Learning and Organization of Memory for Evolving Patterns

Oskar H. Schnaack

*Max Planck Institute for Dynamics and Self-organization, Am Faßberg 17, 37077 Göttingen, Germany
and Department of Physics, University of Washington,
3910 15th Avenue Northeast, Seattle, Washington 98195, USA*

Luca Peliti⬤

Santa Marinella Research Institute, 00058 Santa Marinella, Italy

Armita Nourmohammad⬤*

*Max Planck Institute for Dynamics and Self-organization, Am Faßberg 17, 37077 Göttingen, Germany;
Department of Physics, University of Washington,
3910 15th Avenue Northeast, Seattle, Washington 98195, USA;
and Fred Hutchinson Cancer Research Center,
1100 Fairview Avenue North, Seattle, Washington 98109, USA*

Storing memory for molecular recognition is an efficient strategy for responding to external stimuli. Biological processes use different strategies to store memory. In the olfactory cortex, synaptic connections form when stimulated by an odor and establish an associative distributed memory that can be retrieved upon reexposure to the same odors. In contrast, the immune system encodes specialized memory by diverse receptors that can recognize a multitude of evolving pathogens. Despite the mechanistic differences between memory storage in the olfactory system and the immune system, these processes can still be viewed as different information encoding strategies. Here, we develop analytical and numerical techniques for a generalized Hopfield network to probe the utility of distinct memory strategies against both static and dynamic (evolving) patterns. We find that while classical Hopfield networks with distributed memory can efficiently encode a memory of static patterns, they are inadequate against evolving patterns. To follow an evolving pattern, we show that a Hopfield network should use a higher learning rate, which can in turn distort the energy landscape associated with the stored memory attractors. Specifically, we observe the emergence of narrow connecting paths between memory attractors that lead to misclassification of evolving patterns. We demonstrate that compartmentalized networks with specialized subnetworks are the optimal solutions to memory storage for evolving patterns. We postulate that evolution of pathogens may be the reason for the immune system to be encoded in a focused memory, in contrast to the distributed memory used in the olfactory cortex that interacts with mixtures of static odors. Our approach offers a principled framework to study learning and memory retrieval in out-of-equilibrium dynamical systems.

Subject Areas: Biological Physics, Statistical Physics

I. INTRODUCTION

Storing memory for molecular recognition is an efficient strategy for sensing and response to external stimuli in biology. Apart from the cortical memory in the nervous system, memory is also an integral part of the immune

*To whom all correspondence should be addressed.
armita@uw.edu

response, present in a broad range of organisms from the CRISPR-Cas system in bacteria [1–3] to adaptive immunity in vertebrates [4–6]; CRISPR is an acronym for "clustered regularly interspaced short palindromic repeats" and Cas stands for "CRISPR associated protein". In all of these systems, an encounter with a pattern is encoded as a memory at the molecular level, and is later retrieved and activated in response to a similar stimulus, be it a pathogenic reinfection or a reexposure to a pheromone. Despite this high-level similarity, the immune system and the synaptic nervous system utilize vastly distinct molecular mechanisms for storage and retrieval of their memory.

Memory storage, and in particular, associative memory in the hippocampus and olfactory cortex, has been a focus of theoretical and computational studies in neuroscience

[7–11]. In the case of the olfactory cortex, the input is a combinatorial pattern produced by olfactory receptors which recognize the constituent monomolecules of a given odor. An odor signal is transmitted from the receptors to the olfactory cortex, which serves as a pattern recognition device and enables an organism to distinguish orders-of-magnitude more odors compared to the number of olfactory receptors [12–14]. The synaptic connections in the cortex are formed as they are costimulated by a given odor pattern, thus forming a distributed associative memory that can be retrieved in future exposures [7–11,15–18]. Notably, the distributed representation of odor stimuli has been demonstrated through optical imaging experiments in the piriform cortex, which is a subregion of the olfactory cortex [17,18]. While this description of the olfactory memory may be incomplete, as it does not include the role of the anterior olfactory nucleus [19], it still highlights the significance of distributed associative memory in accurately determining the identity of odor stimuli.

Immune memory is encoded very differently from associative memory in the nervous system. Immune receptors are extremely diverse and can specifically recognize pathogenic molecules without the need for a distributed and combinatorial encoding. In vertebrates, for example, the adaptive immune system consists of tens of billions of diverse B and T cells that can recognize and mount specific responses against the multitude of pathogens [5]. Immune cells activated in response to a pathogen can differentiate into memory cells, which are long-lived and can more efficiently respond to reinfections. Unlike the distributed memory in the olfactory cortex, the receptors encoding immune memory are specialized for a given pathogen class. However, within the same class, they can recognize evolved variants of a primary pathogen, in response to which memory was originally generated [5,20–25].

Learning and encoding of memory in the nervous system has inspired the development of efficient algorithms in machine learning with artificial neural networks [26–29]. In one class of such networks, input patterns trigger interactions among encoding nodes. Such an ensemble of interacting nodes can keep a robust distributed memory, since their coactivation enables the network to reconstruct a memory from even an incomplete pattern. This mode of memory storage resembles the coactivation of synaptic connections in a cortex. Energy-based models, such as Hopfield neural networks with Hebbian update rules [30], are among such systems, in which memory is stored as the network's energy minima [31]. These algorithms are effective in disentangling signal from noise, which makes them highly efficient in encoding static inputs with noise. Although some specialized machine-learning approaches allow for learning dynamically evolving inputs [26,27], we still lack a general framework for learning evolving patterns, relevant for many real-life applications [27].

Inspired by the ability of the immune memory in recognizing evolving patterns (pathogens), we propose a flexible model of learning with neural networks that can interpolate between the specialized and the distributed memory strategies used by the immune system and the nervous system. We formulate this problem as an optimization task to find a strategy (i.e., learning rate and network structure) that confers the highest accuracy for memory retrieval from the static and the dynamic (evolving) patterns.

In contrast to the static case, we show that a distributed memory in the style of a classical Hopfield model [31] fails to efficiently work for evolving patterns. We show that the optimal learning rate should increase with faster evolution of patterns, so that a network can follow the dynamics of the evolving patterns. This heightened learning rate tends to carve narrow connecting paths (mountain passes) between the memory attractors of a network's energy landscape, through which patterns can equilibrate in and be associated with a wrong memory. Importantly, we demonstrate that the problem of memory retrieval for continuously evolving patterns is distinct from that of the noisy patterns [32]. Unlike noise, evolution can systematically eliminate shared features among patterns, making it difficult to retrieve a pattern from an associative memory over time. To overcome this misclassification, we demonstrate that specialized memory compartments emerge in a neural network as an optimal solution to efficiently recognize and retrieve a memory of out-of-equilibrium evolving patterns.

Our results suggest that evolution of pathogenic patterns may be one of the key reasons for the immune system to encode a focused (compartmentalized) memory, as opposed to the distributed memory used in the olfactory system, for which molecular mixtures largely present static patterns.

Our approach provides a flexible framework to characterize the utility of different memory strategies inspired by the distinct organization of memory in the olfactory system and the immune system. However, it should be noted that such a top-down approach inevitably ignores many mechanistic and biological details, including the interaction between the adaptive and innate immunity in responding to pathogens [5] and the role of anterior olfactory nucleus in odor recognition [19]. Nonetheless, the proposed model can guide our biological intuition and offer a principled analytical framework to study learning and memory generation in out-of-equilibrium dynamical systems.

II. RESULTS

A. Model of working memory for evolving patterns

To probe memory strategies for different types of stimuli, we propose a generalized energy-based model of associative memory, in which a Hopfield-like neural network is able to learn and subsequently recognize binary patterns. This neural network is characterized by an energy landscape, and memory is stored as the network's energy
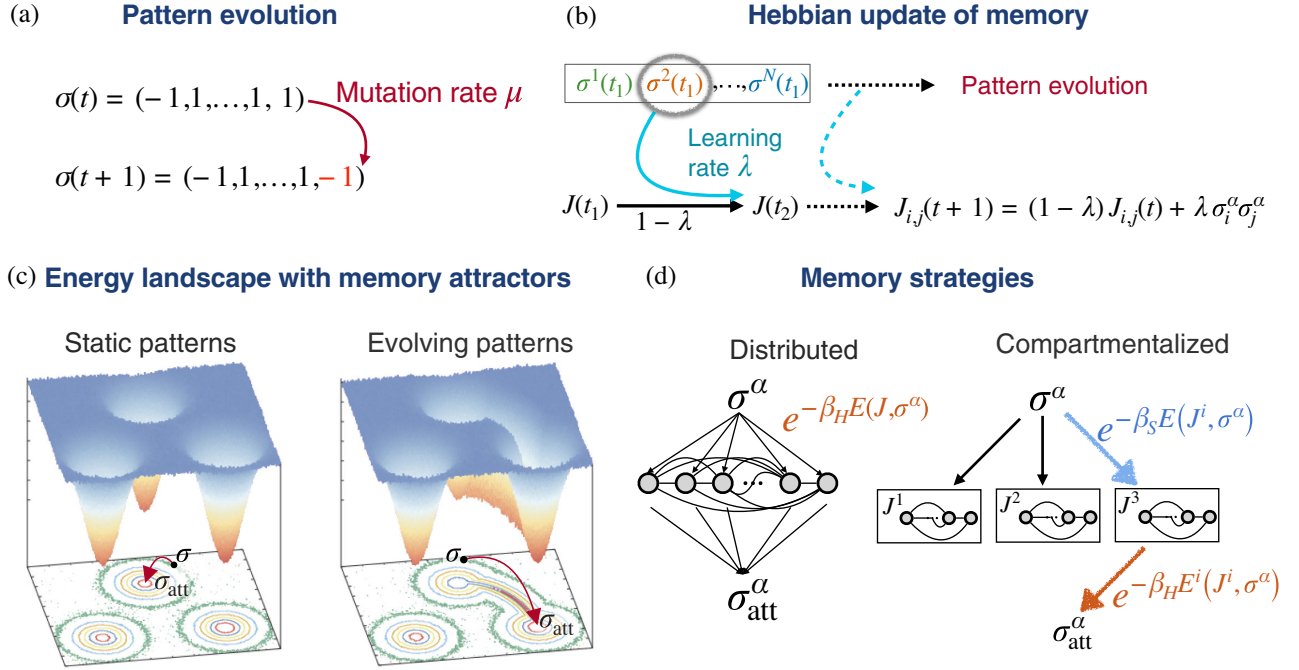
FIG. 1. Model of working memory for evolving patterns. (a) The targets of recognition are encoded by binary vectors $\{\sigma\}$ of length L. Patterns can evolve over time with a mutation rate μ denoting the fraction of spin flips in a pattern per network update event. (b) Hebbian learning rule is shown for a network J, which is presented a set of N patterns $\{\sigma^\alpha\}$ (colors) over time. At each step, one pattern σ^α is randomly presented to the network, and the network is updated with learning rate λ [Eq. (2)]. (c) The energy landscape for networks with distributed memory with optimal learning rate for static (left) and evolving (right) patterns are shown. The equipotential lines are shown in the bottom 2D plane. The energy minima correspond to memory attractors. For static patterns (left), equilibration in the network's energy landscape drives a pattern toward its associated memory attractor, resulting in an accurate reconstruction of the pattern. For evolving patterns (right), the heightened optimal learning rate of the network results in the emergence of connecting paths (mountain passes) between the energy minima. The equilibration process can drive a pattern through a mountain pass toward a wrong memory attractor resulting in pattern misclassification. (d) A network with distributed memory (left) is compared to a specialized network with multiple compartments (right). To find an associative memory, a presented pattern σ^α with energy $E(J,\sigma^\alpha)$ in network J equilibrates with inverse temperature β_H in the network's energy landscape and falls into an energy attractor σ_{att}^α. Memory retrieval is a two-step process in a compartmentalized network (right): First, the subnetwork J^i is chosen with a probability $P_i \sim \exp[-\beta_S E^i(J^i,\sigma^\alpha)]$, where β_S is the inverse temperature for this decision. Second, the pattern equilibrates within the subnetwork and falls into an energy attractor σ_{att}^α.

minima. We encode the target of recognition (stimulus) in a binary vector σ (pattern) with L entries: $\sigma = (\sigma_1, ..., \sigma_L)$ with $\sigma_i = \pm 1$, $\forall\, i$ [Fig. 1(a)]. To store associative memory, we define a fully connected network represented by an interaction matrix $J = (J_{i,j})$ of size $L \times L$ and use a Hopfield-like energy function (Hamiltonian) to describe pattern recognition [31] [Fig. 1(c)]

$$E_J(\sigma) = -\frac{1}{2L}\sum_{ij} J_{i,j}\sigma_i\sigma_j \equiv -\frac{1}{2}\langle\sigma|J|\sigma\rangle. \tag{1}$$

Here, we use a shorthand notation to denote the normalized (scaled) pattern vector by $|\sigma\rangle \equiv (1/\sqrt{L})\sigma$, its transpose by $\langle\sigma|$, resulting in a normalized scalar product $\langle\sigma|\sigma'\rangle \equiv (1/L)\sum_i \sigma_i\sigma_i'$ and a matrix product $\langle\sigma|J|\sigma\rangle \equiv (1/L)\sum_{i,j}\sigma_i J_{i,j}\sigma_j$.

The network undergoes a learning process, during which different patterns are presented sequentially and in random order [Fig. 1(b)]. As pattern σ^α is presented, the interaction matrix J is updated according to the following Hebbian update rule [33]:

$$J_{i,j} \to J'_{i,j} = \begin{cases} (1-\lambda)J_{i,j} + \lambda\sigma_i^\alpha\sigma_j^\alpha, & \text{if } i \neq j; \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Here, λ is the learning rate. In this model, the memorized patterns are represented by energy minima associated with the matrix J. We consider the case where the number N of different pattern classes is below the Hopfield capacity of the network (i.e., $N \lesssim 0.14L$; see Refs. [31,34,35]). Apart from this Hebbian learning rule, we also consider other learning protocols, including the Storkey learning rule [36], the gradient-descent learning rule [37], and sparse Hebbian learning (Appendix D). As we discuss throughout the manuscript, we find that our main conclusions are qualitatively insensitive to the choice of the learning rule, and

therefore, we limit our analysis in the main text to the Hebbian learning rule in Eq. (2).

