# Neuron-Astrocyte Associative Memory

Leo Kozachkov[1,4], Jean-Jacques Slotine[1,2], Dmitry Krotov[3]

[1]Department of Brain and Cognitive Sciences, MIT
[2]Department of Mechanical Engineering, MIT
[3]MIT-IBM Watson AI Lab, IBM Research
[4]IBM Research

## Abstract

Astrocytes, the most abundant type of glial cell, play a fundamental role in memory. Despite most hippocampal synapses being contacted by an astrocyte, there are no current theories that explain how neurons, synapses, and astrocytes might collectively contribute to memory function. We demonstrate that fundamental aspects of astrocyte morphology and physiology naturally lead to a dynamic, high-capacity associative memory system. The neuron-astrocyte networks generated by our framework are closely related to popular machine learning architectures known as Dense Associative Memories or Modern Hopfield Networks. Adjusting the connectivity pattern, the model developed here leads to a family of associative memory networks that includes a Dense Associative Memory and a Transformer as two limiting cases. In the known biological implementations of Dense Associative Memories, the ratio of stored memories to the number of neurons remains constant, despite the growth of the network size. Our work demonstrates that neuron-astrocyte networks follow a superior memory scaling law, outperforming known biological implementations of Dense Associative Memory. Our model suggests an exciting and previously unnoticed possibility that memories could be stored, at least in part, within the network of astrocyte processes rather than solely in the synaptic weights between neurons.

Not all brain cells are neurons. It is estimated that about half of the cells in the human brain are glial cells (from “glue" in Greek) [1]. Glial cells have long been known to play an important role in homeostatic brain functions, such as regulating blood flow [2] – thus contributing to hemodynamic signals such as those measured in fMRI [3] – and removing synaptic debris. Converging lines of recent evidence strongly suggest that they are also directly involved in learning, memory, and cognition [4–10]. Among glial cells, *astrocytes* are particularly important for brain function.

They serve a crucial role in directly sensing neural activity and, in turn, regulating synaptic strength and plasticity [4, 5, 11–14]. In addition to sensing neural activity, astrocytes are also important targets of neuromodulatory signals such as norepinephrine and acetylcholine emerging from potentially distant brain structures such as the brain stem [15].

Of particular relevance to the computational neuroscience community are the recent findings that 1) astrocytes are necessary for forming long-term memories [6, 16–18] and 2) astrocytes respond to neural activity on timescales spanning many orders of magnitude, from several hundred milliseconds to minutes [14, 19, 20]. Despite extensive evidence establishing the importance of neuron-astrocyte interactions for long-term memory function, computational theories of these interactions are still in their infancy.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

**What Shapes Astrocytic Computation?** The core proposal of this paper is that astrocytes compute, with their computations shaped by tunable signaling pathways. We focus on associative computations, where neurons, synapses, and astrocytes collaborate to store and retrieve memories.

In this model, astrocytic $Ca^{2+}$ flux coefficients store memories, and neuron-synapse-astrocyte interactions retrieve them. This extends previous work suggesting memories are stored in synaptic weights [21, 22], offering a new perspective where synaptic weights "emerge" from neuron-astrocyte interactions.

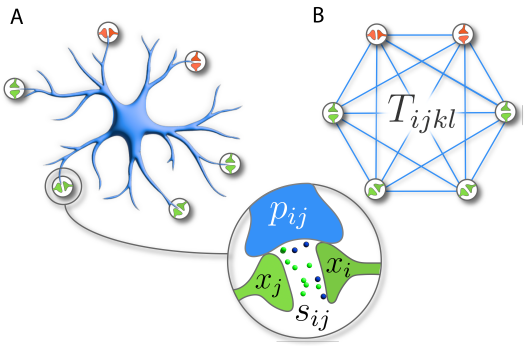



**Figure 1:** A) An abstracted version of an astrocyte, showing the astrocyte processes and the synapses. B) Our mathematical idealization of the mini-circuit defined by a single astrocyte.

# 1 Neuron-Astrocyte Model

Astrocytes, with a central soma and branching processes, envelop nearby synapses to form *tripartite synapses* [23]. A single astrocyte can connect to over $10^6$ synapses [24], and astrocyte networks create non-overlapping regions in the brain [25]. Astrocytes detect synaptic neurotransmitters, triggering intracellular calcium ($Ca^{2+}$ ) increases, which may lead to the release of gliotransmitters that regulate neural activity in a feedback loop. Astrocytes also communicate through calcium transport and gap junctions [26]. This paper focuses on key aspects of astrocyte biology:

- Astrocytes connect to millions of synapses, forming tripartite synapses [23].
- They detect and regulate neural activity through gliotransmitter release [27].
- Tripartite synapses interact via astrocytic calcium transport [26].

**Neural, Synaptic, and Astrocyte Dynamics** Neural, synaptic, and astrocytic dynamics can be captured by a set of coupled equations. The membrane voltage $x_i$ of each neuron $i$ evolves following a standard rate recurrent neural network model [28, 29], with timescale $\tau_n$ and leak rate $\lambda$:

$$\tau_n \, \dot{x}_i = -\lambda \, x_i + \sum_{j=1}^{N} g(s_{ij}) \, \phi(x_j) + b_i \tag{1}$$

Here, $b_i$ represents input to neuron $i$, $\phi(x_j)$ transforms membrane voltages into firing rates, and $g(s_{ij})$ denotes the strength of the synapse between neurons $i$ and $j$. With fixed $s_{ij}$, this reduces to a standard recurrent network. The synaptic strength $s_{ij}$ is dynamic, influenced by both pre- and post-synaptic activity, following synaptic facilitation governed by the equation:

$$\tau_s \, \dot{s}_{ij} = -\alpha \, s_{ij} + f(x_i, x_j, p_{ij}) + c_{ij} \tag{2}$$

Synaptic dynamics depend on neuronal activity, astrocytic interactions ($p_{ij}$), and bias $c_{ij}$. Astrocytes modulate synaptic plasticity via calcium ($Ca^{2+}$ )-dependent gliotransmitters [14], influencing the function $f$. Astrocyte calcium dynamics are described by:

$$\tau_p \, \dot{p}_{ij} = -\gamma \, p_{ij} + \sum_{k,l=1}^{N} T_{ijkl} \, \psi(p_{kl}) + \kappa(s_{ij}) + d_{ij} \tag{3}$$

The double sum captures interactions between astrocyte processes, with $T_{ijkl}$ representing inter-process connections and $\psi$ modeling calcium diffusion. Synaptic state $s_{ij}$ influences astrocyte calcium levels through $\kappa$, while the bias $d_{ij}$ sets a baseline tone. Together, these equations describe the intertwined dynamics of neurons, synapses, and astrocytes in a tripartite system.

# 2 Associative Neuron-Astrocyte Model

In section 1, we presented a framework based on neuron-astrocyte communication via tripartite synapses. Depending on the choice of nonlinearities and parameters, this network can exhibit complex behaviors like chaos or limit cycles, which are challenging to analyze generally.

**Examples of Possible Lagrangians and Activations**

$$\mathcal{L}(\mathbf{z}) = \log \sum_{i=1}^{N} e^{z_i} \quad \rightarrow \quad \frac{\partial \mathcal{L}(\mathbf{z})}{\partial \mathbf{z}} = \text{Softmax}(\mathbf{z})$$
$$\mathcal{L}(\mathbf{z}) = \sum_{i=1}^{N} Q(z_i) \quad \rightarrow \quad \frac{\partial \mathcal{L}(\mathbf{z})}{\partial \mathbf{z}} = \left[q(z_1), \ldots, q(z_n)\right]^T \tag{4}$$

**Figure 2:** Examples of possible Lagrangian functions. Here the variable $\mathbf{z}$ is an arbitrary dynamical variable in our model (e.g., astrocyte calcium level). Recall from the main text that activation functions are defined from the Lagrangians as $\frac{\partial \mathcal{L}}{\partial z_i}$. The first Lagrangian provides an example of a "collective" activation functions. The second Lagrangian leads to an element-wise activation function, assuming $\frac{\partial Q}{\partial z_i} = q(z_i)$. Generally, the only mathematical requirement for our Lagrangians is that they must be convex functions.

