Journal of Statistical Physics (2025) 192:67
https://doi.org/10.1007/s10955-025-03447-4

Analytical Methods for Continuous Attractor Neural Networks

Martino Salomone Centonze[1] · Alessandro Treves[2,3] · Elena Agliari[4] · Adriano Barra[5]

Received: 30 August 2024 / Accepted: 10 April 2025 / Published online: 2 May 2025

Abstract

Pyramidal cells that emit spikes when the animal is at specific locations of the environment are known as *place cells*: these neurons are thought to provide an internal representation of space via *cognitive maps*. Here, we consider the Battaglia-Treves neural network model for cognitive map storage and reconstruction, instantiated with McCulloch & Pitts binary neurons. To quantify the information processing capabilities of these networks, we exploit spin-glass techniques, namely the *interpolation method* and the *replica trick*. In particular, in the low-storage regime (i.e., when the number of stored maps scales sub-linearly with the network size and the order parameters self-average around their means), by adapting the Hamilton-Jacobi PDE-approach, we obtain an exact phase diagram in the noise vs inhibition strength plane. In the high-storage regime, by adapting the standard interpolation based on stochastic stability, we find that—for mild inhibition and not too high noise—memorization and retrieval of an extensive number of spatial maps is possible. These results, holding under the replica-symmetry assumption, are recovered, for completeness, also by the replica method and they are corroborated by Monte Carlo simulations. Finally, by leveraging the integral representation of the model (in terms of a bipartite network equipped with highly-selective hidden units), we successfully test its robustness versus various distributions of place fields, including the log-normal distribution observed in recent experiments on bats navigating long tunnels. Additionally, we demonstrate that, by appropriately coupling these hidden units, the network can effectively orient itself even in dynamic environments.

Communicated by Francesco Zamponi.

✉ Adriano Barra
adriano.barra@uniroma1.it

1 Dipartimento di Matematica, Università di Bologna, Bologna, Italy

2 Cognitive Neuroscience, SISSA, Trieste, Italy

3 DSyNC Lab., University of Agder, Kristiansand, Norway

4 Dipartimento di Matematica, Sapienza Università di Roma, Rome, Italy

5 Dipartimento di Scienze di Base ed Applicate per l'Ingegneria, Sapienza Università di Roma, Rome, Italy

1 Introduction

The hippocampus contributes in the brain to spatial recognition and particularly spatial memory. In rodents, hippocampal cells recorded during free foraging and other spatial tasks have turned out to be mainly correlated with the animal's position; this was first observed for pyramidal neurons in the CA1 area a long time ago [1]: each neuron fires when the animal is in one or more regions of the current environment. These neurons were therefore called *place cells*, and the regions of space in which they get active were called *place fields*. In other words, place cells provide an internal representation of the animal position and additional evidence shows that these neural representations play a crucial role in spatial memory [2, 3].

Continuous-attractor neural networks (CANNs) (see e.g., [4]) offer a natural tool for simulating such complex systems [5–7]. In general, in a recurrent attractor neural network, some information (e.g., different locations in a certain environment) are encoded in the firing patterns of neurons and, for a suitable setting of interaction strengths among neurons, these patterns correspond to stationary states (attractors) of network dynamics. Thus, a stimulus which elicits the retrieval of previously-stored information (e.g., a detail of an experienced location) is expected to yield a stationary (or approximatively stationary) pattern for neuronal activity that codifies for further information (e.g., the stored location in relation to others) allowing for its use in behaviour and possibly for its consolidation in long-term storage. However, unlike simple attractor models, such as the Hopfield network (which can operate with very distributed patterns of activity in which e.g. half of the binary units are active), in a CANN the interaction between two units necessarily includes along with an excitatory term—that depends on the similarity between their preferred stimuli (e.g., the proximity of their place fields) – also a (long range) inhibition term—that prevents all cells being active together. This arrangement enables a CANN to hold a continuous family of stationary states, rather than isolated ones. In the case of place cells, stationary states occur in the form of localized bump of activity (also referred to as *coherent states*), peaked at a certain retrieved location. Ideally, retrievable locations span a continuous set of nearby values, although in practice finite size effects are known to impose a discretization or a roughening of the theoretically continuous manifold of attractor states [8, 9]. This way, a CANN is able to update its states (internal representations of stimuli) smoothly under the drive of an external input: several mathematical formulations for the generation of place-cells have been introduced in the past, as well as others describing the CANN that could be realized with populations of place cells [5, 8, 10]. Much work has focused on the deviations expected in real networks from the idealized models [8, 9]. Despite the intensive investigations that have been (and are still being[1]) carried on by the statistical-mechanics community on Hebbian architectures for neural networks [4, 32, 33], no rigorous results have been obtained dealing with place cells and one of the goals of the current work is to fill this gap by rigorously analysing a model for a CANN with hippocampal place cells. Specifically, we focus on the model introduced by Battaglia and Treves [10] but here, instead of considering real-valued activities for model neurons as in the original work, units are binary variables, taking value 0 (when resting) or 1 (when firing), namely McCulloch-Pitts neurons.

We recall that the quenched average of the free energy for the standard Battaglia-Treves model has been computed in the thermodynamic limit by using the replica trick at the replica symmetric level of approximation, see e.g. [6, 10]. A purpose of the present paper is to obtain an explicit expression of the free energy for the current version of the model (where neurons

[1] See e.g. [11–31] for a glance on recent findings (with a weak bias toward rigorous approaches).

are binary) by the replica trick, as this is a benchmark in the community, and by adopting Guerra's interpolation, as this is mathematically more controllable.
The work is structured as follows. In Sect. 2 we introduce the model and present our major analytical findings. In particular, in Sect. 2.1 we discuss the Battaglia-Treves model on the circular manifold, in Sect. 2.2 we introduce the observables we need to build a theory for this model, in Sect. 2.3 we show the output of the theory and, in Sect. 2.4, we extend the dual representation of Hebbian networks as restricted Boltzmann machines [34–38] (equipped with single-map selectively firing hidden neurons) to the case.
Next, Sect. 3 tackles the methodological aspects underlying our investigation: in Sect. 3.1 we address the (mathematically simpler) low-storage regime, by adapting to the case the Hamilton-Jacobi approach [39, 40] (see also the works by Mourrat and Chen for a sharper mathematical control [41–44]) and in Sect. 3.2 we address the more challenging high-storage regime, by adapting to the case the standard interpolation technique based on stochastic stability [45–47].
Then, Sect. 4 is entirely dedicated to numerical simulations, where we generalize the Battaglia-Treves model in two different directions. First, we show its robustness w.r.t. the way place fields are distributed along the environment and, thus, w.r.t. the way maps are coded and stored. Specifically, we prove that, moving from uniform to log-normal distributions (as those recently seen in experiments with bats [48]), the computational capabilities of the model survive. Then, we consider the restricted Boltzmann machine representation of the Battaglia-Treves model and let its hidden neurons interact with each other so to create preferential paths to be followed by the network while exploring the environment.
Finally, conclusions and outlooks are presented in Sect. 5.
Two appendices, Appendix A and Appendix B, close the paper and provide mathematical details. In particular, Appendix A is dedicated to the Battaglia-Treves model: in Appendix A.1 we report the replica-trick analysis while Appendix A.2 contains long proofs related to the interpolation techniques. Conversely, in Appendix B we summarize how the interpolation techniques used in this paper do work on the paradigmatic Hopfield model in order to make the reader with a background in Neuroscience acquainted with these analytical approaches: the low storage is investigated via the Hamilton-Jacobi approach in Appendix B.1, whereas the high storage is investigated via the stochastic stability in Appendix B.2.

2 The Model: From Definitions to Computational Capabilities

The Battaglia-Treves model was introduced to describe the behavior of pyramidal neurons in a rodent hippocampus. In the original model [5, 10] neuronal activity is represented by continuous variables linearly activated by their input and thresholded at zero, while here we consider N McCulloch-Pitts neurons, whose activity is denoted as $s_i \in \{0, 1\}$ for $i \in (1, ..., N)$, in such a way that, when $s_i = +1$ ($s_i = 0$), the i-th neuron is spiking (quiescent). As anticipated in Sect. 1, the model is designed in such a way that a neuron fires when the rodent is in a certain region of the environment.
The latter is described by a manifold equipped with a metric and uniformly partitioned into N regions of (roughly) equal size, centered in $\vec{r}_i$, for $i \in (1, ..., N)$, which are the cores of their place fields. When the agent modeling the rodent happens to be in the i-th core place field, the corresponding i-th neuron is expected to fire along with other neurons with their core place fields close by. Note that the behavior of different neurons is not independent: in fact, when the distance $|\vec{r}_i - \vec{r}_j|$ between $\vec{r}_i$ and $\vec{r}_j$ is small enough, the reciprocal influence between

neurons i and j is strong and they are likely to be simultaneously active. This is captured by the definition of the interaction matrix (i.e., the synaptic coupling in a neural network jargon) $\boldsymbol{J}$, whose element J_{ij} represents the interaction strength between the neurons (i, j) and is taken proportional to a given kernel function $\mathcal{K}(|\vec{r}_i - \vec{r}_j|)$ depending on the distance between the related place fields:

$$J_{ij} \propto \mathcal{K}(|\vec{r}_i - \vec{r}_j|). \tag{1}$$

Notice that the core place fields in any given environment are assumed to have been already learnt or assigned, thus their coordinates do not vary with time (i.e., they are *quenched* variables in a statistical mechanics jargon), while the state $\boldsymbol{s} = (s_1, s_2, ..., s_N)$ of the neurons (i.e. their neural activity) is a dynamical quantity.

Following [10], and as standard for Hebbian neural networks [12, 49–51], we consider multiple arrangements of place fields, also referred to as *charts* or *maps* corresponding to distinct environments, e.g., different experimental rooms in which a rodent has been left to forage [52], such that each neuron participates in each chart and, thus, contributes to the collective recognition of several charts. The synaptic matrix, accounting for K charts, is obtained by summing up many terms like the one in (1), that is

$$J_{ij} \propto \sum_{\mu=1}^{K} \mathcal{K}(|\vec{r}_i^{\,\mu} - \vec{r}_j^{\,\mu}|), \tag{2}$$

where $\vec{r}_i^{\,\mu}$ represents the core of the i-th place field in the μ-th map. The kernel function $\mathcal{K}$ is taken the same for each chart and the variation from one chart to another is only given by a different arrangement of core place fields $\{\vec{r}_i^{\,\mu}\}_{i=1,...,N}^{\mu=1,...,K}$, in practice a reshuffling. Remarkably, here, different *charts* play a role similar to that played by different *patterns* in the standard Hopfield model [49, 50] and, along the same lines, we assume that the stored maps are statistically independent [52]: in other words, the position of the core place field associated to a certain neuron in one map, say $\vec{r}_i^{\,\mu}$, is uncorrelated to the core place field associated to the same neuron in any other map, say $\vec{r}_i^{\,\nu}$, at least when the network storage K/N is small enough.

2.1 Coherent States, Cost Function, and Other Basic Definitions

Let us now describe more explicitly our framework: for illustrative purposes, here we restrict ourselves to a one-dimensional manifold given by the unit circle S^1 (as often done previously, see e.g., [6, 7, 9, 53]), in such a way that the position of the i-th place field in the μ-th chart is specified by the angle $\theta_i^\mu \in [-\pi, \pi]$ w.r.t. a reference axes and it can thus be expressed by the unit vector $\vec{\eta}_i^{\,\mu}$ defined as

$$\eta_i^\mu = \left(\cos\theta_i^\mu, \sin\theta_i^\mu\right), \tag{3}$$

where we dropped the arrow on top to lighten the notation.
In this coding, the trivial states $\boldsymbol{s} = (1, .., 1)$ and $\boldsymbol{s} = (0, .., 0)$, beyond being biologically unrealistic, do not carry information as they lead to a degenerate activation in such a way that the physical location of the animal cannot be represented by the network.
As opposed to these trivial states, we define

Definition 1 *(Coherent state)* Given a set of N McCulloch-Pitts neurons constituting a CANN, their state $\boldsymbol{s} = (s_1, ..., s_N) \in \{0, 1\}^N$ is said to be a coherent state centered around

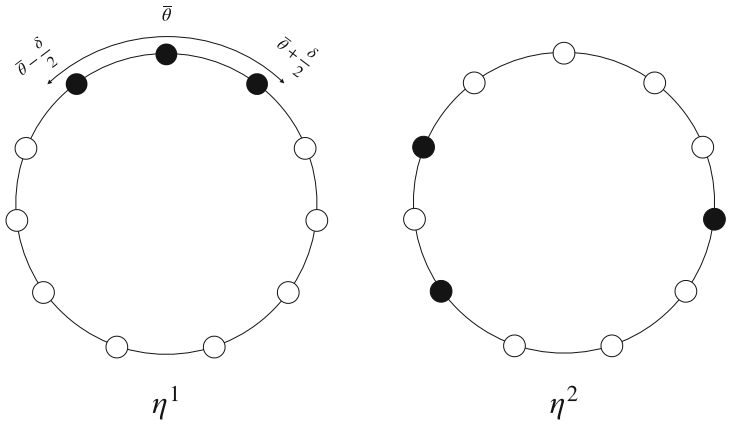

Fig. 1 Two examples of maps η^1 and η^2. A coherent state is shown in η^1 as all neurons that lie within the $[\bar{\theta} - \delta/2, \bar{\theta} + \delta/2]$ interval are activated (here displayed as black dots), while the others are quiescent (in white dots). The same firing pattern of neurons, that looks coherent in the first map η^1, looks disordered in the other map η^2. Note that the centers of the place fields are scattered roughly uniformly along the unitary circle S^1 and that the width of all the place fields is roughy the same in this first scenario

$\bar{\theta} \in [0, 2\pi]$ with width δ if, for each $i = 1, .., N$:

$$s_i = \begin{cases} 1, & |\theta_i^1 - \bar{\theta}| \leq \delta/2, \\ 0, & |\theta_i^1 - \bar{\theta}| > \delta/2, \end{cases} \tag{4}$$

where θ_i^1 is the coordinate of the i-th place cell in the first map η^1 that we used as an example. In other words, a coherent state in a CANN plays the same role of a retrieval state in the Hopfield model [50].

Notice that the network is able to store several place fields for multiple maps and to retrieve a single place field by relaxing on the related coherent state of neural activities: as shown in Figs. 1 and 2, such a coherent state for $\mu = 1$ looks random in others maps (e.g. $\mu = 2$ in the picture), provided that these are independent and that their number in the memory is not too large.

Following [10], keeping in mind that the kernel has to be a function of a distance among place field cores on the manifold and that the latter is the unitary circle, we now make a specific choice for the interaction kernel that will be implemented in the network: to take advantage from the *Hebbian experience* [32], the interacting strength between neurons will be written as[2]

$$J_{ij} = \frac{1}{N} \sum_{\mu=1}^{K} (\eta_i^\mu \cdot \eta_j^\mu). \tag{5}$$

Notice that the Hebbian kernel (5) is a function of the relative euclidean distance of the i, j neuron's coordinates θ_i, θ_j in each map μ: to show this, one can simply compute the dot product as $\eta_i^\mu \cdot \eta_j^\mu = \cos(\theta_i^\mu)\cos(\theta_j^\mu) + \sin(\theta_i^\mu)\sin(\theta_j^\mu) = \cos(\theta_i^\mu - \theta_j^\mu)$.

[2] We can assume that the place field cores cover roughly uniformly the embedding space, namely the angles $\bar{\theta}^\mu$ are uniformly distributed along unitary circle (sampled from $\mathcal{U}_{[-\pi,\pi]}$), yet, there is no need to introduce a prior for them as, given the rotational invariance of the CANN kernel, a uniform dislocation of place fields is automatically fulfilled.