With the update rule in Eq. (2), the network develops energy minima as associative memory close to each of the previously presented pattern classes σ^α ($\alpha \in \{1, \ldots, N\}$) [Fig. 1(c)]. Although the network also has minima close to the negated patterns, i.e., to $-\sigma^\alpha$, they do not play any role in what follows (Appendix C). To find an associative memory, we let a presented pattern σ^α equilibrate in the energy landscape, whereby we accept spin flips $\sigma^\alpha \to \tilde{\sigma}^\alpha$ with a probability $\min(1, e^{-\beta_H(E_J(\tilde{\sigma}) - E_J(\sigma))})$, where β_H is the inverse equilibration (Hopfield) temperature (Appendix A). In the low-temperature regime (i.e., high β_H), equilibration in networks with working memory drives a presented pattern σ^α toward a similar attractor $\sigma^\alpha_{\text{att}}$ reflecting the memory associated with the corresponding energy minimum [Fig. 1(c)]. This is measured similarly by the overlap $q^\alpha \equiv |\langle \sigma^\alpha_{\text{att}} | \sigma^\alpha \rangle|$ and determines the accuracy of the associative memory.

Unlike the classical cases of pattern recognition by Hopfield networks, we assume that patterns can evolve over time with a mutation rate μ per site per network update event [Fig. 1(a)]. Therefore, the expected number of spin flips in a given pattern between two encounters is $N\mu L \equiv \mu_{\text{eff}} L$, since two successive encounters of the same pattern are on average separated by $N - 1$ encounters (and updates) of the network with the other patterns. We work in the regime where the mutation rate μ is small enough such that the evolved patterns stemming from a common ancestor $\sigma^\alpha(t_0)$ at time t_0 (i.e., the members of the class α) remain more similar to each other than to members of the other classes (i.e., $\mu N L \ll L/2$).

The special case of static patterns ($\mu_{\text{eff}} = 0$) can reflect distinct odor molecules, for which associative memory is stored in the olfactory cortex. On the other hand, the distinct pattern classes in the dynamic case ($\mu_{\text{eff}} > 0$) can be attributed to different types of evolving pathogens (e.g., influenza, HIV, etc.), and the patterns within a class as different variants of a given pathogen. In our model, we use the mutation rate as an order parameter to characterize the different phases of memory strategies in biological systems.

B. Optimal learning rate for evolving patterns

In the classical Hopfield model ($\mu_{\text{eff}} = 0$), the learning rate λ is set to very small values for the network to efficiently learn the patterns [33]. For evolving patterns, the learning rate should be tuned so the network can efficiently update the memory retained from the prior learning steps. At each encounter, the overlap $q^\alpha(t; \lambda) = |\langle \sigma^\alpha_{\text{att}}(t; \lambda) | \sigma^\alpha(t) \rangle|$ between a pattern $\sigma^\alpha(t)$ and the corresponding attractor for the associated energy minimum $\sigma^\alpha_{\text{att}}(t; \lambda)$ determines the accuracy of pattern recognition; the parameter λ explicitly indicates the dependence of the network's energy landscape on the learning rate. We declare the recognition of a pattern σ^α as successful if and

only if the accuracy of reconstruction (overlap) is larger than a set threshold $q^\alpha(t) \geq 0.8$. However, our results are insensitive to the exact value of this threshold (Appendix C and Supplemental Material [38] Fig. S1). We define a network's performance as the asymptotic accuracy of its associative memory averaged over the ensemble of all (recognized and unrecognized) pattern classes [Fig. 2(a)],

$$\mathcal{Q}(\lambda) \equiv \mathbb{E}[q^\alpha(t; \lambda)]$$
$$\simeq \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T} \frac{1}{N} \sum_{\alpha=1}^{N} |\langle \sigma^\alpha_{\text{att}}(t; \lambda) | \sigma^\alpha(t) \rangle|. \quad (3)$$

The expectation $\mathbb{E}[\cdot]$ is an empirical average over the ensemble of presented pattern classes over time, which in the stationary state approaches the asymptotic average of the argument. The optimal learning rate is determined by maximizing the network's performance $\lambda^* = \text{argmax}_\lambda \mathcal{Q}(\lambda)$.

The optimal learning rate increases with growing mutation rate so that a network can follow the evolving patterns [Fig. 2(b)]. Although it is difficult to analytically calculate the optimal learning rate, we can use an approximate approach and find the learning rate that minimizes the expected energy of the patterns $\mathbb{E}[E_{\lambda,\rho}(J, \sigma)]$, assuming that patterns are shown to the network at a fixed order (Appendix B). In this case, the expected energy is given by

$$\mathbb{E}[E_{\lambda,\rho}(J, \sigma)] = \frac{L - 1}{2} \times \frac{\lambda}{1 - \lambda} \times \frac{1}{\rho^{-2N}(1 - \lambda)^{-N} - 1}, \quad (4)$$

where $\rho^N \equiv (1 - 2\mu)^N = 1 - 2N\mu + \mathcal{O}(\mu^2) \approx 1 - 2\mu_{\text{eff}}$ is the upper bound for the overlap between a pattern and its evolved form when separated by the other $N - 1$ patterns that are shown in between. The expected energy grows slowly with increasing mutation rate (i.e., with decreasing overlap q), and the approximation in Eq. (4) agrees very well with the numerical estimates for the scenario where patterns are shown in a random order [Fig. 2(c)]. In the regime where memory can still be associated with the evolved patterns ($\mu_{\text{eff}} \ll 0.5$), the minimization of the expected energy [Eq. (4)] results in an optimal learning rate

$$\lambda^*(\mu) = \sqrt{8\mu/(N - 1)} \quad (5)$$

that scales with the square root of the mutation rate. Notably, this theoretical approximation agrees well with the numerical estimates [Fig. 2(b)].

C. Reduced accuracy of distributed associative memory against evolving patterns

Despite using an optimized learning rate, a network's accuracy in pattern retrieval $\mathcal{Q}(\lambda)$ decays much faster than the naive expectation solely based on the evolutionary divergence of patterns between two encounters with a given class [i.e., $\mathcal{Q}_0 = (1 - 2\mu)^N \approx 1 - 2\mu_{\text{eff}}$]; see Fig. 2(a).
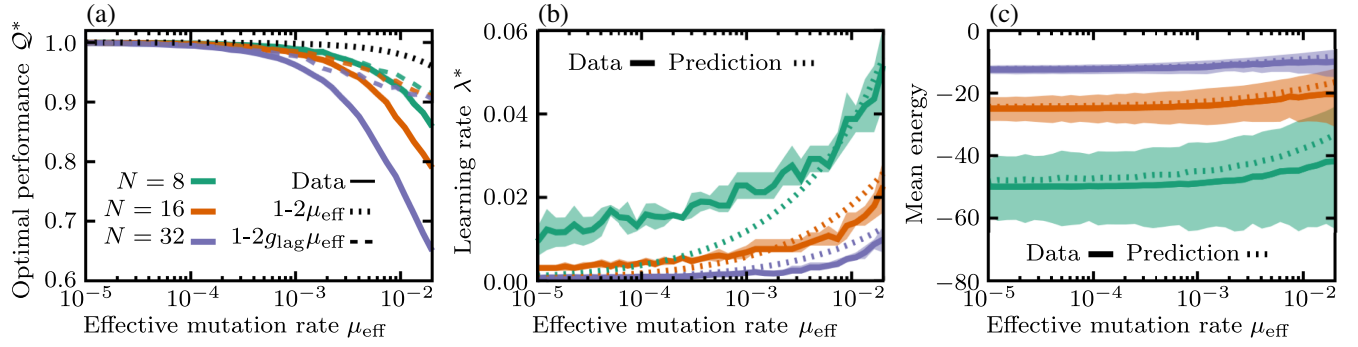
FIG. 2. Reduced performance of Hopfield networks in retrieving memory of evolving patterns. (a) The optimal performance of a network $\mathcal{Q}^* \equiv \mathcal{Q}(\lambda^*)$ [Eq. (3)] is shown as a function of the effective mutation rate $\mu_{\mathrm{eff}} = N\mu$. The solid lines show the simulation results for networks encountering different number of patterns (colors). The black dotted line shows the naive expectation for the performance solely based on the evolutionary divergence of the patterns $\mathcal{Q}_0 \approx 1-2\mu_{\mathrm{eff}}$, and the colored dashed lines show the expected performance after accounting for the memory lag g_{lag}, $\mathcal{Q}_{\mathrm{lag}} \approx 1-2g_{\mathrm{lag}}\mu_{\mathrm{eff}}$; see Fig. S5 in the Supplemental Material [38] for more details. (b) The optimal learning rate λ^* is shown as a function of the effective mutation rate. The solid lines are the numerical estimates, the shaded areas indicate the uncertainty in the estimated optimal learning rate, and dashed lines show the theoretical predictions [Eq. (5)]; see Appendix A and Supplemental Material Fig. S2 [38] for the numerical optimization protocol and the estimation of the uncertainty. (c) The mean energy obtained by simulations of randomly ordered patterns (solid lines) and the analytical approximation [Eq. (4)] for ordered patterns (dotted lines) are shown. Error bars show the standard error from the independent realizations (Appendix A). The color code for the number of presented patterns is consistent across panels, and the length of patterns is set to $L = 800$.

A similar decline in performance can be seen when updating the network with other learning rules (i.e., Storkey learning [36], gradient-descent learning rule [37], and sparse Hebbian learning); see Fig. S3 in the Supplemental Material [38] and Appendix D.

It should be noted that a classical Hopfield network with a small learning rate can accurately retrieve the memory of noisy static patterns by relying on the shared features of the patterns within a class [32] (see Supplemental Material Fig. S4 [38]). However, continuous evolution can systematically eliminate the shared features among patterns of a given class, resulting in a significant reduction in the accuracy of memory retrieval for evolving patterns. There are two reasons for such a reduced accuracy: (i) the lag in the network's memory and (ii) the misclassification of presented patterns.

The memory attractors associated with a given pattern class can lag behind the evolution and reflect only the older patterns presented prior to the most recent encounter of the network with the specified class. We characterize this lag g_{lag} by identifying a previous version of the pattern that has the maximum overlap with the network's energy landscape at a given time t: $g_{\mathrm{lag}} = \mathrm{argmax}_{g \geq 0} \mathbb{E}[\langle \sigma(t-gN)|J(t)|\sigma(t-gN)\rangle]$ (Appendix B). g_{lag} measures time in units of N (i.e., the effective separation time of the pattern of the same class). An increase in the optimal learning rate reduces the time lag and enables the network to follow the evolving patterns more closely (see Supplemental Material Fig. S5 [38]). The accuracy of the memory subject to such a lag decays as $\mathcal{Q}_{\mathrm{lag}} = \rho^{g_{\mathrm{lag}}N} \approx 1-2g_{\mathrm{lag}}\mu_{\mathrm{eff}}$, which is faster than the

naive expectation (i.e., $1-2\mu_{\mathrm{eff}}$); see Fig. 2(a). This memory lag explains the loss of performance for patterns that are still reconstructed by the network's memory attractors [i.e., those with $q^\alpha > 0.8$; see Supplemental Material Fig. S5(a) [38]]. However, the overall performance of the network $\mathcal{Q}(\lambda)$ remains lower than the expectation obtained by taking into account this time lag [Fig. 2(a)]—a discrepancy that leads us to the second root of reduction in accuracy, i.e., pattern misclassification.

As the learning rate increases, the structure of the network's energy landscape changes. In particular, we see that with large learning rates, a few narrow paths emerge between the memory attractors of the networks [Fig. 1(c)]. As a result, the equilibration process for pattern retrieval can drive a presented pattern through the connecting paths toward a wrong memory attractor (i.e., one with a small overlap $\langle \sigma_{\mathrm{att}}|\sigma\rangle$), which leads to pattern misclassification [see Supplemental Material Figs. S1(a), S1(c), and S6(a) [38]]. These paths are narrow, as there are only a few possible spin flips (mutations) that can drive a pattern from one valley to another during equilibration [see Supplemental Material Figs. S1(b), S1(d), S7(a), and S7(c) [38]]. In other words, a large learning rate carves narrow mountain passes in the network's energy landscape [Fig. 1(c)], resulting in a growing fraction of patterns to be misclassified. Importantly, the attractors into which the patterns mistakenly fall are all associated with memory from the previously encountered patterns [see Supplemental Material Figs. S1(a) and S1(c) [38]], and pattern misclassification is not due to the appearance of additional random energy minima.

Interestingly, pattern misclassification occurs even in the absence of mutations for networks with an increased learning rate [see Supplemental Material Fig. S6(a) [38]]. This suggests that mutations only indirectly contribute to the misclassification of memory, as they necessitate a larger learning rate for the networks to optimally operate, which in turn results in the emergence of mountain passes in the energy landscape.

To understand the memory misclassification, particularly for patterns with moderately low (i.e., nonrandom) energy [Fig. 2(c)], we use spectral decomposition to characterize the relative positioning of patterns in the energy landscape (see Appendix C). The vector representing each pattern $|\sigma\rangle$ can be expressed in terms of the network's eigenvectors $\{\Phi^i\}$, $|\sigma\rangle = \sum_i m_i |\Phi^i\rangle$, where the overlap $m_i \equiv \langle \Phi^i | \sigma \rangle$ is the ith component of the pattern in the network's coordinate system. During equilibration, we flip individual spins in a pattern and accept the flips based on their contribution to the recognition energy. We can view these spin flips as rotations of the pattern in the space spanned by the eigenvectors of the network. Stability of a pattern depends on whether these rotations could carry the pattern from its original subspace over to an alternative region associated with a different energy minimum.

There are two key factors that modulate the stability of a pattern in a network. The dimensionality of the subspace in which a pattern resides, i.e., support of a pattern by the network's eigenvectors, is one of the key determining factors for pattern stability. We quantify the support of a pattern σ using the participation ratio $\pi(\sigma) = (\sum_i m_i^2)^2 / \sum_i m_i^4$ [39,40] that counts the number of eigenvectors that substantially overlap with the pattern. A small support $\pi(\sigma) \approx 1$ indicates that the pattern is spanned by only a few eigenvectors and is restricted to a small subspace, whereas a large support indicates that the pattern is orthogonal to only a few eigenvectors. As the learning rate increases, patterns lie in lower-dimensional subspaces supported by only a few eigenvectors [see Supplemental Material Figs. S7(b) and S7(d) [38]]. This effect is exacerbated by the fact that the energy gap between the eigenstates of the network also broaden with increasing learning rate (see Supplemental Material Fig. S8 [38]). The combination of a smaller support for patterns and a larger energy gap in networks with increased learning rate leads to the destabilization of patterns by enabling the spin flips during equilibration to drive a pattern from one subspace to another through the mountain passes carved within the landscape; see Appendix C [Eq. (C14)] and Supplemental Material Fig. S9 [38] for the exact analytical criteria for pattern stability. The change in the structure of the energy landscape by increasing the learning rate also leads to larger differences in the depth of the energy minima across the landscape. This effect can be seen as the increase of energy variance with the learning rate in Fig. 2(c). The differences in depth across energy minima

enable patterns to transition from one valley to another during equilibration.

