Risk-utility tradeoff shapes memory strategies for evolving patterns

Oskar H Schnaack,[1,2] Luca Peliti,[3] and Armita Nourmohammad[1,2,4,*]

[1] *Max Planck Institute for Dynamics and Self-organization, Am Faßberg 17, 37077 Göttingen, Germany*
[2] *Department of Physics, University of Washington, 3910 15th Ave Northeast, Seattle, WA 98195, USA*
[3] *Santa Marinella Research Institute, 00058 Santa Marinella, Italy*
[4] *Fred Hutchinson Cancer Research Center, 1100 Fairview ave N, Seattle, WA 98109, USA*

(Dated: October 28, 2021)

Keeping a memory of evolving stimuli is ubiquitous in biology, an example of which is immune memory for evolving pathogens. However, learning and memory storage for dynamic patterns still pose challenges in machine learning. Here, we introduce an analytical energy-based framework to address this problem. By accounting for the tradeoff between utility in keeping a high-affinity memory and the risk in forgetting some of the diverse stimuli, we show that a moderate tolerance for risk enables a repertoire to robustly classify evolving patterns, without much fine-tuning. Our approach offers a general guideline for learning and memory storage in systems interacting with diverse and evolving stimuli.

I. INTRODUCTION

Biological systems, ranging from the brain to the immune system, store memory of molecular interactions to efficiently recognize and respond to stimuli. Memory encoding in biological networks has also inspired a growing host of algorithms for learning and memory storage in image and pattern recognition by artificial neural networks [1–3]. A critical step in these algorithms is to find regularities in data to associate related patterns with each other. As such, these learning algorithms often assume that the set of training data comes from a stationary distribution that represents the regularities necessary for pattern recognition in data well.

Memory recognition, however, is not limited to static patterns and can be desirable when classifying evolving stimuli that drive the system out of equilibrium. One such example is the adaptive immune system in which memory can effectively recognize evolved variants of previously encountered pathogens [4–8]. In a recent work, we have demonstrated that distributed learning strategies, which are desirable for pattern recognition in the stationary setup, can fail to reliably learn and classify dynamically evolving patterns [9]. Specifically, we showed that to follow evolving patterns, an energy-based Hopfield-like neural network [10] should use a higher learning rate, which in turn, can distort the energy landscape associated with the stored memory attractors, leading to pattern misclassification. To remedy this problem, we proposed compartmentalized networks as the optimal solutions to memory storage for evolving patterns [9].

Irrespective of the network structure, an increase in learning rate is necessary for a network to follow, recognize, and store effective memory of evolving patterns [9]. Increasing learning rate leads to a risky strategy, as the memory repertoire begins to reflect only the most recently encountered pattern while effectively destroying the memory of the prior encounters. Here, we present an analytical approach to explore how the tradeoff between utility and risk can determine learning strategies of a repertoire in keeping a memory of evolving patterns. We show that a moderate risk tolerance enables a repertoire to store an effective and robust memory. Our approach puts forward a guideline for optimal learning and memory storage for systems interacting with multiple evolving pattern classes without much fine-tuning.

II. MODEL

To probe memory strategies, we define a repertoire $\mathcal{M}$ that can store stimuli (patterns) ψ and later utilize them to recognize newly presented stimuli. We consider the space of possible binary patterns of length L, $\{\psi^\alpha\}$ with $\alpha \in 1, \ldots, 2^L$ enumerating all unique patterns, with entries $\psi_i^\alpha = \pm 1$, for $i \in 1, \ldots, L$. The memory repertoire $\mathcal{M}$ associates normalized weights m^α ($\sum m^\alpha = 1$) to all patterns it encounters. The non-zero weights reflect the relative importance of the stored memory from different stimuli, and patterns that are not encountered are associated with a zero weight. After an encounter with a given stimulus ψ^β at time t, all weights are updated according to a Hebbian learning rule $m^\alpha(t) = (1-\lambda)m^\alpha(t-1) + \delta_{\alpha,\beta}\lambda$, where λ is the learning rate [11]. Thus, the weight $m^\alpha(t+\tau)$ associated with pattern α, which was previously encountered at time t, decays in an approximate exponential form over time,

$$m^\alpha(t+\tau) = \lambda(1-\lambda)^\tau \approx \lambda e^{-\lambda\tau}. \tag{1}$$

To determine the memory performance for a given stimulus χ, we define the affinity $\mathcal{A}(\mathcal{M}, \chi)$ as the overlap between the patterns stored in the memory repertoire $\mathcal{M}$

* Correspondence should be addressed to Armita Nourmohammad: armita@uw.edu

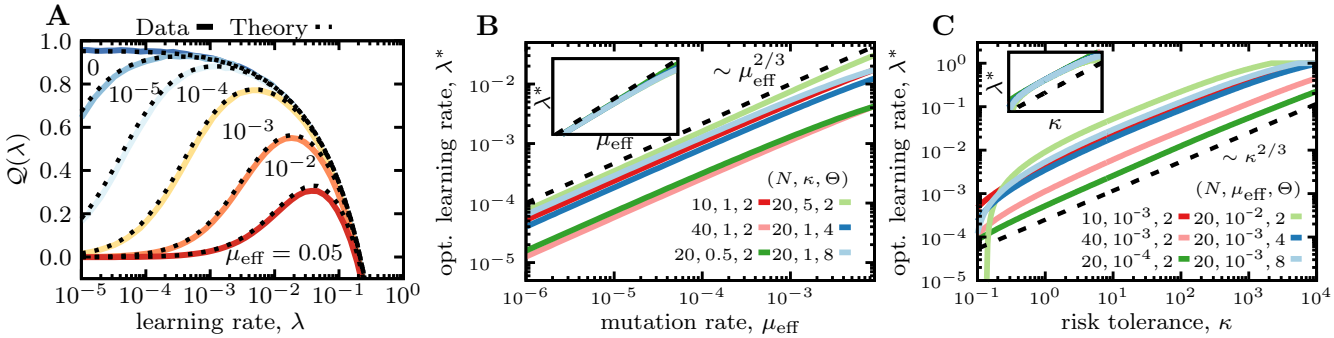

FIG. 1. **Optimal memory strategy for evolving patterns. (A)** The objective function $\mathcal{Q}(\lambda)$ (Eq. 3) is shown as a function of the learning rate, for $\kappa = 1$ and $\Theta = 2$. Analytical approximations (dashed lines) are compared to numerical estimates (full lines), for different mutation rates $\mu_{\rm eff}$ (colors). **(B, C)** The optimal learning rate (Eq. 5) is shown as a function of the mutation rate $\mu_{\rm eff}$ (B) and the risk tolerance κ (C), for different sets of parameters indicated in the legend. Insets show collapsed plots after rescaling according to Eq. 5.

and the pattern χ,

$$\mathcal{A}(\mathcal{M}, \chi) = \mathcal{A}_0 \sum_{\psi^\alpha \in \mathcal{M}} m^\alpha |\langle \psi^\alpha | \chi \rangle|^\Theta - \mathcal{A}_{\rm rand}, \quad (2)$$

Here, we used a short-hand notation to denote the normalized pattern vector by $|\psi\rangle \equiv \psi/\sqrt{L}$, its transpose by $\langle\psi|$, and a normalized scalar product $\langle\psi|\chi\rangle \equiv (\sum_i \psi_i \chi_i)/L$. We use the shape parameter Θ to modulate the dependency of the affinity function on pattern overlap. The scaling parameter $\mathcal{A}_0$ is model dependent and sets the unit of affinity in a given system, yet its precise value does not impact our analysis. The offset $\mathcal{A}_{\rm rand}$ is chosen such that the expected affinity of random patterns remains zero. The choice of affinity as a measure of performance is inspired by memory retrieval in biological systems, where recognition is mediated through biophysical interactions.

We characterize the response of a memory repertoire to N independently evolving pattern classes Ψ^c (with $c \in 1, \ldots, N$). Each class Ψ^c denotes a set $\{\psi\}_c$ of patterns generated over time through evolution. Patterns within each class evolve (mutate) by random spin-flips of rate μ. At each time step t, a pattern from a randomly chose class is presented to the memory repertoire, resulting in an effective observed mutation rate of $\mu_{\rm eff} = N\mu$ per expected encounter with the same pattern class. To simplify, we use $\Psi^c(t)$ to denote the evolved pattern from class c that is generated and presented to the repertoire at time t. Since pattern classes are orthogonal to each other (up to finite-size effects), the expected overlap between patterns presented at different times follows $\left\langle \langle \Psi^c(t) | \Psi^{c'}(t+\tau) \rangle \right\rangle = \delta_{c,c'} \rho^\tau + \mathcal{O}(1/\sqrt{L})$, where $\rho = (1 - 2\mu)$ measures the similarity between evolved patterns.

Interestingly, for $\Theta = 2$ this model is equivalent to the energy function of the classical Hopfield network with Hebbian learning rule [9, 10, 12, 13] (Appendix B). This correspondence enables us to simulate the memory repertoire efficiently and to test analytic predictions with numerical experiments; see Appendix C for numerical method.

III. RESULTS

A. Optimal learning for evolving patterns with risk-return tradeoff

We seek to find an optimal strategy to set the learning rate such that the stored memory in a repertoire can be reliably retrieved for patterns evolving at a specified rate μ. Learning and updates of memory repertoires over time impact both the expected affinity and the variance of the affinity across patterns. A high learning rate can sustain a high affinity in a repertoire for the most recently presented patterns. However, keeping up with the latest trend can be risky as the variance of the affinity across patterns can increase, with older patterns suffering most from this tradeoff. To account for this effect, we optimize an objective function $\mathcal{Q}(\lambda; \mu)$ that balances the risk-utility tradeoff by maximizing the mean (utility) $\langle \mathcal{A} \rangle$ and minimizing the standard deviation (risk) $\sigma_{\mathcal{A}}$ of the affinity across patterns,

$$\mathcal{Q}(\lambda; \mu) = \langle \mathcal{A} \rangle - \frac{1}{\kappa} \sigma_{\mathcal{A}}. \quad (3)$$

Here, κ measures the risk tolerance of the repertoire. Such a risk-utility analysis was initially introduced in economics [14, 15], but has since been used to characterize tradeoffs in biological and evolutionary processes [16–22].