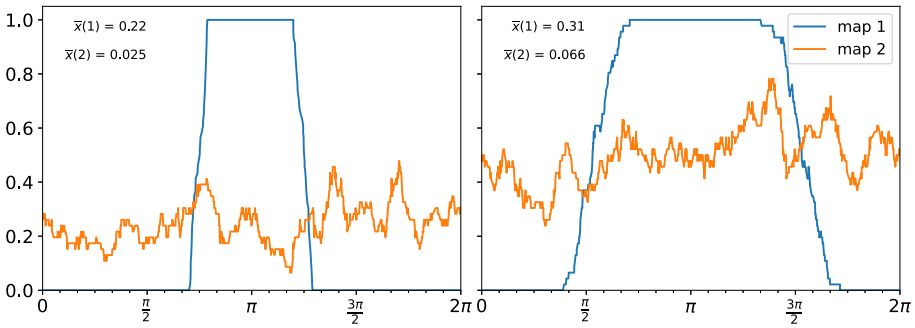

Fig. 2 Neuronal activity in the first two maps $\mu = 1, 2$. The neurons coordinates $\theta_i^\mu \in [0, 2\pi]$ in the given map are displayed along the horizontal axis, while the neuronal activity is shown on the vertical one and computed as the spatial average of the neuronal states in a spatial window of fixed length. *Left*: we provide the model with the initial state $\mathbf{s}_0$ (a *bump* in the map $\mu = 1$ centered around $\theta = \pi$): note that in other maps (e.g. $\mu = 2$, in orange) this input appears as a random state. This Cauchy condition for the neural dynamics (i.e. the stochastic process (7) driven by the Hamiltonian (6)) allows the network to evolve toward a stationary state $\mathbf{s}_{out}$, that, in the map $\mu = 1$, is the coherent state shown on the *right* (while it still appears random in other maps as, e.g., $\mu = 2$ in orange). The value of the map overlap $\bar{x}$ is also reported in each panel: coherent states have higher overlap than random states, as expected

We can now define the Cost function (or Hamiltonian, or energy in a physics jargon) of the model as given by the next

Definition 2 *(Cost function)* Given N binary neurons $\boldsymbol{s} = (s_1, ..., s_N) \in \{0, 1\}^N$, K charts $\boldsymbol{\eta} = (\eta^1, ..., \eta^K)$ with $\eta^\mu \in [-1, +1]^N$ for $\mu \in (1, ..., K)$ encoded with the specific kernel (5) and a free parameter $\lambda \in \mathbb{R}^+$ to tune the global inhibition within the network, the Hamiltonian for chart reconstruction reads as[3]

$$H_N(\boldsymbol{s}|\boldsymbol{\eta}) = -\sum_{i<j}^{N,N} J_{ij} s_i s_j + \frac{(\lambda - 1)}{N} \sum_{i<j}^{N,N} s_i s_j$$
$$\approx -\frac{1}{2N} \sum_{\mu=1}^{K} \sum_{i,j=1}^{N,N} (\eta_i^\mu \cdot \eta_j^\mu) s_i s_j + \frac{(\lambda - 1)}{2N} \sum_{i,j=1}^{N,N} s_i s_j. \tag{6}$$

Notice that the factor N^{-1} in front of the sums ensures that the Hamiltonian is linearly extensive in the thermodynamic limit $N \to \infty$ and the factor $1/2$ is inserted in order to count only once the contribution of each couple of neurons. Also, the hyper-parameter λ tunes a source of inhibition acting homogeneously among all pairs of neurons and prevents the network from collapsing onto a fully firing state $\boldsymbol{s} = (1, ..., 1)$. In fact, for $\lambda \gg 1$ the last term at the r.h.s. of Eq. (6) prevails, global inhibition dominates over local excitation and the most energetically-favorable configuration is the fully inhibited one (where all the neurons are quiescent $\boldsymbol{s} = (0, .., 0)$); in the opposite limit, for $\lambda \ll 1$, the most energetically-favorable configuration is the totally excitatory one (where all the neurons are firing $\boldsymbol{s} = (1, ..., 1)$).

Once a Hamiltonian is provided, it is possible to construct a Markov process for the neural dynamics by introducing a source of noise $\beta \in \mathbb{R}^+$ (i.e., the temperature in a physics jargon)

[3] The symbol '$\approx$' in Eq. (6) becomes an exact equality in the thermodynamic limit, $N \to \infty$, where, splitting the summation as $\sum_{i<j} = 1/2 \sum_{i,j}^N + \sum_{i=1}^N$, the last term, being sub-linear in N, can be neglected.

in the following master equation

$$\mathbf{P}_{t+1}(\boldsymbol{s}|\boldsymbol{\eta}) = \sum_{\boldsymbol{s}'} W_\beta(\boldsymbol{s}, \boldsymbol{s}'|\boldsymbol{\eta})\mathbf{P}_t(\boldsymbol{s}'|\boldsymbol{\eta}), \tag{7}$$

where $\mathbf{P}_t(\boldsymbol{s}|\boldsymbol{\eta})$ represents the probability of finding the system in a configuration of neural activities $\boldsymbol{s}$ at time t, being $\boldsymbol{\eta}$ the set of K charts. $W_\beta(\boldsymbol{s}, \boldsymbol{s}'|\boldsymbol{\eta})$ represents the transition rate, from a state $\boldsymbol{s}'$ to a state $\boldsymbol{s}$ and it is chosen in such a way that the system is likely to lower the energy (6) along its evolution (see e.g. [4] for details): this likelihood is tuned by the parameter β such that for $\beta \to 0^+$ the dynamics is a pure random walk in the neural configuration space (and any configuration is equally likely to occur), while for $\beta \to +\infty$ the dynamics becomes a steepest descend toward the minima of the Hamiltonian and the latter (in this deterministic limit) plays as the Lyapounov function for the dynamical process [4]. Remarkably, the symmetry of the interaction matrix, i.e. $J_{ij} = J_{ji}$, is enough for detailed balance to hold such that the long-time limit of the stochastic process (7) relaxes to the following Boltzmann-Gibbs distribution

$$\lim_{t\to\infty} \mathbf{P}_t(\boldsymbol{s}|\boldsymbol{\eta}) = \mathbf{P}(\boldsymbol{s}|\boldsymbol{\eta}) = \frac{e^{-\beta H_N(\boldsymbol{s}|\boldsymbol{\eta})}}{Z_N(\beta, \lambda, \boldsymbol{\eta})} \tag{8}$$

where $Z_N(\beta, \lambda, \boldsymbol{\eta})$ is the normalization factor, also referred to as *partition function*, as stated by the next

Definition 3 *(Partition function)* Given the Hamiltonian $H_N(\boldsymbol{s}|\boldsymbol{\eta})$ of the model (6) and the control parameters β and λ ruling the neuronal dynamics, the associated partition function $Z_N(\beta, \lambda, \boldsymbol{\eta})$ reads as

$$Z_N(\beta, \lambda, \boldsymbol{\eta}) = \sum_{\{\boldsymbol{s}\}}^{2^N} e^{-\beta H_N(\boldsymbol{s}|\boldsymbol{\eta})}. \tag{9}$$

2.2 Order Parameters, Control Parameters and Other Thermodynamical Observables

We now proceed by introducing the macroscopic observables useful to describe the computational capabilities of the system under investigation.

Definition 4 *(Order parameters)* In order to assess the quality of the retrieval of a given chart $\boldsymbol{\eta}^\mu$, it is useful to introduce the two-dimensional *order parameter* x_μ, $\mu \in (1, ..., K)$ defined as

$$x_\mu = \frac{1}{N}\sum_{i=1}^{N} \eta_i^\mu s_i \in [-1/\pi, +1/\pi]^2. \tag{10}$$

Another order parameter to consider, that does not depend on the place fields, is the mean activity of the whole population of neurons, defined as

$$m = \frac{1}{N}\sum_{i=1}^{N} s_i \in [0, 1]. \tag{11}$$

Finally, the last order parameter to introduce is the two-replica overlap (whose usage is required solely when investigating the high-storage regime), defined as

$$q_{ab} = \frac{1}{N}\sum_{i=1}^{N} s_i^a s_i^b \in [0, 1], \tag{12}$$

where a and b are the replica indices, namely they label different copies of the system, characterized by the same realization of maps (i.e., by the same quenched disorder in a spin-glass jargon).

As a consistency check, we notice that the module of x_μ, that is $|x_\mu| = \sqrt{x_\mu \cdot x_\mu}$, is strictly positive for the coherent states, while for the trivial states (i.e., the pure excitatory and the pure inhibitory configurations), it is vanishing. When all the neurons are quiescent this holds straightforwardly, while, in the case where all neurons are firing, x_μ is just the sum of each coordinate vector η_i^μ in the given map and this quantity is close to zero (and exactly zero in the thermodynamic limit $N \to \infty$) when the coordinate vectors associated to each neuron in the given map are statistically independent and the angles θ_i^μ sampled from a uniform distribution in the interval $[-\pi, \pi]$.
Conversely, for a coherent state, the vector x_μ is just the sum of the coordinates η_i^μ on the sites where the neurons are active, i.e., $x_\mu \equiv (x_\mu^{(1)}, x_\mu^{(2)}) = \frac{1}{N}\sum_{i|s_i=1}(\cos\theta_i^\mu, \sin\theta_i^\mu)$. In the thermodynamic limit we can rewrite it as an integral, namely

$$x_\mu^{(1)} \underset{N\to\infty}{\sim} \frac{1}{2\pi}\int_{-\delta/2}^{\delta/2} d\theta \cos\theta = \frac{1}{\pi}\sin\delta/2,$$
$$x_\mu^{(2)} \underset{N\to\infty}{\sim} \frac{1}{2\pi}\int_{-\delta/2}^{\delta/2} d\theta \sin\theta = 0,$$

hence we get $|x_\mu| = \frac{1}{\pi}\sin\delta/2$. The latter is positive for $0 < \delta < 2\pi$ and reaches its maximum at $\delta = \pi$, corresponding to $|x_\mu| \underset{\delta=\pi}{=} \frac{1}{\pi} \sim 0.32$. This situation represents a coherent state of size $\delta = \pi$.

Definition 5 *(Boltzmann and quenched averages)* Given a function $f(\boldsymbol{s})$, depending on the neuronal configuration $\boldsymbol{s}$, the Boltzmann average, namely the average over the distribution (8), is denoted as $\omega(f(\boldsymbol{s}))$ and defined as

$$\omega(f(\boldsymbol{s})) = \frac{\sum_{\{s\}}^{2^N} f(\boldsymbol{s}) e^{-\beta H_N(\boldsymbol{s}|\boldsymbol{\eta})}}{\sum_{\{s\}}^{2^N} e^{-\beta H_N(\boldsymbol{s}|\boldsymbol{\eta})}}.$$

Further, given a function $g(\boldsymbol{\eta})$, depending on the realization of the K maps, we introduce the quenched average, namely the average over the realizations of the maps, that is denoted as $\mathbb{E}_\eta[g(\boldsymbol{\eta})]$ or as $\langle g(\boldsymbol{\eta})\rangle_\eta$ according to the context, and it is defined as

$$\mathbb{E}_\eta[g(\boldsymbol{\eta})] \equiv \langle g(\boldsymbol{\eta})\rangle_\eta = \int_{-\pi}^{\pi}\left[\prod_{i,\mu=1}^{N,K}\frac{d\theta_i^\mu}{2\pi}\right] g(\boldsymbol{\eta}(\boldsymbol{\theta})). \tag{13}$$

Finally, we denote with the brackets $\langle\cdot\rangle$ the average over both the Boltzmann-Gibbs distribution and the realization of the maps, that is

$$\langle\cdot\rangle = \mathbb{E}_\eta[\omega(\cdot)].$$

Definition 6 *(Replica symmetry)* We assume that, in the thermodynamic limit $N \to \infty$, all the order parameters self-average around their mean values, denoted by a bar, that is

$$\lim_{N\to\infty} \mathbf{P}(x_\mu) = \delta\left(x_\mu - \overline{x}_\mu\right), \quad \forall \mu \in (1, ..., K), \tag{14}$$

$$\lim_{N\to\infty} \mathbf{P}(m) = \delta\left(m - \overline{m}\right), \tag{15}$$

$$\lim_{N\to\infty} \mathbf{P}(q_{12}) = \delta\left(q_{12} - \overline{q}_2\right), \tag{16}$$

$$\lim_{N\to\infty} \mathbf{P}(q_{11}) = \delta\left(q_{11} - \overline{q}_1\right). \tag{17}$$

This assumption is the so-called *replica symmetric approximation* in spin-glass jargon, see e.g., [46, 54] or *concentration of measure* in probabilistic terms [55].

Definition 7 *(Storage capacity)* The storage capacity of the network is denoted as α and defined as

$$\alpha = \lim_{N\to\infty} \frac{K}{N} \tag{18}$$

The regime where the number of stored charts scales sub-linearly with the network size, i.e. where $\alpha = 0$, is referred to as low-storage, while the regime where the number of stored charts scales linearly with the network size, i.e. where $\alpha > 0$, is referred to as high-storage.

The storage capacity α, along with the noise level β and the inhibition parameter λ, constitute the control parameters of the system under study and, by tuning their values, the mean values of the order parameters $\overline{x}, \overline{m}, \overline{q}_1, \overline{q}_2$ change accordingly.
To quantify their evolution in the (α, β, λ) space, we further introduce the main quantity for our investigation, that is

Definition 8 *(Free energy)* The intensive (quenched) free-energy of the Battaglia-Treves model equipped with McCulloch & Pitts neurons is defined as

$$A_N(\beta, \lambda) = \frac{1}{N}\mathbb{E}_\eta \ln Z_N(\beta, \lambda, \boldsymbol{\eta}) = \frac{1}{N}\mathbb{E}_\eta \ln \sum_{\{s\}}^{2^N} \exp\left(-\beta H_N(\boldsymbol{s}|\boldsymbol{\eta})\right) \tag{19}$$

and, in the thermodynamic limit, we write

$$A(\alpha, \beta, \lambda) = \lim_{N\to\infty} A_{N,K}(\beta, \lambda). \tag{20}$$

The explicit knowledge of the free energy in terms of the control parameters α, β, λ and of the (expectation of the) order parameters $\overline{x}, \overline{m}, \overline{q}_1, \overline{q}_2$ is the main focus of the present investigation. In fact, once its explicit expression is obtained, we can extremize the free energy w.r.t. the order parameters to obtain a set of self-consistent equations for their evolution in the space of the control parameters: the study of their solutions allows us to sketch the phase diagram of the model and thus to know *a priori* in which regions in the (α, β, λ) space, the charts can be successfully retrieved by the place cells as we now discuss.

2.3 Phase Diagram of the Battaglia-Treves Model with McCulloch & Pitts Neurons

Whatever the route, replica trick or interpolation method, the main purpose of this analysis is to obtain an explicit expression for the quenched free energy of the model in terms of its order and control parameters: this result, obtained in two spatial dimensions for the sake of simplicity, can be summarized by the next

Theorem 1 *In the thermodynamic limit, the replica-symmetric quenched free-energy of the Battaglia-Treves model, equipped with McCulloch-Pitts neurons as defined in Eq. (2), can be expressed in terms of the expectations of the order parameters $\overline{m}, \overline{x}, \overline{q}_1, \overline{q}_2$ and of the control parameters α, β, λ, as follows*

$$\begin{aligned} A^{RS}(\alpha,\beta,\lambda) = &-\frac{\beta}{2}(1-\lambda)\overline{m}^2 - \frac{\beta}{2}\overline{x}^2 - \frac{\alpha\beta}{2}\frac{\overline{q}_1 - \frac{\beta}{d}(\overline{q}_1-\overline{q}_2)^2}{\left[1-\frac{\beta}{d}(\overline{q}_1-\overline{q}_2)\right]^2} \\ &-\frac{\alpha d}{2}\ln\left[1-\frac{\beta}{d}(\overline{q}_1-\overline{q}_2)\right] + \\ &+\frac{\alpha\beta}{2}\frac{\overline{q}_2}{1-\frac{\beta}{d}(\overline{q}_1-\overline{q}_2)} + \mathbb{E}_\eta \int d\mu(z)\ln \\ &\left[1+\exp\left(\beta(1-\lambda)\overline{m} + \beta\overline{x}\cdot\eta + \beta\frac{\frac{\alpha}{2}+\sqrt{\frac{\alpha\overline{q}_2}{d}}\,z}{1-\frac{\beta}{d}(\overline{q}_1-\overline{q}_2)}\right)\right], \end{aligned} \tag{21}$$

where the superscript RS reminds us that $A^{RS}(\alpha,\beta,\lambda)$ is the replica symmetric approximation of the true free energy.[4] *The mean values $\overline{m}$, $\overline{q}_1$, $\overline{q}_2$, $\overline{x}$ appearing in the above expression have to extremize the free-energy hence their values are obtained as solutions of the constraint $\nabla_{\overline{x},\overline{m},\overline{q}_1,\overline{q}_2} A^{RS}(\alpha,\beta,\lambda) = 0$, resulting in the following self-consistency equations*

$$\langle x\rangle = \overline{x} = \int d\mu(z)\,\langle \eta\,\sigma\,(\beta h(z))\rangle_\eta, \tag{22}$$

$$\langle q_{11}\rangle = \overline{q}_1 = \overline{m} = \int d\mu(z)\,\langle \sigma\,(\beta h(z))\rangle_\eta, \tag{23}$$

$$\langle q_{12}\rangle = \overline{q}_2 = \int d\mu(z)\,\langle \sigma^2\,(\beta h(z))\rangle_\eta, \tag{24}$$

where $\sigma(t) = \frac{1}{1+e^{-t}}$ is the sigmoid function, $d\mu(z)$ is the Gaussian measure for $z \sim \mathcal{N}(0,1)$ and $h(z)$ represents the internal field acting on neurons and reads as

$$h(z) = (1-\lambda)\overline{m} + \overline{x}\cdot\eta + \frac{\frac{\alpha}{2}+\sqrt{\frac{\alpha\overline{q}_2}{d}}\,z}{1-\frac{\beta}{d}(\overline{q}_1-\overline{q}_2)}. \tag{25}$$

A proof of this theorem, based on Guerra's interpolation, can be found in Sect. 3.2, further, in Appendix A.1, we report an independent derivation based on the usage of the replica trick. Finally, we note that the low storage solution provided in Sect. 3.1 can be obtained simply by setting $\alpha = 0$ in the above expression (21).