D. Compartmentalized learning and memory storage

Hopfield-like networks can store accurate associative memory for static patterns. However, these networks fail to perform and store retrievable associative memory for evolving patterns (e.g., pathogens), even when learning is done at an optimal rate (Fig. 2). To overcome this difficulty, we propose to store memory in compartmentalized networks, with C subnetworks of size L_c (i.e., the number of nodes in a subnetwork). Each compartment (subnetwork) can store a few of the total N pattern classes without interference from the other compartments [Fig. 1(d)].

Recognition of a pattern σ in a compartmentalized network involves a two-step process [Fig. 1(d)]: First, we choose a subnetwork J^i associated with compartment i with a probability $P_i \sim \exp[-\beta_S E(J^i, \sigma)]$, where β_S is the inverse temperature for this decision. A larger inverse temperature β_S implies that when associating a presented pattern with a memory compartment, the system is more sensitive to differences in recognition energy across compartments. The inverse temperature β_S can be thought of as the efficacy of information processing during decision-making [41]. In a previous work on decision-making for immune response, we have argued that inverse temperature can be viewed as the accumulated evidence for a given infection, e.g., through contact between the immune memory cells and the replicating pathogens [24]. In this view, a larger inverse temperature β_S corresponds to a larger sampling size (i.e., encounter rate with pathogens) for memory cells as they accumulate information about the infection, and results in a higher sensitivity of immune cells to their cognate pathogens. Indeed, such a correspondence between the inverse temperature in thermodynamics and the effect of sample size has been previously introduced in the context of statistical inference [42,43].

Once the compartment is chosen, we follow the recipe for pattern retrieval in the energy landscape of the associated subnetwork, whereby a pattern equilibrates into a memory attractor.

On average, each compartment stores a memory for $N_c = N/C$ pattern classes. To keep the networks with different number of compartments C comparable, we scale the size of each compartment L_c to keep $C \times L_c = $ constant, which keeps the (Hopfield) capacity of the network $\alpha = N_c/L_c$ invariant under compartmentalization. Moreover, the mutation rate experienced by each subnetwork scales with the number of compartments $\mu_c = C\mu$, which keeps the effective mutation rate $\mu_{\text{eff}} = N_c \mu_c$ invariant under compartmentalization. As a result, the optimal learning rate [Eq. (5)] scales with the number of compartments C as $\lambda_c^* = \sqrt{8\mu_c/(N_c - 1)} \approx C\lambda_1^*$ (Fig. 3). However, since updates are restricted to subnetworks of size L_c at a time, the expected number of updates within a network

$L_c\lambda_c$ remains invariant under compartmentalization. Lastly, since the change in energy by a single spin flip scales as $\Delta E \sim 1/L_c$, we introduce the scaled Hopfield temperature $\beta_{H_c} \equiv C\beta_H$ to make the equilibration process comparable across networks with different number of compartments. No such scaling is necessary for β_S.

By restricting the networks to satisfy the aforementioned scaling relationships, we are left with two independent variables, i.e., (i) the number of compartments C and (ii) the learning rate λ_c, which define a memory strategy $\{C, \lambda_c\}$. A memory strategy can then be optimized to achieve the maximum accuracy for retrieving an associative memory for evolving patterns with a given effective mutation rate μ_{eff}.

E. Phases of learning and memory production

Pattern retrieval can be stochastic due to the noise in choosing the right compartment from the C subnetworks (tuned by the inverse temperature β_S), or the noise in equilibrating into the right memory attractor in the energy landscape of the chosen subnetwork (tuned by the Hopfield inverse temperature β_{H_c}). We use mutual information to quantify the accuracy of pattern-compartment association, where larger values indicate a more accurate association; see Appendix A and Fig. 4. The optimal performance $\mathcal{Q}^*$ determines the overall accuracy of memory retrieval, which depends on both finding the right compartment and equilibrating into the correct memory attractor. The amplitudes of intra- versus intercompartment stochasticity determine the optimal strategy $\{C^*, \lambda_c^*\}$ used for learning and retrieval of patterns with a specified mutation rate. Varying the corresponding inverse temperatures (β_{H_c}, β_S) results in three distinct phases of optimal memory storage.

1. Small intra- and intercompartment noise $(\beta_{H_c} \gg 1, \beta_S \gg 1)$

In this regime, neither the compartment choice nor the pattern retrieval within a compartment are subject to strong noise. As a result, networks are functional with working memory, and the optimal strategies can achieve the highest overall performance. For small mutation rates, we see that all networks perform equally well and can achieve almost perfect performance, irrespective of their number of compartments [Figs. 3(a), 4(a), and 4(b)]. As the mutation rate increases, networks with a larger number of compartments show a more favorable performance, and the 1-to-1 specialized network, in which each pattern is stored in a separate compartment (i.e., $N = C$), reaches the optimal performance $1 - 2\mu_{\text{eff}}$ [Figs. 3(a), 4(c), and 4(d)]. As predicted by the theory, the optimal learning rate for compartmentalized networks scales with the mutation rate as $\lambda_c^* \sim \mu_c^{1/2}$, except for the 1-to-1 network in which $\lambda_c^* \to 1$ and subnetworks are steadily updated upon an encounter with a pattern [Fig. 3(b)]. This rapid update is expected
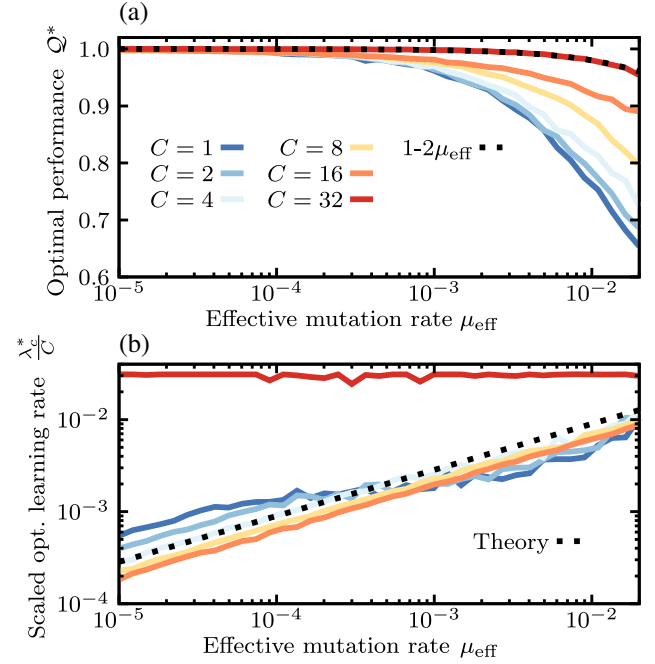

FIG. 3. Compartmentalized memory storage. (a) The optimal performance is shown as a function of the effective mutation rate [similar to Fig. 2(a)] for networks with different number of compartments C (colors) ranging from a network with distributed memory $C = 1$ (blue) to a 1-to-1 compartmentalized network $C = N$ (red). (b) The optimal (scaled) learning rate λ_c/C is shown as a function of the effective mutation rate for networks with different number of compartments [colors according to (a)]. Full lines show the numerical estimates, and the dashed line is from the analytical approximation $\lambda_c^* = \sqrt{8\mu_c/(N_c - 1)} \approx C\lambda_1^*$. The scaled learning rates collapse on the analytical approximation for all networks except for the 1-to-1 compartmentalized network (red), where the maximal learning rate $\lambda \approx 1$ is used, and each compartment is fully updated upon an encounter with a new version of a pattern. The number of presented patterns is set to $N = 32$. We keep $L \times C = \text{const}$, with $L = 800$ used for the network with $C = 1$.

since there is no interference between the stored memories in the 1-to-1 network, and a steady update can keep the stored memory in each subnetwork close to its associated pattern class without disrupting the other energy minima.

2. Small intra- and large intercompartment noise $(\beta_{H_c} \gg 1, \beta_S \ll 1)$

In this regime, there is low noise for equilibration within a compartment but a high level of noise in choosing the right compartment. The optimal strategy in this regime is to store patterns in a single network with a distributed memory, since identifying the correct compartment is difficult due to noise [Figs. 4(b) and 4(d)]. For static patterns, this strategy corresponds to the classical Hopfield model with a high accuracy [Figs. 2(a), 4(a), and 4(b)]. On the other hand, for evolving patterns this strategy results in
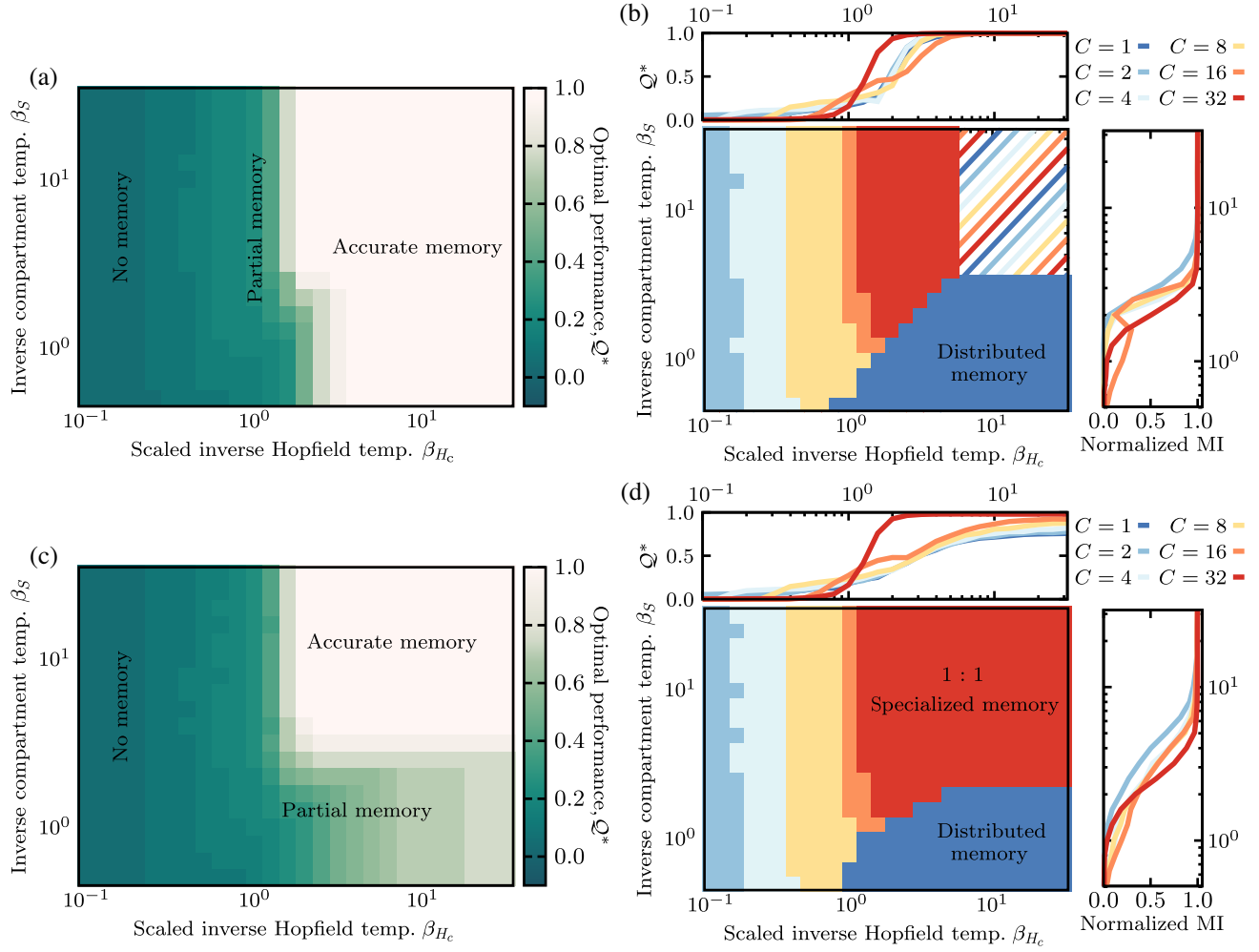
FIG. 4. Phases of learning and memory production. Different optimal memory strategies are shown. (a) The heat map shows the optimal memory performance $\mathcal{Q}^*$ as a function of the (scaled) Hopfield inverse temperature $\beta_{H_c} = \beta_H C$ and the inverse temperature associated with compartmentalization β_S for networks learning and retrieving a memory of static patterns ($\mu = 0$); colors are indicated in the color bar. The optimal performance is achieved by using the optimal strategy (i.e., learning rate λ_c^* and the number of compartments c^*) for networks at each value of β_{H_c} and β_S. The three phases of accurate, partial, and no memory are indicated. (b) The heat map shows the memory strategies for the optimal number of compartments (colors as in the legend) corresponding to the memory performance shown in (a). We limit the optimization to the possible number of compartments indicated in the legend to keep N/C an integer. The dashed region corresponds to the case where all strategies perform equally well. Regions of distributed memory ($C = 1$) and the 1-to-1 specialized memory ($C = N$) are indicated. The top panel shows the optimal performance $\mathcal{Q}^*$ of different strategies as a function of the Hopfield inverse temperature β_{H_c}. The right panel shows the mutual information $\mathrm{MI}(\Sigma, \mathcal{C})$ between the set of pattern classes $\Sigma \equiv \{\sigma^\alpha\}$ and the set of all compartments $\mathcal{C}$ normalized by the entropy of the compartments $H(\mathcal{C})$ as a function of the inverse temperature β_S; see Appendix A. This normalized mutual information quantifies the ability of the system to assign a pattern to the correct compartment. (c),(d) Similar to (a),(b) but for evolving patterns with the effective mutation rate $\mu_{\mathrm{eff}} = 0.01$. The number of presented patterns is set to $N = 32$ (all panels). Similar to Fig. 3, we keep $L \times C = \mathrm{const}$, with $L = 800$ used for networks with $C = 1$.

a partial memory [Figs. 4(c) and 4(d)] due to the reduced accuracy of the distributed associative memory, as shown in Fig. 2(a). Interestingly, the transition between the optimal strategy with highly specific (compartmentalized) memory for evolving patterns in the first regime and the generalized (distributed) memory in this regime is very sharp [Fig. 4(d)]. This sharp transition suggests that depending on the noise in choosing the compartments β_S, an optimal strategy either stores memory in a 1-to-1 specialized

fashion ($C = N$) or in a distributed generalized fashion ($C = 1$), but no intermediate solution (i.e., quasispecialized memory with $1 < C < N$) is desirable.