We can analytically evaluate both the mean and the variance of the affinity by using explicit expansion of the encounter history or by evaluating the cumulant-generating function for the distribution of affinities, valid in the large-N limit (i.e, many patterns); see Appendix A and Fig. S1. The expression for the nth cumulant of the

affinity follows,

$$c_n = a_0^n \frac{\lambda^n \rho^{\Theta n}}{N\left(1-(1-\lambda)^n \rho^{\Theta n}\right)} + \mathcal{O}\left(N^{-2}\right) \tag{4}$$

where $a_0 = \mathcal{A}_0 - \mathcal{A}_{\text{rand}}$, and the mean and the variance are given by $n = 1$ and $n = 2$, respectively. For static patterns ($\mu = 0 \leftrightarrow \rho = 1$), the mean affinity becomes independent of the learning rate and reaches the naïve expectation a_0/N. For evolving patterns, a repertoire with a maximal learning rate $\lambda = 1$ can achieve the highest mean affinity $a_0\rho^{\Theta}/N$ by storing a memory of the most recent pattern with the maximal affinity a_0, while treating the other patterns as random.

For a broad range of evolutionary rates, the optimum of the objective function $\mathcal{Q}(\lambda;\mu)$ is achieved for intermediate values of learning rate λ, where the mean and variance of affinity across patterns are balanced; see Fig. 1A for analytical and numerical results. The optimal learning rate $\lambda^*(\mu) = \max_\lambda \mathcal{Q}(\lambda;\mu)$ that maximizes the objective function scales with different model parameters as,

$$\lambda^* = \frac{2}{N}\left(2\kappa\,\Theta\,\mu_{\text{eff}}\right)^{2/3} + \mathcal{O}\left(\frac{\mu_{\text{eff}}}{N}\right). \tag{5}$$

The analytical scaling relation in Eq. 5 is in excellent agreement with numerical simulations (Fig. 1B,C). As the evolutionary rate of patterns μ_{eff} increases, the optimal learning rate grows so that the repertoire closely follows the patterns' evolution (Fig. 1B). As the shape parameter Θ increases, the affinity function becomes more peaked around the recently stored patterns (Eq. 2), resulting in an increase in the optimal learning rate to keep the memory focused on the more recent (less evolved) encounters. The learning rate scales inversely with the number of patterns N for the repertoire to evenly distribute the resources (allocate memory). Repertoires with higher risk tolerance κ use larger learning rates (Fig. 1C) and store a risky but a high-affinity memory against recent encounters. As repertoire becomes more risk-avert (small κ), it stores a more equitable memory across patterns but at a loss for affinity. In the limit of no risk tolerance ($\kappa \to 0$), the repertoire stops learning ($\lambda = 0$) and adopts a risk-free but impractical strategy where the memory has zero affinity for all patterns. This tradeoff can be depicted by a Pareto front in the affinity-risk space, along which one cannot increase the mean affinity without increasing the risk or vice versa. Fig. 2 shows these Pareto fronts parametrized by scaled mean affinity and risk, for different mutation rates (colors) and by varying the risk tolerance κ along each line. The combinations of risk and affinity values that lie below the Pareto front are inaccessible, and those that lie above are sub-optimal solutions for a memory repertoire.

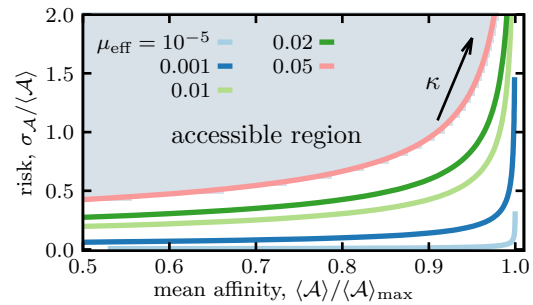

FIG. 2. **Pareto front for risk-utility tradeoff of memory for evolving patterns.** Optimized objective function $\mathcal{Q}(\lambda^*)$ (Eq. 3) describes a Pareto front in the scaled affinity-risk plane by varying the risk tolerance κ along each line, for different mutation rates (colors). To present the Pareto front in a dimensionless space, the mean affinity is scaled by its maximal value $\langle\mathcal{A}\rangle_{\max} = \frac{a_0}{N}$, and the standard deviation as risk is scaled by the mean affinity. Only the combinations of risk and affinity values that lie above the Pareto front (gray shading) are accessible to a memory repertoire. Other parameters: $N = 200$ and $\Theta = 2$. Results for other shape parameters are shown in Fig. S2.

B. Discrimination of random and stored patterns

One key goal of memory usage is to recognize and accurately classify presented patterns with a memory stored from prior encounters. A misclassification could have dire consequences. For example, in the case of immune system, if memory response is not triggered by a secondary infection (i.e., *false negative*), the host would pay a cost by enduring sickness and having to mount a novel response and re-store a new memory. *False positive* responses are also costly, as they can be associated with autoimmunity if mounted against self-antigens [6], or they can interfere with novel responses without preventing the disease, e.g., in the case of original antigenic sin against viruses like influenza [23].

Memory strategies optimized to operate under different risk tolerance κ can yield varying levels of pattern misclassification. We characterize the discrimination accuracy of a repertoire by quantifying the rate by which it recognizes evolved patterns associated with previously stored memory (true positive), or randomly generated patterns without any prior encounter history (false positive). To do so, we need to characterize the distribution of affinities for patterns with prior encounter history, and for novel (random) patterns.

The recognition affinity of a memory repertoire for recurring patterns depends on the history of pattern encounters and on the learning rate λ. For example, in the case where patterns from a given class Ψ^c are presented at time points $[t_1, \ldots, t_n]$, the affinity $\mathcal{A}(\Psi^c(T))$ for an evolved pattern from the same class shown at time

$T > t_n$ can be expressed as

$$\mathcal{A}(\Psi^c(T)) = a_0 \sum_{i=1}^{n} \lambda e^{-\lambda(T-t_i-1)} \rho^{\Theta(T-t_i)} \tag{6}$$

Here, the factor $e^{-\lambda(T-t_i-1)}$ accounts for the exponential decay for the affinity of a memory stored at time t_i from its maximum level λ, due to updates in the repertoire (Eq. 1). $\rho^{\Theta(T-t_i)}$ accounts for the decay in the overlap between the memory stored at time t_i and the presented patterns in the future, due to evolution.

Patterns are presented to the repertoire in random order, and the time $\tau_i = t_i - t_{i-1}$ between consecutive encounters with the same class is exponentially distributed with a mean of N steps (i.e., number of classes), $p(\tau) \approx e^{-N\tau}/N$. The distribution of affinities $P_\Psi(\mathcal{A})$ for patterns of a given class can be derived by convolving the affinity function in Eq. 6 with the exponential waiting time distribution for pattern history. Although the exact form of this affinity distribution is difficult to evaluate, we expect it to be from an exponential family. Indeed, the Gamma distribution is a good approximation to the distribution of the affinities (Fig. S3). We quantify the accuracy of this approximation by the Kullback-Leibler distance $D_{\mathrm{KL}}(P_\Psi(\mathcal{A})\|\Gamma_\mathcal{A})$ between the true affinity distribution $P_\Psi(\mathcal{A})$ and a Gamma distribution $\Gamma_\mathcal{A}$ with the same mean and variance. By using an Edgeworth expansion [24, 25] with the cumulants of the affinity distribution in Eq. 4, we can show that in the limit of small learning and mutation rates the Kullback-Leibler distance between the two distributions is small, with $D_{\mathrm{KL}}(P_\Psi(\mathcal{A})\|\Gamma(\mathcal{A})) = (25/27)\,(2\Theta\mu_{\mathrm{eff}}/\kappa)^{2/3} + \mathcal{O}(\Theta\mu_{\mathrm{eff}}) \ll 1$ (Appendix A 4). This result can also be intuitively understood, since the Gamma distribution arises in processes for which the waiting times between events are relevant.

The distribution of affinities for random patterns (i.e., not belonging to any of the presented classes) $P_0(\mathcal{A})$ can be similarly characterized. The law of large numbers suggest that the overlap between unrelated patterns should be normally distributed with mean zero and variance $1/(4L)$. The overlap between the memory repertoire and a random pattern is the sum of the normally distributed random overlaps weighed by the weights $\{m^\alpha\}$, associated with the stored patterns (Appendix A 5 and Fig. S4). Thus, the distribution of affinities for random patterns $P_0(\mathcal{A})$ is well-approximated by a Gamma distribution with mean $\propto L^{-\Theta/2}$ and variance $\propto L^{-\Theta}$. This distribution should be contrasted to that of the affinities for patterns with prior encounter histories $P_\Psi(\mathcal{A})$, the statistics of which primarily depends on the number of patterns and the learning rate, with mean $\propto 1/N$, and variance $\propto \lambda/N$.

The sensitivity (fraction of true positives to all patterns associated with memory) and specificity (fraction of false negatives to all random patterns) of a repertoire in classifying patterns depends on an affinity cut-off for distinguishing between familiar and random patterns. A receiver operating characteristic (ROC) curve shows the relationship between sensitivity and specificity of a classification in a repertoire for every possible affinity cut-off. The area under the ROC curve (AUROC), which measures the discriminative ability of the repertoire, depends on the evolutionary rate μ_{eff} of patterns it encounters, its risk tolerance κ, and the parameter Θ that determines the shape of the affinity function (Eq. 4). These parameters also determine the optimal learning rate λ^* (Eq. 5), which sets the memory strategy of a repertoire in a given evolutionary setup.

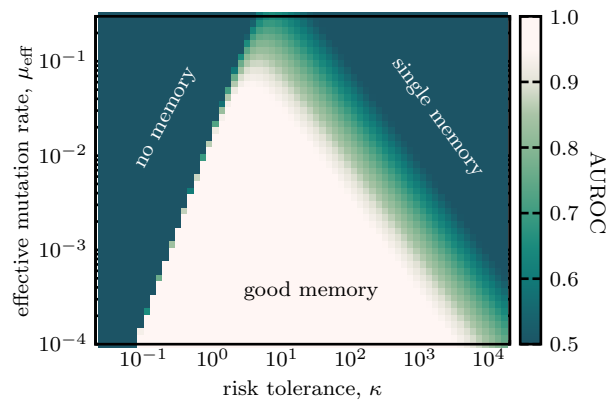

FIG. 3. **Three phases of memory discrimination.** The phase diagram shows the discrimination ability of repertoires (AUROC) between familiar patterns with prior encounter history and random patterns, based on their respective affinities. AUROC is estimated using numerical approximations to the affinity distributions $P_\Psi(\mathcal{A})$ and $P_0(\mathcal{A})$ for familiar and random patterns (Appendix A). Each point in the phase diagram shows AUROC for a repertoire optimized with a given risk tolerance κ for patterns with a specified mutation rate μ_{eff}. Other parameters: $L = 200$, $N = 40$, and $\Theta = 2$. Results for other shape parameters Θ are shown in Fig. S6. Full simulation results are shown in Fig. S7.