Corollary 1 *In order to numerically solve the self-consistent Eqs. (22)–(24), it is convenient to introduce the quantity $C = \frac{\beta}{d}(\overline{q}_1 - \overline{q}_2)$, make the change of variables $(\overline{x}, \overline{q}_1, \overline{q}_2) \to$*

[4] As neural networks are spin glasses in a statistical mechanical jargon, Parisi's replica symmetry breaking is expected to occur in the low noise and high storage limit [32]. Yet, in this first investigation we confine ourselves in providing an exhaustive picture of the RS scenario, in agreement with the existing literature [6, 10].

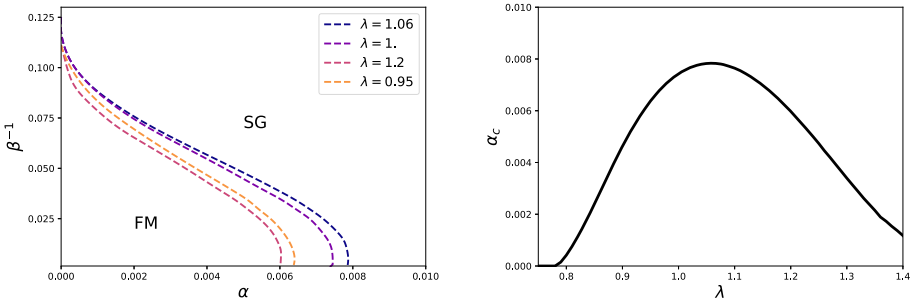

Fig. 3 *Left*: Lower region of the phase diagram of the model in the load α and noise β^{-1} plane, at various inhibition strength λ as shown in the legend. We recognize the presence of two phases, the Spin Glass phase (SG), where $\overline{x} = 0$, $(\overline{q}_1)^2 \ll \overline{q}_2 \leq \overline{q}_1$, and the Ferromagnetic phase (FM), with $\overline{x} > 0$, $\overline{q}_2 = \overline{q}_1$ (i.e. the region where coherent states collectively appear). Note that the paramagnetic phase, which would be in the upper region of the diagram, is not shown for the sake of clearness, to emphasize the boundary between the SG and FM phases, which differs from what is seen in neural networks with linear threshold units [56, 57]). *Right*: Critical load as a function of the global inhibition strength λ: the maximum value of $\alpha_c \sim 0.0078$ is found at $\lambda_{\alpha_c} = 1.06$

$(\overline{x}, C, \overline{q}_2)$ and write them as

$$\overline{x} = \Omega_d \int d\mu(z) \int_{-1}^{1} dt\, t \left(1 - t^2\right)^{\frac{d-3}{2}} \sigma\left(\beta h(z,t)\right), \tag{26}$$

$$\overline{q}_2 = \Omega_d \int d\mu(z) \int_{-1}^{1} dt \left(1 - t^2\right)^{\frac{d-3}{2}} \sigma^2\left(\beta h(z,t)\right) \tag{27}$$

$$C \equiv \frac{\beta}{d}(\overline{q}_1 - \overline{q}_2) = \Omega_d \frac{1 - C}{\sqrt{\alpha \overline{q}_2 d}} \int d\mu(z)\, z \int_{-1}^{1} dt \left(1 - t^2\right)^{\frac{d-3}{2}} \sigma\left(\beta h(z,t)\right), \tag{28}$$

where we posed

$$h(z,t) \equiv (1-\lambda)\left(\frac{d}{\beta} C + \overline{q}_2\right) + t\overline{x} + \frac{\frac{\alpha}{2} + \sqrt{\frac{\alpha \overline{q}_2}{d}}\, z}{1 - C}, \tag{29}$$

and we used the solid angle in $d-$dimensions $\Omega_d = \dfrac{\Gamma\left(\frac{d}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{d-1}{2}\right)}$.

Proof See the appendix, Sect. A.2 □

The self-consistent Eqs. (26)–(28), with the internal field defined according to (29), are solved numerically to draw a phase diagram for the model as reported in Fig. 3 (left panel). In particular, we find that the model exhibits three different phases. In the *paramagnetic* phase (PM) -not shown- the model does not retrieve any map, hence $\overline{x}_\mu = 0$ for any $\mu = 1, .., K$ and the replicas are totally uncorrelated: in this regime noise prevails over signals and the network dynamics is ergodic. However, lowering the noise, the ergodic region breaks and the network may enter a spin glass (SG) region (if the amount of stored maps is too large) or a ferromagnetic (FM) region, namely a retrieval region where charts are spontaneously reinstated by the network.
The maximum capacity of the model—see Fig. 3 (right panel)—can be estimated numerically by studying the $\beta \to \infty$ limit of the self-consistency equations, which leads to the following.

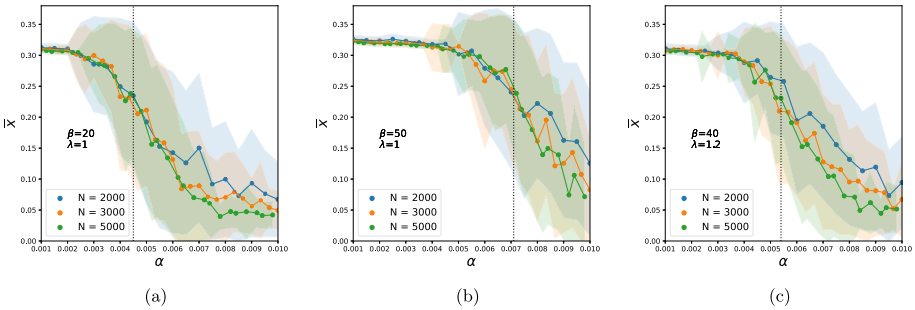

Fig. 4 Finite-Size-Scaling simulations of neural dynamics by Monte Carlo (MC) runs for three different choices of the control parameters: $\beta = 20$ and $\lambda = 1$ *(a)*, $\beta = 50$ and $\lambda = 1$ *(b)*, $\beta = 40$ and $\lambda = 1.2$ *(c)*. Each scenario has been investigated by enlarging the network size from $N = 2000$ to $N = 5000$: the error bands (indicated by the shadowed regions) set to ± 1 standard deviation around the experimental points. Each MC run is the average over 50 samples. The dotted line indicates the theoretical transition line, as shown in the phase diagram provided in Fig. 3 and they sit close to the inflection points of $\overline{x}(\alpha)$

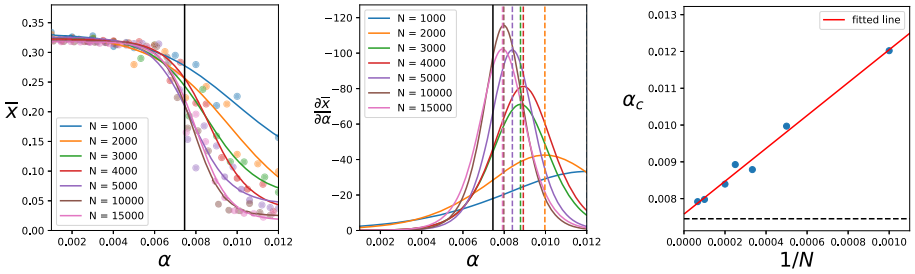

Fig. 5 Finite-Size-Scaling simulations of neural dynamics by MC runs at $\beta = 100$ and $\lambda = 1$ and different sizes, namely $N = \{1000, 2000, 3000, 4000, 5000, 10,000, 15,000\}$, with K varied accordingly to span up to 30 points in the $\alpha \in [0.001, 0.012]$ range: this includes the theoretical critical value $\alpha_c = 0.0075$ where the FM-SG transition is expected (indicated as a vertical black line in the left and center plots and as a dashed horizontal line in the right plot). *Left*: for each choice (N, K), we average over 20 MC independent runs (circles) and fit these experimental data versus a sigmoidal function to highlight a flex that represents our related estimate of α_c at which the transition occurs. *Center*: for each curve, we extract the derivative in such a way that α_c can be easily determined as the extremal points (dashed vertical lines). *Right*: these estimates for α_c (circles) are depicted as a function of $1/N$ and fitted by a linear function (solid line), to get an asymptotic extrapolation of the critical load in the thermodynamic limit. The R^2 coefficient of the fit is 0.987; the intercept of the line with the origin (where $1/N = 0$) is obtained at $\alpha_c \sim 0.0076$, that is in good agreement with the theoretical prediction of $\alpha_c = 0.0075$ at $\lambda = 1$

Corollary 2 *In the noiseless limit $\beta \to \infty$ limit, the self-consistency Eqs. (26)–(28) become*

$$\overline{x} = \frac{1}{2\pi} \int_0^\pi d\theta \, \cos\theta \, \text{erf}\left(\frac{g(\theta)}{\sqrt{2}}\right), \tag{30}$$

$$\overline{q}_2 = \frac{1}{2} + \frac{1}{2\pi} \int_0^\pi d\theta \, \text{erf}\left(\frac{g(\theta)}{\sqrt{2}}\right), \tag{31}$$

$$C = \frac{1 - C}{\sqrt{2\pi^3 \alpha \overline{q}_2 d}} \int_0^\pi d\theta \, \exp\left(-\frac{g(\theta)^2}{2}\right), \tag{32}$$

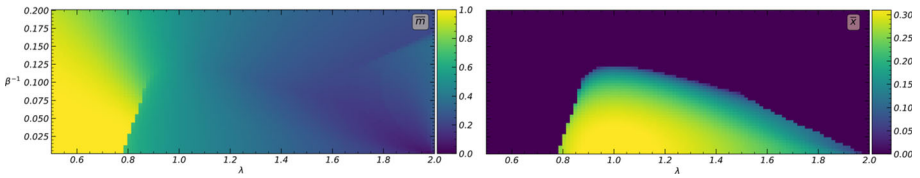

Fig. 6 Phase diagram of the model in terms of the inhibition strength λ and noise β, in the low-storage regime $\alpha = 0$. We show the overall magnetization of the network $\overline{m}$ (*left panel*) and the module of the neural activity $\overline{x}$ (*right panel*)

where we posed

$$g(\theta) = \sqrt{\frac{d}{\alpha \overline{q}_2}} \left[\frac{\alpha}{2} + (1 - C) \left((1 - \lambda) \overline{q}_2 + \overline{x} \cos\theta \right) \right]. \tag{33}$$

Proof In the $\beta \to \infty$ limit, the sigmoid function collapses to the Heaviside function Θ, i.e. $\sigma(\beta h) \overset{\beta \to \infty}{\longrightarrow} \Theta(h)$. The Gaussian integral appearing in the self-consistency Eqs. (26)–(28) can then be restricted to the domain where the internal field $h(z)$ is positive, that is the interval $h(z) \in [-g(t), \infty)$, where $g(t)$ is defined in Eq. (33): after a trivial rescaling of the integrals, we obtain the expressions (30)–(32) in terms of the error function $\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x dz\, e^{-z^2}$. Note that, as standard [32], we used $\frac{Cd}{\beta} + \overline{q}_2 \overset{\beta \to \infty}{\longrightarrow} \overline{q}_2$. □

These equations can be solved numerically and, in particular, we find that there exists a critical value for α, referred to as α_c, such that above that threshold no positive solution for $\overline{x}$ exists. The numerical estimate of this critical storage $\alpha_c(\lambda)$ as a function of λ is reported in Fig. 3 (right panel); the maximal critical load is found to be $\alpha_c \sim 0.0078$ at $\lambda = 1.06$, suggesting that the network works at its best for mild values of global inhibition.
These theoretical results are finally compared to numerical outcomes as reported in Figs. 4 and 5 displaying overall a very good agreement between theory and simulations.

In the low-storage regime (i.e., $\alpha = 0$) the phase diagram only depends on two control parameters: the neural noise β and the inhibition strength λ. In order to work out a phase diagram for the network in this regime it is enough to simplify the self-consistency Eqs. (22)–(23) by setting $\alpha = 0$ therein, thus obtaining

$$\overline{x} = \frac{1}{\pi} \int_{-1}^{1} dt\, \frac{t}{\sqrt{1 - t^2}}\, \sigma\left(\beta(1 - \lambda)\overline{m} + \beta t \overline{x}\right), \tag{34}$$

$$\overline{m} = \frac{1}{\pi} \int_{-1}^{1} \frac{dt}{\sqrt{1 - t^2}}\, \sigma\left(\beta(1 - \lambda)\overline{m} + \beta t \overline{x}\right) \tag{35}$$

These equations are solved numerically to trace the evolution of $\overline{m}$ and $\overline{x}$ in the β, λ plane and results are presented in Fig. 6: as expected, and as anticipated in the previous section, the limits $\lambda \ll 1$ and $\lambda \gg 1$ both lead to "paramagnetic" phases where $\overline{x} = 0$ and the computational capabilities of the network are lost. In contrast, in the region close to $\lambda = 1$ a retrieval phase with $\overline{x} > 0$ and $\overline{m} \sim 0.5$ naturally appears.

As within the numerical exploration we can inspect finite size effects, we further conducted extensive simulations to check the existence of spontaneous map transitions—as those originally discussed in [6]—also in the present model, with a number of spontaneous transitions that gets exponentially suppressed in N, as expected. In Fig. 7 we present an example, relative to a network with $N = 300$, $K = 2$, $\beta = 10$ and $\lambda = 1$.

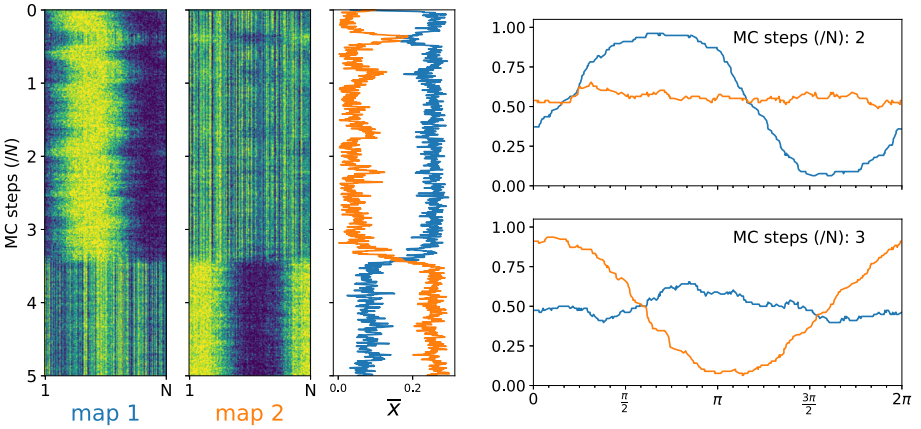

Fig. 7 (*Left*): MC runs to inspect the presence of spontaneous transitions between $K = 2$ uncorrelated maps in a network away from the thermodynamic limit (here $N = 300$) as discussed in [6]. One transition, from map 1 (blue curve) to map 2 (orange curve), is observed within $5N$ MC steps. Each yellow dot represents an active neuron, while blue dots represent inactive neurons. The evolution of $\overline{x}$ is also shown, as a function of the number of MC steps (normalized to the network size N), for the 2 considered maps. (*Right*): two snapshots of the network activity are shown, depicting its state after $2N$ MC steps (*top*) and $3N$ MC steps (*bottom*)

The coherent states are very large in our setting (roughly one half of the length of the maps, i.e. $\delta = 1/2$) and the correlation between coherent states that inhabit two different maps can thus reach values of the order of $\sim 30\%$, much higher that those reported in [6] where the Authors focused on much smaller values of δ.