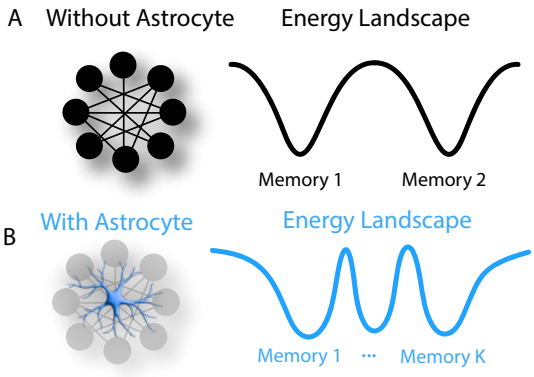



**Figure 3:** Energy landscapes of a neuron-only vs. neuron-astrocyte network. The neuron-astrocyte network stores more memories.

To simplify analysis, we focus on a limiting case where the system demonstrates associative memory. Like other models of biological associative memory, this requires symmetries in the governing equations. Under certain conditions, the neuron-astrocyte model admits a global energy (Lyapunov) function that decreases monotonically and is bounded from below. This guarantees that the system's dynamics converge to fixed-point attractors, corresponding to "memories" stored in the weight matrices. The model thus functions as an energy-based Dense Associative Memory [30, 31], and we show that the presence of a single astrocyte can increase the memory capacity of a neural circuit by a factor of $N$.

We follow the general formulation of energy-based associative memories [32–34], selecting three Lagrangians to define the layers (neurons, synapses, astrocytes) and their activation functions. These are: neural Lagrangian $\mathcal{L}^{[n]}$, synaptic Lagrangian $\mathcal{L}^{[s]}$, and astrocyte process Lagrangian $\mathcal{L}^{[p]}$, which can be arbitrary differentiable functions of their dynamical variables. Details are in Appendix A.From these Lagrangians, we derive three terms in the overall energy function of the neuron-astrocyte system: $E^{[n]}$, $E^{[s]}$, and $E^{[p]}$, using a Legendre transformation. The activation functions are the partial derivatives of the Lagrangians with respect to their corresponding dynamical variables (see Appendix A and Figure 2). Additional contributions to the total energy describe interactions between neurons, synapses, and astrocytes: $E^{[ns]}$, $E^{[ps]}$, and $E^{[pp]}$.

The total energy of the neuron-astrocyte model is:

$$E = E^{[n]} + E^{[s]} + E^{[p]} + E^{[ns]} + E^{[ps]} + E^{[pp]} \tag{5}$$

From this formalism [32–34], the dynamical equations can be written as the negative gradient:

$$\begin{cases} \tau_n \, \dot{x}_i = -\dfrac{\partial E}{\partial \phi_i} = -\lambda \, x_i + \displaystyle\sum_{j=1}^{N} g_{ij} \phi_j \\ \tau_s \, \dot{s}_{ij} = -2\dfrac{\partial E}{\partial g_{ij}} = -\alpha \, s_{ij} + \phi_i \phi_j \; + \; \psi_{ij} \\ \tau_p \, \dot{p}_{ij} = -2\dfrac{\partial E}{\partial \psi_{ij}} = -\gamma \, p_{ij} + \displaystyle\sum_{k,l=1}^{N} T_{ijkl} \, \psi_{kl} \; + \; g_{ij} \end{cases} \tag{6}$$

The equations exhibit symmetry in both parameters and degrees of freedom (e.g., $T_{ijkl} = T_{klij}$), which ensures the existence of a global energy function, aiding mathematical tractability. While real biology may break some symmetries, making analysis harder, we use this energy-based model to theoretically establish memory storage capabilities. The non-symmetric model, studied numerically in section 3, shows similar properties despite lacking the energy-based formulation. Unlike Hopfield networks [35], the symmetry in $\mathbf{T}$ reflects natural calcium diffusion symmetry.

The first two equations in (6) resemble the approach by Dong and Hopfield [36], which models both neural dynamics and synaptic plasticity with a single energy function. Our system differs due to the inclusion of astrocytic processes, which interact with synapses and each other. Using the Lagrangian formalism, it can be shown (see Appendix B) that the energy decreases: i.e., $\frac{dE}{dt} \leq 0$. If each Lagrangian has a positive semidefinite Hessian, the dynamical equations (6) lead to a fixed point since the energy is bounded from below (via the invariant set theorem [37]). Thus, starting from an initial state, the network converges to fixed points, representing associative memory.

## 2.1 Connection to Dense Associative Memory

Energy-based neuron-astrocyte networks are governed by nonlinear differential equations (6) that lead to fixed-point attractors, assuming certain conditions on the Lagrangians are met. These fixed points $x_i^*$, $s_{ij}^*$, and $p_{ij}^*$ correspond to the local minima of the energy function (5), independent of the time scales $\tau_n$, $\tau_s$, and $\tau_p$. While the kinetics of the model depend on these time scales, neurons generally operate on faster time scales than synaptic plasticity and astrocyte processes, $\tau_n \ll \tau_s, \tau_p$. Although we use an "unbiological" choice of time scales for convenience, the final result accurately represents fixed-point locations for the biologically relevant regime $\tau_n \ll \tau_s, \tau_p$. In Appendix C we show that the fixed points of the neuron-astrocyte network coincide with those of the effective neuron-only system. Astrocytes enable four-body neuron interactions, unlike conventional models where synapses connect two neurons [29]. This results in an “effective” four-neuron synapse, which integrates information across distant neurons.

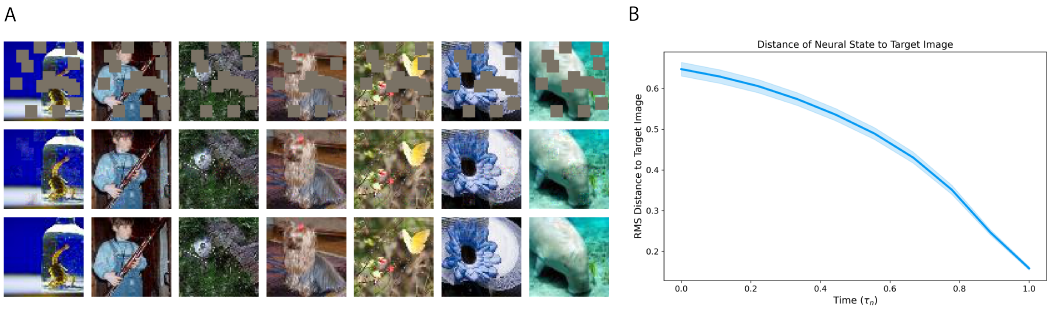



**Figure 4:** A) Error-correcting capabilities of the neuron-astrocyte network, trained with backpropagation, demonstrated with images from the Tiny ImageNet dataset [38]. Top row: masked input; middle row: network’s final state; bottom row: ground-truth image. B) Root-mean-squared distance of the network state to the ground-truth as a function of time (standard error across 64 images).