The phase diagram for the discrimination ability of a repertoire defines three distinct regions determined by a combination of risk tolerance κ and the evolutionary rate of patterns μ_{eff}, for a specified shape parameter Θ (Fig. 3): (i) A triangular region at the center of the phase diagram with AUROC $\simeq 1$ indicates the range of parameters for which the repertoire can efficiently discriminate between familiar and novel patterns. Within this region, as the evolutionary rate of patterns increases, the range for risk tolerance that enables a repertoire to efficiently discriminate between familiar and random patterns narrows, and the repertoire has to fine-tune its learning rate to match the faster evolution of the patterns (Fig. S5). (ii) When risk tolerance κ is large, the repertoire updates rapidly and only keeps a memory of the most recent encounters, resulting in an inefficient discrimination between familiar and novel patterns with AUROC $\simeq 0.5$. (iii) For risk avert strategies (small κ), the memory effectively shuts down and the repertoire cannot anymore

discriminate between random and familiar patterns. It should be noted that a change in the affinity shape parameter Θ can shift the exact boundaries between these phases, but it does not impact the overall structure of the phase diagram (Fig. S6).

Taken together, a moderate risk tolerance ($\kappa\Theta = \mathcal{O}(1)$) enables a repertoire to store an effective and a robust memory and to operate without fine tuning. We further confirm these results with simulations for $\Theta = 2$ in Fig. S7.

IV. DISCUSSION

Devising a learning strategy to store functional memory for evolving stimuli is an open problem with potential applications in many fields. Inspired by immune memory that reliably recognizes evolving threats, we present an analytical energy-based memory model against evolving patterns that captures the risk-utility tradeoff in memory repertoires.

We found that risk-tolerant repertoires adopt faster learning rates and keep a memory of the recent patterns with high affinity, at the risk of disregarding older patterns. On the other hand, risk-avert repertoires effectively shut down their learning to minimize the variance in their recognition affinity for different patterns, at the cost of having a low affinity for all patterns. A moderate risk tolerance enables a repertoire to achieve a desirable balance between risk and affinity (utility) to robustly classify evolving patterns and associate them with prior memory. This risk-utility tradeoff defines Pareto front for accessible memory strategies by a repertoire (Fig. 2).

One key role of memory storage is to reliably discriminate between familiar and novel stimuli. The discrimination ability of an optimal memory repertoire depends on its risk tolerance and the evolutionary rate of the patterns that it encounters. Interestingly, as the evolutionary rate of the presented patterns increases, the range of risk tolerance that allows a repertoire to retain a functional memory narrows and approaches a regime where risk and utility are equally valued. The classification criteria in our analysis are solely based on the affinity of the memory repertoire to a presented pattern. This stands in contrast to energy-based Hopfield-like networks that achieve recognition by retrieving associative memory stored in the networks' energy minima [10].

We have used the fluctuations (standard deviation) of the memory's affinity over the ensemble of presented patterns to measure the risk of misclassification by the stored storage. While variance might be more suitable in other settings [26, 27], using standard deviation as a measure of risk keeps the risk tolerance κ dimensionless and comparable across different systems. Nonetheless, we expect the Pareto front's overall structure for the risk-utility tradeoff and the phase diagram for discrimination ability of the repertoires to remain qualitatively intact, irrespective of the exact choices made for the risk function.

Although our model is inspired by immune memory repertoires, the introduced approach is general enough that can be applied to other models of memory. In particular, incorporating risk and utility in training of artificial neural networks can guide these machine learning efforts to store efficient of memory for evolving signals [28].

ACKNOWLEDGEMENT

This work has been supported by the NSF CAREER award (grant No: 2045054), DFG grant (SFB1310) for Predictability in Evolution and the MPRG funding through the Max Planck Society. O.H.S also acknowledges funding from Georg-August University School of Science (GAUSS) and the Fulbright foundation.

Appendix A: Statistics of the affinity distribution

1. The average affinity of a repertoire for presented patterns

In the main text (Eq. 6) we express the affinity of a pattern class at time T in terms of the prior encounters with the patterns of the same class at times $[t_1, ..., t_n]$ with $t_{n-1} < t_n < T$. To compute the statistics (mean and variance) of a repertoire's affinity, we first express the affinity function in terms of the times $\hat{t}_i = T - t_{n-i+1}$, passed since the ith encounter, when counting backward in time. This allows us to write the affinity as

$$\mathcal{A}(\Psi^c(T)) = a_0 \lambda \sum_{i=1}^{n} e^{-\lambda \hat{t}_i - 1} \rho^{2\hat{t}_i}. \tag{S1}$$

Similar to forward times t_i, the reverse times $\hat{t}_i$ are separated by exponentially distributed independent waiting times $\hat{\tau}_i = \hat{t}_i - \hat{t}_{i-1}$,

$$p(\hat{\tau}) = \frac{(N-1)^{\hat{\tau}-1}}{N^{\hat{\tau}}} \approx \frac{1}{N} e^{-N\tau}. \tag{S2}$$

Given the relationship $\sum_{j=1}^{i} \hat{\tau}_j = \hat{t}_i$, we can express the affinity of the patterns in Eq. S1 in terms of the statistically independent $\hat{\tau}_i$ as,

$$\mathcal{A}(\Psi) = a_0 \lambda \sum_{g=1}^{n} (1-\lambda)^{\sum_{g'=1}^{g} \hat{\tau}_{g'} - 1} \rho^{\Theta \sum_{g'=1}^{g} \hat{\tau}_{g'}}. \tag{S3}$$

Here, we can interpret each term in the sum as the contribution of a memory from the corresponding prior generation to the affinity at the current time. These contributions decay due to the updates (learning) in the repertoire with rate $(1-\lambda)$ and due to evolution of the patterns with rate ρ. The expected affinity contribution $\langle \mathcal{A}_g \rangle$ from the gth generation follows,

$$\begin{aligned}
\langle \mathcal{A}_g \rangle &= a_0 \lambda \sum_{\hat{\tau}_1} \cdots \sum_{\hat{\tau}_g} \left(\prod_{i=1}^{g} p(\hat{\tau}_i) \right) (1-\lambda)^{\sum_{i=1}^{g} \hat{\tau}_i - 1} \rho^{\Theta \sum_{i=1}^{g} \hat{\tau}_i} \\
&= a_0 \lambda (1-\lambda)^{-1} \prod_{i=1}^{g} \left(\sum_{\hat{\tau}_i} p(\hat{\tau}_i) (1-\lambda)^{\hat{\tau}_i} \rho^{\Theta \hat{\tau}_i} \right) \\
&= a_0 \lambda (1-\lambda)^{-1} \left(\frac{(1-\lambda)\rho^{\Theta}}{N + (1-\lambda)\rho^{\Theta} - N(1-\lambda)\rho^{\Theta}} \right)^{g}
\end{aligned} \tag{S4}$$

where we used the fact that the time windows $\hat{\tau}_i$ are independent from each other. The expected affinity follows from adding up the contributions from all generations,

$$\langle \mathcal{A} \rangle = \sum_{g=1}^{\infty} \langle \mathcal{A}_g \rangle = a_0 \frac{\lambda \rho^{\Theta}}{N\left(1 - (1-\lambda)\rho^{\Theta}\right)}. \tag{S5}$$

As mentioned in the main text, this result immediately shows that the system reaches the maximal mean affinity of $a_0/N\rho^{\Theta}$ for the maximal learning rate $\lambda = 1$. Moreover, the mean affinity becomes independent of the learning rate when the patterns are static ($\rho = 1$).

2. The variance of repertoire's affinity across presented patterns

To calculate the variance of the affinity (Eq. S4 in the main text) we need to account for the covariance $\mathbf{cov}(\mathcal{A}_{g_1}, \mathcal{A}_{g_2})$ between the contributions from different generations. Because the covariance is symmetric, we can write the variance of the affinity as

$$\mathbf{var}\left(\mathcal{A}\right) = \sum_{g=1}^{\infty} \mathbf{var}(\mathcal{A}_g) + 2 \sum_{g_1=1}^{\infty} \sum_{g_2=g_1+1}^{\infty} \mathbf{cov}\left(\mathcal{A}_{g_1}, \mathcal{A}_{g_2}\right). \tag{S6}$$

First we calculate the variance of the individual terms $\mathcal{A}_g$. Following the notation introduced in Eq. S4, we find

$$\begin{aligned}
\mathbf{var}(\mathcal{A}_g) =& \mathbf{var}\left(a_0\lambda(1-\lambda)^{-1}\prod_{i=1}^{g}\left((1-\lambda)^{\hat{\tau}_i}\rho^{\Theta\hat{\tau}_i}\right)\right) \\
=& \mathbf{var}\left(a_0\lambda(1-\lambda)^{-1}\prod_{i=1}^{g-1}\left((1-\lambda)^{\hat{\tau}_i}\rho^{\Theta\hat{\tau}_i}\right)\ \left((1-\lambda)^{\hat{\tau}_g}\rho^{\Theta\hat{\tau}_g}\right)\right) \qquad \text{(S7)} \\
=& \left\langle(1-\lambda)^{\hat{\tau}_g}\rho^{\Theta\hat{\tau}_g}\right\rangle^2 \mathbf{var}\left(a_0\lambda(1-\lambda)^{-1}\prod_{i=1}^{g-1}\left((1-\lambda)^{\hat{\tau}_i}\rho^{\Theta\hat{\tau}_i}\right)\right) \\
&+\left\langle a_0\lambda(1-\lambda)^{-1}\prod_{i=1}^{g-1}\left((1-\lambda)^{\hat{\tau}_i}\rho^{\Theta\hat{\tau}_i}\right)\right\rangle^2 \mathbf{var}\left((1-\lambda)^{\hat{\tau}_g}\rho^{\Theta\hat{\tau}_g}\right) \qquad \text{(S8)} \\
=& \left(a_0\lambda(1-\lambda)^{-1}\right)^{-2}\left(\langle\mathcal{A}_1\rangle^2\,\mathbf{var}\left(\mathcal{A}_{g-1}\right)+\langle\mathcal{A}_{g-1}\rangle^2\,\mathbf{var}\left(\mathcal{A}_1\right)\right) \qquad \text{(S9)} \\
=& g\left(a_0\lambda(1-\lambda)^{-1}\right)^{-(2g-2)}\langle\mathcal{A}_1\rangle^{2g-2}\,\mathbf{var}\left(\mathcal{A}_1\right). \qquad \text{(S10)}
\end{aligned}$$

Because the time intervals $\hat{\tau}_i$ are independent from each other, we could use error propagation to get from Eq. S7 to Eq. S8. Moreover, since all time time intervals follow the same distribution (Eq. S2), we could insert the statistics of $\mathcal{A}_g$ with the correct normalization $\left(a_0\lambda(1-\lambda)^{-1}\right)^{-2}$ to arrive at Eq. S9. Finally, we performed $g-1$ further iterations to get a result that only depends on the statistics of the first contribution $\mathcal{A}_1$ in Eq. S10. Thus, it is sufficient to calculate the variance of only one generation,

$$\mathbf{var}(\mathcal{A}_1) = a_0^2\rho^{2\Theta}\left(\frac{1}{N-(1-\lambda)^2(N-1)\rho^{2\Theta}}-\frac{1}{\left(N-(1-\lambda)(N-1)\rho^{\Theta}\right)^2}\right). \qquad \text{(S11)}$$