3. Large intracompartment noise ($\beta_{H_c} < 1$)

In this regime, there is a high level of noise in equilibration within a network, and memory cannot be reliably retrieved [Figs. 4(a) and 4(c)], regardless of the

compartmentalization temperature β_S. However, the impact of the equilibration noise β_{H_c} on the accuracy of memory retrieval depends on the degree of compartmentalization. For the 1-to-1 specialized network ($C = N$), the transition between the high and the low accuracy is smooth and occurs at $\beta_{H_c} = 1$, below which no memory attractor can be found. As we increase the equilibration noise (decrease β_{H_c}), the networks with distributed memory ($C < N$) show two-step transitions, with a plateau in the range of $1/N_c \lesssim \beta_{H_c} \lesssim 1$. Similar to the 1-to-1 network, the first transition at $\beta_{H_c} \approx 1$ results in the reduced accuracy of the networks' memory retrieval. At this transition point, the networks' learning rate λ_c approaches its maximum value 1 (see Supplemental Material Fig. S10 [38]), which implies that the memory is stored (and steadily updated) for only $C < N$ patterns (i.e., one pattern per subnetwork). Because of the invariance of the networks' mean energy under compartmentalization, the depth of the energy minima associated with the stored memory in each subnetwork scales as N/C, resulting in deeper and more stable energy minima in networks with smaller number of compartments C. Therefore, as the noise increases (i.e., β_{H_c} decreases), we observe a gradient in transition from partial retrieval to a no-memory state at $\beta_H \approx 1/N_c$, with the most compartmentalized network (larger C) transitioning the fastest, reflecting the shallowness of their energy minima.

Taken together, the optimal strategy leading to working memory depends on whether a network is trained to learn and retrieve dynamic (evolving) patterns or static patterns. Specifically, we see that the 1-to-1 specialized network is the unique optimal solution for storing working memory for evolving patterns, whereas the distributed generalized memory (i.e., the classical Hopfield network) performs equally well in learning and retrieval of memory for static patterns. The contrast between these outcomes can shed light on the distinct strategies used by different biological systems to encode memory.

III. DISCUSSION

Storing and retrieving memory from prior molecular interactions is an efficient scheme to sense and respond to external stimuli. Here, we introduce a flexible energy-based neural network model that can adopt different memory strategies, including distributed memory, similar to the classical Hopfield network, or compartmentalized memory. The learning rate and the number of compartments in a network define a memory strategy, and we probe the efficacy of different strategies for static and dynamic patterns. We find that Hopfield-like networks with distributed memory are highly accurate in storing associative memory for static patterns, even when patterns are noisy. However, these networks fail to reliably store retrievable associative memory for evolving patterns, even when learning is done at an optimal rate.

To achieve high accuracy, we show that a retrievable memory for evolving patterns should be compartmentalized, where each pattern class is stored in a separate subnetwork. In addition, we find a sharp transition between the different phases of working memory (i.e., compartmentalized and distributed memory), suggesting that intermediate solutions (i.e., quasispecialized memory) are suboptimal against evolving patterns.

The contrast between these memory strategies is reflective of the distinct encoding of memory in the adaptive immune system and in the olfactory cortex. Although some organisms use chemical deception to mimic other odors [44,45], the constituent odor molecules can still be assumed to be static. Consistently, the memory of such static odor molecules is stored in a distributed fashion in the olfactory cortex [7–11,15–18]. Notably, recording from a large number of neurons has shown that the identity of odors, irrespective of their intensity, is encoded by unique and distributed ensembles of neurons in the piriform region of the olfactory cortex [18]. The resulting distributed memory allows for a robust and accurate retrieval of a stimulus identity across a broad range of odor concentrations [46]— a feature that is critical for olfactory behavior.

The adaptive immune system, on the other hand, interacts with pathogenic epitopes that constantly evolve. Consistently, the adaptive immune system allocates distinct immune cells (i.e., compartments) to store a memory for different types of pathogens (e.g., different evolved variants of influenza or HIV) [5,20–25]—a strategy that resembles that of the 1-to-1 specialized networks. In this case, the two-step process of finding the right compartment and then equilibrating within the compartment can be thought of as recognizing an infecting pathogen with the correct memory cell, and then adjusting the structural conformation of the memory receptor to form strong chemical bonds with the target epitope on the pathogen.

Although distributed memory fails to distinguish evolving pattern classes from each other, it can still discriminate between preencountered and random patterns [see Supplemental Material Figs. S1(a) and S1(c) [38]]. Indeed, in a related work, we have shown that the statistics of the recognition energy in memory repertoires can be used to classify familiar and novel patterns [47].

It is conceivable that memory allocation may be a limiting factor in some biological systems, resulting in a cost for adding extra compartments to the network, which is currently missing from our model. In this case, we expect the transition between the performance of the distributed memory and the 1-to-1 specialized strategy to be a smoother function of the patterns' evolutionary rate, and that a partially compartmentalized network to be the optimal strategy, dependent on the exact form of the cost function. Nevertheless, if the cost inflicted by the reduced memory performance of a partially compartmentalized system outweighs the (biological) cost of allocating

additional memory compartments, the 1-to-1 memory strategy can remain the optimal solution.

This appears to be the case for the immune system, where the harm caused by an uncontrolled infection outweighs the cost of storing a highly diverse and specialized memory repertoire. It is estimated that the adaptive immune system in humans can roughly generate 10^{17} unique B-cell receptors [48], which is much larger than the total of about 10^{10} B cells circulating in a human body. B-cell clones within individuals follow a long-tail distribution [49,50], indicating the variability in their extent of clonal expansion in response to different infections. Although it is not clear how sequence diversity translates to immune function, the diverse and preferentially expanded repertoires of immune receptors can efficiently mount specific responses against the multitude of infecting pathogens [5]. Similar to other biophysical interactions, immune-pathogen recognition is cross-reactive. Nonetheless, an immune receptor is only efficient in recognizing pathogenic epitopes that are within a limited antigenic distance, and therefore, a given memory cell can respond only to close evolutionary variants of a given pathogen and not across pathogens. As such, the encoding of memory in the adaptive immune system seems to be consistent with the 1-to-1 specialized memory.

One of the features of our model is that memory is updated upon presentation of evolved patterns to the network. For 1-to-1 compartmentalized networks, this update resembles maturation of memory B cells in the immune system in response to reinfections. Antibody-secreting B cells can specialize through a process of affinity maturation, which is a form of somatic Darwinian evolution with mutation and selection within an individual to enhance the affinity of B-cell receptors to pathogens [51]. Several rounds of mutation and selection can increase the binding affinities of receptors up to (10–10 000)-fold [52]. Stored immune memory cells can also initiate affinity maturation during a secondary or later responses to new variants of a pathogen [53,54], during which new receptors are formed that are specific to the new variant. Analysis of B-cell repertoires in HIV patients has shown accumulation of such mutations in B-cell lineages over years of infection, which are likely to have stemmed from memory responses [55,56]. Similar features are reported for immune response to vaccination against evolving viruses like influenza [57,58].

The increase in the optimal learning rate in anticipation of patterns' evolution significantly changes the structure of the energy landscape for associative memory. In particular, we find the emergence of narrow connectors (mountain passes) between the memory attractors of a network, which destabilize the equilibration process and significantly reduce the accuracy of memory retrieval. Indeed, tuning the learning rate as a hyperparameter is one of the challenges of current machine-learning algorithms with deep neural networks (DNNs) [28,29]. The goal is to navigate the trade-off between the speed (i.e., rate of convergence) and accuracy without overshooting during optimization. It will be interesting to see how the insights developed in this work can inform rational approaches to choose an optimal learning rate in optimization tasks with DNNs.

Machine-learning algorithms with DNNs [28] and modified Hopfield networks [59–62] are able to accurately classify hierarchically correlated patterns, where different objects can be organized into an ultrametric tree based on some specified relations of similarity. For example, faces of cats and dogs have the oval shape in common but they branch out in the ultrametric tree according to the organism-specific features, e.g., whiskers in a cat, and the cat branch can then further diversify based on the breed-specific features. A classification algorithm can use these hierarchical relations to find features common among members of a given subtype (cats) that can distinguish them from another subtype (dogs). Although evolving patterns within each class in our model are correlated, the random evolutionary dynamics of these patterns does not build a hierarchical structure where a pattern class branches in two subclasses that share a common ancestral root. Therefore, the optimal memory strategies found here for evolving patterns are distinct from those of the hierarchically correlated patterns. It will be interesting to see how our approaches can be implemented in DNNs to classify dynamic and evolving patterns.

ACKNOWLEDGMENTS

APPENDIX A: COMPUTATIONAL PROCEDURES

In the following, we describe the numerical procedure used in this manuscript.

The code used to produce and analyze the data, as well as some example notebooks to generate data for smaller systems can be accessed through Ref. [63].

1. Initialization of the network

A network J (with elements J_{ij}) is presented with N random (orthogonal) patterns $|\sigma^\alpha\rangle$ (with $\alpha = 1, \ldots, N$) with entries $\sigma_i^\alpha \in \{-1, 1\}$ reflecting the N pattern classes. For a network with C compartments (with $1 \leq C \leq N$), we initialize each subnetwork J^s at time t_0 as $J_{i,j}^s(t_0) = [1/(N/C)] \sum_{\alpha \in \mathcal{A}_s} \sigma_i^\alpha \sigma_j^\alpha$ and $J_{ii}^s(t_0) = 0$; here, $\mathcal{A}_s$ is a set of N/C randomly chosen (without replacement) patterns

initially assigned to the compartment (subnetwork) s. We then let the network undergo an initial learning process. At each step, an arbitrary pattern σ^ν is presented to the network, and a subnetwork J^s is chosen for an update with a probability

$$P_s = \frac{\exp\{-\beta_S E[J^s(t), \sigma^\nu(t)]\}}{\sum_{r=1}^C \exp\{-\beta_S E[J^s(t), \sigma^\nu(t)]\}}, \quad \text{(A1)}$$

where the energy is defined as

$$E(J^s(t), \sigma^\nu(t)) = \frac{-1}{2L} \sum_{i,j} J_{i,j}^s(t) \sigma_i^\nu(t) \sigma_j^\nu(t)$$

$$\equiv \frac{-1}{2} \langle \sigma^\nu(t) | J^s(t) | \sigma^\nu(t) \rangle, \quad \text{(A2)}$$

and β_S is the inverse temperature associated with choosing the right compartment. We then update the selected subnetwork J^s using the Hebbian update rule

$$J_{i,j}^s(t+1) = \begin{cases} (1-\lambda) J_{i,j}^s(t) + \lambda \sigma_i^\nu \sigma_j^\nu, & \text{if } i \neq j; \\ 0 & \text{otherwise.} \end{cases} \quad \text{(A3)}$$

For dynamic patterns, the presented patterns undergo evolution with mutation rate μ, which reflects the average fraction of flipped spins in a given pattern per network update event (Fig. 1). For noisy patterns, the actual patterns remain unchanged, but the network is presented with noisy versions of these patterns. Here, the noise amplitude reflects the average fraction of flipped spins between presented and actual patterns.

Our goal is to study the memory retrieval problem in a network that has reached its steady state. The state of a network $J(t_n)$ at the time step n can be traced back to the initial state $J(t_0)$ as

$$J(t_n) = (1-\lambda)^n J(t_0) + \lambda \sum_{i=1}^n (1-\lambda)^{n-i} |\sigma(t_i)\rangle \langle \sigma(t_i)|. \quad \text{(A4)}$$

The contribution of the initial state $J(t_0)$ to the state of the network at time t_n decays as $(1-\lambda)^n$ [Eq. (A4)]. Therefore, we choose the number of steps to reach the steady state as $n_{\text{stat}} = \max\{10N, 2C\text{ceil}[\log 10^{-5} / \log(1-\lambda)]\}$. This criteria ensures that $(1-\lambda)^{n_{\text{stat}}} \leq 10^{-5}$ and the memory of the initial state $J(t_0)$ is removed from the network $J(t)$. We then use this updated network to collect the statistics for memory retrieval. To report a specific quantity from the network (e.g., the energy), we pool the n_{stat} samples collected from each of the 50 realizations.

2. Pattern retrieval from associative memory

Once the trained network approaches a stationary state, we collect the statistics of the stored memory.

To find a memory attractor $\sigma_{\text{att}}^\alpha$ for a given pattern σ^α we use a Metropolis algorithm in the energy landscape $E(J^s, \sigma^\alpha)$ [Eq. (A2)]. To do so, we make spin flips in a presented pattern $\sigma^\alpha \to \tilde{\sigma}^\alpha$ and accept a spin flip with probability

$$P(\sigma^\alpha \to \tilde{\sigma}^\alpha) = \min(1, e^{-\beta_H \Delta E}), \quad \text{(A5)}$$

where $\Delta E = E(J^s, \tilde{\sigma}^\alpha) - E(J^s, \sigma^\alpha)$, and β_H is the inverse (Hopfield) temperature for pattern retrieval in the network (see Fig. 1). We repeat this step for 2×10^6 steps, which is sufficient to find a minimum of the landscape (see Supplemental Material Fig. S5 [38]).

For systems with more than one compartment C, we first choose a compartment according to Eq. (A1) and then perform the Metropolis algorithm within the associated compartment.

After finding the energy minima, we update the systems for $n_{\text{stat}}' = \max[2 \times 10^3, n_{\text{stat}}]$ steps. At each step, we present patterns as described above and collect the statistics of the recognition energy $E(J^s(t), \sigma^\alpha(t))$ between a presented pattern σ^α and the memory compartment $J^s(t)$ assigned according to Eq. (A1). These measurements are used to describe the energy statistics (see Fig. 2 and Supplemental Material Fig. S6 [38]) of the patterns and the accuracy of pattern-compartment association [Figs. 4(b) and 4(d)]. After the n_{stat}' steps, we again use the Metropolis algorithm to find the memory attractors associated with the presented patterns. We repeat this analysis for 50 independent realizations of the initializing pattern classes $\{\sigma^\alpha(t_0)\}$ for each set of parameters $\{L, N, C, \lambda, \mu, \beta_S, \beta_H\}$.