2.4 The Dual Representation of the Model: Hidden Map-Selective Cells

Before deepening the interpolation approaches, it is convenient to provide the integral representation of the Battaglia-Treves model, as this will be used in the stochastic stability analysis of the high-storage regime in Sect. 3.2 and will be generalized to allow the model to cope with a dynamical environment in Sect. 4.2. Indeed, an equivalent mathematical model can be given in terms of a network of N McCulloch&Pitts neurons interacting with K real-valued hidden variables that are highly selective, namely they are expected to fire if rat transits within a given place field. These cells are distinct from the standard neurons $\boldsymbol{s}$ previously introduced; instead, they naturally emerge as highly selective chart cells.
To show this duality of representation (see also Fig. 8), it is enough to point out that the partition function of the model can be written as

$$Z_N(\beta, \lambda, \boldsymbol{\eta}) = \sum_{\{s\}} \exp\left(\frac{\beta}{2N}\sum_{i,j}^{N,N}\sum_{\mu=1}^{K}\eta_i^\mu\eta_j^\mu s_i s_j - \frac{\beta(\lambda-1)}{2N}\sum_{i,j}^{N,N} s_i s_j\right) \tag{36}$$

$$= \sum_{\{s\}}\int d\mu(z_\mu)\exp\left(\frac{\beta}{\sqrt{N}}\sum_{i}^{N}\sum_{\mu}^{K}\eta_i^\mu s_i z_\mu - \frac{\beta(\lambda-1)}{2N}\sum_{i,j}^{N,N} s_i s_j\right), \tag{37}$$

where in the last line, by using the Hubbard-Stratonovich transform (where the Gaussian measure is $d\mu(z_\mu) = \prod_\mu \frac{d^d z_\mu}{\sqrt{2\pi}}\exp\left(-\frac{z_\mu^2}{2}\right)$), we introduced K hidden variables that could

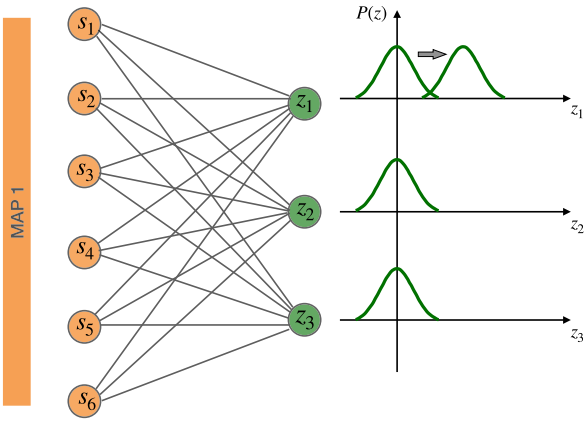

Fig. 8 Dual representation of the Battaglia-Treves model in terms of a bipartite network equipped with highly selective hidden neurons: these are one per stored chart, such that the retrieval of, say, η^1 by the s triggers the corresponding z_1 cell to fire too. Note that chart elements η_i^μ (rescaled by the $\sqrt{N}$ factor) now work as weights in this bipartite representation of the BT model: see the first term at the r.h.s. of (38)

play the formal role of chart cells z_μ, $\mu \in (1, ..., K)$, one per chart, such that the exponent in Eq. (37) can be thought of as an effective Hamiltonian as stated by the next

Proposition 1 *Having introduced K real-valued z_μ, $\mu \in (1, ..., K)$, hidden neurons equipped with a standard Gaussian prior (i.e. the measure $d\mu(z_\mu)$), the Battaglia-Treves model admits a dual representation in terms of a bipartite neural network whose weights are the map elements and whose cost function reads as*

$$H_N(s, z|\boldsymbol{\eta}) = -\frac{1}{\sqrt{N}} \sum_i^N \sum_\mu^K \eta_i^\mu s_i z_\mu + \frac{\lambda - 1}{2N} \sum_{i,j}^{N,N} s_i s_j = -\sqrt{N} \sum_\mu^K x_\mu z_\mu + (\lambda - 1) N m^2. \tag{38}$$

We highlight that the chart-selective neurons $\boldsymbol{z}$ account for individual charts, namely, their mean values are zero unless the conjugated order parameters $x_\mu \neq 0$ for some z_μ: in this case, as the overall field experienced by the generic z_μ is proportional $\sqrt{N} x_\mu$, $x_\mu > 0$ this triggers a response in z_μ, as shown in Fig. 8.

3 Guerra's Techniques for Place Cells

In this section we discuss the underlying mathematical methodologies and we provide two statistical-mechanic techniques, both based on Guerra's interpolation [47], to solve for the free energy of the model under study: the former, meant to address the low-load regime, is the Hamilton-Jacobi approach and it is based on the so-called *mechanical analogy* [39, 40]. This analogy lies in the observation that the free energy in Statistical Mechanics plays as an action in Analytical Mechanics, namely it obeys a Hamilton-Jacobi PDE in the space of the couplings: as a result, its explicit expression can be obtained as a solution of the Hamilton-Jacobi equation, that is by leveraging tools typical of Analytical Mechanics rather than Statistical Mechanics. The latter, meant to cover the high-load regime, is the standard one-body interpolation based on stochastic stability [46, 58]: here we interpolate between the free energy of the original model and the free energy of a suitably designed one-body model, whose solution is straightforward as one-body models enjoy trivial factorized probability distributions. To connect these two extrema of the interpolation we use the fundamental

theorem of calculus, where, crucially, the assumption of replica symmetry makes the integral analytical.
These techniques are exploited in Sects. 3.1 and 3.2, respectively.
For the reader with a main background in Neuroscience rather than spin glasses, we refer to the Appendix B, where we used the same interpolation schemes at work with the standard Hopfield model: Appendix B.1 deals with the low storage via the Hamilton-Jacobi interpolation, whereas Appendix B.2 faces the high storage via the stochastic stability.

3.1 Phase Diagram in the Low-Storage Regime

We start our analytical investigation by addressing the low-storage regime (where no spin-glass know-how is required) and we exploit the *mechanical analogy*. We first introduce the interpolating parameters: a variable $t \in \mathbb{R}^+$ to mimic *time* and three *spatial coordinates* $(y, \vec{z}) \in \mathbb{R} \times \mathbb{R}^2$ that we use to define the Guerra action $f(t, y, \vec{z})$.[5] Then, once checked that $f(t = \beta, y = 0, \vec{z} = 0)$ coincides with the free energy $A(\beta, \lambda)$, we show that $f(t, y, \vec{z})$ plays as an action in this spacetime, namely that it obeys a Hamilton-Jacobi PDE. Finally, we find the PDE solution by integrating the Lagrangian coupled to this action over time which, in turn, provides an explicit expression also for the original free energy of the Battaglia-Treves model confined to the low-storage.

As the low-storage regime is defined by $\alpha = 0$, we consider the case where the model only stores and retrieves one map, say η^1, with no loss of generality [4, 59]. In this setting, the Hamiltonian reduces to that provided in the next

Definition 9 *(One-map Cost function)* The Battaglia-Treves Hamiltonian (6) for map retrieval, equipped with N McCulloch-Pitts neurons $\boldsymbol{s} = (s_1, ..., s_N) \in \{0, 1\}^N$ and a free parameter $\lambda \in \mathbb{R}^+$ to tune the global inhibition within the network, at work with solely one map $\boldsymbol{\eta}^1$, reads as

$$H_N(\boldsymbol{s}|\boldsymbol{\eta}) = -\frac{1}{2N}\sum_{i,j=1}^{N,N}(\eta_i^1 \cdot \eta_j^1)s_i s_j + \frac{\lambda - 1}{2N}\sum_{i,j=1}^{N,N} s_i s_j = -\frac{N}{2}(x_1)^2 + (\lambda - 1)\frac{N}{2}m^2, \tag{39}$$

where, in the right-most side, we used the order parameters x_1 and m, as defined in Eqs. (10)–(11).

Definition 10 *(Guerra Action)* Following the mechanical analogy, we introduce a fictitious time $t \in \mathbb{R}^+$ and a $(1 + 2)$ fictitious space $(y, \vec{z}) \in \mathbb{R} \times \mathbb{R}^2$ that we use to define the interpolating free energy (or *Guerra action* [58]) $f(t, y, \vec{z})$ as

[5] Note that $\vec{z} = (z_1, z_2)$ is a bi-dimensional vector in the present two-dimensional embedding of a one-dimensional manifold but, should we work in K dimensions, $\vec{z}$ would be a K dimensional vector.

$$f(t,y,\vec{z}) = \frac{1}{N}\mathbb{E}_\eta \left[\ln \sum_s \exp\left(-t\left(\frac{1}{2}\sum_{ij}^{N,N} \eta_i^1 \eta_j^1 s_i s_j - \frac{\lambda-1}{2}\sum_{ij}^{N,N} s_i s_j \right) + y\sqrt{1-\lambda}\sum_i^N s_i + \vec{z}\cdot\sum_i^N \eta_i s_i \right) \right]. \tag{40}$$

Remark 1 Note that, by setting $t = -\beta$ and $(y,\vec{z}) = (0,\mp$ in the above Guerra action, the latter coincides with the original free-energy (20) of the Battaglia-Treves model in the low-storage limit.

Proposition 2 *(Hamilton-Jacobi PDE) The Guerra action (40), related to the model described by the Hamiltonian (39), obeys by construction the following Hamilton-Jacobi equation in the* $3+1$ *space-time* $(y,\vec{z},t)$

$$\begin{cases} \dfrac{\partial f(t,y,\vec{z})}{\partial t} + \dfrac{1}{2}[\nabla f(t,y,\vec{z})]^2 + \mathcal{V}(t,y,\vec{z}) = 0 & \text{(41a)} \\ \mathcal{V}(t,y,\vec{z}) = \dfrac{1}{2}\left[\langle x_1^2\rangle - \langle x_1\rangle^2\right] + \dfrac{1-\lambda}{2}\left[\langle m^2\rangle - \langle m\rangle^2\right] & \text{(41b)} \end{cases}$$

where the gradient reads as $\nabla f(t,y,\vec{z}) \equiv \left(\frac{\partial f(t,y,\vec{z})}{\partial y}, \frac{\partial f(t,y,\vec{z})}{\partial \vec{z}}\right)$.

Proof The proof works by direct evaluation of the derivatives (w.r.t. y, $\vec{z}$ and t) of the Guerra's action (40).
The $t-$derivative of Guerra action $f(t,y,\vec{z})$, can be obtained straightforwardly as

$$\frac{\partial f(t,y,\vec{z})}{\partial t} = -\frac{1}{2}\langle x_1^2\rangle + \frac{\lambda-1}{2}\langle m^2\rangle, \tag{42}$$

while its gradient is

$$\nabla f(t,y,\vec{z}) \equiv \left(\frac{\partial f(t,y,\vec{z})}{\partial y}, \frac{\partial f(t,y,\vec{z})}{\partial \vec{z}}\right) = \left(\langle x_1\rangle, \sqrt{1-\lambda}\,\langle m\rangle\right). \tag{43}$$

□

Remark 2 Note that the Hamilton-Jacobi equation can also be written as

$$\frac{\partial f(t,y,\vec{z})}{\partial t} + \mathcal{H}(t,y,\vec{z}) = 0,$$

where the effective Hamiltonian $\mathcal{H}(t,y,\vec{z}) \equiv \mathcal{T}(t,y,\vec{z}) + \mathcal{V}(t,y,\vec{z})$ is the sum of a kinetic term $\mathcal{T}(t,y,\vec{z}) = \frac{1}{2}[\nabla f(t,y,\vec{z})]^2$ and a potential term $\mathcal{V}(t,y,\vec{z}) = \frac{1}{2}\left[\langle x_1^2\rangle - \langle x_1\rangle^2\right] + \frac{1-\lambda}{2}\left[\langle m^2\rangle - \langle m\rangle^2\right]$.
Accordingly, the effective Lagrangian $\mathcal{L}(t,y,\vec{z}) = \mathcal{T}(t,y,\vec{z}) - \mathcal{V}(t,y,\vec{z})$ is the difference between the kinetic and potential terms.

Remark 3 Note that, as the potential $\mathcal{V}(t,y,\vec{z})$ is the sum of two variances—namely those of two order parameters that are self-averaging as $N \to \infty$, see Definition 6—we have that $\lim_{N\to\infty} \mathcal{V}(t,y,\vec{z}) = 0$ hence, in this asymptotic regime of interest in Statistical Mechanics, the above PDE describes a free motion whose trajectories are Galilean. For the sake of consistency with the brackets, clearly $\overline{x}_1 \equiv \langle x_1\rangle$ and $\overline{m} \equiv \langle m\rangle$.

To evaluate the explicit form of the Guerra action we need the following

Theorem 2 *The solution of the Hamilton-Jacobi PDE can be written as the Cauchy condition $f(t=0, y, \vec{z})$ plus the integral of the Lagrangian $\mathcal{L}(t, y, \vec{z})$ over time, namely*

$$f(t, y, \vec{z}) = f(t=0, y_0, \vec{z}_0) + \int_0^t dt'\, \mathcal{L}(t, y, \vec{z}). \tag{44}$$

Explicitly and in the thermodynamic limit the solution reads as

$$\begin{aligned} f(t, y, \vec{z}) = & \mathbb{E}_\eta \ln \left\{ 1 + \exp\left[(y(t) - v_y t)\sqrt{1-\lambda} + (\vec{z}(t) - \vec{v}_z t)\cdot \eta \right] \right\} \\ & + \frac{1-\lambda}{2} \langle m \rangle^2 t + \frac{1}{2} \langle x_1 \rangle^2 t. \end{aligned} \tag{45}$$

Corollary 3 *The expression of the free energy of the Battaglia-Treves model in the low-storage regime can be obtained simply by setting $t = -\beta$ and $y = 0$, $\vec{z} = \vec{0}$ in the expression of the Guerra action $f(t, y, \vec{z})$ provided in Eq. (45), namely*

$$\begin{aligned} A(\alpha = 0, \beta, \lambda) = f(t = -\beta, y = 0, \vec{z} = \mp) = \mathbb{E}_\eta \ln \left(1 + \exp\left(\beta\,(1-\lambda)\,\langle m \rangle + \beta \eta \cdot \langle x_1 \rangle\right)\right) \\ - (1-\lambda)\frac{\beta}{2}\langle m \rangle^2 - \frac{\beta}{2}\langle x_1 \rangle^2. \end{aligned} \tag{46}$$

It is a straightforward exercise to prove that, in order to extremize the above free energy w.r.t. the order parameters, their mean values $\langle x_1 \rangle$ and $\langle m \rangle$ have to obey the self-consistency Eqs. (34)–(35).

Proof To solve for the problem posed in Proposition 2, we are left with two calculations to perform: the evaluation of the Cauchy condition $f(t=0, y_0, \vec{z}_0)$ and the integral of the Lagrangian over time.
The evaluation of the Cauchy condition is straightforward since, at $t = 0$, neurons do not interact (see Eq. (40)) and we get

$$\begin{aligned} f(t=0, y_0, \vec{z}_0) = & \frac{1}{N} \mathbb{E}_\eta \ln \prod_i \sum_{s_i = 0,1} \exp\left(y_0 \sqrt{1-\lambda}\, s_i + \vec{z}_0 \cdot \eta_i s_i \right) \\ = & \mathbb{E}_\eta \ln \left(1 + \exp\left(y_0 \sqrt{1-\lambda} + \vec{z}_0 \cdot \eta \right) \right). \end{aligned} \tag{47}$$

Evaluating the integral of the Lagrangian over time is trivial too (as it is just the multiplication of the kinetic energy times the time as the potential is null in the thermodynamic limit) and returns

$$\int_0^t dt'\, \mathcal{L} = \frac{1-\lambda}{2} \langle m \rangle^2 t + \frac{1}{2} \langle x_1 \rangle^2 t. \tag{48}$$

Hence, plugging (47) and (48) into (44), we get

$$\begin{aligned} f(t, y, \vec{z}) = & \mathbb{E}_\eta \ln \left(1 + \exp\left((y(t) - v_y t)\sqrt{1-\lambda} + (\vec{z}(t) - \vec{v}_z t)\cdot \eta \right) \right) \\ & + \frac{1-\lambda}{2} \langle m \rangle^2 t + \frac{1}{2} \langle x_1 \rangle^2 t, \end{aligned} \tag{49}$$

where we used $y_0 = y(t) - v_y t$ and $\vec{z}_0 = \vec{z}(t) - \vec{v}_z t$ as the trajectories are Galilean. Also, the velocities are

$$v_y = \frac{df(t, y, \vec{z})}{dy} = \sqrt{1-\lambda}\,\langle m \rangle, \tag{50}$$

$$\vec{v}_z = \frac{df(t, y, \vec{z})}{d\vec{z}} = \langle x_1 \rangle. \tag{51}$$

Finally, the original free energy is recovered by setting $t = -\beta$ and $y = \vec{z} = 0$ in (49). □

3.2 Phase Diagram in the High-Storage Regime

In the high-load regime ($\alpha > 0$), we can not rely on the mechanical analogy as the concentration of measure argument used to neglect the potential in the Hamilton-Jacobi PDE no longer works.[6] Thus, here, we exploit the classical one-parameter interpolation based on stochastic stability [47], adapted to deal with neural networks, see e.g. [46, 58].

As standard in the high-storage investigation, we assume that a finite number of charts (actually just one) is retrieved and this map, say $\boldsymbol{\eta}^1$, plays as the signal, while the remaining ones (i.e., those with labels $\nu \neq 1$) play as quenched noise against the formation of the coherent state for the retrieval of η^1.

In this setting, in order to solve for the free energy, the idea is to introduce an interpolating parameter $t \in [0, 1]$ and an interpolating free energy $\mathcal{A}(t)$ such that, when $t = 1$, the interpolating free energy recovers the free energy of the original model, i.e., $\mathcal{A}(t = 1) = A(\alpha, \beta, \lambda)$, while, when $t = 0$, the interpolating free energy recovers the free energy of an "easy" one-body system (where neurons interact with a suitably-constructed external field but their activity is no longer affected by the state of the other neurons). The main theorem we use in this section is the fundamental theorem of calculus that plays as the natural bridge between these two extrema, as

$$A(\alpha, \beta, \lambda) = \mathcal{A}(t = 1) = \mathcal{A}(t = 0) + \int_0^1 ds \left[\frac{d}{dt}\mathcal{A}(t)\right]\Big|_{t=s} \tag{52}$$

$$\mathcal{A}(t) = \lim_{N\to\infty} \frac{1}{N} \mathbb{E}_\eta \ln \mathcal{Z}_N(t), \tag{53}$$

where $\mathcal{Z}_N(t)$ is the interpolating partition function that will be defined hereafter.
We emphasize that, for this analysis, it is by far convenient to deal with the representation of the Battaglia-Treves model in terms of a bipartite network with a layer equipped with chart cells as provided by Eq. 38 in Proposition 1.