To evaluate the variance of the affinity (Eq. S6), we still need to calculate the covariance between the contributions from different generations. Similar to Eq. S7, we use,

$$\mathcal{A}_g = \mathcal{A}_{g-i}\cdot\prod_{j=g-i}^{g}\left((1-\lambda)^{\hat{\tau}_j}\rho^{\Theta\hat{\tau}_j}\right). \qquad \text{(S12)}$$

which entails the following relationship for the covariance between $\mathcal{A}_g$ and $\mathcal{A}_{g-i}$,

$$\begin{aligned}
\mathbf{cov}\left(\mathcal{A}_g,\mathcal{A}_{g-i}\right) &= \mathbf{cov}\left(\mathcal{A}_{g-i}\cdot\prod_{j=g-i}^{g}\left((1-\lambda)^{\hat{\tau}_j}\rho^{\Theta\hat{\tau}_j}\right),\mathcal{A}_{g-i}\right) = \mathbf{var}(\mathcal{A}_{g-i})\left\langle\prod_{j=g-i}^{g}\left((1-\lambda)^{\hat{\tau}_j}\rho^{\Theta\hat{\tau}_j}\right).\right\rangle \\
&= \left(a_0\lambda(1-\lambda)^{-i}\right)^{-1}\mathbf{var}(\mathcal{A}_{g-i})\langle\mathcal{A}_1\rangle^i. \qquad \text{(S13)}
\end{aligned}$$

Here, we again used the fact that all time intervals are independent and that they follow the same distribution in Eq. S2. Using the expressions for variance of individual terms (Eq. S10) and the covariance between them (Eq. S13), we can characterize the variance of the affinity (Eq. S6) as,

$$\mathbf{var}(\mathcal{A}) = a_0^2\left[\frac{\lambda(N-1)\rho^{2\Theta}\left(N-(1-\lambda)(N-1)\rho^{\Theta}\right)^2\left(\lambda N-(1-\lambda)\rho^{\Theta}(\lambda(N+2)-2)\right)}{N^2\left(1-(1-\lambda)\rho^{\Theta}\right)\left(N-(1-\lambda)(N-2)\rho^{\Theta}\right)^2\left(N-(1-\lambda)^2(N-1)\rho^{2\Theta}\right)}\right]. \qquad \text{(S14)}$$

Note that the expression in Eq. S14 is accurate up to terms of order $a_0^2\mathcal{O}(N^{-1}L^{-\Theta})$, which arise from the (negligible) overlap between patterns of difference classes due to the finite size of the patterns; see Appendix A 5.

3. Cumulant-generating function for the affinity distribution

To characterize the distribution of the affinities, we can rely on the corresponding cumulant-generating function. The cumulant-generating function $R_X(q)$ of a random variable X is defined as the logarithm of the moment-generating function

$$R_X(q) = \log\left\langle e^{qX}\right\rangle. \qquad \text{(S15)}$$

If we define a new random variable $Y = \sum_i X_i$ as the sum of independent random variables, its cumulant-generating function is given by $\mathcal{R}_Y(q) = \sum_i R_{X_i}(q)$. To use this relation for the affinity function, we need to rewrite the affinity of a pattern class (Eq. S1) as a sum of independent random variables. In Eq. S1 we calculate the affinity as the sum over the contributions of past encounters. In that view, the times of the past encounters $\hat{t}_i$ are random variables while the contributions to the affinity at those times are deterministic.

We now change the point of view and write the affinity function as sum of contributions from all time, instead of only the past encounters (Eq. S1),

$$\mathcal{A}(\Psi^c) = \sum_{\hat{t}=1} \mathcal{A}^c(\Psi^{c'}(\hat{t})) \tag{S16}$$

Here, $\hat{t}$ is the reverse time with the origin at the final (current) time point, as defined in Eq. S1. In this case, the encounter of a network with pattern $\Psi^{c'}(\hat{t})$ from class c' at (reverse) time $\hat{t}$ is accounted for by the contribution $\mathcal{A}^c(\Psi^{c'}(\hat{t}))$ to the network's affinity against the pattern Ψ^c. We then find

$$\mathcal{A}^c(\Psi^{c'}(\hat{t})) = \begin{cases} 0 & \text{if } c \neq c' \\ a_0 \lambda (1-\lambda)^{\hat{t}-1} \rho^{\Theta\,\hat{t}} =: E(\hat{t}) & \text{if } c = c' \end{cases} \tag{S17}$$

which implies that a repertoire's affinity against a pattern from a given class is only determined by the prior history of the repertoire's encounters with patterns of the same class, and the memory of these prior encounters decay over time according to Eq. S1.

The repertoire encounters a pattern of a specific class at a given time point with probably $1/N$. As a result, the distribution of affinity contributions from a given time point $\mathcal{A}^c(\Psi^{c'}(\hat{t}))$ can be expressed as,

$$P[\mathcal{A}^c(\Psi^{c'}(\hat{t}))] = \left(1 - \frac{1}{N}\right) \delta\left(\mathcal{A}^c(\Psi^{c'}(\hat{t})) - 0\right) + \frac{1}{N} \delta\left(\mathcal{A}^c(\Psi^{c'}(\hat{t})) - E(\hat{t})\right), \tag{S18}$$

where $E(\hat{t})$ is defined in Eq. S17. As a result, the expectation value $\langle \mathcal{A}(\Psi^c(T)) \rangle$ for the affinity of the repertoire against a pattern from class c (Eq. S16) can be evaluated as,

$$\langle \mathcal{A}(\Psi^c) \rangle = \sum_{\hat{t}=1} \int \mathrm{d}\mathcal{A}_{\hat{t}}^{c,c'} \mathcal{A}_{\hat{t}}^{c,c'} P[\mathcal{A}_{\hat{t}}^{c,c'}]. \tag{S19}$$

where we used the shorthand notation $\mathcal{A}_{\hat{t}}^{c,c'} \equiv \mathcal{A}^c(\Psi^{c'}(\hat{t}))$. Fig. S3 shows an agreement between simulations and the expected affinities estimated with this procedure.

We can now express the cumulant-generating function of the affinity distribution $P[\mathcal{A}^c(\Psi^{c'}(\hat{t}))]$ as the sum of the cumulant-generating functions of the independent terms $\mathcal{A}^c(\Psi^{c'}(\hat{t})$. With the definition in Eq. S15 we can evaluate the cumulant-generating functions $R_{\hat{t}}(q, \hat{t})$ for the contributions of each time point $\mathcal{A}^c(\Psi^{c'}(\hat{t})$ to the affinity function,

$$\begin{aligned} R_{\hat{t}}(q, \hat{t}) = \log \left\langle e^{q \mathcal{A}_{\hat{t}}(\hat{t})} \right\rangle &= \log \left((1 - \frac{1}{N}) + \frac{1}{N} e^{qE(\hat{t})} \right) \\ &\approx -\frac{1}{N} + \frac{1}{N} e^{qE(\hat{t})} + \mathcal{O}\left(N^{-2}\right). \end{aligned} \tag{S20}$$

The cumulant-generating function $\mathcal{R}(q)$ of the affinity distribution $P[\mathcal{A}]$ can be expressed as the sum of the cumulant-generating functions of the independent contributions from each time point (Eq. S20), which entail,

$$\begin{aligned} \mathcal{R}(q) = \sum_{\hat{t}=1} R(q, \hat{t}) &= \sum_{\hat{t}=1} \left[-\frac{1}{N} + \frac{1}{N} e^{qE(\hat{t})} \right] + \mathcal{O}\left(N^{-2}\right) \\ &= \sum_{\hat{t}=1} \left[-\frac{1}{N} + \frac{1}{N} \sum_{n=0} \frac{q^n E(\hat{t})^n}{n!} \right] + \mathcal{O}\left(N^{-2}\right) \\ &= \frac{1}{N} \sum_{n=1} \frac{q^n}{n!} a_0^n \sum_{\hat{t}=0} \lambda^n \rho^{\Theta n} \left((1-\lambda)\rho^{\Theta}\right)^{n\hat{t}} + \mathcal{O}\left(N^{-2}\right) \\ &= \frac{1}{N} \sum_{n=1} \frac{q^n}{n!} \times \frac{a_0^n \lambda^n \rho^{n\Theta}}{1 - (1-\lambda)^n \rho^{\Theta n}} + \mathcal{O}\left(N^{-2}\right). \end{aligned} \tag{S21}$$

where we have substituted the exponential function with an infinite sum, used the expression in Eq. S17 for $E(\hat{t})$, and performed the resulting geometric sum.

We can now evaluate the nth cumulant c_n of the affinity function as,

$$c_n = \left.\frac{\mathrm{d}^n \mathcal{R}(q)}{\mathrm{d}q^n}\right|_{q=0} = a_0^n \frac{\lambda^n \rho^{n\Theta}}{N\left(1-(1-\lambda)^n \rho^{\Theta n}\right)} + \mathcal{O}\left(N^{-2}\right). \tag{S22}$$

While the first cumulant is equal to the mean affinity in Eq. S5, obtained from the direct calculation of the moments, the second cumulant differs from the variance in Eq. S14. These differences arise due to the expansion of the logarithm in Eq. S20, which is only valid for large N. However, both results describe well the behavior of the variance in the regime that we are interested in (Fig. S1), and therefore, it is warranted to use the simplified form in Eq. 2 for our analysis.

To characterize the scaling relation between the cumulants c_n and the model parameters, we can expand Eq. S22 for the case where the learning rate is close to its optimal value $\lambda^* \sim \mu_{\text{eff}}^{2/3}$ (Eq. 5), which entails,

$$c_n = a_0^n \frac{\lambda^{n-1}}{nN} + \mathcal{O}\left(\frac{\lambda^{n-2}\mu_{\text{eff}}}{nN^2}\right), \tag{S23}$$

and thus, to the leading order, the mean affinity scales as $c_1 \sim a_0/N$ and the variance scales as $c_2 \sim a_0^2\lambda/(2N)$.

4. Approximating the affinity distribution with a Gamma distribution

As mentioned in the main text, the interpretation of the affinity as a sum of samples from an exponential function separated by exponentially distributed times motivates us to model the affinity distribution as a Γ-distribution. To corroborate this choice, we evaluate the Kullback-Leibler divergence (D_{KL}) between the distribution of patterns affinities in the repertoire $P_\Psi(\mathcal{A})$ and a Gamma distribution with matching mean and variance. Since we do not have an analytical expression for $P_\Psi(\mathcal{A})$ we use an Edgeworth approximation [25] to evaluate the Kullback-Leibler divergence between the two distribution, by relying on the cumulants of $P_\Psi(\mathcal{A})$ (Eq. S22).