## 2.2 Storing Memories in Astrocyte Networks

Given $K$ memory patterns $\xi^\mu$ (with index $\mu = 1...K$), the task is to store these patterns in the neural-astrocyte network, enabling the system to retrieve them. The tensor $\mathbf{T}$ is chosen as:

$$T_{ijkl} \equiv \sum_{\mu=1}^{K} \xi_i^\mu \, \xi_j^\mu \, \xi_k^\mu \, \xi_l^\mu \tag{7}$$

This yields an effective neuron-only theory with quartic interactions, akin to Dense Associative Memory models [30, 31]. Dense Associative Memories extend traditional Hopfield Networks [29] by introducing higher-order terms in the energy function, leading to superior storage capacity and representational power [30]. They are also related to attention mechanisms in Transformers [32, 39] and are used in state-of-the-art energy-based models [40, 41].

**Memory Capacity of a Neuron-Astrocyte Network** An insightful question is how many memories the model can store *per compute unit*. Using a conservative definition of compute units, the neuron-

astrocyte model (6) has approximately $N^2$ compute units in the large $N$ limit:

$$N \text{ neurons} + N^2 \text{ synapses} + N^2 \text{ processes} \sim N^2 \text{ compute units}$$

The storage capacity $K^{\max}$ of the Dense Associative Memory model with quartic energy is $K^{\max} \sim N^3$ [30], and thus, the number of memories per compute unit grows linearly:

$$\frac{K^{\max}}{\text{Number of compute units}} \sim N$$

This surpasses other Dense Associative Memory implementations, such as Krotov and Hopfield's [32], where the memory per compute unit remains constant, making neuron-astrocyte networks a promising candidate for biological Dense Associative Memory. The memory storage in our model is attributed to the tensor $\mathbf{T}$, which describes astrocyte process interactions and calcium or molecular transport. Memories can be stored via a Hebbian-like plasticity rule (7), though more sophisticated rules are also possible. Future experiments may confirm this storage mechanism.

**How Many Astrocyte Parameters are Needed?** The Hebbian-like rule (7) requires all-to-all connectivity between astrocyte processes. If we wish to store $K$ memories with $N$ independent bits, we need on the order of $KN$ parameters. For $K = N$, the required number of connections is:

$$KN = rN^2 \quad \Longrightarrow \quad r = \frac{K}{N}$$

Thus, achieving linear storage capacity ($K = N$) allows us to ignore process-to-process connectivity. For supralinear storage, connecting astrocyte processes increases capacity. The number of memories stored depends on $r$, which can potentially be determined experimentally. Interestingly, detailed entries of tensor $T_{ijkl}$ are not required for neuron-astrocyte models to perform meaningful computations. It is shown in Appendix D that setting $T_{ijkl} = 1$ in (3) leads to a model that approximates a transformer's self-attention mechanism [42].

# 3 Simulations

We conducted two experiments: one using the energy-based equations (6) with the Hebbian-like rule (7), and another using backpropagation-through-time (BPTT) without symmetry constraints. The first validates our theoretical claims, while the second shows that strong symmetry, though sufficient, is not necessary for associative memory in biological systems.

**Energy-Based Experiments** We tested the memory storage scheme (7) on the CIFAR10 dataset. Figure 6 shows retrieval of four memories after encoding $K = 25$. As predicted, the network converged to fixed points corresponding to stored memories. Training details are in Appendix E.

**Backpropagation-Based Experiments** To show that symmetry is not required, we trained the network on a self-supervised task using Tiny ImageNet [38], where the network reconstructed masked images using BPTT. Results are shown in Figure 4, with more details in Appendix F. The energy-based model can also be trained with BPTT, using methods like recurrent backpropagation [43, 44].

# 4 Discussion

We have introduced a biologically-inspired model that describes the interactions between neurons, synapses, and astrocytes. In our model, astrocytes are able to adaptively control synaptic weights in an online fashion. Theoretical analysis has demonstrated that this model can exhibit associative memory and is closely related to the Dense Associative Memory family of models with supralinear memory capacity, as well as to transformers. We have shown that, through the choice of the connectivity tensor $T_{ijkl}$, our neuron-astrocyte model can be smoothly dialed between operating as a transformer and operating as a Dense Associative Memory network. This opens up the possibility for exploring novel architectures "in-between" transformers and Dense Associative Memories. Furthermore, we have presented a simple algorithm for memory storage and have provided numerical evidence of our models' effectiveness, such as successfully storing and retrieving CIFAR10 and ImageNet images.

In broader terms, this work proposes that memories can, at least in part, be stored within the molecular machinery of astrocytes. This contrasts with the prevailing neuroscience viewpoint that memories are stored in the synaptic weights between neurons. To experimentally validate this claim, one would need to selectively interfere with the ability of $Ca^{2+}$ to diffuse intracellularly through astrocytes. Our model predicts that hindering this diffusion would significantly impair memory recall. Our model is flexible enough to accommodate many different types of process-to-process coupling patterns, which could presumably be fit to match physiological data. For example, it is possible to enforce "nearest-neighbor" coupling between astrocyte processes (which can be achieved by e.g., imposing a block-diagonal structure on the tensor $T$ such that $T_{ijkl} = 0$ if processes $ij$ and $kl$ are not spatially close to each other), while still guaranteeing convergence of our model to an equilibrium point.

While our focus has been on a mini-circuit consisting of a single astrocyte interacting with multiple nearby synapses, astrocytes also extensively communicate with each other through chemical gap junctions. Exploring the implications of this intercellular coupling will be the subject of future research.

Key ideas in machine learning and AI drew initial inspiration from neuroscience, including neural networks, convolutional nets, threshold linear (ReLu) units, and dropout. Yet it is debatable whether neuroscience research from the last fifty years has significantly influenced or informed machine learning. Astrocytes, along with other biological structures such as dendrites [45] and neuromodulators [46] may offer a fresh source of inspiration for building state-of-the-art AI systems.

# References


[1] Eric R Kandel, James H Schwartz, Thomas M Jessell, Steven Siegelbaum, A James Hudspeth, Sarah Mack, et al. *Principles of neural science*, volume 4. McGraw-hill New York, 2000.

[2] Alexandra Witthoft and George Em Karniadakis. A bidirectional model for communication in the neurovascular unit. *Journal of Theoretical Biology*, 311:80–93, October 2012. ISSN 0022-5193. doi: 10.1016/j.jtbi.2012.07.014.

[3] James Schummers, Hongbo Yu, and Mriganka Sur. Tuned Responses of Astrocytes and Their Influence on Hemodynamic Signals in the Visual Cortex. *Science*, 320(5883):1638–1643, June 2008. doi: 10.1126/science.1156120.

[4] Yu Mu, Davis V Bennett, Mikail Rubinov, Sujatha Narayan, Chao-Tsung Yang, Masashi Tanimoto, Brett D Mensh, Loren L Looger, and Misha B Ahrens. Glia accumulate evidence that actions are futile and suppress unsuccessful behavior. *Cell*, 178(1):27–43, 2019.

[5] Ksenia V Kastanenka, Rubén Moreno-Bote, Maurizio De Pittà, Gertrudis Perea, Abel Eraso-Pichot, Roser Masgrau, Kira E Poskanzer, and Elena Galea. A roadmap to integrate astrocytes into systems neuroscience. *Glia*, 68(1):5–26, 2020.

[6] Adi Kol, Adar Adamsky, Maya Groysman, Tirzah Kreisel, Michael London, and Inbal Goshen. Astrocytes contribute to remote memory formation by modulating hippocampal–cortical communication during learning. *Nature neuroscience*, 23(10):1229–1239, 2020.

[7] Yves Agid and Pierre Magistretti. *Glial Man: A Revolution in Neuroscience*. Oxford University Press, 2020.