Definition 11 *(Interpolating partition function)* The interpolating partition function $\mathcal{Z}_N(t)$ for the binary Battaglia-Treves model in the high-storage regime can be expressed as

$$\mathcal{Z}_N(t) = \sum_{\{s\}}^{2^N} \int d^d\mu(z_\mu)\ \exp\left(-\beta \mathcal{H}(t)\right), \tag{54}$$

where $\mathcal{H}(t)$ is the interpolation Hamiltonian (*vide infra*) and the Gaussian measure $d^d\mu(z_\mu) = \left(\prod_{\mu>1} \frac{d^d z_\mu}{\sqrt{2\pi}} \exp\left(-\frac{z_\mu^2}{2}\right)\right)$ is due to the application of the Hubbard-Stratonovich

[6] Indeed, while each single variance could still be negligible (vanishing at a rate $1/N$ in the thermodynamic limit), now we are summing over an extensive number of them (as K grows linearly with N in the high-storage setting, that is for $\alpha > 0$).

transformation to the Boltzmann-Gibbs measure, i.e.

$$\sum_{\{s\}}^{2^N} \exp\left(\frac{1}{2}\frac{\beta}{N}\left(\sum_{i,\mu>1}\eta_i^\mu s_i\right)^2\right) = \sum_{\{s\}}^{2^N}\int d^d\mu(x)\ \exp\left(\sqrt{\frac{\beta}{N}}\sum_{i,\mu>1}\eta_i^\mu\cdot z_\mu s_i\right). \tag{55}$$

The interpolating partition function is, in turn, based on the interpolating Hamiltonian $\mathcal{H}(t)$: the latter is a functional that combines two Hamiltonians, the *true* Hamiltonian (isolated by setting $t = 1$ in the interpolation procedure) and a *new* Hamiltonian (recovered by setting $t = 0$ instead). The new Hamiltonian has to be constructed *ad hoc* with two requisites: it should allow for an analytical treatment of its free energy (typically it has to be a sum of one-body models) and the effective field that it produces on the neuron s_i has to best mimic the true post synaptic potential (and this is achieved by using linear combinations of independent random variables).

Definition 12 *(Interpolating Hamiltonian)* Given an interpolating parameter $t \in [0, 1]$ and the real-valued function $\phi(t)$, the interpolation Hamiltonian $\mathcal{H}(t)$ reads as

$$-\beta\mathcal{H}_N(t) = \frac{t}{2}N\beta x_1^2 + \frac{t}{2}N\beta\,(1-\lambda)\,m^2 + \sqrt{t}\sqrt{\frac{\beta}{N}}\sum_{i,\mu>1}\eta_i^\mu s_i z_\mu + N\phi(t), \tag{56}$$

$$\begin{aligned} N\phi(t) = {} & \phi_1(t)\sum_{\mu>1}x_\mu^2 + \phi_2(t)\sum_\mu \rho_\mu x_\mu + \phi_3(t)\sum_i h_i s_i + \phi_4(t)\sum_i \eta_i^1 s_i \\ & + \phi_5(t)\sum_i s_i^2 + \phi_6(t)\sum_i s_i. \end{aligned} \tag{57}$$

The auxiliary functions $\phi_i(t)$ must obey $\phi_i(t = 1) = 0$; the specific functional form of $\phi(t)$, as well as its derivation, are provided in Appendix A.2.

Remark 4 Note that, in the integral representation of the partition function achieved by the Hubbard-Stratonovich transformation (see (55)), K hidden, selectively-firing variables z_μ naturally arise within the theory and, as a consequence, their related overlaps must be introduced as auxiliary order parameters, namely

$$p_{12} = \frac{1}{K}\sum_{\mu=1}^{K} z_\mu^1 z_\mu^2, \tag{58}$$

$$p_{11} = \frac{1}{K}\sum_{\mu=1}^{K} z_\mu^1 z_\mu^1. \tag{59}$$

However, these "chart overlaps" do not deserve a dedicated definition as they are dummy variables that will disappear in the final expression of the free energy.

Definition 13 *(Generalized average)* The generalized average $\langle\cdot\rangle_t$ is defined as

$$\langle\cdot\rangle_t = \mathbb{E}_\eta[\omega_t(\cdot)],$$

where $\omega_t(\cdot)$ is the Boltzmann average stemming from the interpolating Boltzmann factor $\exp(-\beta\mathcal{H}(t))$ defined in Eq. (56) together with the interpolating partition function $\mathcal{Z}_N(t)$ defined in Eq. (54), while the operator $\mathbb{E}_\eta$ still averages over the quenched maps.

In the following, if not otherwise specified, we refer to the generalized averages simply as $\langle\cdot\rangle$ in order to lighten the notation. By a glance at (52) we see that we have to evaluate the Cauchy condition $\mathcal{A}(t=0)$—that is straightforward as in $t=0$ the Gibbs measure is factorized over the neural activities—and integrate its derivative $\frac{d\mathcal{A}(t)}{dt}$—that is more cumbersome. Starting from the evaluation of the t-derivative of the interpolating free energy $\mathcal{A}(t)$, we state the next

Proposition 3 *Under the RS assumption and in the thermodynamic limit, the $t-$derivative of $\mathcal{A}$ can be written as*

$$\frac{d\mathcal{A}_{RS}}{dt} = -\frac{\beta}{2}(1-\lambda)\overline{m}^2 - \frac{\beta}{2}\overline{x}^2 - \frac{\alpha\beta}{2d}\left(\overline{p}_1\overline{q}_1 - \overline{p}_2\overline{q}_2\right). \tag{60}$$

Proof The proof works by direct calculation and by requiring that $\lim_{N\to\infty}\mathcal{P}(q_{11}) = \delta\left(q_{11}-\bar{q}_1\right)$ and the same for $q_{12},\ p_{11},\ p_{12}$. See Appendix A.2 for details.

For the Cauchy condition we can state the next

Proposition 4 *The Cauchy condition $\mathcal{A}(t=0)$, related to the expression of the interpolating free energy $\mathcal{A}(t)$ provided in Eq. (53), in the thermodynamic limit reads as*

$$\begin{aligned}\mathcal{A}(t=0) =& \frac{\alpha d}{2}\ln\frac{2\pi}{1-2\phi_1(0)} + \frac{\alpha}{2}\mathbb{E}_\rho\left(\frac{\phi_2^2(0)\rho^2}{1-2\phi_1(0)}\right)\\ &+ \mathbb{E}_{\eta^1}\int d\mu(z)\ln\left(1+\exp\left(\phi_3(0)z+\phi_4(0)\eta_i^1+\phi_5(0)+\phi_6(0)\right)\right).\end{aligned} \tag{61}$$

Proof The proof is a direct trivial calculation (as at $t=0$ there are no interactions among neural activities) with the usage of the self-averaging assumption of the order parameters. See Appendix A.2 for details.

These results merge together thanks to the fundamental theorem of calculus (see (52)) that we use to state the next

Theorem 3 *In the thermodynamic limit, the replica-symmetric quenched free-energy of the Battaglia-Treves model, equipped with McCulloch-Pitts neurons as defined in Eq. (2), can be expressed in terms of the (mean values of the) order parameters $\overline{m}, \overline{x}, \overline{q}_1, \overline{q}_2, \overline{p}_1, \overline{p}_2$ and of the control parameters α, β, λ, as follows*

$$\begin{aligned}A_{RS} =& -\frac{\beta}{2}(1-\lambda)\overline{m}^2 - \frac{\beta}{2}\overline{x}^2 - \frac{\alpha\beta}{2d}\left(\overline{p}_1\overline{q}_1-\overline{p}_2\overline{q}_2\right) - \frac{\alpha d}{2}\ln\left(1-\frac{\beta}{d}(\overline{q}_1-\overline{q}_2)\right)\\ &+ \frac{\alpha\beta}{2}\frac{\overline{q}_2}{1-\frac{\beta}{d}(\overline{q}_1-\overline{q}_2)} +\\ &+ \mathbb{E}_\eta\int d\mu(z)\ln\left(1+\exp\left(\beta(1-\lambda)\overline{m}+\beta\overline{x}\cdot\eta+\frac{\alpha\beta}{2d}(\overline{p}_1-\overline{p}_2)+\sqrt{\frac{\alpha\beta}{d}\overline{p}_2}\,z\right)\right).\end{aligned} \tag{62}$$

The chart cell's overlaps p_1 and p_2 can be substituted by their saddle-point values, obtained by differentiating A_{RS} w.r.t. q_1 and q_2, i.e. $\frac{\partial A^{RS}}{\partial q_1}=0$, $\frac{\partial A^{RS}}{\partial q_2}=0$, namely

$$\overline{p}_2 = \frac{\beta\overline{q}_2}{\left(1-\frac{\beta}{d}(\overline{q}_1-\overline{q}_2)\right)^2}, \tag{63}$$

$$\overline{p}_1 - \overline{p}_2 = \frac{d}{1-\frac{\beta}{d}(\overline{q}_1-\overline{q}_2)}. \tag{64}$$

Proof See Appendix A.2. □

Corollary 4 *By inserting the expressions for the chart overlaps* $\overline{p}_1$ *and* $\overline{p}_2$ *reported at the r.h.s. of eq.s* (63), (64) *within the expression of the replica-symmetric quenched free-energy* (62) *we obtain the r.h.s. of Eq.* (21) *and thus Theorem* 1 *is proved.*

Remark 5 For completeness, in Appendix A.1, we carry out the same analysis using the standard replica-trick: it is instructive to compare how the two procedures work. For instance, it is evident how the request of replica symmetry within the replica trick actually coincides with the self-averaging property within the interpolation scheme. Furthermore, by looking for the extremal points of such a free energy, it is a simple exercise to obtain the same self-consistencies for the order parameters reported in Eqs. (22)–(24).

4 From Rats on a Circular Track to Bats in the Tunnel

The model so far investigated was originally developed to mimic real neurons in rats exploring (or rather foraging in) small boxes, or short tracks, each of which could be idealized as being represented by an unrelated chart, in which each neuron has a place field of standard size. We now turn it into a model inspired by recent experiments with bats [48], of the neural representation of a single extended environment, expressed by place fields of widely different sizes, from very long to very short. For simplicity, we stick again to 1D environments.

Indeed, recent experiments recording CA1 place cells either in rats running on long tracks [60, 61] or bats flying in long tunnels [48] have evidenced that individual place cells can have multiple place fields, of remarkable variability in size and peak firing rate. In the following, extending the formalism developed above, we consider a mathematical model which assigns this variability to distinct portions of the environment, representing them with a sequence of gross, or fine-grained, charts. Alongside *traditional* place cells, which participate in every chart, and hence have multiple place fields of variable width, the model envisages *chart* cells, which are active only when the current position is within their chart.

Instead of covering the whole space with a single map, or chart, we combine a number $K = \alpha N$ of maps η^μ to split the embedding space $I_L = [0, L]$ (the one-dimensional interval that represents the environment) in patches, by distributing the place field's centers at different locations $\overline{r}^\mu \in I_L$ and assigning them different sizes, denoted by the parameters σ^μ (*vide infra*).

The characteristic width of the coherent states that are reconstructed by the model (previously indicated by δ), beyond being field-dependent now, has also to be kept much smaller than the length of the tunnel, i.e. $\delta \ll L$ in order to reconstruct the position of the animal with good accuracy.

Secondly, driven by the duality (between the original place cells, now with multiple fields along L, and the chart cells, acting as hidden units), we show that by suitably coupling these hidden variables we can turn the original model for chart reconstruction into a behavioral model, namely into a model for spatial navigation too, both externally (see Sect. 4.1) or internally (see Sect. 4.2) driven.

Indeed, the environment sensed by the animal is now dynamical (i.e. it is no longer the same constant space perceived when confined in a small cage) hence, in Sect. 4.1, we mimic the input perceived by the bat by introducing a time-dependent external field, that is moving at its same speed and guides the animal through the tunnel: see Fig. 9.

In Sect. 4.2, instead, we assume that the bat has learnt how place fields are dislocated within such a dynamical environment and we show how, by introducing a simple coupling among

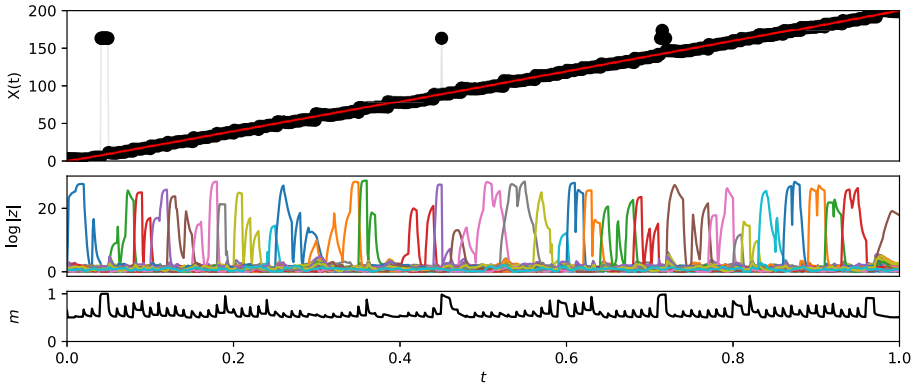

Fig. 9 Simulation of the bat traveling in the tunnel driven by the external field $h(t)$. *Top* The motion of the bat follows the field, as expected, almost everywhere apart from a small number of points where the model fails to reconstruct the position (i.e. the upper black points away from the red straight line). *Middle* Chart cell firing patterns during the motion of the animal in the tunnel are highly correlated with the position of the animal. *Bottom* The overall activity of the model is also shown as a function of time and the points where chart reconstruction fails correspond to higher level of neural activity. The simulation features $N = 8000$ visible neurons $\mathbf{s}$ and $K = 40$ chart cells $\mathbf{z}$, at a temperature of $\beta^{-1} = 10^{-3}$

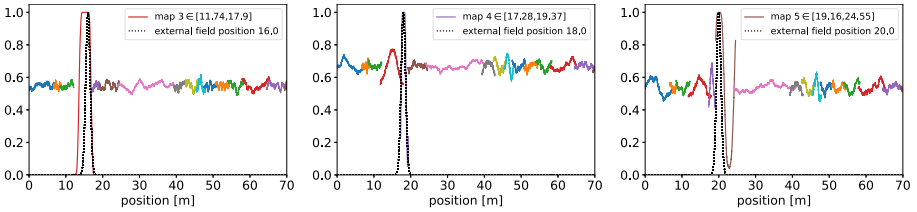

Fig. 10 Three snapshots of the reconstruction process that allows the network to recognize the position of the external Gaussian field $\mathbf{h}$ (indicated with the dotted black line). The model activity is shown for each chart (for a total of $K = 40$ charts) distributed along the tunnel at different positions

their corresponding chart cells, such a minimal generalization of the Battaglia-Treves model can account for autonomous motion within the environment. This generalization can be viewed as a variant of the Battaglia-Treves model, driven more by mathematical considerations than by neuroscience. In fact, its primary goal is to examine the robustness and adaptability of the reference framework across different scenarios.

4.1 Numerical Experiments Part One: Externally-Driven Motion

We assume that the space along the tunnel is uniformly covered with maps, but each map has a different width σ^μ, with distribution $\rho(\sigma^\mu)$. According with experiments [48], we assume that $\rho(\sigma^\mu)$ is *log-normal*, while the density of neurons in each chart, $\rho\left(r_i^\mu|\bar{r}^\mu;\sigma^\mu\right)$, is uniform. These assumptions are summarized as

$$\rho\left(\sigma^\mu|\bar{r}^\mu,\sigma\right) = \log\mathcal{N}(\bar{r}^\mu,\sigma), \tag{65}$$

$$\rho\left(r_i^\mu|\bar{r}^\mu;\sigma^\mu\right) = \mathcal{U}_{I^\mu}, \tag{66}$$

where $I^\mu = \left[\bar{r}^\mu - \frac{\sigma^\mu}{2}, \bar{r}^\mu + \frac{\sigma^\mu}{2}\right]$ is the interval centered around $\bar{r}^\mu$ and width σ^μ; the parameter σ appearing in Eq. (65) is a free parameter of the model. Note that each map is periodic in its domain, I^μ, given the definition of the map coordinates

$$\eta^\mu(r|\bar{r}^\mu, \sigma^\mu) = \left(\cos\left(\frac{2\pi}{\sigma^\mu}(r - \bar{r}^\mu)\right), \sin\left(\frac{2\pi}{\sigma^\mu}(r - \bar{r}^\mu)\right)\right), \tag{67}$$

where the chart centers $\bar{r}^\mu$ are distributed according to a prior $\mathcal{P}(\bar{r}^\mu)$. Hereafter, we call t the time[7] such that the sensory input of the animal in the tunnel is represented by a time dependent external field $h(t)$, which is added to the bare Hamiltonian H of the model as $H(t) = H + \sum_i h_i(t)s_i$. This new term in the energy, i.e. $\sum_i h_i(t)s_i$, accounts for the motion of the bat and it produces a bias in the update equations for the neural dynamics (see the stochastic process coded in Eq. (7)), which can be explicitly written as

$$P(\mathbf{z}|\mathbf{s}) = \prod_\mu \mathcal{N}(\sqrt{N}x_\mu(\mathbf{s}), \mathbb{I}_d\, \beta^{-1}), \tag{68}$$

$$P(\mathbf{s} = 1|\mathbf{z}) = \prod_i \sigma\left(\frac{\beta}{\sqrt{N}}\sum_\mu \eta_i^\mu \cdot z_\mu + \beta h_i\right). \tag{69}$$

Making explicit the update rule for the visible neurons $\boldsymbol{s}$ (those that perceive the external field) we write

$$s_i^{t+1} = \Theta\left(\sigma\left(\beta\mathcal{V}_i(\mathbf{s}^t)\right) - u_i\right), \quad u_i \in \mathcal{U}_{[0,1]}, \tag{70}$$

$$\mathcal{V}_i(\mathbf{s}^t) = \frac{1}{2}\sum_j J_{ij}s_j^t + h_i(t), \tag{71}$$

where, in order to sample from a Markov process, $u_i \in \mathcal{U}_{[0,1]}$ is a random number uniformly distributed in the interval [0, 1], $\Theta(x)$ is the Heaviside step function and $\mathcal{V}_i(\mathbf{s}^t)$ is the overall post-synaptic potential (which depends on the synaptic matrix (5) and is comprehensive of the external field $h_i(t)$ too) acting on the i-th neuron s_i.