When calculating the mean performance $\mathcal{Q}$ of a strategy (see Figs. 2–4 and Supplemental Material Fig. S10 [38]), we set the overlap between the attractor and pattern $q^\alpha = |\langle \sigma_{\text{att}}^\alpha | \sigma^\alpha \rangle|$ equal to zero when the patterns are not recognized ($q^\alpha < 0.8$). As a result, the systems can achieve only a nonzero performance if they recognize some of the patterns. This choice eliminates the finite-size effect of a random overlap of approximately $1/\sqrt{L}$ between an attractor and a pattern (see Supplemental Material Fig. S5 [38]). This correction is especially important when comparing systems with different subnetwork sizes ($L_c \equiv L/C$) in the $\beta_H < 1$ regime (see Fig. 4 and Supplemental Material Fig. S10 [38]), where random overlaps for small L_c could otherwise result in a larger mean performance compared to larger systems that correctly reconstruct a fraction of the patterns.

3. Optimization procedure

We use grid search to find the optimal learning rate for a given parameter set ($\{L, N, C, \mu, \beta_S, \beta_H\}$). In the first step, we use the full range of learning rates ($\lambda \in (0, 1]$) to get a rough estimate $\tilde{\lambda}^*$ for the optimal learning rate. We then run simulations for a grid with 60 points between $\lambda = 10^{-4}$ and $3\tilde{\lambda}^*$ to estimate the optimal learning rate λ^* more precisely.

The reported optimal learning rates in Fig. 2(b) are obtained by averaging over a minimum of 200 independent repetitions of such a grid search. Through this process, we gather statistics from at least $200 \times 2 \times N$ pattern retrieval events, and for each case, we evaluate the overlap between the presented pattern and the attractor $\langle \sigma_{\text{att}} | \sigma \rangle$.

Despite the large number of realizations used for this optimization, the estimated performances along the grid can still be noisy, especially for small mutation and learning rates (μ and λ); note that a single unsuccessful recognition can lead to fluctuations in the results (see Supplemental Material Fig. S2 [38]). These fluctuations could lead to a noisy estimate of the optimal learning rate. We characterize the statistics of the optimal learning rate in the following way: For each λ on the grid, we calculate the mean performance of all the n recorded patterns ($\mathcal{Q} = (1/n) \sum_{i=1}^{n} q^i$). We also record the number n_0 of misclassified patterns (patterns with $q^i < 0.8$). Since the overlap q for misclassified patterns is close to zero and for recognized patterns it is close to 1 (Supplemental Material Fig. S1 [38]), misclassifying an additional pattern would reduce the performance $\mathcal{Q}$ by $1/n$. Assuming that the number of misclassified patterns for a given learning rate is Poisson distributed, the expected fluctuations in performance due to pattern misclassification should be given by $\sqrt{n_0}/n$. This error goes to zero in the limit of a large number of realizations n. To characterize the statistics of the optimal learning rate [Fig. 2(b)], we pull all the grid realizations whose performances $\mathcal{Q}$ lie within the error range $\sqrt{2 \max[1, n_0]}/n$ of the highest recorded performance. We use the median of the learning rates associated with these realizations as the optimal learning rate λ^*, and their standard deviation as the confidence interval indicated by the shaded area in Fig. 2(b).

4. Accuracy of pattern-compartment association

We use the mutual information $\text{MI}(\Sigma, \mathcal{C})$ between the set of pattern classes $\Sigma \equiv \{\sigma^\alpha\}$ and the set of all compartments $\mathcal{C}$ to quantify the accuracy in associating a presented pattern with the correct compartment,

$$
\begin{aligned}
\text{MI}(\Sigma, \mathcal{C}) &= H(\mathcal{C}) - H(\mathcal{C}|\Sigma) \\
&= -\sum_{c \in \mathcal{C}} P(c) \log P(c) \\
&\quad - \left[-\sum_{\sigma^\alpha \in \Sigma} P(\sigma^\alpha) \sum_{c \in \mathcal{C}} P(c|\sigma^\alpha) \log P(c|\sigma^\alpha) \right].
\end{aligned} \tag{A6}
$$

Here, $H(\mathcal{C})$ and $H(\mathcal{C}|\Sigma)$ are the entropy of the compartments and the conditional entropy of the compartments given the presented patterns, respectively. If chosen randomly, the entropy associated with choosing a compartment is $H_{\text{random}}(\mathcal{C}) = \log C$. The mutual information [Eq. (A6)] measures the reduction in the entropy of compartments due to the association between the patterns and the compartments measured by the conditional entropy $H(\mathcal{C}|\Sigma)$. Figures 4(b) and 4(d) show the mutual information $\text{MI}(\Sigma, \mathcal{C})$ scaled by its upper bound $H(\mathcal{C})$ in order to compare networks with a different number of compartments.

APPENDIX B: ESTIMATING ENERGY AND OPTIMAL LEARNING RATE FOR WORKING MEMORY

1. Approximate solution for optimal learning rate

The optimal learning rate is determined by maximizing the network's performance $\mathcal{Q}(\lambda)$ [Eq. (2)] against evolving patterns with a specified mutation rate:

$$
\lambda^* = \text{argmax}_\lambda \mathcal{Q}(\lambda). \tag{B1}
$$

We can numerically estimate the optimal learning rate as defined by Eq. (B1); see Figs. 2 and 3. However, the exact analytical evaluation of the optimal learning rate is difficult, and we use an approximate approach and find the learning rate that minimizes the expected energy of the patterns in the stationary state $\mathbb{E}[E_{\lambda,\rho}(J, \sigma)]$, assuming that patterns are shown to the network at a fixed order. Here, the subscripts explicitly indicate the learning rate of the network λ and the evolutionary overlap of the pattern ρ. To evaluate an analytical approximation for the energy, we first evaluate the state of the network $J(t)$ at time step t, given all the prior encounters of the networks with patterns shown at a fixed order

$$
\frac{1}{L}(J(t) + \mathbb{1}) = \lambda \sum_{j=1}^{\infty} (1 - \lambda)^{(j-1)} |\sigma(t - j)\rangle \langle \sigma(t - j)| \tag{B2}
$$

$$
= \lambda \sum_{i=1}^{\infty} (1 - \lambda)^{(i-1)N} \underbrace{\sum_{\alpha=1}^{N} (1 - \lambda)^{\alpha-1} |\sigma^\alpha(t - \alpha - (i-1)N)\rangle \langle \sigma^\alpha(t - \alpha - (i-1)N)|}_{\text{sum over } N \text{ pattern classes}} \tag{B3}
$$

$$
\underbrace{\phantom{= \lambda \sum_{i=1}^{\infty} (1 - \lambda)^{(i-1)N}}}_{\text{sum over time (generations,} i)}
$$

$$
= \lambda \sum_{\alpha=1}^{N} \underbrace{\sum_{i=0}^{\infty} (1 - \lambda)^{(\alpha-1)+iN} |\sigma^\alpha(t - \alpha - iN)\rangle \langle \sigma^\alpha(t - \alpha - iN)|}_{\text{sum over time}}. \tag{B4}
$$

$$
\underbrace{\phantom{= \lambda \sum_{\alpha=1}^{N} \sum_{i=0}^{\infty}}}_{\text{sum over patterns}}
$$

Here, we refer to the (normalized) pattern vector from the class α presented to the network at time step t by $|\sigma^\alpha(t)\rangle \equiv (1/\sqrt{L})\sigma^\alpha(t)$. Without loss of generality, we assume that the last pattern presented to the network at time step $t-1$ is from the first pattern class $|\sigma^1(t-1)\rangle$, which enables us to split the sum in Eq. (B2) into two

separate summations over pattern classes and N time-step generations [Eq. (B3)]. Adding the identity matrix 1 on the left-hand side of Eq. (B2) assures that the diagonal elements vanish, as defined in Eq. (A3).

The mean energy of the patterns, which in our setup is the energy of the pattern from the Nth class at time t, follows

$$
\begin{aligned}
\mathbb{E}[E_{\lambda,\rho}(J,\sigma)] &= \mathbb{E}\left[-\frac{1}{2}\langle\sigma^N(t)|J(t)|\sigma^N(t)\rangle\right] \\
&= \mathbb{E}\left[-\frac{L-1}{2}\lambda\sum_{\alpha=1}^{N}\sum_{i=0}^{\infty}(1-\lambda)^{(\alpha-1)+iN}\langle\sigma^N(t)|\sigma^\alpha(t-\alpha-iN)\rangle\langle\sigma^\alpha(t-\alpha-iN)|\sigma^N(t)\rangle\right].
\end{aligned} \tag{B5}
$$

Since the pattern families are orthogonal to each other, we can express the overlap between patterns at different times as $\langle\sigma^\alpha(t_1)|\sigma^\nu(t_2)\rangle = \delta_{\alpha,\nu}(1-2\mu)^{|t_2-t_1|} \equiv \delta_{\alpha,\nu}\rho^{|t_2-t_1|}$ and simplify the energy function in Eq. (B5),

$$
\begin{aligned}
\mathbb{E}[E_{\lambda,\rho}(J,\sigma)] &= -\frac{L-1}{2}\lambda\sum_{i=0}^{\infty}(1-\lambda)^{(N-1)+iN}\rho^{2(N+iN)} \\
&= -\frac{L-1}{2}\lambda(1-\lambda)^{(N-1)}\rho^{2N}\sum_{i=0}^{\infty}((1-\lambda)^N\rho^{2N})^i \\
&= -\frac{L-1}{2}\lambda\frac{(1-\lambda)^{(N-1)}\rho^{2N}}{1-(1-\lambda)^N\rho^{2N}}.
\end{aligned} \tag{B6}
$$

Since accurate pattern retrieval depends on the depth of the energy valley for the associative memory, we use the expected energy of the patterns as a proxy for the performance of the network. We can find the approximate optimal learning rate that minimizes the expected energy by setting $\partial\mathbb{E}[E_{\lambda,\rho}(J,\sigma)]/\partial\lambda = 0$, which results in

$$
\begin{aligned}
(1-2\mu)^{2N} &= (1-\lambda^*)^{-N}(1-N\lambda^*) \\
\Rightarrow 1-4N\mu + \mathcal{O}(\mu^2) &= 1+\frac{1}{2}(N-N^2)(\lambda^*)^2 + \mathcal{O}(\lambda^3); \\
\Rightarrow \lambda^*(\mu) &\simeq \sqrt{8\mu/(N-1)},
\end{aligned} \tag{B7}
$$

where we use the fact that both the mutation rate μ and the learning rate λ are small, and therefore, expand Eq. (B7) up to the leading orders in these parameters.

In addition, Eq. (B7) establishes an upper bound for the learning rate $\lambda < (1/N)$. Therefore, our expansion in mutation rate [Eq. (B7)] is only valid for $8\mu < (1/N)$, or

equivalently, for $\mu_{\text{eff}} = N\mu < 12.5\%$; the rates used in our analyses lie far below these upper bounds.

2. Lag of memory against evolving patterns

The memory attractors associated with a given pattern class can lag behind the evolution and reflect only the older patterns presented prior to the most recent encounter of the network with the specified class. As a result, the upper bound for the performance of a network $\mathcal{Q}_{\text{lag}} = \rho^{g_{\text{lag}}N} \approx 1-2g_{\text{lag}}\mu_{\text{eff}}$ is determined by both the evolutionary divergence of patterns between two encounters μ_{eff} and the number of generations g_{lag} by which the stored memory lags. We measure g_{lag} in units of generations; one generation is defined as the average time between a network's encounter with the same pattern class, i.e., N. We characterize this lag g_{lag} by identifying the past pattern (at time $t - g_{\text{lag}}N$) that has the maximum overlap with the network's energy landscape at given time t:

$$
\begin{aligned}
g_{\text{lag}} &= \text{argmax}_{g\geq 0}\mathbb{E}[\langle\sigma(t-gN)|J(t)|\sigma(t-gN)\rangle] \\
&\equiv \text{argmin}_{g\geq 0}\mathbb{E}[E_{\text{lag}}(g)],
\end{aligned} \tag{B8}
$$

where we introduce the expected lagged energy $\mathbb{E}[E_{\text{lag}}(g)]$. Here, the vector $|\sigma(t)\rangle$ refers to the pattern σ presented to the network at time t, which can be from any of the pattern classes. Because of the translational symmetry in time in the stationary state, the lagged energy can also be expressed in terms of the overlap between a pattern at time t and the network at a later time $t + gN$. We evaluate the lagged energy by substituting the expression for the network's state $J(t + gN)$ from Eq. (B2), which entails

$$\frac{2}{L-1}\mathbb{E}[E_{\text{lag}}(g)] = -\frac{1}{L-1}\mathbb{E}[\langle\sigma(t)|J(t+gN)|\sigma(t)\rangle] \tag{B9}$$

$$= -\mathbb{E}\left[\frac{1}{L-1}(1-\lambda)^{Ng}\langle\sigma(t)|J(t)|\sigma(t)\rangle + \lambda\sum_{j=0}^{gN-1}(1-\lambda)^{gN-1-j}\langle\sigma(t)|\sigma(t+j)\rangle^2\right] \tag{B10}$$

$$= \frac{2}{L-1}(1-\lambda)^{Ng}\mathbb{E}[E_{\lambda,\rho}(J,\sigma)] - \lambda\sum_{i=0}^{g-1}\sum_{\alpha=0}^{N-1}(1-\lambda)^{gN-1-Ni-\alpha}\langle\sigma^N(t)|\sigma^{N-\alpha}(t+Ni+\alpha)\rangle^2 \tag{B11}$$

$$= \frac{2}{L-1}(1-\lambda)^{Ng}\mathbb{E}[E_{\lambda,\rho}(J,\sigma)] - \lambda\sum_{i=0}^{g-1}(1-\lambda)^{gN-1-Ni}\rho^{2Ni} \tag{B12}$$

$$= -\lambda\left(\frac{(1-\lambda)^{N(g+1)}\rho^{2N}}{1-(1-\lambda)^N\rho^{2N}} + \frac{(1-\lambda)^{N(g+1)-1} - (1-\lambda)^{N-1}\rho^{2Ng}}{(1-\lambda)^N - \rho^{2N}}\right). \tag{B13}$$

Here, we use the expression of the network's matrix J in Eq. (A4) to arrive at Eq. (B10), and then follow the procedure introduced in Eq. (B3) to arrive at the double summation in Eq. (B11). We then use the equation for pattern overlap $\langle\sigma^\alpha(t_1)|\sigma^\nu(t_2)\rangle = \delta_{\alpha,\nu}\rho^{|t_2-t_1|}$ to reduce the sum in Eq. (B12) and arrive at the result in Eq. (B13) by evaluating the geometric sum and substituting the empirical average for the energy $\mathbb{E}[E_{\lambda,\rho}(J,\sigma)]$ from Eq. (B6).