[8] Jun Nagai, Xinzhu Yu, Thomas Papouin, Eunji Cheong, Marc R Freeman, Kelly R Monk, Michael H Hastings, Philip G Haydon, David Rowitch, Shai Shaham, et al. Behaviorally consequential astrocytic regulation of neural circuits. *Neuron*, 109(4):576–596, 2021.

[9] Linda Maria Requie, Marta Gómez-Gonzalo, Michele Speggiorin, Francesca Managò, Marcello Melone, Mauro Congiu, Angela Chiavegato, Annamaria Lia, Micaela Zonta, Gabriele Losi, et al. Astrocytes mediate long-lasting synaptic regulation of ventral tegmental area dopamine neurons. *Nature Neuroscience*, 25(12):1639–1650, 2022.

[10] Peter Rupprecht, Sian N Duss, Denise Becker, Christopher M Lewis, Johannes Bohacek, and Fritjof Helmchen. Centripetal integration of past events in hippocampal astrocytes regulated by locus coeruleus. *Nature Neuroscience*, pages 1–13, 2024.

[11] Rahul Srinivasan, Ben S Huang, Sharmila Venugopal, April D Johnston, Hua Chai, Hongkui Zeng, Peyman Golshani, and Baljit S Khakh. Ca2+ signaling in astrocytes from ip3r2-/- mice in brain slices and during startle responses in vivo. *Nature neuroscience*, 18(5):708–717, 2015.

[12] Alexey Semyanov and Alexei Verkhratsky. Astrocytic processes: from tripartite synapses to the active milieu. *Trends in neurosciences*, 44(10):781–792, 2021.

[13] Kyungchul Noh, Woo-Hyun Cho, Byung Hun Lee, Dong Wook Kim, Yoo Sung Kim, Keebum Park, Minkyu Hwang, Ellane Barcelon, Yoon Kyung Cho, C Justin Lee, et al. Cortical astrocytes modulate dominance behavior in male mice by regulating synaptic excitatory and inhibitory balance. *Nature Neuroscience*, 26(9):1541–1554, 2023.

[14] Roberta de Ceglia, Ada Ledonne, David Gregory Litvin, Barbara Lykke Lind, Giovanni Carriero, Emanuele Claudio Latagliata, Erika Bindocci, Maria Amalia Di Castro, Iaroslav Savtchouk, Ilaria Vitali, et al. Specialized astrocytes mediate glutamatergic gliotransmission in the cns. *Nature*, pages 1–10, 2023.

[15] Ciaran Murphy-Royal, ShiNung Ching, and Thomas Papouin. A conceptual framework for astrocyte function. *Nature Neuroscience*, pages 1–9, 2023.

[16] Akinobu Suzuki, Sarah A Stern, Ozlem Bozdagi, George W Huntley, Ruth H Walker, Pierre J Magistretti, and Cristina M Alberini. Astrocyte-neuron lactate transport is required for long-term memory formation. *Cell*, 144(5):810–823, 2011.

[17] António Pinto-Duarte, Amanda J Roberts, Kunfu Ouyang, and Terrence J Sejnowski. Impairments in remote memory caused by the lack of type 2 ip3 receptors. *Glia*, 67(10):1976–1989, 2019.

[18] Wenfei Sun, Zhihui Liu, Xian Jiang, Michelle B Chen, Hua Dong, Jonathan Liu, Thomas C Südhof, and Stephen R Quake. Spatial transcriptomics reveal neuron–astrocyte synergy in long-term memory. *Nature*, pages 1–8, 2024.

[19] Maria Amalia Di Castro, Julien Chuquet, Nicolas Liaudet, Khaleel Bhaukaurally, Mirko Santello, David Bouvier, Pascale Tiret, and Andrea Volterra. Local ca2+ detection and modulation of synaptic release by astrocytes. *Nature neuroscience*, 14(10):1276–1284, 2011.

[20] Jillian L. Stobart, Kim David Ferrari, Matthew J. P. Barrett, Chaim Glück, Michael J. Stobart, Marc Zuend, and Bruno Weber. Cortical Circuit Activity Evokes Rapid Astrocyte Calcium Signals on a Similar Timescale to Neurons. *Neuron*, 98(4):726–735.e4, May 2018. ISSN 0896-6273. doi: 10.1016/j.neuron.2018.03.050.

[21] German Barrionuevo and Thomas H Brown. Associative long-term potentiation in hippocampal slices. *Proceedings of the National Academy of Sciences*, 80(23):7347–7351, 1983.

[22] Henry Markram, Joachim Lübke, Michael Frotscher, and Bert Sakmann. Regulation of synaptic efficacy by coincidence of postsynaptic aps and epsps. *Science*, 275(5297):213–215, 1997.

[23] Gertrudis Perea, Marta Navarrete, and Alfonso Araque. Tripartite synapses: Astrocytes process and control synaptic information. *Trends in Neurosciences*, 32(8):421–431, August 2009. ISSN 0166-2236. doi: 10.1016/j.tins.2009.05.001.

[24] Nicola J Allen and Cagla Eroglu. Cell biology of astrocyte-synapse interactions. *Neuron*, 96(3): 697–708, 2017.

[25] M. M. Halassa, T. Fellin, H. Takano, J.-H. Dong, and P. G. Haydon. Synaptic Islands Defined by the Territory of a Single Astrocyte. *Journal of Neuroscience*, 27(24):6473–6477, June 2007. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.1419-07.2007.

[26] Alfonso Araque, Giorgio Carmignoto, Philip G. Haydon, Stéphane H. R. Oliet, Richard Robitaille, and Andrea Volterra. Gliotransmitters Travel in Time and Space. *Neuron*, 81(4): 728–739, February 2014. ISSN 0896-6273. doi: 10.1016/j.neuron.2014.02.007.

[27] Maurizio De Pittà, Nicolas Brunel, and Andrea Volterra. Astrocytes: Orchestrating synaptic plasticity? *Neuroscience*, 323:43–61, 2016.

[28] Hugh R. Wilson and Jack D. Cowan. Excitatory and Inhibitory Interactions in Localized Populations of Model Neurons. *Biophysical Journal*, 12(1):1–24, 1972. ISSN 0006-3495. doi: 10.1016/S0006-3495(72)86068-5.

[29] John J Hopfield. Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the national academy of sciences*, 81(10):3088–3092, 1984.

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

[31] Dmitry Krotov. A new frontier for hopfield networks. *Nature Reviews Physics*, pages 1–2, 2023.

[32] Dmitry Krotov and John J Hopfield. Large associative memory problem in neurobiology and machine learning. In *International Conference on Learning Representations*, 2021.

[33] Dmitry Krotov. Hierarchical Associative Memory. https://arxiv.org/abs/2107.06446v1, July 2021.

[34] Benjamin Hoover, Duen Horng Chau, Hendrik Strobelt, and Dmitry Krotov. A universal abstraction for hierarchical hopfield networks. In *The Symbiosis of Deep Learning and Differential Equations II*, 2022.

[35] J J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the United States of America*, 79 (8):2554–2558, April 1982. ISSN 0027-8424.

[36] Dawei W Dong and John J Hopfield. Dynamic properties of neural networks with adapting synapses. *Network: Computation in Neural Systems*, 3(3):267, 1992.

[37] Jean-Jacques E Slotine and Weiping Li. *Applied nonlinear control*, volume 199. Prentice hall Englewood Cliffs, NJ, 1991.

[38] Jiayu Wu, Qixiang Zhang, and Guoxi Xu. Tiny imagenet challenge. *Technical report*, 2017.

[39] Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Lukas Gruber, Markus Holzleitner, Thomas Adler, David Kreil, Michael K Kopp, et al. Hopfield networks is all you need. In *International Conference on Learning Representations*, 2021.