In brief, Edgeworth series expands a probability density function around a normal distribution in terms of its cumulants, and it provides a true asymptotic expansion with controlled error [25]. To use the Edgeworth series, we transform the data to have mean zero and variance one, resulting in a modified probability density function for affinities $\hat{P}_\Psi(\mathcal{A})$. The second leading order approximation to $\hat{P}_\Psi(\mathcal{A})$ is given by

$$\hat{P}_\Psi(\mathcal{A}) \approx \phi(\mathcal{A})\left(1 + \frac{1}{3!}\hat{c}_3 He_3(\mathcal{A}) + \frac{1}{4!}\hat{c}_4 He_4(\mathcal{A}) + \frac{10}{6!}\hat{c}_3^2 He_6(\mathcal{A})\right) := \phi(\mathcal{A})\left(1 + u_{P_\Psi}\right), \tag{S24}$$

where $\phi(\mathcal{A})$ is a standard normal distribution, He_n is a Hermite polynomial of order n, and $\hat{c}_n = c_n/c_2^{n/2}$ is the nth normalized cumulant [24, 25]. As seen in Fig. S3, the approximation in Eq. S24 describes the distribution and especially the bulk of the affinity distribution very well.

To evaluate the Kullback-Leibler divergence $D_{\text{KL}}(\hat{P}_\Psi||\hat{\Gamma}_\Psi)$ between the modified affinity distribution $\hat{P}_\Psi(\mathcal{A})$ and a Gamma distribution $\hat{\Gamma}$ with matching mean and variance, we use an Edgeworth expansion for both of these distributions,

$$\begin{aligned} D_{\text{KL}}(\hat{P}_\Psi(\mathcal{A})||\hat{\Gamma}_\Psi(\mathcal{A})) &= \int_{\mathcal{A}} \hat{P}(\mathcal{A}) \log \frac{\hat{P}_\Psi(\mathcal{A})}{\hat{\Gamma}_\Psi(\mathcal{A})} = \int_{\mathcal{A}} \phi(\mathcal{A})\left(1+u_{P_\Psi}\right) \log \frac{\phi(\mathcal{A})\left(1+u_{P_\Psi}\right)}{\phi(\mathcal{A})\left(1+u_{\Gamma_\Psi}\right)} \\ &= \int_{\mathcal{A}} \phi(\mathcal{A}) \left[\left(u_{P_\Psi} + \frac{1}{2}u_{P_\Psi}^2\right) - \left(u_{\Gamma_\Psi} + u_{P_\Psi}u_{\Gamma_\Psi} - \frac{1}{2}u_{\Gamma_\Psi}^2\right)\right] + \mathcal{O}(u^3) \end{aligned} \tag{S25}$$

where u_{Γ_Ψ} arises from the Edgeworth expansion of the Gamma distribution, in analogy to Eq. S24. A similar approach has previously been used to approximate the Kullback-Leibler divergence between two distributions [29, 30].

To evaluate the integral in Eq. S25, we use the orthogonality of the Hermit polynomials, i.e.,

$$\int_x \phi(x) He_n(x) He_m(x) = n!\delta_{n,m} \tag{S26}$$

with $He_0 = 1$. As a result, all the linear terms in Eq. S25 vanish and only the squared terms with equal polynomial orders He_n contribute. We thus obtain,

$$\begin{aligned} D_{\mathrm{KL}}(\hat{P}_\Psi(\mathcal{A})||\hat{\Gamma}_\Psi(\mathcal{A})) =& \frac{1}{2}\left(\frac{1}{3!}\hat{c}_3^2 + \frac{1}{4!}\hat{c}_4^2 + \frac{100}{6!}\hat{c}_3^4\right) + \frac{1}{2}\left(\frac{1}{3!}\hat{\gamma}_3^2 + \frac{1}{4!}\hat{\gamma}_4^2 + \frac{100}{6!}\hat{\gamma}_3^4\right) \\ & - \left(\frac{1}{3!}\hat{c}_3\hat{\gamma}_3 + \frac{1}{4!}\hat{c}_4\hat{\gamma}_4 + \frac{100}{6!}\hat{c}_3^2\hat{\gamma}_3^2\right) \\ =& \frac{1}{2}\left[\frac{1}{3!}\left(\hat{c}_3 - \hat{\gamma}_3\right)^2 + \frac{1}{4!}\left(\hat{c}_4 - \hat{\gamma}_4\right)^2 + \frac{100}{6!}\left(\hat{c}_3^2 - \hat{\gamma}_3^2\right)^2\right]. \end{aligned} \tag{S27}$$

where $\hat{\gamma}_n$ is the nth cumulant of the modified Gamma distribution $\hat{\Gamma}_\Psi$.

By substituting the cumulant of the affinity distribution (Eq. S22) and that of the Gamma distribution, we arrive at an approximation for the Kullback-Leibler distance between the two distributions,

$$D_{\mathrm{KL}}(P_\Psi(\mathcal{A})||\Gamma_\Psi(\mathcal{A})) = \frac{25}{27}\left(2\Theta\mu_{\mathrm{eff}}\kappa\right)^{2/3} + \mathcal{O}(\Theta\mu_{\mathrm{eff}}) \ll 1. \tag{S28}$$

indicating that the Gamma distribution is a good approximation for the true distribution of the affinities $P_\Psi(\mathcal{A})$ for small mutation rate μ_{eff} as long as $\Theta\kappa$ is of order 1. However, for risk tolerant strategies (large κ), this approximation fails. In this regime the repertoire only learns one pattern effectively, resulting in a bi-modal distribution of the affinities with one mode reflecting the low-affinity recognition and the other, the higher affinity of the latest stored pattern. This bimodal distribution cannot be approximated by a Gamma distribution.

5. Affinity of random patterns

The affinity of a random pattern χ (i.e., patterns unrelated to the previously encountered classes Ψ^c) is determined by summing over the overlaps $|\langle\psi^\alpha|\chi\rangle|^\Theta$ between the random pattern χ and all previously stored patterns in the memory repertoire $\psi^\alpha \in \mathcal{M}$, weighed by their contributions to memory m^α. It should be noted that the mean affinity of random patterns is set to be zero and therefore, the affinity shift $\mathcal{A}_{\mathrm{rand}}$, defined in Eq. 2, can be evaluated as,

$$\mathcal{A}_{\mathrm{rand}} = \left\langle \mathcal{A}_0 \sum_{\psi^\alpha\in\mathcal{M}} m^\alpha |\langle\psi^\alpha|\chi\rangle|^\Theta \right\rangle \tag{S29}$$

$$= \mathcal{A}_0 \left\langle |\langle\psi|\chi\rangle|^\Theta \right\rangle = \mathcal{A}_0 \frac{2^{\Theta/2}\Gamma\left[\frac{1+\Theta}{2}\right]}{\sqrt{\pi L^\Theta}}, \tag{S30}$$

where we use the fact that random pattern χ is independent of the stored patterns ψ^α, and that on average their overlap $\langle\psi|\chi\rangle$ is a normally distributed random variable with mean zero and variance $\sigma^2_{\mathrm{rand}} = \frac{1}{4L}$, according to the law of large numbers.

To evaluate the variance of random patterns, we first introduce a basis that spans all the stored patterns in a repertoire ψ^α that have non-zero weights m^α at a given time point t. One choice for this basis is to use the N directions corresponding to the pattern classes $\Psi^c(t)$ at time t. However, these N vectors do not fully span all the previously presented patterns, due to the evolutionary divergence of these patterns over time. To account for this remaining subspace, we introduce N' auxiliary directions $\Phi^c(t)$ (with $c' = 1, \ldots, N'$). In principle, the space encompassing all possible the patterns is L-dimensional. However, the space of all stored patterns is typically more restricted (i.e., $N + N' < L$) due to the relatively fast updates in the repertoire such that it only keeps a memory of patterns that are similar to the current states of the presented classes $\Psi^c(t)$. Using the set of basis vectors $\{\Psi^c(t), \Phi^{c'}(t)\}$, we can express any stored pattern in the repertoire as,

$$\langle\psi^\alpha| = \sum_{c=1}^{N} \langle\psi^\alpha|\Psi^c(t)\rangle \langle\Psi^c(t)| + \sum_{c'=1}^{N'} \langle\psi^\alpha|\Phi^{c'}(t)\rangle \langle\Phi^{c'}(t)|. \tag{S31}$$

As a result, the overlaps between a random pattern χ and all the stored patterns in the repertoire in Eq. S29 can be

expressed as,

$$\sum_{\psi^\alpha\in\mathcal{M}} m^\alpha |\langle\psi^\alpha|\chi\rangle|^\Theta = \sum_{\psi^\alpha\in\mathcal{M}} m^\alpha \left|\sum_{c=1}^{N}\langle\Psi^c(t)|\chi\rangle\ \langle\psi^\alpha|\Psi^c(t)\rangle + \sum_{c'=1}^{N'}\langle\Phi^{c'}(t)|\chi\rangle\ \langle\psi^\alpha|\Phi^{c'}(t)\rangle\right|^\Theta \tag{S32}$$

$$\approx \sum_{c=1}^{N}\left(\sum_{\psi^\alpha\in\mathcal{M}} m^\alpha |\langle\psi^\alpha|\Psi^c(t)\rangle\ \langle\Psi^c(t)|\chi\rangle|^\Theta\right) + \sum_{c'=1}^{N'}\left(\sum_{\psi^\alpha\in\mathcal{M}} m^\alpha \left|\langle\psi^\alpha|\Phi^{c'}(t)\rangle\ \langle\Phi^{c'}(t)|\chi\rangle\right|^\Theta\right) \tag{S33}$$

$$\equiv \sum_{c=1}^{N} M_c\ |\langle\Psi^c(t)|\chi\rangle|^\Theta + \sum_{c'=1}^{N'} M'_{c'}\ \left|\langle\Phi^{c'}(t)|\chi\rangle\right|^\Theta. \tag{S34}$$

To arrive at Eq. S33, we assumed that the stored patterns ψ^α have non-vanishing overlaps with only one of the bases in the set $\{\Psi^c, \Phi^{c'}\}$. As a result, in the expansion of the expression to the power Θ in Eq. S32, all the cross terms associated with different bases vanish, and the expression can be simply written as the sum of independent terms. The final form in Eq. S34 expresses the overlap of a random pattern with the repertoire as a sum over the overlaps with the bases $\{\Psi^c(t), \Phi^{c'}(t)\}$, with effective weights $M_c(t) = \sum_\alpha m^\alpha |\langle\psi^\alpha|\Psi^c(t)\rangle|^\Theta$, and $M'_{c'}(t) = \sum_\alpha m^\alpha \left|\langle\psi^\alpha|\Phi^{c'}(t)\rangle\right|^\Theta$. The effective weight M_c associated with the presented pattern classes $\Psi^c(t)$ follows,

$$M_c(t) = \sum_\alpha m^\alpha |\langle\psi^\alpha|\Psi^c(t)\rangle|^\Theta = \frac{\mathcal{A}\left(\Psi^c(t)\right) + \mathcal{A}_{\text{rand}}}{\mathcal{A}_0} \approx \frac{\mathcal{A}\left(\Psi^c(t)\right)}{a_0}, \tag{S35}$$

where we used the approximation $a_0 = \mathcal{A}_0 - \mathcal{A}_{\text{rand}} = \mathcal{A}_0 + \mathcal{O}(L^{-\Theta/2})$.