We probe this lagged memory by looking at the performance $\mathcal{Q}$ for patterns that are correctly associated with their memory attractors (i.e., those with $\langle\sigma_{\text{att}}|\sigma\rangle > 0.8$). As shown in Fig. S5 of the Supplemental Material [38], for a broad parameter regime, the mean performance for these correctly associated patterns agrees well with the theoretical expectation $\mathcal{Q}_{\text{lag}} = \rho^{g_{\text{lag}}N}$, which is lower than the naive expectation $\mathcal{Q}_0$.

APPENDIX C: STRUCTURE OF THE ENERGY LANDSCAPE FOR WORKING MEMORY

1. Formation of mountain passes in the energy landscape of memory for evolving patterns

As shown in Fig. 1, large learning rates in networks with memory for evolving patterns result in the emergence of narrow connecting paths between the minima of the energy landscape. We refer to these narrow connecting paths as mountain passes. In pattern retrieval, the Monte Carlo search can drive a pattern out of one energy minimum into another minimum and potentially lead to pattern misclassification.

We use two features of the energy landscape to probe the emergence of the mountain passes. First, we show that if a pattern is misclassified, it has fallen into a memory attractor associated with another pattern class and not a spuriously made energy minima. To do so, we compare the overlap of the attractor with the original pattern $|\langle\sigma_{\text{att}}^\alpha|\sigma^\alpha\rangle|$ (i.e., the

reconstruction performance of the patterns) with the maximal overlap of the attractor with all other patterns $\max_{\nu\neq\alpha}|\langle\sigma_{\text{att}}^\alpha|\sigma^\nu\rangle|$. Indeed, as shown in Fig. S5(a) of the Supplemental Material [38], for evolving patterns, the memory attractors associated with 99.4% of the originally stored patterns have either a large overlap with the correct pattern or with one of the other previously presented patterns; 71.3% of the patterns are correctly classified [stable patterns in sector I in Fig. S5(a) of the Supplemental Material [38]], whereas 28.1% of them are associated with a secondary energy minima after equilibration [unstable patterns in sector II in Fig. S5(a) of the Supplemental Material [38]]. A very small fraction of patterns ($< 1\%$) fall into local minima given by the linear combinations of the presented patterns [sector IV in Fig. S5(a) of the Supplemental Material [38]]. These minima are well known in the classical Hopfield model [64,65]. Moreover, we see that equilibration of a random pattern (i.e., a pattern orthogonal to all the presented classes) in the energy landscape leads to memory attractors for one of the originally presented pattern classes. The majority of these random patterns lie in sector II of Fig. S5(a) of the Supplemental Material [38]; i.e., they have a small overlap with the network since they are orthogonal to the originally presented pattern classes, and they fall into one of the existing memory attractors after equilibration.

Second, we characterize the possible paths for a pattern to move from one valley to another during equilibration using a Monte Carlo algorithm with the Metropolis acceptance probability,

$$\rho(\sigma \to \sigma') = \min\left(1, e^{-\beta(E(J,\sigma')-E(J,\sigma))}\right). \tag{C1}$$

We estimate the number of beneficial spin flips (i.e., open paths) that decrease the energy of a pattern at the start

of equilibration [Supplemental Material Fig. S1(b) [38]]. The average number of open paths is smaller for stable patterns compared to the unstable patterns that are misclassified during retrieval [Supplemental Material Fig. S1(b) [38]]. However, the distributions for the number of open paths largely overlap for stable and unstable patterns. Therefore, the local energy landscapes of stable and unstable patterns are quite similar, and it is difficult to discriminate between them solely based on the local gradients in the landscape. Figure S7(a) in the Supplemental Material [38] shows that the average number of beneficial spin flips grows with the mutation rate of the patterns, but this number is comparable for stable and unstable patterns. Moreover, the unstable stored patterns (blue) have far fewer open paths available to them during equilibration compared to random patterns (red) that are presented to the network for the first time [Supplemental Material Figs. S1(b) and S7(a) [38]]. Notably, on average, half of the spin flips reduce the energy for random patterns, irrespective of the mutation rate. This indicates that even though previously presented pattern classes are statistically distinct from random patterns, they can still become unstable, especially in networks which are presented with evolving patterns.

It should be noted that the evolution of the patterns only indirectly contribute to the misclassification of memory, as it necessitates a larger learning rate for the networks to optimally operate, which in turn results in the emergence of mountain passes. To clearly demonstrate this effect, Figs. S1(c), S1(d), and S7(d) in the Supplemental Material [38] show the misclassification behavior for networks trained to store memory for static patterns while using a larger learning rate that is optimized for evolving patterns. Indeed, we see that pattern misclassification in this case is consistent with the existence of mountain passes in the network's energy landscape.

2. Spectral decomposition of the energy landscape

We use spectral decomposition of the energy landscape to characterize the relative positioning and the stability of patterns in the landscape. As shown in Figs. S1 and S6 in the Supplemental Material [38], destabilization of patterns due to equilibration over mountain passes occurs in networks with high learning rates, even for static patterns. Therefore, we focus on how the learning rate impacts the spectral decomposition of the energy landscape in networks presented with static patterns. This simplification will enable us to analytically probe the structure of the energy landscape, which we will compare with the numerical results for evolving patterns.

We can represent the network J (of size $L \times L$) that stores a memory of N static patterns with N nontrivial eigenvectors $|\Phi^i\rangle$ with corresponding eigenvalues Γ_i and $N - L$ degenerate eigenvectors $|\Psi^k\rangle$ with corresponding trivial eigenvalues $\gamma_k = \gamma = -1$:

$$J = \sum_{i=1}^{N} \Gamma_i |\Phi^i\rangle\langle\Phi^i| + \sum_{k=1}^{L-N} \gamma_k |\Psi^k\rangle\langle\Psi^k|. \quad \text{(C2)}$$

The nontrivial eigenvectors span the space of the presented patterns, for which the recognition energy can be expressed by

$$E(J, \sigma^\alpha) = -\frac{1}{2} \sum_{i=1}^{N} \Gamma_i \langle\sigma^\alpha|\Phi^i\rangle\langle\Phi^i|\sigma^\alpha\rangle. \quad \text{(C3)}$$

An arbitrary configuration χ, in general, can have components orthogonal to the N eigenvectors $|\Phi^i\rangle$, as it points to a vertex of the hypercube and should be expressed in terms of all the eigenvectors $\{\Phi^1, ..., \Phi^N, \Psi^1, ..., \Psi^{L-N}\}$:

$$E(J, \chi) = -\frac{1}{2}\left(\underbrace{\sum_{i=1}^{N} \Gamma_i \langle\chi|\Phi^i\rangle\langle\Phi^i|\chi\rangle}_{\text{stored patterns}} + \underbrace{\sum_{k=1}^{L-N} \gamma\langle\chi|\Psi^k\rangle\langle\Psi^k|\chi\rangle}_{\text{trivial space}}\right).$$
$$\text{(C4)}$$

Any spin flip in a pattern (e.g., during equilibration) can be understood as a rotation in the eigenspace of the network [Eq. (C4)]. As a first step in characterizing these rotations, we remind ourselves of the identity

$$|\chi\rangle = \sum_{i=1}^{N} \langle\Phi^i|\chi\rangle|\Phi^i\rangle + \sum_{k=1}^{L-N} \langle\Psi^k|\chi\rangle|\Psi^k\rangle, \quad \text{(C5)}$$

with the normalization condition

$$\sum_{i=1}^{N} (\langle\Phi^i|\chi\rangle)^2 + \sum_{k=1}^{L-N} (\langle\Psi^k|\chi\rangle)^2 = 1. \quad \text{(C6)}$$

In addition, since the diagonal elements of the network are set to $J_{ii} = 0$ [Eq. (A3)], the eigenvalues should sum to zero, or alternatively,

$$\sum_{i=1}^{N} \Gamma_i = -\sum_{k=1}^{L-N} \gamma_k = L - N. \quad \text{(C7)}$$

To asses the stability of a pattern σ^ν, we compare its recognition energy $E(J, \sigma^\nu)$ with the energy of the rotated pattern after a spin flip $E(J, \tilde{\sigma}^\nu)$. To do so, we first consider a simple scenario, where we assume that the pattern σ^ν has a large overlap with one dominant nontrivial eigenvector Φ^A (i.e., $\langle\sigma^\nu|\Phi^A\rangle^2 = m^2 \approx 1$). The other components of the pattern can be expressed in terms of the remaining $N - 1$ nontrivial eigenvectors with mean-squared overlap $1 - m^2/N - 1$. The expansion of the recognition energy

for the presented pattern is restricted to the N nontrivial directions [Eq. (C4)] resulting in

$$E(J, \sigma^\nu) = -\frac{1}{2}\left(m^2\Gamma_A + \sum_{i \neq A}\frac{1 - m^2}{N - 1}\Gamma_i\right)$$

$$= -\frac{1}{2}(m^2\Gamma_A + (1 - m^2)\tilde{\Gamma}), \qquad \text{(C8)}$$

where $\tilde{\Gamma} = [1/(N - 1)]\sum_{i \neq A}\Gamma_i = [(N\bar{\Gamma} - \Gamma_A)/(N - 1)]$ is the mean eigenvalue for the nondominant directions.

A spin flip ($|\sigma^\nu\rangle \to |\tilde{\sigma}^\nu\rangle$) can rotate the pattern out of the dominant direction Φ^A and reduce the squared overlap by ϵ^2. The rotated pattern $|\tilde{\sigma}^\nu\rangle$, in general, lies in the L-dimensional space and is not restricted to the N-dimensional (nontrivial) subspace. We first take a mean-field approach in describing the rotation of the pattern after a spin flip. Because of the normalization condition [Eq. (C6)], the loss in the overlap with the dominant direction should result in an average increase in the overlap with the other $L - 1$ eigenvectors by $\epsilon^2/(L - 1)$. The energy of the rotated pattern after a spin flip $E(J, \tilde{\sigma}^\nu)$ can be expressed in terms of all the L eigenvectors [Eq. (C4)],

$$E(J, \tilde{\sigma}^\nu) = -\frac{1}{2}\left[(m^2 - \epsilon^2)\Gamma_A + \sum_{i \neq A}\left(\frac{1 - m^2}{N - 1} + \frac{\epsilon^2}{L - 1}\right)\Gamma_i\right.$$

$$\left. + \sum_k \frac{\epsilon^2}{L - 1}\gamma_k\right]$$

$$= E(J, \sigma^\nu) + \frac{\epsilon^2}{2}\left[\Gamma_A - \frac{1}{L - 1}\left(\sum_{i \neq A}\Gamma_i + \sum_k\gamma_k\right)\right]$$

$$\text{(C9)}$$

$$= E(J, \sigma^\nu) + \frac{\epsilon^2}{2}\Gamma_A\left(1 + \frac{1}{L - 1}\right), \qquad \text{(C10)}$$

where in Eq. (C10) we use the fact that the eigenvalues should sum up to zero. On average, a spin flip $|\sigma^\nu\rangle \to |\tilde{\sigma}^\nu\rangle$ increases the recognition energy by $E(J, \tilde{\sigma}^\nu) - E(J, \sigma^\nu) = (\epsilon^2/2)\Gamma_A[1 + \mathcal{O}(L^{-1})]$. This is consistent with the results shown in Figs. S5(b), S5(d), S6(a), and S6(d) in the Supplemental Material [38], which indicate that the majority of the spin flips keep a pattern in the original energy minimum, and only a few of the spin flips drive a pattern out of the original attractor.

In the analysis above, we assume that the reduction in overlap with the dominant eigenvector ϵ^2 is absorbed equally by all the other eigenvectors (i.e., the mean-field approach). In this case, the change in energy is equally distributed across the positive and the negative eigenvalues [Γ's and γ's in Eq. (C9)], resulting in an overall increase in the energy due to the reduced overlap with the dominant direction $|\Phi^A\rangle$. The destabilizing spin flips are associated with atypical changes that rotate a pattern onto a secondary

nontrivial direction $|\Phi^B\rangle$ (with positive eigenvalue Γ_B), as a result of which the total energy could be reduced. To better characterize the conditions under which patterns become unstable, we introduce a perturbation to the mean-field approach used in Eq. (C10). We assume that a spin flip results in a rotation with a dominant component along a secondary nontrivial direction $|\Phi^B\rangle$. Specifically, we assume the reduced overlap ϵ^2 between the original pattern $|\sigma^\nu\rangle$ and the dominant direction $|\Phi^A\rangle$ is distributed in an imbalanced fashion between the other eigenvectors, with a fraction p projected onto a new (nontrivial) direction $|\Phi^B\rangle$, while all the other $L - 2$ directions span the remaining $(1 - p)\epsilon^2$. In this case, the energy of the rotated pattern is given by

$$E(J, \tilde{\sigma}^\nu) = -\frac{1}{2}\left[(m^2 - \epsilon^2)\Gamma_A + \left(\frac{1 - m^2}{N - 1} + p\epsilon^2\right)\Gamma_B\right.$$

$$+ \sum_{i \neq A, B}\left(\frac{1 - m^2}{N - 1} + \frac{(1 - p)\epsilon^2}{L - 2}\right)\Gamma_i$$

$$\left. + \sum_k\frac{(1 - p)\epsilon^2}{L - 2}\gamma_k\right]$$

$$= E(J, \sigma^\nu) + \frac{\epsilon^2}{2}[\Gamma_A - p\Gamma_B + \mathcal{O}(L^{-1})]. \qquad \text{(C11)}$$

Therefore, a spin flip is beneficial if $\Gamma_A < p\Gamma_B$. To further concretize this condition, we estimate the typical loss ϵ^2 and gain $p\epsilon^2$ in the squared overlap between the pattern and its dominating directions due to rotation by a single spin flip.