[40] Benjamin Hoover, Yuchen Liang, Bao Pham, Rameswar Panda, Hendrik Strobelt, Duen Horng Chau, Mohammed Zaki, and Dmitry Krotov. Energy transformer. *Advances in Neural Information Processing Systems*, 36, 2024.

[41] Benjamin Hoover, Hendrik Strobelt, Dmitry Krotov, Judy Hoffman, Zsolt Kira, and Duen Horng Chau. Memory in plain sight: A survey of the uncanny resemblances between diffusion models and associative memories. *arXiv preprint arXiv:2309.16750*, 2023.

[42] Leo Kozachkov, Ksenia V Kastanenka, and Dmitry Krotov. Building transformers from neurons and astrocytes. *Proceedings of the National Academy of Sciences*, 120(34):e2219150120, 2023.

[43] Fernando J Pineda. Generalization of back-propagation to recurrent neural networks. *Physical review letters*, 59(19):2229, 1987.

[44] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. Deep equilibrium models. *Advances in Neural Information Processing Systems*, 32, 2019.

[45] Kwabena Boahen. Dendrocentric learning for synthetic intelligence. *Nature*, 612(7938):43–50, 2022.

[46] Yuhan Helena Liu, Stephen Smith, Stefan Mihalas, Eric Shea-Brown, and Uygar Sümbül. Biologically-plausible backpropagation through arbitrary timespans via local neuromodulators. *Advances in Neural Information Processing Systems*, 35:17528–17542, 2022.

[47] Danil Tyulmankov, Ching Fang, Annapurna Vadaparty, and Guangyu Robert Yang. Biological learning in key-value memory networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 22247–22258. Curran Associates, Inc., 2021.

[48] Wei Wang and Jean-Jacques E Slotine. On partial contraction analysis for coupled nonlinear oscillators. *Biological cybernetics*, 92(1):38–53, 2005.

[49] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.

# A Definitions of Lagrangians and Energy

As described in the main text, the Lagrangians are: a neural Lagrangian $\mathcal{L}^{[n]}$, a synaptic Lagrangian $\mathcal{L}^{[s]}$, and an astrocyte process Lagrangian $\mathcal{L}^{[p]}$. In general these scalar functions can be arbitrary (differentiable) functions of the corresponding dynamical variables. The activation functions are defined as partial derivatives of the Lagrangians

$$\underbrace{\mathcal{L}^{[n]}(\mathbf{x}) \;\rightarrow\; \phi_i \equiv \frac{\partial \mathcal{L}^{[n]}}{\partial x_i}}_{\text{Neural Lagrangian}}, \qquad \underbrace{\mathcal{L}^{[s]}(\mathbf{s}) \;\rightarrow\; g_{ij} \equiv \frac{\partial \mathcal{L}^{[s]}}{\partial s_{ij}}}_{\text{Synaptic Lagrangian}}, \qquad \underbrace{\mathcal{L}^{[p]}(\mathbf{p}) \;\rightarrow\; \psi_{ij} \equiv \frac{\partial \mathcal{L}^{[p]}}{\partial p_{ij}}}_{\text{Astrocyte Process Lagrangian}} \tag{8}$$

One possible choice of these functions is additive: summing each contribution from all the individual computational elements (e.g., individual neurons), which results in activation functions that depend only on individual computational elements – for instance, $\phi(x_i) = \tanh(x_i)$. More general choices of the Lagrangians allow for "collective" activation functions, which depend on the dynamical degrees of freedom of several or all the computational elements in a given layer, for example a softmax.

From the Lagrangians (8), we may derive via a Legendre transform three terms in the overall energy function of the neuron-astrocyte system, corresponding to three layer energies,

$$E^{[n]} + E^{[s]} + E^{[p]} = \underbrace{\lambda\Big[\sum_{i=1}^{N} x_i \phi_i - \mathcal{L}^{[n]}\Big]}_{\text{Neural Energy}} + \underbrace{\frac{\alpha}{2}\Big[\sum_{i,j=1}^{N} s_{ij} g_{ij} - \mathcal{L}^{[s]}\Big]}_{\text{Synaptic Energy}} + \underbrace{\frac{\gamma}{2}\Big[\sum_{i,j=1}^{N} p_{ij}\psi_{ij} - \mathcal{L}^{[p]}\Big]}_{\text{Astrocyte Process Energy}} \tag{9}$$

where for simplicity of the presentation we dropped the input signals, $b_i = c_{ij} = d_{ij} = 0$. The remaining contributions to the total energy of the system describe the interactions between these three layers. These contributions describe the synapse-mediated interactions between the neurons $E^{[ns]}$, the interactions between the processes and the synapses $E^{[ps]}$, and the interactions between the individual processes inside the astrocyte $E^{[pp]}$,

$$\begin{aligned} E^{[ns]} + E^{[ps]} + E^{[pp]} = -\Bigg[ & \frac{1}{2}\sum_{i,j=1}^{N} g_{ij}(\mathbf{s})\phi_i(\mathbf{x})\phi_j(\mathbf{x}) \\ & + \frac{1}{2}\sum_{i,j=1}^{N} \psi_{ij}(\mathbf{p}) g_{ij}(\mathbf{s}) \\ & + \frac{1}{4}\sum_{i,j,k,l=1}^{N} T_{ijkl}\, \psi_{ij}(\mathbf{p})\psi_{kl}(\mathbf{p})\Bigg] \end{aligned} \tag{10}$$

The overall energy function of the neuron-synapse-astrocyte model can now be written as the sum of these six terms

$$E = E^{[n]} + E^{[s]} + E^{[p]} + E^{[ns]} + E^{[ps]} + E^{[pp]} \tag{11}$$

As mentioned previously, the energy-based equations have a large amount of symmetry–both in the parameters and the dynamical degrees of freedom. Specifically, $s_{ij} = s_{ji}$, $g_{ij} = g_{ji}$, $p_{ij} = p_{ji}$, $\psi_{ij} = \psi_{ji}$, and $T_{ijkl} = T_{klij}$, $T_{ijkl} = T_{jikl}$, $T_{ijkl} = T_{ijlk}$. These symmetries, are needed for the existence of the global energy function for our neuron-astrocyte network, which leads to mathematical tractability. In real biology some (or all) of these symmetries might be broken, and the analytical tractability might be more difficult or even impossible. We use the energy-based model to establish theoretically the memory storage capabilities of our model. The non-symmetric model is studied numerically in section 3.

# B Proof of Decreasing Energy Function

The overall time derivative of the energy function may be written as

$$\frac{dE}{dt} = \sum_{i=1}^{N} \frac{\partial E}{\partial x_i}\, \dot{x}_i + \sum_{i,j=1}^{N} \frac{\partial E}{\partial s_{ij}}\, \dot{s}_{ij} + \sum_{i,j=1}^{N} \frac{\partial E}{\partial p_{ij}}\, \dot{p}_{ij}$$

which may be expressed using the chain rule as

$$\begin{aligned} \frac{dE}{dt} &= \sum_{i,j=1}^{N} \frac{\partial E}{\partial \phi_i}\, \frac{\partial \phi_i}{\partial x_j}\, \dot{x}_j + \sum_{i,j,k,l=1}^{N} \frac{\partial E}{\partial g_{ij}} \frac{\partial g_{ij}}{\partial s_{kl}}\, \dot{s}_{kl} + \sum_{i,j,k,l=1}^{N} \frac{\partial E}{\partial \psi_{ij}} \frac{\partial \psi_{ij}}{\partial p_{kl}}\, \dot{p}_{kl} \\ &= \sum_{i,j=1}^{N} \frac{\partial E}{\partial \phi_i}\, \frac{\partial^2 \mathcal{L}^{[n]}}{\partial x_i \partial x_j}\, \dot{x}_j + \sum_{i,j,k,l=1}^{N} \frac{\partial E}{\partial g_{ij}} \frac{\partial^2 \mathcal{L}^{[s]}}{\partial s_{ij} \partial s_{kl}}\, \dot{s}_{kl} + \sum_{i,j,k,l=1}^{N} \frac{\partial E}{\partial \psi_{ij}} \frac{\partial^2 \mathcal{L}^{[p]}}{\partial p_{ij} \partial p_{kl}}\, \dot{p}_{kl} \end{aligned} \tag{12}$$