When the shape parameter is $\Theta = 2$, the effective weights define a normalized set (i.e., $\sum_{c=1}^{N} M_c + \sum_{c'=1}^{N'} M'_{c'} = 1$). However, for sharper affinity functions ($\Theta > 2$), the relationship between the effective weights is bounded from above by $\sum_{c=1}^{N} M_c + \sum_{c'=1}^{N'} M'_{c'} \leq 1$, and for broader affinity functions ($\Theta < 2$), the weights are bounded from below as, $1 \leq \sum_{c=1}^{N} M_c + \sum_{c'=1}^{N'} M'_{c'}$. Here, we discuss the case with $\Theta > 2$, which entails

$$\sum_{c'=1}^{N'} M'_{c'}(t) \leq 1 - \sum_{c=1}^{N} M_c(t) \approx 1 - N\frac{\langle\mathcal{A}\rangle}{a_0}. \tag{S36}$$

This upper bound implies that when the patterns are well represented in the repertoire (i.e., $\langle\mathcal{A}\rangle \approx \mathcal{A}_{\max} = a_0/N$; see Eq. S5), the weights of the auxiliary bases shrink to zero, i.e., $N' = 0$. This result is in line with the eigen-decomposition analysis of the generalized Hopfield network in Appendix C of ref. [9].

For the cases that $M'_{c'} > 0$, we will consider a mean field approximation, where all the auxiliary directions are equally important (i.e., $M'_{c'} = M', \forall c'$) and that on average the memory stored in the auxiliary directions has a comparable weight to that of the bases spanned by the presented patterns, (i.e., $\mathbb{E}_c[M_c] = \mathbb{E}_{c'}[M'_{c'}] = M'$, where $\mathbb{E}_\star[\cdot]$ denotes expectation over the argument in the subscript). These approximations together with Eq. S36 define an upper bound for the number of auxiliary bases,

$$N' \leq \frac{a_0}{\langle\mathcal{A}\rangle}\left(1 - N\frac{\langle\mathcal{A}\rangle}{a_0}\right). \tag{S37}$$

Using these relationships, we can now evaluated the variance of the affinities across random patterns χ, as,

$$\mathbf{var}_{\text{rand}} = \mathbf{var}\left[\mathcal{A}_0 \sum_{\psi^\alpha \in \mathcal{M}} m^\alpha |\langle \psi^\alpha | \chi \rangle|^\Theta - \mathcal{A}_{\text{rand}}\right]_\chi \tag{S38}$$

$$= \mathcal{A}_0^2\, \mathbf{var}\left[\sum_{c=1}^{N} M_c\, |\langle \Psi^c(t)|\chi\rangle|^\Theta + \sum_{c'=1}^{N'} M'_{c'}\, \left|\langle \Phi^{c'}(t)|\chi\rangle\right|^\Theta\right] \tag{S39}$$

$$= \mathcal{A}_0^2 \sum_{c=1}^{N} (M_c)^2\, \mathbf{var}\left(|\langle \Psi^c|\chi\rangle|^\Theta\right) + \mathcal{A}_0^2 \sum_{c'}^{N'} (M'_{c'})^2\, \mathbf{var}\left(\left|\langle \Phi^{c'}|\chi\rangle\right|^\Theta\right) \tag{S40}$$

$$= \mathcal{A}_0^2 \left(\sum_{c=1}^{N} (M_c)^2 + \sum_{c'}^{N'} (M'_{c'})^2\right) \mathbf{var}\left(|\langle \psi|\chi\rangle|^\Theta\right) \tag{S41}$$

$$\leq \mathcal{A}_0^2 \left(\frac{N}{a_0^2}\left(\mathbf{var}(\mathcal{A}) + \langle \mathcal{A}\rangle^2\right) + \frac{1}{a_0}\langle \mathcal{A}\rangle \left(1 - \frac{N}{a_0}\langle \mathcal{A}\rangle\right)\right) \times \frac{2^\Theta\left(\sqrt{\pi}\Gamma\left[\frac{1}{2}+\Theta\right] - \Gamma\left[\frac{1+\Theta}{2}\right]^2\right)}{\pi L^\Theta}, \tag{S42}$$

where, we first used Eq. S30 to express the affinity function in Eq. S38 in terms of the bases in Eq. S39. Given that the projections of a random pattern along different bases are statistically independent from each other, we expressed the variance of the sum of these contributions in Eq. S39 as the sum of their variances in Eq. S40. Next we used the fact that the overlaps of a random pattern with all the bases (i.e., $\langle \Psi^c|\chi\rangle$, $\langle \Phi^{c'}|\chi\rangle$, $\forall c, c'$) are Gaussian random variates with mean zero and variance $L/4$. Thus, the variance of these overlaps to the power Θ are equal for all bases, which we simply expressed as $\mathbf{var}\left[|\langle \psi|\chi\rangle|^\Theta\right]$ in Eq. S41. We then evaluated the term $\mathbf{var}\left[|\langle \psi|\chi\rangle|^\Theta\right]$ from the underlying Gaussian distribution, and used the definition of the effective weight M_c in Eq. S35 and the inequality in Eq. S37 to substitute $\sum_{c=1}^{N}(M_c)^2 + \sum_{c'=1}^{N'}(M'_{c'})^2$ and arrive at the upper bound of the variance for the affinity for random patterns in Eq. S42. The results in Eq. S42 match very well with simulations for the affinity function with the shape parameter $\Theta = 2$ (Fig S4).

Appendix B: Mapping between immune memory repertoires and the Hopfield model

The Hopfield network is among the most frequently used models for associative memory [10]. A classical Hopfield network describes a fully connected graph with interaction matrix J_{ij}. A binary pattern ψ of length L presented to the network J is assigned an energy $E(\psi, J)$,

$$E(\psi, J) = -\frac{1}{2L}\sum_{i,j}^{L} J_{i,j}\psi_i\psi_j. \tag{S1}$$

Hopfiled networks can learn and store an associative memory of the presented patterns as the minima of the energy landscape [10]. One way to construct such network is by Hebbian learning, whereby the network is updated with a learning rate λ upon an encounter with a pattern $\psi(t)$ at time t,

$$J_{i,j}(t+1) = \begin{cases} (1-\lambda)\, J_{i,j}(t) + \lambda\, \psi_i(t)\psi_j(t), & \text{if } i \neq j; \\ 0, & \text{otherwise.} \end{cases} \tag{S2}$$

In the supplementary information of ref. [9], we show that this learning rule can be expressed as

$$J(t+1) = (1-\lambda)J(t) + \lambda\left(L\,|\psi(t)\rangle\langle\psi(t)| - \mathbb{1}\right). \tag{S3}$$

Using this expression for the network $J(T)$ at time T with an encounter history with patterns $\psi(t)$ (for $t \leq T$), we

can evaluate the energy $E(\chi, J(T))$ of an arbitrary pattern χ presented to the network as,

$$\begin{aligned} E(\chi, J(T)) &= -\frac{1}{2}\langle\chi|J(T)|\chi\rangle \\ &= -\frac{L}{2}\sum_{t=1}^{T-1}\lambda(1-\lambda)^{T-1-t}\langle\chi|\psi(t)\rangle\langle\psi(t)|\chi\rangle + \frac{1}{2}\sum_{t=1}^{T-1}\lambda(1-\lambda)^{T-1-t}\langle\chi|\mathbb{1}|\chi\rangle \\ &= -\frac{L}{2}\sum_{t=1}^{T-1}\lambda(1-\lambda)^{T-1-t}|\langle\psi(t)|\chi\rangle|^2 + \frac{1}{2} + \mathcal{O}\left((1-\lambda)^T\right) \\ &= -\frac{L}{2}\sum_{t=1}^{T-1} m_t\,|\langle\psi(t)|\chi\rangle|^2 + \frac{1}{2} + \mathcal{O}\left((1-\lambda)^T\right) \end{aligned} \tag{S4}$$

with the time-dependent weights, $m_t = \lambda(1-\lambda)^{T-1-t}$. Note that the correction terms vanish in the limit of large time T and that the weights sum up to one, i.e., $\sum m_t = 1$. In this limit, the energy function $E(\chi, J)$ corresponds to the affinity function given in Eq. 2 of the main text, for the choice of the shape parameter $\Theta = 2$, the energy scale $\mathcal{A}_0 = -L/2$, and the random energy $\mathcal{A}_{\text{rand}} = -1/2$.

Appendix C: Numerical methods

As discussed in Section B, for the shape parameter $\Theta = 2$ the affinity function in Eq. 2 maps onto the energy function of a Hopfield network with Hebbian learning [9, 10, 12], which is easily tractable with numerical techniques. Specifically, using a Hopfield model for simulating the repertoire problem with $\Theta = 2$ has the advantage that we only need to keep track of the interaction matrix $J_{i,j}$ of size L^2 as opposed to all the 2^L memory wights m^α. This dramatic reduction in complexity makes the numerical simulations for $\Theta = 2$ highly efficient.

For the simulations, we use the same approach as in [9]. We initialize the interaction matrix J of size $L \times L$ with all entries set to zero $J_{i,j} = 0$ and choose N independent random patterns of size L.

Before collecting any data, we first update the network until it reaches a quasi-stationary state so that it no longer depends on the initial condition. Since the stored memory of a pattern within the network decays as $(1-\lambda)^s$ with the number the number update steps s since the original encounter, we update the network following the initialization by $s_{\text{stat.}} = \text{ceil}\left(\frac{\log 10^{-5}}{\log(1-\lambda)}\right)$ steps for the network to reach a quasi-stationary state. This criterion ensures that $(1-\lambda)^{s_{\text{stat.}}} \leq 10^{-5}$ and the memory of the initial state is removed. Moreover, this criterion implies that in the quasi-stationary state the memory weights become normalized i.e., $1 - \sum_\alpha m^\alpha(s_{\text{stat.}}) < 10^{-5}$, after starting from a no memory initial state of $\sum_\alpha m^\alpha(0) = 0$. During each step of this preparation phase, we also evolve the patterns, whereby we flip the spins of patterns within each class at rate μ. We then randomly choose one of the patterns to present to the network and update the network according to the learning rule in Eq. S2.