Let us consider a rotation $|\sigma^\nu\rangle \to |\tilde{\sigma}^\nu\rangle$ by a flip in the nth spin of the original pattern $|\sigma^\nu\rangle$. This spin flip reduces the original overlap of the pattern $m = \langle\sigma^\nu|\Phi^A\rangle$ with the dominant direction $|\Phi^A\rangle$ by the amount $(2/\sqrt{L})\Phi_n^A$, where Φ_n^A is the nth entry of the eigenvector $|\Phi^A\rangle$. Since the original overlap is large (i.e., $m \simeq 1$), all entries of the dominant eigenvector are approximately $\Phi_i^A \simeq 1/\sqrt{L}$, $\forall i$, resulting in a reduced overlap of the rotated pattern $\langle\sigma^\nu|\Phi^A\rangle = m - (2/L)$. Therefore, the loss in the squared overlap ϵ^2 by a spin flip is given by

$$\epsilon^2 = \langle\sigma^\nu|\Phi^j\rangle^2 - \langle\tilde{\sigma}^\nu|\Phi^j\rangle^2$$

$$= m^2 - \left(m^2 - 4\frac{m}{L} + 4\frac{1}{L^2}\right)$$

$$= 4\frac{m}{L} + \mathcal{O}\left(\frac{1}{L^2}\right). \qquad \text{(C12)}$$

Similarly, we can derive the gain in the squared overlap $p\epsilon^2$ between the pattern $|\sigma^\nu\rangle$ and the new dominant direction $|\Phi^B\rangle$ after a spin flip. Except for the direction $|\Phi^A\rangle$, the expected squared overlap between the original pattern (prior to the spin flip) and any of the nontrivial eigenvectors

including $|\Phi^B\rangle$ is $\langle\sigma^\nu|\Phi^B\rangle^2 = [(1-m^2)/(N-1)]$. The flip in the nth spin of the original pattern increases the overlap of the rotated pattern with the new dominant direction $|\Phi^B\rangle$ by $2(\Phi_n^B/\sqrt{L})$, where Φ_n^B should be of the order of $\sqrt{1/L}$. Therefore, the largest gain in overlap due to a spin flip is given by

$$
p\epsilon^2 = \langle\tilde{\sigma}^\nu|\Phi^B\rangle^2 - \langle\sigma^\nu|\Phi^B\rangle^2
$$

$$
\simeq \left(\frac{1-m^2}{N-1} + 4\sqrt{\frac{1-m^2}{N-1}}\frac{\Phi_n^B}{\sqrt{L}} + 4\frac{(\Phi_n^B)^2}{L}\right) - \frac{1-m^2}{N-1}
$$

$$
= \sqrt{\frac{1-m^2}{N-1}}\frac{\Phi_n^B}{\sqrt{L}} + \mathcal{O}\left(\frac{1}{L^2}\right). \qquad (C13)
$$

By using the results from Eqs. (C12) and (C13), we can express the condition for beneficial spin flips to drive a pattern over the carved mountain passes during equilibration [Eq. (C11)],

$$
\epsilon^2\Gamma_A < \epsilon^2 p\Gamma_B \rightarrow \frac{\Gamma_A}{\Gamma_B} < \sqrt{\frac{1-m^2}{m^2}}\frac{1}{\sqrt{\alpha}}\Phi_n^B, \qquad (C14)
$$

where $\alpha = N/L$. This result suggests that the stability of a pattern depends on how the ratio of the eigenvalues associated with the dominant projections of the pattern before and after the spin flip Γ_A/Γ_B compare to the overlap m of the original pattern with the dominant eigenvector Φ^A and the change due to the spin flip Φ_n^B.

So far, we have constrained our analysis to patterns that have a dominant contribution to only one eigenvector Φ^A. To extend our analysis to patterns which are instead constrained to a small subspace $\mathcal{A}$ spanned by nontrivial eigenvalues, we define the squared pattern overlap with the subspace $m_\mathcal{A}^2 = \sum_{a\in\mathcal{A}}\langle\sigma^\nu|\Phi^a\rangle^2$ and a weighted average eigenvalue in the subspace $\Gamma_\mathcal{A} = \sum_{a\in\mathcal{A}}\langle\sigma^\nu|\Phi^a\rangle^2\Gamma_a$. As a result, the difference in the energy of a pattern before and after a spin flip [Eq. (C11)] can be extended to $E(J,\sigma^\nu) - E(J,\tilde{\sigma}^\nu) = (\epsilon^2/2)[\Gamma_\mathcal{A} - p\Gamma_B + \mathcal{O}(L^{-1})]$. Similarly, the stability condition in Eq. (C14) can be extended to $(\Gamma_\mathcal{A}/\Gamma_B) < \sqrt{[(1-m_\mathcal{A}^2)/m_\mathcal{A}^2](1/\sqrt{\alpha})}\Phi_n^B$. Patterns that are constrained to larger subspaces are more stable, since the weighted average eigenvalue for their containing subspace $\Gamma_\mathcal{A}$ is closer to the mean of all eigenvalues $\bar{\Gamma} = 1 - N/L$ (law of large numbers). Therefore, in these cases a much larger eigenvalue gap (or a broader eigenvalue spectrum) is necessary to satisfy the condition for pattern instability.

Figure S9 in the Supplemental Material [38] compares the loss in energy with the original dominant direction $\epsilon^2\Gamma_A$ to the maximal gain in any of the other directions $\epsilon^2 p\Gamma_B$ to test the pattern stability criteria presented in Eq. (C14). To do so, we identify a spin flip n in a secondary direction B that confers the maximal energy gain: $\epsilon^2 p\Gamma_B \approx \max_{n,B}\sqrt{(1-m^2)/(N-1)}(\Phi_n^B/\sqrt{L})\Gamma_B$. In Figs. S9(a)

and S9(c) of the Supplemental Material [38], we specifically focus on the subset of patterns that show a large (squared) overlap with the one dominant direction (i.e., $m > 0.85$). Given that evolving patterns are not a constraint to the $\{\Phi\}$ (nontrivial) subspace, we find a smaller fraction of these patterns to fulfill the condition for such a large overlap m [see Supplemental Material Fig. S9(a) [38]] compared to the static patterns [Supplemental Material Fig. S9(c) [38]]. Nonetheless, we see that the criteria in Eq. (C14) can be used to predict the stability of patterns in a network for both static and evolving patterns; note that here we use the same learning rate for both the static and evolving patterns.

We then relax the overlap condition by including all patterns that have a large overlap with a subspace $\mathcal{A}$ spanned by up to ten eigenvectors (i.e., $m_\mathcal{A}^2 = \sum_{a\in\mathcal{A}}\langle\sigma|\Phi^\alpha\rangle^2 > 0.85$). For these larger subspaces, the transition between stable and unstable patterns is no longer exactly given by Eq. (C14). However, the two contributions $\epsilon^2\Gamma_A$ and $\epsilon^2 p\Gamma_B$ still clearly separate the patterns into stable and unstable classes for both static and evolving patterns [Supplemental Material Figs. S9(b) and S9(d) [38]]. The softening of this condition is expected, as in this regime we can no longer assume that a single spin flip can reduce the overlap with all the eigenvectors in the original subspace. As a result, the effective loss in the overlap becomes smaller than ϵ^2 and patterns become unstable below the dotted line in Supplemental Material Figs. S9(b) and S9(d) [38].

As the learning rate increases, the gap between the eigenvalues Γ_B/Γ_A (Supplemental Material Fig. S8 [38]) becomes larger; note that the inverse of this gap sets the lower bound for destabilization condition in Eq. (C14). At the same time, with increasing learning rate, patterns become more constrained to smaller subspaces [Supplemental Material Figs. S6(c) and S6(d) [38]]. As a result of these two effects, more patterns satisfy the instability criteria in Eq. (C14). These patterns are misclassified, as they fall into a wrong energy minimum by equilibrating through the mountain passes carved in the energy landscape of networks with large learning rates.

APPENDIX D: ALTERNATIVE LEARNING RULES

In the main text, we focus on the standard Hebbian learning rule [Eq. (2)], both for the entire network and for individual compartments. Hopfield networks are among the most studied models for learning. Thus, it is no surprise that many other learning rules have been developed for these networks [37].

Here, we discuss a number of these rules and assess their impact on our results. We focus on local and incremental learning rules, in which the updates of the weights depend only on one pattern at a time. Other learning rules, such as the Kraut-Mézard class [66], use the information from all

patterns during each round of update, which makes them unrealistic for biological systems. With this condition in mind, we introduce and discuss the consequences of the Storkey learning rule [36], the gradient-descent learning rule [37], and the sparse Hebbian learning rule.

1. Storkey learning

Introduced by Storkey in 1997, this rule had the goal of increasing the capacity of Hopfield networks [36]. Indeed, Hopfield networks that are trained with this rule have a capacity of $L/\sqrt{2\ln L}$ [36], which is significantly larger than the capacity of $L/(2\ln L)$ reached with the conventional Hebbian learning [35].

This rule was originally designed for a learning phase in which all patterns are known. In that case, the update to the interaction matrix $J_{i,j}$ is given by

$$\Delta J_{i,j} = \begin{cases} \frac{1}{N}(\sigma_i - f_{i,j})(\sigma_j - f_{ji}), & \text{if } i \neq j; \\ 0 & \text{otherwise,} \end{cases} \quad \text{(D1)}$$

where $f_{i,j}$ is the local field on the spin σ_i except for the contribution from σ_j,

$$f_{i,j} = \frac{1}{L-2}\left(\sum_k J_{i,k}\sigma_k - J_{i,i}\sigma_i - J_{i,j}\sigma_j\right). \quad \text{(D2)}$$

To use this learning rule for consecutive encounters with (evolving) patterns, we add a learning rate λ to Eq. (D1) and then use the rule

$$\Delta J_{i,j} = \begin{cases} \lambda(\sigma_i - f_{i,j})(\sigma_j - f_{ji}), & \text{if } i \neq j; \\ 0 & \text{otherwise.} \end{cases} \quad \text{(D3)}$$

2. Gradient descent

The discussion of gradient descent in the Hopfield model follows from Ref. [37].

Gradient descent is the foundation of many optimization problems. With respect to the Hopfield model, we want to construct energy minima associated with the stored memory that are as deep as possible. For any given pattern σ, this is achieved if $W\sigma = \sigma$, where $W_{i,j} = [1/(L-1)]J_{i,j}$ is a normalized interaction matrix. In other words, we want to minimize the distance $D^{(p)}(\sigma, W\sigma)$ between the pattern σ and its projection $W\sigma$. Here, $D^{(p)}(\cdot,\cdot)$ is the distance measure for a general L^p-norm, which for the L^2-norm follows

$$D^{(2)}(\sigma, W\sigma) = \sum_i \left(\sigma_i - \sum_j W_{i,j}\sigma_j\right)^2. \quad \text{(D4)}$$

The derivative of the L^2 distance with respect to the element $W_{i,j}$ is given by

$$\frac{d}{dW_{i,j}}D^{(2)}(\sigma, W\sigma) = -2\left(\sigma_i - \sum_k W_{i,k}\sigma_k\right)\sigma_j. \quad \text{(D5)}$$

Therefore, we can define the gradient-descent learning rule toward the energy minimum with learning rate (step size) λ that is consistent with minimization of the L^2 distance as

$$\Delta J_{i,j} = \begin{cases} \lambda\left(\sigma_i - \sum_k W_{i,k}\sigma_k\right)\sigma_j, & \text{if } i \neq j; \\ 0 & \text{otherwise.} \end{cases} \quad \text{(D6)}$$

Interestingly, when using an L^1-norm distance, the gradient-descent learning rule is equivalent to the original Hebbian learning rule [Eq. (2)]. This can be shown by first evaluating the derivative of the distance for the L^1-norm with respect to the element $W_{i,j}$ as

$$\begin{aligned} \frac{d}{dW_{i,j}}D^{(1)}(\sigma, W\sigma) &= -\frac{(\sigma_i - \sum_k W_{i,k}\sigma_k)}{\sqrt{(\sigma_i - \sum_k W_{i,k}\sigma_k)^2}}\sigma_j \\ &= -\text{sign}\left(\sigma_i - \sum_k W_{i,k}\sigma_k\right)\sigma_j \\ &= -\sigma_i\sigma_j, \end{aligned} \quad \text{(D7)}$$

where we use the fact that $|\sigma_i| \geq |\sum_k W_{i,k}\sigma_k|$, which results in $\text{sign}(\sigma_i - \sum_k W_{i,k}\sigma_k) = \text{sign}(\sigma_i) = \sigma_i$. Therefore, for the L^1-norm distance, we recover the original Hebbian learning rule with $\Delta J_{i,j} = \lambda\sigma_i\sigma_j$.

3. Sparse Hebbian learning

Machine-learning algorithms often enforce sparsity to regularize neural networks to avoid overfitting [29]. While a direct translation of such a regularization to the Hopfield model is nontrivial, we can enforce sparsity on the interaction matrix $J_{i,j}$. To achieve a sparsity of $X\%$ in the interaction matrix $J_{i,j}$, we use the standard Hebbian learning rule [Eq. (2)] and set entries $J_{i,j}$ with absolute values smaller than $(100 - X)\%$ of all entries to zero.

4. Performance of networks with alternative learning rules

To characterize the impact of learning rules on our results, we perform the same optimization procedure as for the standard Hebbian learning in the main text. It should be noted that these alternative learning protocols are substantially more complex than the standard Hebbian learning, which limits our simulations to networks of maximum size $L = 100$ and $N = 8$ (in contrast to $L = 800$ and $N = 32$ in the main text). As we increase the ratio of N/L, we observe stronger finite-size effects. Still, we stay far below the capacity of the network and have no reason to expect any qualitative changes in the outcomes for networks of larger size.

In Fig. S3(a) of the Supplemental Material [38], we compare the performance of the networks trained with the standard Hebbian learning to that of the alternative models, i.e., the Storkey, the gradient descent, and sparse Hebbian learning rules, in recognizing patterns evolving with a range of effective mutation rates μ_{eff}. For small sparsity (10%), the networks trained with sparse Hebbian learning perform similar to those with the standard Hebbian learning. However, when sparsity is large (50%), the sparse networks appear to lose the exact position of the minima. As a result, the system's performance systematically decays even for very slowly evolving patterns (small μ_{eff}). The Storkey and the gradient-descent learning rules perform slightly better than the Hebbian learning for evolving patterns. This slight increase in performance is most likely a consequence of the increased capacity. A similar effect is seen in Fig. 2, as emptier (larger) networks perform better at a fixed effective mutation rate. However, the gain due to capacity is negligible compared to the reduction in performance of all networks with increasing mutation rate.