The second line follows from the definition of the Lagrangians (8). Plugging the dynamics defined in equations (6) into this last expression, we get the desired result, provided that the Lagrangians are all convex (i.e., have positive semi-definite Hessians)

$$\frac{dE}{dt} = -\left[ \tau_n \sum_{i,j=1}^{N} \dot{x}_i\, \frac{\partial^2 \mathcal{L}^{[n]}}{\partial x_i \partial x_j}\, \dot{x}_j + \frac{\tau_s}{2} \sum_{i,j,k,l=1}^{N} \dot{s}_{ij}\, \frac{\partial^2 \mathcal{L}^{[s]}}{\partial s_{ij}\, \partial s_{kl}}\, \dot{s}_{kl} + \frac{\tau_p}{2} \sum_{i,j,k,l=1}^{N} \dot{p}_{ij}\, \frac{\partial^2 \mathcal{L}^{[p]}}{\partial p_{ij} \partial p_{kl}}\, \dot{p}_{kl} \right] \leq 0 \tag{13}$$

# C Effective Energy

The fixed points of the synaptic and astrocyte dynamics in (6) are defined by:

$$\begin{cases} \psi_{ij} = -\phi_i \phi_j \\ g_{ij} = \sum\limits_{k,l=1}^{N} T_{ijkl} \phi_k \phi_l \end{cases}$$

For fixed $\phi_i$, these equations uniquely determine $s_{ij}$ and $p_{ij}$ when $g$ and $\psi$ are strictly monotonic. Substituting this into the neural dynamics equation, we get:

$$\tau_n \dot{x}_i = -x_i + \sum_{j,k,l=1}^{N} T_{ijkl}\, \phi_j \phi_k \phi_l \tag{14}$$

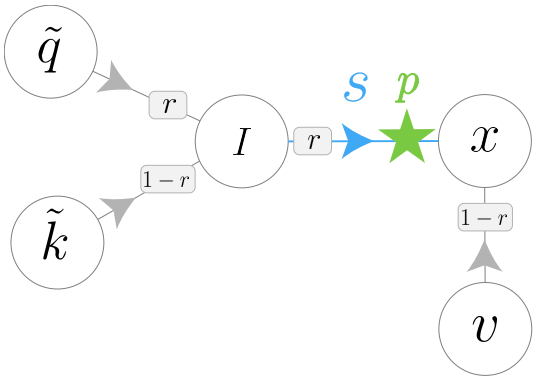
The corresponding effective energy is:

$$E^{\text{eff}} = \Big[ \sum_{i=1}^{N} x_i \phi_i - \mathcal{L}^{[n]} \Big] - \frac{1}{4} \sum_{i,j,k,l=1}^{N} T_{ijkl}\, \phi_i \phi_j \phi_k \phi_l \tag{15}$$

These equations capture the essence of our argument: the fixed points of the neuron-astrocyte network coincide with those of the effective neuron-only system. Astrocytes enable four-body neuron interactions, unlike conventional models where synapses connect two neurons [29]. This results in an "effective" four-neuron synapse, which integrates information across distant neurons.

# D Proof of Neuron-Astrocyte Equilibration to Transformer Output

**Neuron-Astrocyte Transformer Architecture** The aim of this section is to demonstrate that a simple selection of the astrocyte process-to-process weights $T_{ijkl} = 1$ is sufficient, along with a specialized architecture (Figure 5), to produce interesting computations in the general neuron-astrocyte network equations (1), (2), (3). We consider a single group of $N$ neurons, where the state




**5:** Dynamic, stable neuron-astrocyte architecture which implements the self-attention operation in transformers.

of the $i$-th neuron in this group is denoted by $x_i$. These neurons receive inputs from another group of $M$ neurons, where the state of the $j$-th neuron in this group is denoted $I_j$. The synaptic connection between neuron $I_j$ and neuron $x_i$ is represented by $s_{ij}$. The $x_i$ neurons also receive input from another group of $N$ neurons, whose state we denote by $v_i$, for reasons that will become clear later on. The dynamical equations for the $x_i$ layer are given by

$$\tau_n \dot{x}_i = -x_i + r \sum_{j=1}^{M} s_{ij} I_j + (1-r) v_i \tag{16}$$

where $r = \{0, 1\}$ stands for "read", and is a global parameter controlling whether the network is in "read" or "write" mode. Biologically, global coordination of this kind may be achieved by neuromodulators (e.g., acetylcholine) [47]. We additionally assume that the $I_j$ neurons receive strong input from two $M$-dimensional neural populations which we denote as $\tilde{q}_j$ and $\tilde{k}_j$ (again for reasons that will become clear shortly), so that the state of neuron $I_j$ is given by

$$I_j = r\, \tilde{q}_j + (1-r)\, \tilde{k}_j \tag{17}$$

The synaptic weights $s_{ij}$ are modulated by an astrocyte and evolve according to the following dynamical equations:

$$\tau_s \dot{s}_{ij} = -p_{ij} s_{ij} + c_{ij} \tag{18}$$

where $p_{ij}$ represents the state of the astrocyte process $ij$, and $c_{ij}$ is a fixed bias term. This set of synaptic equations can be associated with equations (2) by setting

$$\alpha = 0, \quad \text{and} \quad f(s_{ij}, x_i, x_j, p_{ij}) = -p_{ij} s_{ij}$$

The astrocyte dynamics are described by simple diffusive equations:

$$\tau_p \dot{p}_{ij} = \sum_{k=1}^{N} \sum_{l=1}^{M} [p_{kl} - p_{ij}] \quad \text{with} \quad \sum_{i=1}^{N} \sum_{j=1}^{M} p_{ij}(0) > 0 \tag{19}$$

The inequality is to ensure that the total amount of $Ca^{2+}$ initially in the astrocyte is positive. Biologically, even $Ca^{2+}$ concentrations inside individual processes are positive $p_{ij}(0) \geq 0$, but, mathematically, we will only use the positivity of the total amount of calcium inside the astrocyte. Similar to the synaptic variables, this set of astrocyte equations can be associated with the astrocyte equations (3) by setting

$$\psi(p_{ij}) = p_{ij}, \quad \gamma = NM, \quad T_{ijkl} = 1, \quad \kappa(s_{ij}) = 0 \quad \text{and} \quad d_{ij} = 0$$

Before establishing a connection with transformer networks, we will describe the dynamical properties of Equations (16), (18), and (19). Specifically, we will demonstrate that, during the reading phase, the neurons $x_i$ converge to an equilibrium point determined solely by the input neurons $I_j$, the initial $Ca^{2+}$ concentration in the astrocyte, and the synaptic bias terms $c_{ij}$. Following this, we will illustrate how a judicious and biologically plausible selection of input neuron states, initial $Ca^{2+}$ levels, and synaptic biases enables the neurons $x_i$ to mimic the output of the self-attention mechanism in transformers.