After reaching the quasi-stationary state of memory, we collect data from the network over 10^4 steps. During this process, the patterns evolve with rate μ and are presented at a random order to the network. We record the affinity of each presented pattern before updating the network. At each step we also record the affinity of a randomly generated pattern that belongs to none of the previously encountered pattern classes. For each learning rate λ, we repeat this process for 50 independent initial sets of pattern classes $\{\Psi^c\}$. Overall, we perform a total of 5×10^5 affinity measurements on both the previously encountered and the random patterns.

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

[2] A. Soltoggio, K. O. Stanley, and S. Risi, Born to learn: The inspiration, progress, and future of evolved plastic artificial neural networks, Neural Networks **108**, 48 (2018).

[3] 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, Physics Reports **810**, 1 (2019), arXiv: 1803.08823.

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

[5] R. Barrangou and L. A. Marraffini, CRISPR-Cas systems: Prokaryotes upgrade to adaptive immunity, Molecular cell **54**, 234 (2014).

[6] G. Altan-Bonnet, T. Mora, and A. M. Walczak, Quantitative immunology for physicists, Physics Reports **849**, 1 (2020).

[7] 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).

[8] O. H. Schnaack and A. Nourmohammad, Optimal evolutionary decision-making to store immune memory, eLife **10**, e61346 (2021).

[9] O. H. Schnaack, L. Peliti, and A. Nourmohammad, Learning and organization of memory for evolving patterns, arXiv:2106.02186 [physics] (2021), arXiv: 2106.02186.

[10] J. J. Hopfield, Neural networks and physical systems with emergent collective computational abilities, Proc. Natl. Acad. Sci. U.S.A. **79**, 2554 (1982).

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

[12] M. Mezard, J. P. Nadal, and G. Toulouse, Solvable models of working memories, J. Physique **47**, 1457 (1986).

[13] 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).

[14] R. E. Steuer, *Multiple criteria optimization: theory, computation, and application*, Wiley series in probability and mathematical statistics (Wiley, New York, 1986).

[15] A. Sen, Markets and Freedoms: Achievements and Limitations of the Market Mechanism in Promoting Individual Freedoms, Oxf. Econ. Pap. **45**, 519 (1993).

[16] O. Shoval, H. Sheftel, G. Shinar, Y. Hart, O. Ramote, A. Mayo, E. Dekel, K. Kavanagh, and U. Alon, Evolutionary Trade-Offs, Pareto Optimality, and the Geometry of Phenotype Space, Science **336**, 1157 (2012).

[17] R. Schuetz, N. Zamboni, M. Zampieri, M. Heinemann, and U. Sauer, Multidimensional Optimality of Microbial Metabolism, Science **336**, 601 (2012).

[18] Y. Hart, H. Sheftel, J. Hausser, P. Szekely, N. B. Ben-Moshe, Y. Korem, A. Tendler, A. E. Mayo, and U. Alon, Inferring biological tasks using Pareto analysis of high-dimensional data, Nat Methods **12**, 233 (2015).

[19] P. Szekely, Y. Korem, U. Moran, A. Mayo, and U. Alon, The Mass-Longevity Triangle: Pareto Optimality and the Geometry of Life-History Trait Space, PLoS Comput Biol **11**, e1004524 (2015).

[20] L. F. Seoane and R. Solé, Phase transitions in Pareto optimal complex networks, Phys. Rev. E **92**, 032807 (2015).

[21] A. Tendler, A. Mayo, and U. Alon, Evolutionary tradeoffs, Pareto optimality and the morphology of ammonite shells, BMC Syst Biol **9**, 12 (2015).

[22] L. Koçillari, P. Fariselli, A. Trovato, F. Seno, and A. Maritan, Signature of Pareto optimization in the Escherichia coli proteome, Sci Rep **8**, 9141 (2018).

[23] S. Cobey and S. E. Hensley, Immune history and influenza virus susceptibility, Current Opinion in Virology **22**, 105 (2017).

[24] P. Hall, *The Bootstrap and Edgeworth Expansion*, Springer Series in Statistics (Springer New York, New York, NY, 1992).

[25] S. Blinnikov and R. Moessner, Expansions for nearly Gaussian distributions, Astron. Astrophys. Suppl. Ser. **130**, 193 (1998).

[26] M. Schweizer, Mean-Variance Hedging for General Claims, Ann. Appl. Probab. **2**, 10.1214/aoap/1177005776 (1992).

[27] K. B. Toft, On the Mean-Variance Tradeoff in Option Replication with Transactions Costs, The Journal of Financial and Quantitative Analysis **31**, 233 (1996).

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

[29] J.-J. Lin, N. Saito, and R. A. Levine, Edgeworth approximation of the Kullback-Leibler distance towards problems in image analysis, University of California, Davis, Tech. Rep (1999).

[30] J. Inglada and G. Mercier, A New Statistical Similarity Measure for Change Detection in Multitemporal SAR Images and Its Extension to Multiscale Change Analysis, IEEE Trans. Geosci. Remote Sensing **45**, 1432 (2007).

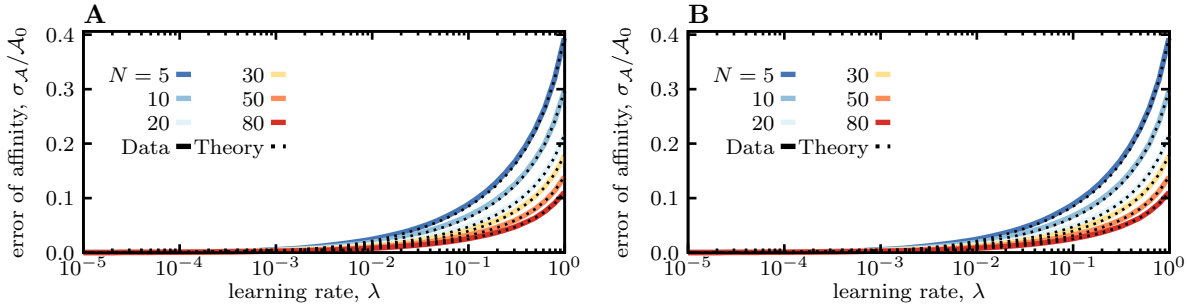

FIG. S1. **Standard deviation of affinities for pre-encountered patterns**. Solid lines shown the numerical estimates for the standard deviation of pattern affinities ($\sigma_{\mathcal{A}}$) divided by the scale of the affinity ($\mathcal{A}_0$) for different numbers of patterns N (colors), using Hopfield model to simulate of repertoires. Dotted lines show the analytical estimates, using **(A)** the full solution in Eq. S14, and **(B)** the approximation in Eqs. 4 & S22. Simulation parameters: $L = 100$, $\mu_{\text{eff}} = 0.01$, $\Theta = 2$.

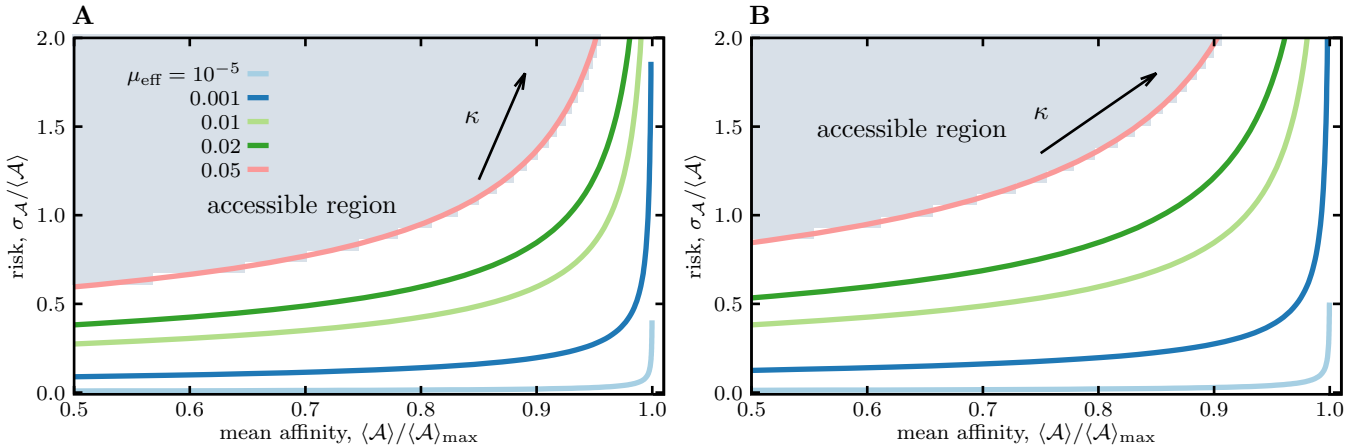

FIG. S2. **Pareto front for risk-utility tradeoff of memory for evolving patterns.** Similar to Fig. 2 in the main text, the risk-affinity Pareto front for the optimized objective function $\mathcal{Q}(\lambda^*)$ (Eq. 3) is shown for different mutation rates (colors) by varying the risk tolerance κ along each line, for the shape parameters **(A)** $\Theta = 4$, and **(B)** $\Theta = 8$. In both cases, $N = 200$.