The explicitly time-dependent field, $h_i(t)$ enables the movement of the bump of neural activity representing the position of the bat along the tunnel if the external field follows the bat's motion. For the sake of simplicity, we assume its motion to be Galilean with constant velocity v, in such a way that $h(t)$ can be modeled as a Gaussian bump traveling with constant velocity v and represented in the place cells coordinate system as

$$h_i(t) = \exp\left(-\frac{(r_i^\nu - vt)^2}{2\delta_0^2}\right),$$

with δ_0^2 the size of the bump; r_i^ν is the coordinate relative to the $\nu-$th map, η^ν; the index ν is chosen with the following criterion (see Fig. 10): a collection of external fields $h^\mu(t)$ is computed for each map μ, then the map with the highest average activity (that is, the map with the highest overlap with the external field) is chosen and the relative index is indicated as ν.[8]

[7] In this context, t is meant as a time and should not be confused with the fictitious-time variable appearing in Sect. 3.1 or with the interpolating parameter appearing in Sect. 3.2.

[8] Note that the quantity $\frac{1}{N}\sum_i h_i(t)$ can be approximated, in the thermodynamic limit $N \to \infty$, by the integral

$$\int_{\bar{r}^\nu - \frac{\sigma^\nu}{2}}^{\bar{r}^\nu + \frac{\sigma^\nu}{2}} dr\, h(r,t) \sim \sqrt{2\pi}\frac{\delta_0}{\sigma^\nu}$$

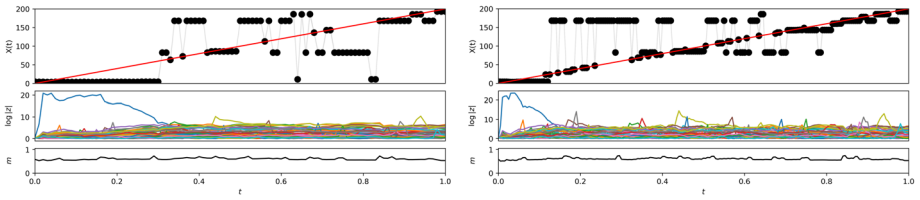

Fig. 11 Examples of failures in reconstruction of the position of the bat along the tunnel during its motion, caused by a low ratio $R = M/M_0$, respectively $R = 1$ (left panel) or $R = 2$ (right panel)

The velocity v of the traveling bump introduces a new time scale in the dynamics of the system; this time scale, written as $\tau = L/(M_0 v)$, is the time that the bump has to spend covering the length of the tunnel L, divided by the number M_0 of discrete-time realizations of the field itself (hence, in the time window $M_0\tau$, the Gaussian bump moves in a given number of steps M_0 from the origin of the tunnel to its end). In order to allow the network to stabilize within a given fixed point of the dynamics (70) at each new position at discrete time t, i.e. $r_0 + vt$, we require the number of steps of the dynamics to be $M \gg M_0$. In our simulations we inspect in detail the case where $M_0 = 100$ and $M = 1000$, with a neural noise level fixed at $\beta = 100$: see Figs. 9 and 10.

We run 1000 simulations of this dynamics (each one with a different realization of the chart's representation) and results are shown in Fig. 9, for a ratio of the two involved numbers of steps $R = M/M_0 = 10$. In this setting, the network is able to successfully follow the external field $\mathbf{h}(t)$ and, as the bat flies along the tunnel, all the various place cells sequentially fire. To demonstrate the impact of R on the model's ability to accurately retrieve the sequence of coherent states corresponding to the bat's motion along the tunnel, we performed similar simulations with $R = 1$ (left panel) and $R = 2$ (right panel) in Fig. 11.

4.2 Numerical Experiments Part Two: Self-driven Motion

In the present setting, place cells reconstruct the environment and activate the corresponding chart cells as the animal goes through their several place fields. While simple place cells thus remap the familiar environment, they can not easily drive the movement of the animal within the environment [2] (simply because of the multiplicity of their place fields). However, a coupling between consecutive chart cells -say μ and $\mu + 1$- naturally accounts for a minimal model of bat's dynamics: this coupling can be easily added to the Hamiltonian of the original Battaglia-Treves model in its integral representation.
As we distributed the charts sequentially in the tunnel, such that the bat encounters them one after the other (i.e., $\mu = 1 \to \mu = 2 \to, ... \to \mu = K$), we generalize the Battaglia-Treves Hamiltonian by introducing a coupling between chart cells (i.e., $J_z \sum_{\mu}^{K} z_\mu z_{\mu+1}$) as:

$$H_N(s, z|\boldsymbol{\eta}) = -\frac{1}{\sqrt{N}} \sum_{i}^{N} \sum_{\mu}^{K} \eta_i^\mu s_i z_\mu + \frac{(\lambda - 1)}{2N} \sum_{i,j}^{N,N} s_i s_j + J_z \sum_{\mu}^{K} z_\mu z_{\mu+1}, \tag{72}$$

(as it is the integral on a finite domain of a Gaussian distribution whose center is vt within the interval $[\bar{r}^\nu - \frac{\sigma^\nu}{2}, \bar{r}^\nu + \frac{\sigma^\nu}{2}]$ and whose standard deviation reads $\delta_0 \ll \sigma^\nu$) that is an intensive quantity (as it only depends on the ratio of two intensive parameters δ_0, σ^ν), such that, overall, the external field contribution to the bare Hamiltonian H is *extensive* in N, as it should.

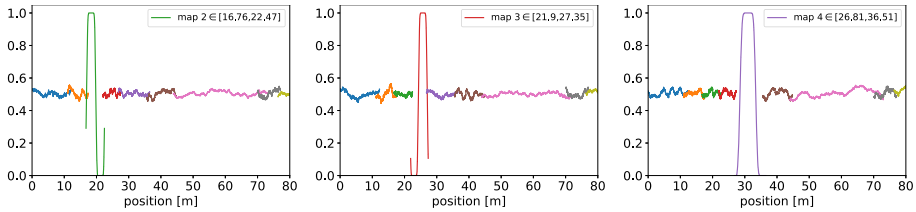

Fig. 12 Three consecutive snapshots of the activity of chart cells simulated during the dynamics generated according to Eq. 72. Note that, in the first panel, the bat retrieves map 2 and thus the corresponding chart cell fires. This, in turn, drives the firing of the next chart cell, related to retrieval of map 3 and so on

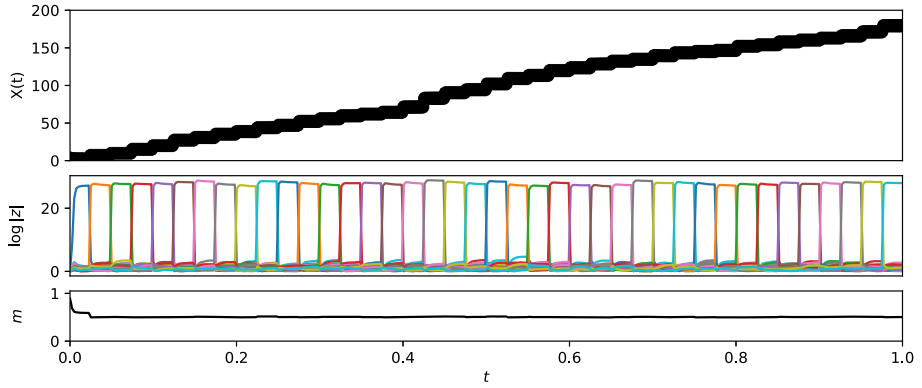

Fig. 13 Simulation of the network moving in the tunnel driven by the coupling between chart cells $z_\mu z_{\mu+1}$ (i.e. the extension of the Battaglia-Treves model provided in (72)). *Top* The position of the animal can now be thought of as moved along by its neural network and no external field $h(t)$ is required. *Middle* Chart cells drive the motion of the animal by activating each other in a sequence; but—note— at a constant rate per cell, not at a constant bat speed. *Bottom* The overall activity of the model remains around $m \sim 0.5$ as expected. The simulation features $N = 8000$ visible neurons $\mathbf{s}$ and $K = 40$ chart cells $\mathbf{z}$, at a temperature of $\beta^{-1} = 10^{-3}$ and with $J_z \sim O(1)$

in the standard hetero-associative way introduced by Amit [62] and Kosko [63].[9]
We end up with a neural network that *knows* the environment and, thus, *drives* the bat correctly along the tunnel: as the charts are crossed, they activate the corresponding neurons $\boldsymbol{z}$, one after the other. They guide the retrieval by the neurons in the visible layer $\boldsymbol{s}$, see Figs. 12 and 13. The update rule of the chart cell simply reads as

$$P(\mathbf{z}^{t+1}|\mathbf{z}^t, \mathbf{s}^t) = \prod_\mu \mathcal{N}(J_z z^t_{\mu+1} + \sqrt{N}\,\overline{x}^\mu(\mathbf{s}^t), \mathbb{I}_d \beta^{-1}). \tag{73}$$

5 Conclusions

Research on place cells, grid cells and, in general, the way hippocampus stores spatial representations of the environment is obviously a central theme in Neuroscience (see e.g. [48,

[9] See also [64, 65] for a recent re-visitation of the emergent computational capabilities of those networks analyzed with the techniques exploited in this paper.

66–68]). Since the AGS milestone in the mid eighties [4, 32, 50], the statistical mechanics of spin glasses played a central role in the mathematical modeling of the collective, emergent features shown by large assemblies of neurons, hence it is not by surprise that computational and analytical investigations on place cells along statistical mechanics lines have been extensive (see e.g. [6, 9]). However, previous studies in this field based their findings on methodological tools (e.g. the so-called replica trick [4, 29, 32]) that, from a mathematical perspective, are heuristic [55] thus raising the quest for confirmation by an independent, alternative, approach that we pursue here. In these regards, since the pioneering Guerra's works on spin glasses [39, 47], interpolation techniques quickly became a mathematical alternative to the replica trick also in neural networks (see e.g. [15, 16, 45, 46, 58, 69]) and these have been the underlying methodological leitmotif of the present paper too[10]: indeed one aim of the present research is to provide a rigorous picture about CANN behavior.

In particular, we studied analytically a variant of the Battaglia-Treves model [10] for chart storage and reinstatement, namely a model that describes the collective emerging capabilities of place cells in rats exploring small boxes (where maps are encoded by uniform coverage of similarly sized place fields). Specifically, our variant of the network is equipped with McCulloch-Pitts neurons [4] and we study it both in the low-storage (where the number of stored charts scales sub-linearly w.r.t. the network size) and high-storage (where the number of stored charts scales linearly w.r.t. the network size) regimes [32]. The former has been inspected by adapting the Hamilton-Jacobi (two-parameters) interpolation [39, 40], while the latter has been tackled by adapting the stochastic stability (one-parameter) interpolation [46, 58]. Further, we have also derived the same results independently via the replica trick in the high-storage limit (these calculations are reported in the Appendix for the high storage regime but turn out to be coherent also with the low storage picture reported in the main text simply by setting $\alpha = 0$). Confined to the replica symmetric level of description (fairly standard in neural networks [28]), as expected, we have obtained full agreement among the results stemming from these methodologies and those available in the Literature. Further, extensive Monte Carlo simulations for finite-size-scaling examined the network behavior away from the thermodynamic limit and provided numerical confirmation of the picture obtained under replica symmetry assumption. Interestingly, working computationally at finite sizes, we also observed the existence of spontaneous transitions among maps finding behaviors very close to those described by Monasson and Rosay [6].

Further, since simulations allow us to inspect scenarios that are too cumbersome for a rigorous analytical treatment, in order to test the robustness of the Battaglia-Treves model w.r.t. the underlying distributions of place fields, we numerically simulated an extension of the model that accounts for hippocampal cells in bats flying in large tunnels [48]. Now, nor place fields are uniformly distributed neither their width is constant: by replacing the original uniform distributions for place field and chart width with those observed in the experiments (that prescribe an exponential distribution of place cells along the tunnel and a log-normal distribution of their width), we showed that the Battaglia-Treves model still succeeds in reproducing the reconstruction of a map by a bat flying within such an environment.

Finally, we looked for a dual representation of the Battaglia-Treves model by exploiting its integral representation. In fact, following its Hebbian structure [34–38], we expect that it exhibits an equivalent (from a mathematical modelling perspective) representation in terms of a bipartite network. The two layers are interpreted as, respectively, visible and hidden. Neurons in the visible layer correspond to those of the original model, while neurons in the

[10] For the sake of completeness, we point out that rigorous methods in the statistical mechanical formalization of neural networks are obviously not confined to Guerra interpolation and we just mention the books [70, 71] as classical milestones in this field, should the reader be interested.

hidden layer are shown to play as chart cells, namely they are selectively firing a stored map or chart, also, the weights among these layers coincide with the chart elements.
This dual representation inspired us to propose a minimal generalization of the model, introducing a coupling among the hidden neurons. This interaction allows the network, when situated in place field μ to transition in a Markovian manner toward the subsequent place field $\mu+1$. Such a feature encapsulates an "anticipation mechanisms", enabling the animal to predict the upcoming place $\mu+1$ while being in place field μ. Consequently, the addition of intra-layer interactions transforms the original model into a behavioral framework, achieving a level of functionality that would be less straightforward to realize by merely adjusting the synaptic weights of the initial model.

Acknowledgements A.B. and M.S.C. acknowledge *Statistical Mechanics of Learning Machines: from algorithmic and information theoretical limits to new biologically inspired paradigms* funded by the Italian Ministry of University and Research (MUR) under the PRIN 2022 (cod. 20229T9EAT) in the framework of European Union - Next Generation EU (CUP: J53D2300364001) E.A. acknowledges financial support from PNRR MUR project PE0000013-FAIR and from Sapienza University of Rome (RM12117A8590B3FA, RM12218169691087). A.B. and E.A. are members of the GNFM-INdAM and INFN (Sezione di Roma1) which are acknowledged too.

Funding Open access funding provided by Università degli Studi di Roma La Sapienza within the CRUI-CARE Agreement.