**Convergence & Synchronization of Astrocyte Processes** To begin, note that the astrocyte equations (19) are autonomous with respect to the neural and synaptic variables. Therefore, we can analyze their convergence properties independently from these variables. In particular, we can show that the astrocyte equations synchronize to the average of their initial conditions. To see this, first note that the total amount of $Ca^{2+}$ in the astrocyte, which we denoted $z$, is conserved throughout the diffusion process

$$z \equiv \sum_{i=1}^{N}\sum_{j=1}^{M} p_{ij} \qquad \Longrightarrow \qquad \dot{z} = \sum_{i=1}^{N}\sum_{j=1}^{M} \dot{p}_{ij} = 0$$

Second, note that this property implies that if the astrocyte processes *synchronize*, i.e., $p_{ij} = p_{kl} = p^*$, then the state of each astrocyte process must converge to the average of the astrocyte initial conditions, because

$$z(t) \;=\; NMp^* \;=\; z(0) \;=\; \sum_{i=1}^{N}\sum_{j=1}^{M} p_{ij}(0) \qquad \Longrightarrow \qquad p^* = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} p_{ij}(0) \;>\; 0 \quad (20)$$

The inequality follows from the assumption in (19), that the total initial amount of $Ca^{2+}$ in the astrocyte is positive. To prove that the astrocyte processes in fact synchronize, one can use a virtual system, as in [48] or a Lyapunov-like function

$$L = \frac{1}{2}\,(p_{ij} - p_{kl})^2 \geq 0$$

for arbitrary indices $ij$ and $kl$. Taking the time derivative of this function, one sees that

$$\dot{L} = (p_{ij} - p_{kl})(\dot{p}_{ij} - \dot{p}_{kl}) = -\frac{NM}{\tau_p}(p_{ij} - p_{kl})^2 = -\frac{2NM}{\tau_p}L \qquad \Longrightarrow \qquad L(t) = L(0)e^{-\frac{2NMt}{\tau_p}}$$

which shows that the astrocyte processes do in fact synchronize (i.e., $|p_{ij} - p_{kl}| \to 0$) exponentially with rate $\frac{NM}{\tau_p}$.

**Convergence of Synapses** Moving on to the synaptic equations (18), we will assume that the astrocyte processes have converged to $p^* > 0$. This assumption is justified because, as the preceding paragraph shows, the converge of the astrocyte process to $p^*$ is *exponential*, meaning that $p_{ij}$ can be brought arbitrarily close to $p^*$ after finite time. Because $c_{ij}$ is a constant, and because $p^*$ is strictly positive, this implies that the synapses simply converge exponentially quickly to the value

$$s_{ij}^* = \frac{c_{ij}}{p^*} \tag{21}$$

**Convergence of Neurons** Following a similar logic, the neural equations (16) converge exponentially. When the network is in its writing phase (i.e., $r = 0$), the neurons converge to the equilibrium point

$$x_i^* = v_i \tag{22}$$

otherwise, when the network is in the reading phase (i.e., $r = 1$), the network converges exponentially to the equilibrium point

$$x_i^* = \sum_{j=1}^{M} s_{ij}^* I_j = \frac{1}{p^*}\sum_{j=1}^{M} c_{ij} I_j = \frac{NM\sum\limits_{j=1}^{M} c_{ij} I_j}{\sum\limits_{i=1}^{N}\sum\limits_{j=1}^{M} p_{ij}(0)} \tag{23}$$

The first equality was obtained by substituting in $s_{ij}^*$ from (21), while the second equality was obtained by subsituting in the value of $p^*$ from (20).

**Transformer Self-Attention** We are now in a position to relate the neural fixed point (23) to the output of the self-attention mechanism in transformers. To establish this connection, we define several important terms. Consider a set of $K_{\text{tok}}$ *tokens*, which are vectors in $\mathbb{R}^D$. As is standard in transformer architectures, these tokens are transformed via three linear mappings into three new sets of vectors known as keys, queries, and values. By collecting these transformed vectors into matrices, we denote

$$K,\ Q \in\ \mathbb{R}^{K_{\text{tok}} \times D} \qquad \text{and} \qquad V\ \in\ \mathbb{R}^{K_{\text{tok}} \times N}.$$

The *self-attention* matrix $A$ associated with these matrices is given by

$$A_{\mu i} = \sum_{\beta=1}^{K_{\text{tok}}} \frac{\exp\left(\sum\limits_{s=1}^{D} Q_{\mu s} K_{\beta s}\right) V_{\beta i}}{\sum\limits_{\sigma=1}^{K_{\text{tok}}} \exp\left(\sum\limits_{s=1}^{D} Q_{\mu s} K_{\sigma s}\right)}$$

An important characteristic of the above self-attention matrix is that it may be approximated via feature maps [49] with the following property

$$\phi(\mathbf{x})^T \phi(\mathbf{y}) \approx \exp(\mathbf{x}^T \mathbf{y})$$

where $\mathbf{x}$ and $\mathbf{y}$ are two vectors. In general, the output dimension of $\phi$, which we denote $M$ (the same $M$ as above) is much larger than the input dimension $D$. To keep notations clean, we define the output of these feature maps (applied column-wise to the matrices K and Q) as

$$\tilde{K}, \tilde{Q}\ \equiv\ \phi(K), \phi(Q)\ \ \in\ \ \mathbb{R}^{K_{tok} \times M}$$

With this notation, we have that

$$A_{\mu i} \approx \sum_{\beta=1}^{K_{\text{tok}}} \frac{\sum\limits_{j=1}^{M} \tilde{Q}_{\mu j} \tilde{K}_{\beta j} V_{\beta i}}{\sum\limits_{\sigma=1}^{K_{\text{tok}}} \sum\limits_{j=1}^{M} \tilde{Q}_{\mu j} \tilde{K}_{\sigma j}}$$

**Neuron-Astrocyte Self-Attention** To make a connection to the fixed point equation (23), we first rearrange the above terms as follows

$$A_{\mu i} \approx \frac{\sum\limits_{j=1}^{M} \left(\sum\limits_{\beta=1}^{K_{\text{tok}}} V_{\beta i} \tilde{K}_{\beta j}\right) \tilde{Q}_{\mu j}}{\sum\limits_{j=1}^{M} \tilde{Q}_{\mu j} \sum\limits_{\sigma=1}^{K_{\text{tok}}} \tilde{K}_{\sigma j}} \tag{24}$$

We then set the bias terms $c_{ij}$ in the synaptic equations as follows:

$$c_{ij} = \frac{1}{M} \sum_{\beta=1}^{K_{\text{tok}}} V_{\beta i} \tilde{K}_{\beta j} \tag{25}$$

Biologically, this corresponds to a simple form of Hebbian learning between two groups of neurons. Within the framework of (23), this can be achieved during the writing phase (i.e., $r = 0$), such that $x_i = v_i = V_{\beta i}$ and $I_j = \tilde{k}_j \equiv \tilde{K}_{\mu j}$ (from (17)). Then, updating $c_{ij}$ by adding the product of these two terms for each $\beta$ represents a simple form of associative Hebbian learning, and yields (25). Assuming $c_{ij}$ is initially zero, we see that

$$\Delta c_{ij} = \frac{1}{M} x_i I_j = \frac{1}{M} V_{\beta i} \tilde{K}_{\beta j} \qquad \Longrightarrow \qquad c_{ij} = \frac{1}{M} \sum_{\beta=1}^{K_{\text{tok}}} V_{\beta i} \tilde{K}_{\beta j}$$

Finally, during the reading phase ($r = 1$) we select an index $\mu$ in the token sequence to run the neuron-astrocyte dynamics forward on. In other words, $c_{ij}$ is fixed across all tokens, but $I_j$ and

$p_{ij}(0)$ change from token to token. For a particular index $\mu$ we instantiate the neurons $I_j$ (17) and the astrocyte processes $p_{ij}$ as follows

