf{s}_a$  or  $\mathbf{s}_b$  stimuli, the above results can be applied to all relevant situations *mutatis mutandis*.