A Detailed Calculations for the Analyzed Battaglia-Treves Model

A.1 Free Energy Calculation Via the Replica Trick

For the sake of methodological completeness, in this appendix we provide a derivation for the free energy of the model also with the old fashioned replica trick, to guarantee that results do coincide with those obtained via the Guerra routes in the main text.
As the Battaglia-Treves model is ultimately a spin glass model, it is rather natural to tackle the evaluation of its free energy by the formula

$$A(\alpha,\beta,\lambda)=\lim_{N\to\infty}\lim_{n\to 0}\frac{Z_N(\beta,\lambda)-1}{nN}, \tag{74}$$

thus with the computational reward of bartering the evaluation of the logarithm of the partition function with its momenta: the price to pay for this reward is the *blind* analytical extension toward the limit of zero replicas $n\to 0$ that must be performed (under the assumptions that the limits commute, i.e. $[\lim_{N\to\infty},\lim_{n\to 0}]=0$).[11]

[11] While this procedure has been proved to be correct for the harmonic oscillator of spin glasses, that is the Sherrington-Kirkpatrick model [72], by constraining the Parisi expression for its free energy among the Guerra [47] and the Talagrand [73] upper and lower bounds (and it is also true that, for that model, $[\lim_{N\to\infty},\lim_{n\to 0}]=0$ [74]), at present, in neural networks we do not have a general prescription for

As a consequence we are left to evaluate the moments of the partition function, often called *replicated partition function*, Z^n that reads

$$Z^n = \sum_{s^a} \exp\left[\frac{\beta}{2N}\sum_{\mu,a}\left(\sum_i \eta_i^\mu s_i^a\right)^2 - \frac{\beta}{2N}(\lambda-1)\sum_a\left(\sum_i s_i^a\right)^2\right]; \tag{75}$$

As standard we assume the network lies in the basin of attraction of one map, say η_i^1, hence we split the signal term (provided by that map) from the background noise (resulting from all the other maps $\nu \neq 1$) by rewriting the replicated partition function as

$$Z^n = \sum_{s^a} \exp\left[\frac{\beta}{2N}\sum_a\left(\sum_i \eta_i^1 s_i^a\right)^2 + \frac{\beta}{2N}\sum_{\mu>1,a}\left(\sum_i \eta_i^\mu s_i^a\right)^2 - \frac{\beta}{2N}(\lambda-1)\sum_a\left(\sum_i s_i^a\right)^2\right]. \tag{76}$$

After inserting the order parameters via the integral representations of the delta functions [4] we end up with

$$\begin{aligned} Z^n = \sum_{s^a}\int & dm^a dr^a d^2x_{\mu a} d^2t_{\mu a} d^2\overline{x}_a d^2\overline{t}_a \\ & \exp\left(iN\sum_a r_a m_a - i\sum_a r_a\sum_i s_i^a + iN\sum_a \overline{t}_a\overline{x}_a - i\sum_a \overline{t}_a\sum_i \eta_i^1 s_i^a + \right. \\ & + i\sum_{a,\mu>1} t_{\mu a}x_{\mu a} - \frac{i}{\sqrt{N}}\sum_{a,\mu>1} t_{\mu a}\sum_i \eta_i^\mu s_i^a + \frac{\beta N}{2}\sum_a \overline{x}_a^2 \\ & \left. + \frac{\beta}{2}\sum_{a,\mu>1} x_{\mu a}^2 - \frac{\beta}{2}(\lambda-1)N\sum_a m_a^2\right) \end{aligned} \tag{77}$$

The expectation over the quenched noise $\eta_{\mu>1}$ is easily computed and it reads

$$\mathbb{E}_{\eta_{\mu>1}} \exp\left[-\frac{i}{\sqrt{N}}\sum_{a,\mu>1} t_{\mu a}\sum_i \eta_i^\mu s_i^a\right] = \exp\left[-\frac{1}{2Nd}\sum_{i,\mu>1}\left(\sum_a t_{\mu a} s_i^a\right)^2 + \mathcal{O}(N^{-2})\right]. \tag{78}$$

We therefore introduce the replica overlap order parameter q_{ab} as

$$q_{ab} = \frac{1}{N}\sum_i s_i^a s_i^b. \tag{79}$$

The quenched replicated partition function then reads

$$\begin{aligned} \langle Z^n\rangle = & \int dm_a dr_a d^2\overline{x}_a d^2\overline{t}_a dq_{ab} dp_{ab} \\ & \exp\left(iN\sum_a r_a m_a + iN\sum_a \overline{t}_a\overline{x}_a + iN\sum_{ab} p_{ab}q_{ab} + \frac{\beta N}{2}\sum_a \overline{x}_a^2 - N\frac{\beta}{2}(\lambda-1)\sum_a m_a^2\right. \end{aligned}$$

broken replica theories a' la Parisi nor we can be sure (in the mathematical sense) that the formula we obtain by the replica trick are ultimately correct, hence the need for alternative approaches.

$$+K\ln\int d^2x_{\mu a}d^2t_{\mu a}\exp\left(i\sum_a t_a x_a-\frac{1}{2d}\sum_{ab}q_{ab}t_at_b+\frac{\beta}{2}\sum_a x_a^2\right)$$
$$+N\left\langle\ln\sum_{s^a}\exp\left(-i\sum_a r_a s^a-i\sum_a\bar{t}_a\eta_i^1 s^a-i\sum_{ab}p_{ab}s^a s^b\right)\right\rangle\Bigg) \tag{80}$$

The Gaussian integrals in the third line of the latter equation can be computed and give

$$K\ln\int d^2x_{\mu a}d^2t_{\mu a}\exp\left(i\sum_a t_a x_a-\frac{1}{2d}\sum_{ab}q_{ab}t_at_b+\frac{\beta}{2}\sum_a x_a^2\right)$$
$$=-\frac{Kd}{2}\ln\det\left(\delta_{ab}-\frac{\beta}{d}q_{ab}\right). \tag{81}$$

In the thermodynamic limit, within the replica trick, the free energy of the Battaglia-Treves in the high storage regime of charts is given by the following extremal condition

$$\lim_{n\to0}\lim_{N\to\infty}\frac{\ln\langle Z^n\rangle}{n}=\lim_{n\to0}extr_{m,r,q,p,\bar{t},\bar{x}}\frac{\Phi(m_a,r_a,q_{ab},p_{ab},\bar{t}_a,\bar{x}_a)}{n}, \tag{82}$$

where

$$\Phi=iN\sum_a r_a m_a+iN\sum_a\bar{t}_a\bar{x}_a+iN\sum_{ab}p_{ab}q_{ab}+\frac{\beta N}{2}\sum_a\bar{x}_a^2-N\frac{\beta}{2}(\lambda-1)\sum_a m_a^2+$$
$$-\frac{Pd}{2}\ln\det\left(\delta_{ab}-\frac{\beta}{d}q_{ab}\right)$$
$$+N\left\langle\ln\sum_{s^a}\exp\left(-i\sum_a r_a s^a-i\sum_a\bar{t}_a\eta_i^1 s^a-i\sum_{ab}p_{ab}s^a s^b\right)\right\rangle \tag{83}$$

We proceed with the *RS* ansatz [72]:

$$q_{ab}=q_1\delta_{ab}+q_2(1-\delta_{ab}) \tag{84}$$
$$p_{ab}=\frac{i}{2}p_1\delta_{ab}+\frac{i}{2}p_2(1-\delta_{ab}) \tag{85}$$
$$m_a=\overline{m} \tag{86}$$
$$\bar{t}_a=i\bar{t} \tag{87}$$
$$\bar{x}_a=i\bar{x} \tag{88}$$
$$r_a=i\frac{\beta}{2}r. \tag{89}$$

The *RS* Φ functional then reads

$$\lim_{n\to0}\frac{\Phi^{RS}}{n}=-\frac{\beta}{2}r\overline{m}-\bar{t}\bar{x}+\frac{1}{2}(p_1q_1-p_2q_2)+\frac{\beta}{2}\bar{x}^2-\frac{\beta}{2}(\lambda-1)\overline{m}^2$$
$$-\lim_{n\to0}\frac{1}{n}\left(\frac{\alpha d}{2}\ln\det\left(\mathbf{1}-\frac{\beta}{d}\mathbf{q}\right)_{RS}-G_{RS}\right), \tag{90}$$

where G_{RS} is

$$G_{RS}=\left\langle\ln\sum_{s^a}\exp\left(-i\sum_a r_a s^a-i\sum_a\bar{t}_a\eta s^a-i\sum_{ab}p_{ab}s^a s^b\right)\right\rangle_\eta. \tag{91}$$

The latter reads

$$G_{RS} = n\left\langle \int d\mu(z) \ln\left(1+\exp\left(\frac{\beta}{2}r+\bar{t}\eta+\sqrt{p_2}z+\frac{1}{2}(p_1-p_2)\right)\right)\right\rangle_\eta, \tag{92}$$

The determinant is

$$\ln\det\left(\mathbf{1}-\frac{\beta}{d}\mathbf{q}\right)_{RS} = n\left(\ln\left(1-\frac{\beta}{d}(q_1-q_2)\right)-\frac{\frac{\beta}{d}q_2}{1-\frac{\beta}{d}(q_1-q_2)}\right)+\mathcal{O}(n^2). \tag{93}$$

Summing all the contributions we end up with

$$\begin{aligned} A_{RS} = &-\frac{\beta}{2}r\overline{m}-\bar{t}\bar{x}-\frac{1}{2}(p_1q_1-p_2q_2)+\frac{\beta}{2}\bar{x}^2-\frac{\beta}{2}(\lambda-1)\overline{m}^2-\frac{\alpha d}{2}\ln\left(1-\frac{\beta}{d}(q_1-q_2)\right)\\ &-\frac{\alpha\beta}{2}\frac{q_2}{1-\frac{\beta}{d}(q_1-q_2)}\\ &+\left\langle\int d\mu(z)\ln\left(1+\exp\left(\frac{\beta}{2}r+\bar{t}\eta+\sqrt{p_2}z+\frac{1}{2}(p_1-p_2)\right)\right)\right\rangle_\eta. \end{aligned} \tag{94}$$

To get rid off the auxiliary integration variables we start to extremize the free energy, namely we perform

$$\frac{\partial A^{RS}}{\partial\overline{m}}=0,$$
$$\frac{\partial A^{RS}}{\partial\bar{x}}=0,$$

from which we obtain

$$\frac{1}{2}r=(1-\lambda)\overline{m} \tag{95}$$
$$\bar{t}=\beta\bar{x}. \tag{96}$$

By inserting the above expressions in the free energy expression at the r.h.s. of Eq. (94), the latter reads

$$\begin{aligned} A_{RS} = &-\frac{\beta}{2}(1-\lambda)\overline{m}^2-\frac{\beta}{2}\bar{x}^2-\frac{1}{2}(p_1q_1-p_2q_2)-\frac{\alpha d}{2}\ln\left(1-\frac{\beta}{d}(q_1-q_2)\right)\\ &+\frac{\alpha\beta}{2}\frac{q_2}{1-\frac{\beta}{d}(q_1-q_2)}+\\ &+\mathbb{E}_\eta\int d\mu(z)\ln\left(1+\exp\left(\beta(1-\lambda)\overline{m}+\beta\bar{x}\eta+\frac{1}{2}(p_1-p_2)+\sqrt{p_2}\,z\right)\right). \end{aligned} \tag{97}$$

Finally we eliminate also the auxiliary parameters p_1 and p_2 by differentiating A_{RS} w.r.t. q_1 and q_2, i.e. $\frac{\partial A^{RS}}{\partial q_1}=0$, $\frac{\partial A^{RS}}{\partial q_2}=0$,

$$\frac{1}{2}p_2=\frac{\alpha\beta^2}{2d}\frac{q_2}{\left(1-\frac{\beta}{d}(q_1-q_2)\right)^2}, \tag{98}$$

$$\frac{1}{2}(p_1-p_2)=\frac{\frac{\alpha\beta}{2}}{1-\frac{\beta}{d}(q_1-q_2)}. \tag{99}$$

By inserting these expression in the r.h.s. of Eq. (97) we end up with

$$A^{RS}(\alpha,\beta,\lambda) = -\frac{\beta}{2}(1-\lambda)\overline{m}^2 - \frac{\beta}{2}\overline{x}^2 - \frac{\alpha\beta}{2}\frac{q_1 - \frac{\beta}{d}(q_1-q_2)^2}{\left(1-\frac{\beta}{d}(q_1-q_2)\right)^2} - \frac{\alpha d}{2}\ln\left(1-\frac{\beta}{d}(q_1-q_2)\right)$$
$$+\frac{\alpha\beta}{2}\frac{q_2}{1-\frac{\beta}{d}(q_1-q_2)} +$$
$$+\left\langle \int d\mu(z)\ln\left(1+\exp\left(\beta(1-\lambda)\overline{m}+\beta\overline{x}\eta+\beta\frac{\frac{\alpha}{2}+\sqrt{\frac{\alpha q_2}{d}}z}{1-\frac{\beta}{d}(q_1-q_2)}\right)\right)\right\rangle_\eta , \tag{100}$$

that is the same expression we obtained by the Guerra's route based on stochastic stability in the high storage limit (and that collapses to the low-storage expression obtained by the Guerras' route based on the mechanical analogy simply by forcing α to be zero): see Theorem 1 in the main text.

A.2 Free Energy Calculation Via the Interpolation Method

Hereafter we prove the results reported in Propositions 3 and 4 in Theorem 3 and in Corollary 1.

Proof of Proposition 3 Let us start with a slightly more general definition of the interpolating Hamiltonian $\mathcal{H}(t)$ provided in Definition 12:

$$-\beta\mathcal{H}(t) = \frac{a_1(t)}{2}N\beta x_1^2 + \frac{a_2(t)}{2}N\beta(1-\lambda)m^2 + a_3(t)\sqrt{\frac{\beta}{N}}\sum_{i,\mu>1}^{N,K}\eta_i^\mu s_i z_\mu + N\phi(t), \tag{101}$$

where we kept the specific interpolation coefficient as general functions $a_i(t),\ i\in(1,2,3)$, and let us perform the evaluation of the t-derivative of $\mathcal{A}(t)$, this is computed as

$$\mathcal{A}'(t) = \lim_{N\to\infty} N^{-1}\left(\frac{a_1'}{2}N\beta\langle x_1^2\rangle + \frac{a_2'}{2}N\beta\,(1-\lambda)\,\langle m^2\rangle + a_3'\sqrt{\frac{\beta}{N}}\sum_{i,\mu>1}\langle\eta_i^\mu s_i x_\mu\rangle + N\langle\phi'\rangle\right), \tag{102}$$

where we used the superscript $'$ to indicate the derivative with respect to t in order to lighten the notation.
In order to evaluate $\langle\eta_i^\mu s_i x_\mu\rangle$ we exchange the expectation over the maps $\mathbb{E}_\eta$ with a d-variate Gaussian measure with the same mean and variance: $\mathcal{N}(0,\mathbf{1}_d/d)$. This is done by noticing that $\langle\eta_i^\mu s_i x_\mu\rangle$ can be written as

$$\sum_{i,\mu>1}\langle\eta_i^\mu s_i x_\mu\rangle = \sum_{i,\mu>1}\mathbb{E}_\eta\left[\eta_i^\mu\omega\left(s_i x_\mu\right)\right] \sim \sum_{i,\mu>1}\mathbb{E}_z\left(z_i^\mu s_i x_\mu\right),$$

with $z_i^\mu\sim\mathcal{N}(0,\mathbf{1}_d/d)$ and noticing that the introduced error, computed as $|\mathbb{E}_\eta\left(\eta_i^\mu s_i x_\mu\right) - \mathbb{E}_z\left(z_i^\mu s_i x_\mu\right)|$, vanishes in the thermodynamic limit by virtue of the Stein lemma [54]. The variance of order $1/d$ is required by the condition $\mathbb{E}_\eta(\eta^2)=1$.

Hence we are able to compute the $\langle \eta_i^\mu s_i x_\mu \rangle$ term by using the Gaussian integration by parts (namely by exploiting the Wick-Isserlis theorem [46]):

$$\sum_{i,\mu>1} \langle \eta_i^\mu s_i x_\mu \rangle = \frac{1}{d} \sum_{i,\mu>1} \mathbb{E}_\eta \sum_{q=1}^{d} \frac{\partial \omega \left(s_i x_\mu^{(q)} \right)}{\partial \eta_i^{\mu,(q)}} = \frac{b}{d} \sqrt{\frac{\beta}{N}} N K \left(\langle q_{11} p_{11} \rangle - \langle q_{12} p_{12} \rangle \right), \tag{103}$$

where, in the first equality, we use the Wick-Isserlis theorem component-wise, with components $x_\mu = \left(x_\mu^{(q)} \right)_{q=1,..,d}$ and $\eta_i^\mu = \left(\eta_i^{\mu,(q)} \right)_{q=1,..,d}$. This quantity is computed by observing that the partial derivative of the Boltzmann average $\omega(y)$ w.r.t. η produces the difference of two averages which, apart from global multiplying factors, corresponds to the average of the squared argument, $\omega(y^2)$, and the square of the Boltzmann average, $\omega^2(y)$. In our case we use this relation component-wise to get

$$\sum_{q=1}^{d} \frac{\partial \omega \left(s_i x_\mu^{(q)} \right)}{\partial \eta_i^{\mu,(q)}} = b \sqrt{\frac{\beta}{N}} \sum_{q=1}^{d} \left[\omega \left((x_\mu^{(q)} s_i)^2 \right) - \omega^2 \left(x_\mu^{(q)} s_i \right) \right]. \tag{104}$$

The first term reads

$$\sum_{q=1}^{d} \omega \left((x_\mu^{(q)} s_i)^2 \right) = \omega \left(\sum_{q=1}^{d} \left(x_\mu^{(q)} \right)^2 (s_i)^2 \right) = \omega \left(\left(x_\mu \right)^2 \ (s_i)^2 \right); \tag{105}$$

in the last equality we used $\left(x_\mu \right)^2 = \sum_{q=1}^{d} \left(x_\mu^{(q)} \right)^2$, that is the definition of the Euclidean norm of x_μ. This term can be rewritten as the scalar product of the first field replica x_μ^1 with itself times the square of the first spin replica s_i^1:

$$\omega \left(\left(x_\mu \right)^2 \ (s_i)^2 \right) = \omega \left(x_\mu^1 \cdot x_\mu^1 \ s_i^1 s_i^1 \right). \tag{106}$$