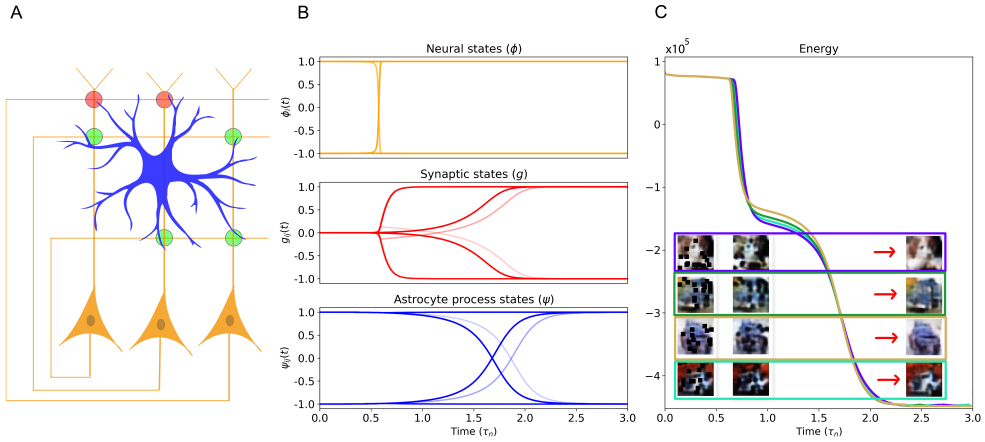
$$I_j = \tilde{q}_j \equiv \tilde{Q}_{\mu j} \qquad \text{and} \qquad p_{ij}(0) = \tilde{Q}_{\mu j} \sum_{\sigma=1}^{K_{tok}} \tilde{K}_{\sigma j} \tag{26}$$

Plugging (25) and (26) into the neural fixed point condition for the reading phase (23), we arrive at the desired result

$$x_i^* = \frac{NM \sum\limits_{j=1}^{M} c_{ij} I_j}{\sum\limits_{i=1}^{N}\sum\limits_{j=1}^{M} p_{ij}(0)} = \frac{\frac{NM}{M} \sum\limits_{j=1}^{M}\sum\limits_{\beta=1}^{K_{\text{tok}}} V_{\beta i}\tilde{K}_{\beta j}\tilde{Q}_{\mu j}}{\sum\limits_{i=1}^{N}\sum\limits_{j=1}^{M} \tilde{Q}_{\mu j}\sum\limits_{\sigma=1}^{K_{\text{tok}}} \tilde{K}_{\sigma j}} = \frac{\frac{NM}{M} \sum\limits_{j=1}^{M}\left(\sum\limits_{\beta=1}^{K_{\text{tok}}} V_{\beta i}\tilde{K}_{\beta j}\right)\tilde{Q}_{\mu j}}{N\sum\limits_{j=1}^{M} \tilde{Q}_{\mu j}\sum\limits_{\sigma=1}^{K_{\text{tok}}} \tilde{K}_{\sigma j}} \approx A_{\mu i}$$

which shows that for a particular choice of parameters and initialization, the neuron-astrocyte network converges to the output of self-attention. In other words, the neural fixed point equation (23) is equal to the self-attention approximation (24).

# E Details of Energy Network Experiments




**6:** A) A schematic for our associative neuron-synapse-astrocyte network. B) The neural, synaptic, and astrocyte process activations during memory retrieval. In this case, the memory item being retrieved is an image of a dog taken from the CIFAR10 dataset. C) Decreasing energy function of the neuron-synapse-astrocyte network as the dynamics evolve. The decreasing energy functions during four different retrievals are shown.

To reduce the dimensionality of the problem, we use a custom autoencoder to encode the 3072-dimensional ($32 \times 32 \times 3$) CIFAR10 images into a smaller, 768 dimensional, latent space. A single CIFAR10 image in this latent space corresponds to a single memory $\xi^\mu$. In addition to being 768-dimensional, this latent space was also binary, so that $\xi^\mu \in [-1, 1]^{768}$. To ensure that the latent space was binary, we wrote a custom autograd function which outputs the sign of the argument during the forward pass, but is linear during the backwards pass. The discrepancy between forward and backward pass induces a small amount of gradient noise in the training process, which is not significant enough to impair learning. For concreteness, in PyTorch this custom activation is given by:

```
class RoundWithGradient(torch.autograd.Function):
    @staticmethod
    def forward(ctx, x):
        return torch.sign(x)

    @staticmethod
```

```
    def backward(ctx, grad_output):
        return grad_output


def round_with_gradient(x):
    return RoundWithGradient.apply(x)
```

To initialize the network, we reasoned (in analogy with traditional Hopfield networks) that the entire system should be initialized close to a stored memory. In our case, this includes all dynamical variables: neuron, synapses, and astrocytes. To do this, we set the time derivatives in (6) equal to zero, clamped the neural state at the corrupted memory $x_0$, and then solved the resulting set of algebraic equations for $p_{ij}(0)$ and $s_{ij}(0)$. Note that the synaptic states and process states are uniquely determined given a fixed neural state, due to the invertibility of $g$ and $\psi$.

# F Details of Backpropagation Experiment

To reduce the dimensionality of the problem, we assume that the the state of the processes does not depend on index $i$, in other words $p_{ij} = p_j$. Biologically, this has the interpretation that the astrocyte processes associated with post-synaptic neuron $i$ are all synchronized. This can be justified by assuming that nearby astrocyte processes are sensitive to inputs arrive at the dendritic tree of neuron $i$, and can rapidly redistribute their $Ca^{2+}$ levels. Similarly, we assume that the weights $T_{ijkl}$ between astrocyte processes $ij$ and $kl$ is only a function of indices $j$ and $l$. We likewise assume that the synapses only receive pre-synaptic input. That is,

$$\begin{aligned}
\tau\, \dot{x}_i &= -x_i + \sum_{j=1}^{N} g_{ij}\phi_j + b_i \\
\tau\, \dot{s}_{ij} &= -s_{ij} + \phi_j + \psi_j \\
\tau\, \dot{p}_j &= -p_j + \sum_{l=1}^{N} T_{jl}\psi_l + s_j
\end{aligned}$$

where $g_{ij} = W_{ij}\tanh(s_{ij})$, $W_{ij}$ is a trainable parameter, and $\psi$ and $\phi$ are both also hyperbolic tangent. To match the dimensionality of the Tiny ImageNet dataset, our network contains $N = 12288 = 64 \times 64 \times 3$ neurons. We numerically integrate the network using Euler integration for 20 timesteps, using a step-size of dt $= 0.1\tau$. We set $\tau = 1$ in our experiment. As described in the main text, we initialized the neurons in the network as the masked images. The synapses and astrocyte processes we initialized at zero. The output of the network was a linear layer followed by a sigmoid function, to ensure valid RGB values. The network was trained using the Adam optimizer with a learning rate of $0.001$, using a batch size of 64 images. We trained on a subset of 5000 images in the TinyImage dataset, which enabled our network to learn quickly.

# NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We do not claim anything other than what is proven.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Please refer to the discussion.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: To the best of our knowledge, the proofs prsented herein are self-contained.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We will provide all the code which generated the figures.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code is available through this GitHub repository .

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes, in the code.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: The results are not statistical, they are theoretically guaranteed.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All code can be run on Google Colab.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: There are no ethical concerns in this work, as far as we know.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: The results are too theoretical at the moment to speculate as to the downstream societal consequences.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper is theoretical.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We are the only creators.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: There are no assets introduced here.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: We did not use crowdsourcing.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: We did not use an IRB or do any experiments on people.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.