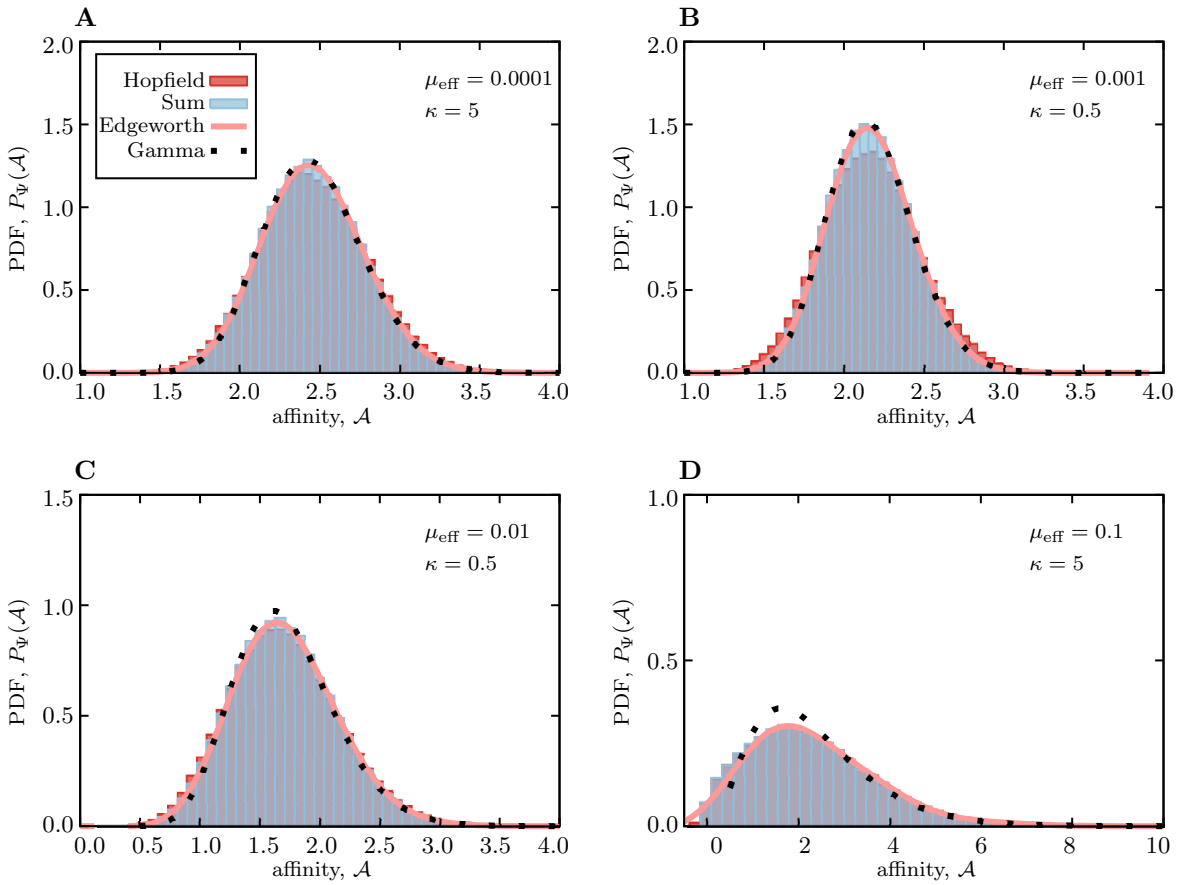

FIG. S3. **Distribution of pattern affinities.** The distribution of affinities between a memory repertoire and patterns from previously enchanted classes are shown for different effective mutation rates μ_{eff} and risk tolerance κ in different panels. The distributions are characterized based on the simulations using Hopfield network (red; see Appendix C), the process for deriving the cumulant-generating function in Eq S19. (blue), the Edgeworth approximation for the probability density function in Eq. S24 (pink), and the Gamma distribution with the matching mean and variance (dotted lines). For small $(\mu_{\text{eff}} \cdot \kappa)$ in (**A-C**) all distributions are comparable and as suggested by the Kullback-Leibler divergence in Eq. S28, the Gamma distribution is a good approximation to the underlying distribution of affinities. For larger $(\mu_{\text{eff}} \cdot \kappa)$ in **(D)**, the optimal learning rate becomes large (Eq. 5), and the repertoire only remembers the most recently encountered patterns, resulting in the break down of the analytical approximations. Simulation parameters: $L = 200$, $N = 40$, and $\Theta = 2$.

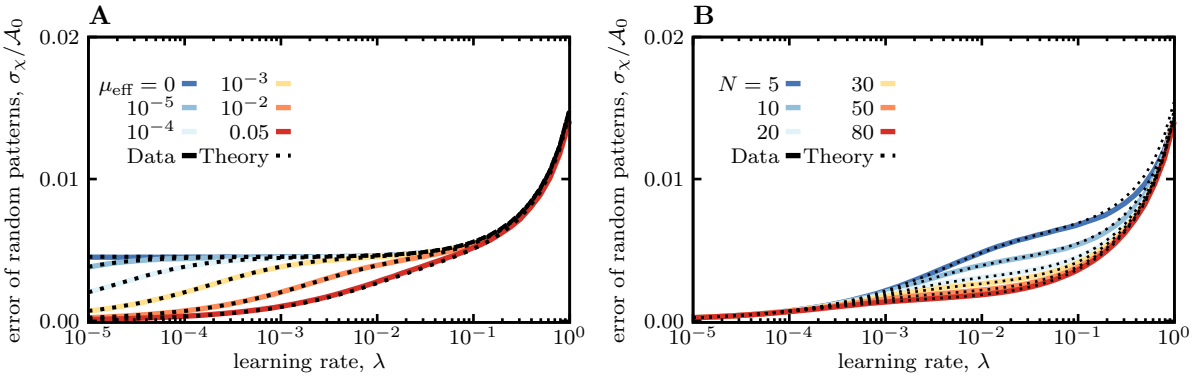

FIG. S4. **Standard deviation of affinities for random patterns**. Solid lines show simulation results for $\Theta = 2$ and dashed lines give the analytic result from Eq. S42. **(A)** shows results for constant number of pattern classes $N = 30$ and for different effective mutation rates $\mu_{\rm eff}$ (colors). **(B)** shows results for constant $\mu_{\rm eff} = 0.01$ and for different numbers of pattern classes N (colors). In both cases we observe, that when the learning rate λ is large and only one pattern class is stored in the system the fluctuations of random affinities are the largest as there is no self averaging over all stored patterns. When the system can learn close to the optimum (maximal affinity for stored patterns with small fluctuations) the prefactor in Eq. S42 reaches the naïve expectation $1/N$ and the fluctuations of random overlaps are averaged over all stored patterns. When the learning rate become small and the system can no longer follow the evolution of the patterns, the system effectively stores many additional states and the random fluctuations are averaged over all these states. As a consequence, the fluctuations σ_χ go to zero. Simulation parameters: $L = 100$, $\Theta = 2$.

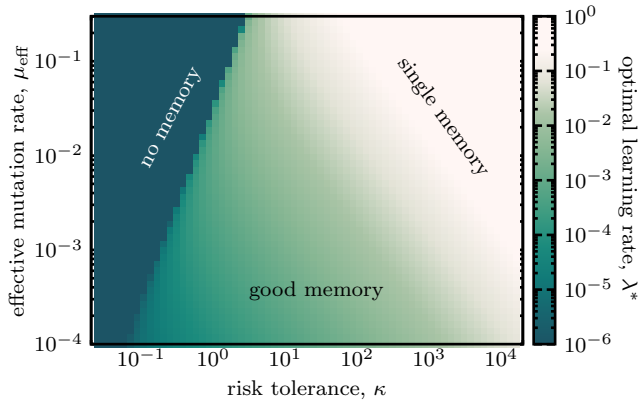

FIG. S5. **Optimal learning rate for three phases of memory.** The optimal learning rate λ^* corresponding to the different strategies in Fig. 3 is shown for different values of risk tolerance and mutation rate. Fast learning with $\lambda \sim 1$ corresponds to the phase of single memory storage (light), where only the memory of most recent encounter is retained. On the other hand, slow learning $\lambda \ll 1$ corresponds to the phase where effectively no memory is stored (dark). The triangle of good memory is associated with intermediate rates of learning. Simulation parameters: $L = 200$, $N = 40$, and $\Theta = 2$.

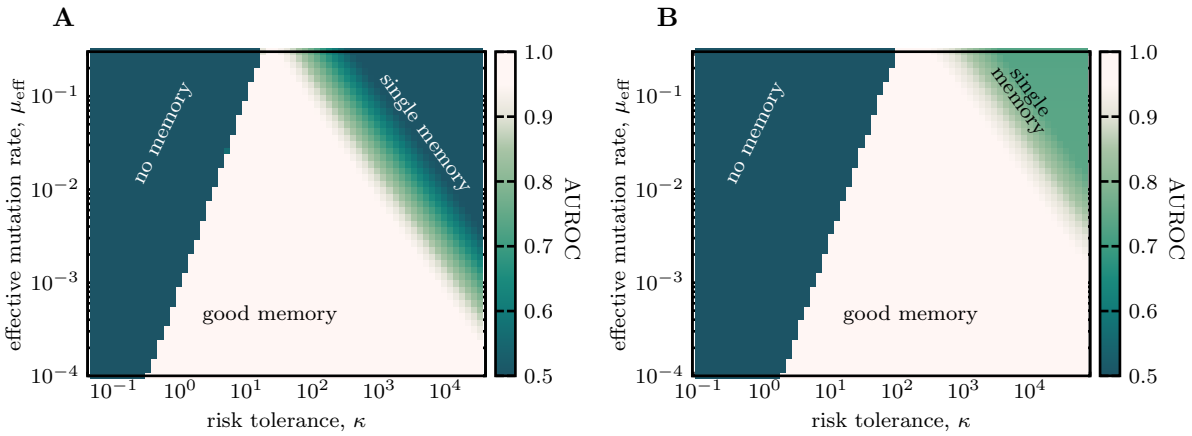

FIG. S6. **Three phases of memory discrimination for different shape parameters of the affinity function.** Similar to Fig. 3 the phase diagram shows the discrimination ability of repertoires (AUROC) between familiar patterns with prior encounter history and random patterns, for different shape parameters of the affinity function **(A)** $\Theta = 4$, and **(B)** $\Theta = 8$. Other parameters: Parameters: $L = 200$, $N = 40$.

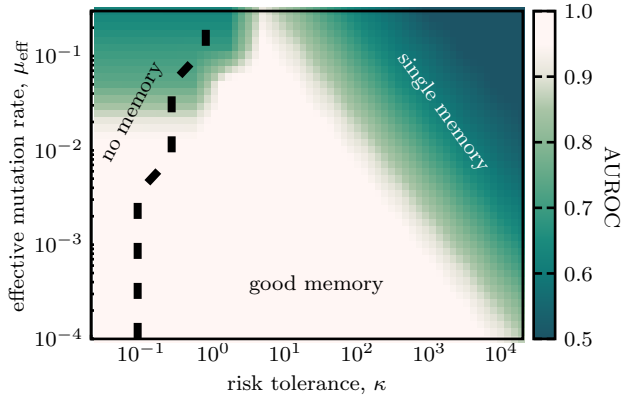

FIG. S7. **Three phases of memory discrimination uncovered with simulations of Hopfield networks.** Similar to Fig. 3 the phase diagram shows discrimination ability of repertoires (AUROC) between familiar patterns with prior encounter history and random patterns. The phase diagram is acquired by direct simulation of memory, using the correspondence between repertoires with shape parameter $\Theta = 2$ and Hopfield network; see Appendix C for numeral technique. The low risk region on the left side of the dashed line is not accessible by simulations, which explains the differences between the simulations and results of the analytic approximation shown in Fig. 3. Parameters: $L = 200$, $N = 40$, and $\Theta = 2$.