Aside from their similar performances, the distortion of the energy landscape for networks following the alternative learning rules are also comparable to that of the standard Hebbian learning. Specifically, we see that with all the alternative learning rules, the misclassified patterns fall into attractors associated with one of the other pattern classes [see Supplemental Material Fig. S3(b) [38]], consistent with the results for the standard Hebbian learning in Fig. S1 (a) of the Supplemental Material [38]. Moreover, alternative learning rules also give rise to network structures in which the average number of open paths (i.e., number of beneficial spin flips during equilibration) is smaller for stable (correctly classified) patterns compared to the unstable (misclassified) patterns, and both are smaller than for random patterns [Supplemental Material Fig. S3(c) [38]]. This result is also similar to that of the standard Hebbian learning in Fig. S1(b) of the Supplemental Material [38]. It should be noted that the larger fluctuations seen in Supplemental Material Figs. S3(b) and S3(c) compared to Figs. S1(a) and S1(b) [38] are due to the smaller system size used for simulations with the alternative learning rules.

In conclusion, the fraction of the correctly reconstructed patterns decay with increasing evolutionary rates due to the emergence of narrow passes in a network's energy landscape, irrespective of the choice of the learning rule. While we cannot exclude that other learning rules might achieve better performance, the results suggest that local (in time) learning rules which act on all network weights cannot learn multiple evolving patterns with maximal performance. Therefore, it is likely that the specialized 1-to-1 memory remains the only strategy that effectively learns and recovers evolving patterns.

[1] S. J. Labrie, J. E. Samson, and S. Moineau, *Bacteriophage Resistance Mechanisms*, Nat. Rev. Microbiol. **8**, 317 (2010).

[2] R. Barrangou and L. A. Marraffini, *CRISPR-Cas Systems: Prokaryotes Upgrade to Adaptive Immunity*, Mol. Cell **54**, 234 (2014).

[3] S. Bradde, A. Nourmohammad, S. Goyal, and V. Balasubramanian, *The Size of the Immune Repertoire of Bacteria*, Proc. Natl. Acad. Sci. U.S.A. **117**, 5144 (2020).

[4] A. S. Perelson and G. Weisbuch, *Immunology for Physicists*, Rev. Mod. Phys. **69**, 1219 (1997).

[5] C. Janeway, P. Travers, M. Walport, and M. Schlomchik, *Immunobiology*, 5th ed., The Immune System in Health and Disease (Garland Science, New York, 2001).

[6] G. Altan-Bonnet, T. Mora, and A. M. Walczak, *Quantitative Immunology for Physicists*, Phys. Rep. **849**, 1 (2020).

[7] L. B. Haberly and J. M. Bower, *Olfactory Cortex: Model Circuit for Study of Associative Memory?*, Trends Neurosci. **12**, 258 (1989).

[8] P. Brennan, H. Kaba, and E. B. Keverne, *Olfactory Recognition: A Simple Memory System*, Science **250**, 1223 (1990).

[9] R. Granger and G. Lynch, *Higher Olfactory Processes: Perceptual Learning and Memory*, Curr. Opin. Neurobiol. **1**, 209 (1991).

[10] L. B. Haberly, *Parallel-Distributed Processing in Olfactory Cortex: New Insights from Morphological and Physiological Analysis of Neuronal Circuitry*, Chem. Senses **26**, 551 (2001).

[11] D. A. Wilson, A. R. Best, and R. M. Sullivan, *Plasticity in the Olfactory System: Lessons for the Neurobiology of Memory*, Neurosci. **10**, 513 (2004).

[12] C. Bushdid, M. O. Magnasco, L. B. Vosshall, and A. Keller, *Humans Can Discriminate More than 1 Trillion Olfactory Stimuli*, Science **343**, 1370 (2014).

[13] R. C. Gerkin and J. B. Castro, *The Number of Olfactory Stimuli that Humans Can Discriminate Is Still Unknown*, eLife **4**, e08127 (2015).

[14] E. J. Mayhew, C. J. Arayata, R. C. Gerkin, B. K. Lee, J. M. Magill, L. L. Snyder, K. A. Little, C. W. Yu, and J. D. Mainland, *Drawing the Borders of Olfactory Space*, bioRxiv, 10.1101/2020.12.04.412254.

[15] K. R. Illig and L. B. Haberly, *Odor-Evoked Activity Is Spatially Distributed in Piriform Cortex*, J. Comp. Neurol. **457**, 361 (2003).

[16] A. Lansner, *Associative Memory Models: From the Cell-Assembly Theory to Biophysically Detailed Cortex Simulations*, Trends Neurosci. **32**, 178 (2009).

[17] D. D. Stettler and R. Axel, *Representations of Odor in the Piriform Cortex*, Neuron **63**, 854 (2009).

[18] B. Roland, T. Deneux, K. M. Franks, B. Bathellier, and A. Fleischmann, *Odor Identity Coding by Distributed Ensembles of Neurons in the Mouse Olfactory Cortex*, eLife **6**, e26337 (2017).

[19] A. J. Aqrabawi and J. C. Kim, *Olfactory Memory Representations Are Stored in the Anterior Olfactory Nucleus*, Nat. Commun. **11**, 1246 (2020).

[20] A. Mayer, V. Balasubramanian, T. Mora, and A. M. Walczak, *How a Well-Adapted Immune System Is Organized*, Proc. Natl. Acad. Sci. U.S.A. **112**, 5950 (2015).

[21] R. Shinnakasu, T. Inoue, K. Kometani, S. Moriyama, Y. Adachi, M. Nakayama, Y. Takahashi, H. Fukuyama, T. Okada, and T. Kurosaki, *Regulated Selection of Germinal-Center Cells into the Memory B Cell Compartment*, Nat. Rev. Immunol. **17,** 861 (2016).

[22] R. Shinnakasu and T. Kurosaki, *Regulation of Memory B and Plasma Cell Differentiation*, Current opinion in immunology **45,** 126 (2017).

[23] A. Mayer, V. Balasubramanian, A. M. Walczak, and T. Mora, *How a Well-Adapting Immune System Remembers*, Proc. Natl. Acad. Sci. U.S.A. **116,** 8815 (2019).

[24] O. H. Schnaack and A. Nourmohammad, *Optimal Evolutionary Decision-Making to Store Immune Memory*, eLife **10,** e61346 (2021).

[25] C. Viant, G. H. J. Weymar, A. Escolano, S. Chen, H. Hartweger, M. Cipolla, A. Gazumyan, and M. C. Nussenzweig, *Antibody Affinity Shapes the Choice between Memory and Germinal Center B Cell Fates*, Cell **183,** 1298 (2020).

[26] R. Polikar and C. Alippi, *Guest Editorial Learning in Nonstationary and Evolving Environments*, IEEE Trans. Neural Netw. Learn. Syst. **25,** 9 (2014).

[27] G. Ditzler, M. Roveri, C. Alippi, and R. Polikar, *Learning in Nonstationary Environments: A Survey*, IEEE Comput. Intell. Mag. **10,** 12 (2015).

[28] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT Press, Cambridge, MA, 2016) [http://www.deeplearningbook.org].

[29] P. Mehta, M. Bukov, C.-H. Wang, A. G. R. Day, C. Richardson, C. K. Fisher, and D. J. Schwab, *A High-Bias, Low-Variance Introduction to Machine Learning for Physicists*, Phys. Rep. **810,** 1 (2019).

[30] D. O. Hebb, *The Organization of Behavior: A Neuropsychological Theory* (Wiley, New York, 1949).

[31] J. J. Hopfield, *Neural Networks and Physical Systems with Emergent Collective Computational Abilities*, Proc. Natl. Acad. Sci. U.S.A. **79,** 2554 (1982).

[32] J. F. Fontanari and R. Meir, *Learning Noisy Patterns in a Hopfield Network*, Phys. Rev. A **40,** 2806 (1989).

[33] M. Mezard, J. P. Nadal, and G. Toulouse, *Solvable Models of Working Memories*, J. Phys. **47,** 1457 (1986).

[34] D. J. Amit, H. Gutfreund, and H. Sompolinsky, *Storing Infinite Numbers of Patterns in a Spin-Glass Model of Neural Networks*, Phys. Rev. Lett. **55,** 1530 (1985).

[35] R. McEliece, E. Posner, E. Rodemich, and S. Venkatesh, *The Capacity of the Hopfield Associative Memory*, IEEE Trans. Inf. Theory **33,** 461 (1987).

[36] A. Storkey, in *Proceedings of the Artificial Neural Networks ICANN'97*, edited by G. Goos, J. Hartmanis, J. van Leeuwen, W. Gerstner, A. Germond, M. Hasler, and J.-D. Nicoud (Springer, Berlin, 1997), Vol. 1327, pp. 451–456.

[37] P. Tolmachev and J. H. Manton, in *Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, 2020* (IEEE, New York, 2020), pp. 1–8.

[38] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevX.12.021063 for 10 additional figures (Figs. S1–S10) that support the results presented in the main text.

[39] J.-P. Bouchaud and M. Mezard, *Universality Classes for Extreme-Value Statistics*, J. Phys. A **30,** 7997 (1997).

[40] B. Derrida, *From Random Walks to Spin Glasses*, Physica (Amsterdam) **107D,** 186 (1997).

[41] P. A. Ortega and D. A. Braun, *Thermodynamics as a Theory of Decision-Making with Information-Processing Costs*, Proc. R. Soc. A **469,** 20120683 (2013).

[42] V. Balasubramanian, *Statistical Inference, Occam's Razor, and Statistical Mechanics on the Space of Probability Distributions*, Neural Comput. **9,** 349 (1997).

[43] C. H. LaMont and P. A. Wiggins, *Correspondence between Thermodynamics and Inference*, Phys. Rev. E **99,** 052140 (2019).

[44] M. Mokkonen and C. Lindstedt, *The Evolutionary Ecology of Deception: The Evolution of Deception*, Biol. Rev. Camb. Philos. Soc. **91,** 1020 (2016).

[45] J. Stökl, A. Strutz, A. Dafni, A. Svatos, J. Doubsky, M. Knaden, S. Sachse, B. S. Hansson, and M. C. Stensmyr, *A Deceptive Pollination System Targeting Drosophilids through Olfactory Mimicry of Yeast*, Curr. Biol. **20,** 1846 (2010).

[46] R. Homma, L. B. Cohen, E. K. Kosmidis, and S. L. Youngentob, *Perceptual Stability during Dramatic Changes in Olfactory Bulb Activation Maps and Dramatic Declines in Activation Amplitudes*, Eur. J. Neurosci. **29,** 1027 (2009).

[47] O. H. Schnaack, L. Peliti, and A. Nourmohammad, *Risk-Utility Tradeoff Shapes Memory Strategies for Evolving Patterns*, arXiv:2110.15008.

[48] G. Isacchini, A. M. Walczak, T. Mora, and A. Nourmohammad, *Deep Generative Selection Models of T and B Cell Receptor Repertoires with soNNia*, Proc. Natl. Acad. Sci. U.S.A. **118,** e2023141118 (2021).

[49] T. Mora, A. M. Walczak, W. Bialek, and C. G. Callan, *Maximum Entropy Models for Antibody Diversity*, Proc. Natl. Acad. Sci. U.S.A. **107,** 5405 (2010).

[50] J. Desponds, T. Mora, and A. M. Walczak, *Fluctuating Fitness Shapes the Clone-Size Distribution of Immune Repertoires*, Proc. Natl. Acad. Sci. U.S.A. **113,** 274 (2016).

[51] C. A. Janeway, P. Travers, M. Walport, and M. Shlomchik, *Immunobiology: The Immune System in Health and Disease*, 6th ed. (Garland Science, New York, 2005).

[52] G. D. Victora and M. C. Nussenzweig, *Germinal Centers*, Annu. Rev. Immunol. **30,** 429 (2012).

[53] M. J. Shlomchik, *Do Memory B Cells Form Secondary Germinal Centers? Yes and No*, Cold Spring Harbor Perspect. Biol. **10,** a029405 (2018).

[54] L. J. McHeyzer-Williams, C. Dufaud, and M. G. McHeyzer-Williams, *Do Memory B Cells Form Secondary Germinal Centers? Impact of Antibody Class and Quality of Memory T-Cell Help at Recall*, Cold Spring Harbor Perspect. Biol. **10,** a028878 (2018).

[55] H.-X. Liao *et al.*, *Co-Evolution of a Broadly Neutralizing HIV-1 Antibody and Founder Virus*, Nature (London) **496,** 469 (2013).

[56] A. Nourmohammad, J. Otwinowski, M. Łuksza, T. Mora, and A. M. Walczak, *Fierce Selection and Interference in B-Cell Repertoire Response to Chronic HIV-1*, Mol. Biol. Evol. **36,** 2184 (2019).

[57] F. Horns, C. Vollmers, C. L. Dekker, and S. R. Quake, *Signatures of Selection in the Human Antibody Repertoire:*

Selective Sweeps, Competing Subclones, and Neutral Drift, Proc. Natl. Acad. Sci. U.S.A. **116**, 1261 (2019).

[58] C. Vollmers, R. V. Sit, J. A. Weinstein, C. L. Dekker, and S. R. Quake, *Genetic Measurement of Memory B-Cell Recall Using Antibody Repertoire Sequencing*, Proc. Natl. Acad. Sci. U.S.A. **110**, 13463 (2013).

[59] N. Parga and M. Virasoro, *The Ultrametric Organization of Memories in a Neural Network*, J. Phys. France **47**, 1857 (1986).

[60] M. A. Virasoro, in *Disordered Systems and Biological Organization* (Springer, Berlin, 1986), pp. 197–204.

[61] H. Gutfreund, *Neural Networks with Hierarchically Correlated Patterns*, Phys. Rev. A **37**, 570 (1988).

[62] M. V. Tsodyks, *Hierarchical Associative Memory in Neural Networks with Low Activity Level*, Mod. Phys. Lett. B **04**, 259 (1990).

[63] https://github.com/StatPhysBio/Working_memory.

[64] D. J. Amit, H. Gutfreund, and H. Sompolinsky, *Spin-Glass Models of Neural Networks*, Phys. Rev. A **32**, 1007 (1985).

[65] J. F. Fontanari, *Generalization in a Hopfield Network*, J. Phys. France **51**, 2421 (1990).

[66] W. Krauth and M. Mezard, *Learning Algorithms with Optimal Stability in Neural Networks*, J. Phys. A **20**, L745 (1987).