In order to deal with the second term, we introduce the distinct spin replicas s_i^1, s_i^2 and the field replicas $x_\mu^{(q),1}, x_\mu^{(q),2}$ component-wise, namely:

$$\sum_{q=1}^{d} \omega^2 \left(x_\mu^{(q)} s_i \right) = \sum_{q=1}^{d} \omega \left(s_i^1 s_i^2 \ x_\mu^{(q),1} x_\mu^{(q),2} \right), \tag{107}$$

which thus reads

$$\sum_{q=1}^{d} \omega \left(s_i^1 s_i^2 \ x_\mu^{(q),1} x_\mu^{(q),2} \right) = \omega \left(s_i^1 s_i^2 \sum_{q=1}^{d} x_\mu^{(q),1} x_\mu^{(q),2} \right) = \omega \left(s_i^1 s_i^2 \ x_\mu^1 \cdot x_\mu^2 \right). \tag{108}$$

Finally, collecting all terms together we are able to write

$$\sum_{i,\mu>1} \mathbb{E}_\eta \sum_{q=1}^{d} \frac{\partial \omega \left(s_i x_\mu^{(q)} \right)}{\partial \eta_i^{\mu,(q)}} = \sum_{i,\mu>1} \mathbb{E}_\eta \left(\omega \left(s_i^1 s_i^1 \ x_\mu^1 \cdot x_\mu^1 \right) - \omega \left(s_i^1 s_i^2 \ x_\mu^1 \cdot x_\mu^2 \right) \right). \tag{109}$$

The latter allows us to introduce the replica overlaps order parameters $Nq_{11} = \sum_i s_i^1 s_i^1$, $Nq_{12} = \sum_i s_i^1 s_i^2$, $Kp_{11} = \sum_\mu x_\mu^1 \cdot x_\mu^1$, $Kp_{12} = \sum_\mu x_\mu^1 \cdot x_\mu^2$, to get

$$\sum_{i,\mu>1} \mathbb{E}_\eta \omega \left(s_i^1 s_i^1 \, x_\mu^1 \cdot x_\mu^1\right) = NK\langle q_{11} p_{11}\rangle, \tag{110}$$

$$\sum_{i,\mu>1} \mathbb{E}_\eta \, \omega \left(s_i^1 s_i^2 \, x_\mu^1 \cdot x_\mu^2\right) = NK\langle q_{12} p_{12}\rangle. \tag{111}$$

The derivative of $\mathcal{A}(t)$, in their terms, reads as

$$\mathcal{A}' = \frac{a_1'}{2}\beta\langle x_1^2\rangle + \frac{a_2'}{2}\beta\,(1-\lambda)\,\langle m^2\rangle + \frac{\beta\alpha}{2d}\, a_3 a_3' \left(\langle q_{11} p_{11}\rangle - \langle q_{12} p_{12}\rangle\right) + \langle \phi'\rangle, \tag{112}$$

where the load of the model $\alpha = \lim_{N\to\infty} K/N$ has been introduced.The latter equation contains only two-points correlation functions, e.g. $\langle x_1^2\rangle, \langle q_{12}\rangle, ..,$ due to the presence of second order interactions in the Hamiltonian: these terms are generally hard to compute and the general strategy of the Guerra's interpolation technique is to balance them out with appropriate counter-terms that can be obtained by differentiating the $\langle \phi(t)\rangle$ functional. For example, the two-point function $\langle x_1^2\rangle$ can be produced by a term of the form $\sum_{\mu>1} x_\mu^2$, while a term like $\sum_i h_i s_i$ (with $h_i \in \mathcal{N}(0,1)$) produces a contribution of the form of $\langle q_{11}\rangle - \langle q_{12}\rangle$ (apart from global factors). This motivates our choice of the $\phi(t)$ functional, which is thus expressed as a sum of one-body and two-bodies terms as follows:

$$\begin{aligned} N\phi(t) =& \phi_1(t)\sum_{\mu>1} x_\mu^2 + \phi_2(t)\sum_\mu \rho_\mu x_\mu + \phi_3(t)\sum_i h_i s_i + \phi_4(t)\sum_i \eta_i^1 s_i \\ &+ \phi_5(t)\sum_i s_i^2 + \phi_6(t)\sum_i s_i, \end{aligned} \tag{113}$$

where the $t-$dependence is delegated to the parameters $\phi_1(t), \phi_2(t), \phi_3(t), \phi_4(t), \phi_5(t), \phi_6(t)$, which have to respect the boundary conditions $\phi_1(1) = \phi_2(1) = \phi_3(1) = \phi_4(1) = \phi_5(1) = \phi_6(1) = 0$. The derivative of $\phi(t)$ then reads

$$\begin{aligned} \phi' =& \phi_1'\alpha\langle p_{11}\rangle + \phi_2\phi_2'\alpha\left(\langle p_{11}\rangle - \langle p_{12}\rangle\right) + \phi_3\phi_3'\left(\langle q_{11}\rangle - \langle q_{12}\rangle\right) + \phi_4'\langle x_1\rangle \\ &+ \phi_5'\langle q_{11}\rangle + \phi_6'\langle m\rangle. \end{aligned} \tag{114}$$

Collecting all the terms, the derivative of $\mathcal{A}(t)$ finally reads

$$\begin{aligned} \mathcal{A}' =& \frac{a_1'}{2}\beta\langle x_1^2\rangle + \frac{a_2'}{2}\beta\,(1-\lambda)\,\langle m^2\rangle + \alpha\beta\frac{a_3 a_3'}{d}\left(\langle q_{11} p_{11}\rangle - \langle q_{12} p_{12}\rangle\right) + \phi_1'\alpha\langle p_{11}\rangle \\ &+ \phi_2\phi_2'\alpha\left(\langle p_{11}\rangle - \langle p_{12}\rangle\right) + \phi_3\phi_3'\left(\langle q_{11}\rangle - \langle q_{12}\rangle\right) + \phi_4'\langle x_1\rangle + \phi_5'\langle q_{11}\rangle + \phi_6'\langle m\rangle. \end{aligned} \tag{115}$$

Next, we treat the order parameters expectations in terms of their fluctuations, under the RS hypothesis (remember Definition 6) these fluctuations are vanishing in the thermodynamic limit $N \to \infty$, namely

$$\langle \Delta^2 x\rangle = \langle (x_1 - \overline{x})^2\rangle \to 0, \tag{116}$$

$$\langle \Delta^2 m\rangle = \langle (m - \overline{m})^2\rangle \to 0, \tag{117}$$

$$\langle \Delta p_1 \Delta q_1\rangle = \langle (q_{11} - \overline{q}_1)(p_{11} - \overline{p}_1)\rangle \to 0, \tag{118}$$

$$\langle \Delta p_2 \Delta \overline{q}_2\rangle = \langle (q_{12} - \overline{q}_2)(p_{12} - \overline{p}_2)\rangle \to 0, \tag{119}$$

where $\overline{x} = \langle x\rangle, \overline{m} = \langle m\rangle$ and $\overline{p}_1 = \langle p_{11}\rangle, \overline{p}_2 = \langle p_{12}\rangle, \overline{q}_1 = \langle q_{11}\rangle, \overline{q}_2 = \langle q_{12}\rangle$.

Collecting the homogeneous terms in equation (115) and systematically eliminating the terms containing the expectation values of the order parameters (10)–(12), we obtain the following system of coupled differential equations to be solve in the interval $t \in (0, 1)$:

$$\phi_6' + \beta(1-\lambda)a_2'\overline{m} = 0, \tag{120}$$
$$\phi_4' + \beta a_1'\overline{x} = 0, \tag{121}$$
$$\phi_1' + \beta\frac{a_3a_3'}{d}\overline{q}_1 + \phi_2\phi_2' = 0, \tag{122}$$
$$\phi_5' + \alpha\beta\frac{a_3a_3'}{d}\overline{p}_1 + \phi_3\phi_3' = 0, \tag{123}$$
$$\phi_2\phi_2' + \beta\frac{a_3a_3'}{d}\overline{q}_2 = 0, \tag{124}$$
$$\phi_3\phi_3' + \alpha\beta\frac{a_3a_3'}{d}\overline{p}_2 = 0. \tag{125}$$

Without loosing generality we can also impose the conditions

$$a_1' = 1, \tag{126}$$
$$a_2' = 1, \tag{127}$$
$$a_3a_3' = 1/2. \tag{128}$$

The solution reads

$$\begin{cases} \phi_1(t) = \dfrac{\beta}{2d}(\overline{q}_1 - \overline{q}_2)(1-t), \\ \phi_2(t) = \sqrt{\dfrac{\beta}{d}\overline{q}_2\,(1-t)}, \\ \phi_3(t) = \sqrt{\dfrac{\alpha\beta}{d}\overline{p}_2\,(1-t)}, \\ \phi_4(t) = \beta\overline{x}\,(1-t), \\ \phi_5(t) = \dfrac{\alpha\beta}{2d}(\overline{p}_1 - \overline{p}_2)(1-t), \\ \phi_6(t) = \beta(1-\lambda)\overline{m}\,(1-t). \end{cases} \qquad \begin{cases} a_1(t) = t, \\ a_2(t) = t, \\ a_3(t) = \sqrt{t}, \end{cases}$$

The boundary conditions are finally derived from the interpolation conditions $\phi(1) = 0$ and $a_1(0) = a_2(0) = a_3(0) = 0$. Once all the auxiliary functions have been made fully explicit, we can sum up all the contributions to the $t-$derivative of $\mathcal{A}$ that then reads

$$\mathcal{A}' = \frac{\beta}{2}(1-\lambda)\left(\langle\Delta^2 m\rangle - \overline{m}^2\right) + \frac{\beta}{2}\left(\langle\Delta^2 x\rangle - \overline{x}^2\right) + \frac{\alpha\beta}{2d}\left(\langle\Delta p_1\Delta q_1\rangle - \overline{p}_1\overline{q}_1\right) - \frac{\alpha\beta}{2d}\left(\langle\Delta p_2\Delta q_2\rangle - \overline{p}_2\overline{q}_2\right). \tag{129}$$

Under the RS assumption, the fluctuations of the order parameters vanish and the $t-$derivative of $\mathcal{A}$ becomes the expression reported in Proposition 3. □

Proof of Proposition 4 We are left with the Cauchy condition to calculate: the interpolating free energy has to be evaluated at $t = 0$ but, before doing that, we write explicitly the partition function at $t = 0$ for the sake of clearness:

$$\mathcal{Z}(t=0) = \sum_{\boldsymbol{s}} \int Dx \ \exp\Big(\phi_1(0) \sum_{\mu>1} x_\mu^2 + \phi_2(0) \sum_{\mu>1} \rho_\mu x_\mu + \phi_3(0) \sum_i h_i s_i + \phi_4(0) \sum_i \eta_i^1 s_i + \phi_5(0) \sum_i s_i^2 + \phi_6(0) \sum_i s_i \Big). \tag{130}$$

After performing the Gaussian integrals in $D^d x = \prod_{\mu>1} D^d x_\mu$, where $Dx_\mu = \frac{d^d x_\mu}{\sqrt{2\pi}} e^{-\frac{x_\mu^2}{2}}$ is the Gaussian measure; we end up with:

$$\mathcal{Z}(t=0) = \left[\prod_{\mu>1} \left(\frac{2\pi}{1-2\phi_1(0)} \right)^{d/2} \exp\left(\frac{1}{2} \frac{\phi_2^2(0)\rho_\mu^2}{1-2\phi_1(0)} \right) \right] \times \prod_i \left[1 + \exp\left(\phi_3(0) h_i + \phi_4(0)\eta_i^1 + \phi_5(0) + \phi_6(0) \right) \right]. \tag{131}$$

We can finally rearrange all the terms as provided by Proposition 4. □

Proof of Corollary 1 The self-consistent Eqs. (22)–(24) can be made explicit simply by applying the following Eqs. (132)–(133) to evaluate the expectation over the map realizations.
We evaluate the quenched average of a function $g(\boldsymbol{\eta})$ whose dependence on $\boldsymbol{\eta}$ occurs via the scalar product $\eta_i^\mu \cdot a$, where a is a two-dimensional vector with module $|a|$ and direction specified by the unitary vector $\hat{a}$, that is, $a = |a|\,\hat{a}$. Then, dropping the scripts μ and i in η_i^μ, without loss of generality, we get[12]

$$\langle g(\eta \cdot a) \rangle_\eta = \int_{-\pi}^{\pi} \frac{d\theta}{2\pi} \, g(|a| \cos\theta) = \frac{1}{\pi} \int_{-1}^{1} \frac{dt}{\sqrt{1-t^2}} \, g(|a|\, t), \tag{132}$$

$$\langle \eta \, g(\eta \cdot a) \rangle_\eta = \int_{-\pi}^{\pi} \frac{d\theta}{2\pi} \, \hat{a} \cos\theta \, g(|a| \cos\theta) = \frac{\hat{a}}{\pi} \int_{-1}^{1} \frac{dt}{\sqrt{1-t^2}} \, t \, g(|a|\, t). \tag{133}$$

The relations provided in Eqs. (3), (13) and (132)–(133) are valid for the 2-dimensional unitary circle but can be easily generalized in d-dimension. Then, the given map η^μ is a unit vector on the (hyper-)sphere S^{d-1} as $\eta^\mu \in \mathbb{R}^d$, $|\eta^\mu| = 1$, and, in spherical coordinates, η^μ is a function of the angles $\Omega = (\theta, \phi, ..)$: $\eta_i^\mu = \eta_i^\mu(\Omega_i^\mu)$. Notice that the dot product of two maps, i.e. $\eta_i^\mu \cdot \eta_j^\mu = \cos\gamma$, is still a function of the relative angle γ between the two unit vectors, hence our requirement for the kernel function of the model is respected also in d-dimension.
Then, we introduce the volume form $d\omega_d$ on S^{d-1}, which in spherical coordinates can be written as

$$d\omega_d = (\sin\theta)^{d-2} \, d\theta d\omega_{d-1}, \quad \theta \in [0, \pi]. \tag{134}$$

The expectation over the maps (13) can therefore be generalized in $d-$dimensions as follows

$$\langle g(\boldsymbol{\eta}) \rangle_{\eta \in S^{d-1}} = \int \left[\prod_{i,\mu=1}^{N,K} \frac{d\omega_i^\mu}{|S^{d-1}|} \right] g(\boldsymbol{\eta}(\boldsymbol{\omega})), \tag{135}$$

[12] These relations can be easily derived by applying the change of variable $t = \cos\theta$ in the integrals, where θ is the angle between the two vectors involved in the scalar product, and using $|\eta| = 1$. Also notice that, because of the identity $\frac{1}{\pi}\int_{-1}^{1} \frac{dt}{\sqrt{1-t^2}} = 1$, we have $\langle \exp(\eta \cdot a) \rangle_\eta \sim 1 + \frac{|a|^2}{4} + \mathcal{O}(|a|^4)$, for $|a| \to 0$.

where the normalization factor $|S^{d-1}|$ is the volume of the sphere, which is computed by integrating (134), and reads

$$|S^{d-1}| = \int d\omega_d = \frac{2\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}\right)},$$

where Γ is the gamma function. Further, the relations (132)–(133) can be generalized as

$$\langle g(\eta \cdot a)\rangle_\eta = \frac{1}{|S^{d-1}|}\int d\omega_d\, g(\eta \cdot a) = \Omega_d \int_{-1}^{1} dt\, \left(1-t^2\right)^{\frac{d-3}{2}}\, g(|a|\, t), \tag{136}$$

$$\langle \eta\, g(\eta \cdot a)\rangle_\eta = \hat{a}\, \Omega_d \int_{-1}^{1} dt\, t\, \left(1-t^2\right)^{\frac{d-3}{2}}\, g(|a|\, t), \tag{137}$$

where we performed the change of variables $t = \cos\theta$ (with θ being the angle between a and η), after which and the factor Ω_d emerges as:

$$\Omega_d = \frac{|S^{d-2}|}{|S^{d-1}|} = \frac{\Gamma\left(\frac{d}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{d-1}{2}\right)}. \tag{138}$$

Notice that for $d = 2$ one has $\Omega_2 = \frac{1}{\pi}$, restoring the relations (132)–(133). Notice also that the series expansion of (136) reads $\langle \exp(\eta \cdot a)\rangle_{\eta \in S^{d-1}} \sim 1 + \frac{|a|^2}{2d} + \mathcal{O}(|a|^4)$. □

B A Gentle Introduction to the Interpolation Method for the Hopfield Net

In order to make the reader from Neuroscience more acquainted with Guerra interpolation approaches, we take this appendix to summarize how these work on the paradigmatic Hopfield model. In particular, we study the low storage regime $\alpha = 0$ with the Hamilton-Jacobi approach while we tackle the high storage regime $\alpha > 0$ with a sum rule based on stochastic stability interpolation, mirroring the approach pursued in the main text for the (generalization of) the Battaglia-Treves model we discussed.[13]
Given P random (i.e. Rademacher) patterns $\xi^\mu,\ \mu \in (1, ..., P)$ of length N and N Ising spins $\sigma_i,\ i \in (1, ..., N)$, the Hamiltonian we study only in this section reads as

$$H_N(\boldsymbol{\sigma}|\boldsymbol{\xi}) = -\frac{1}{2N}\sum_{i,j=1}^{N}\sum_{\mu=1}^{P}\left(\xi_i^\mu \xi_j^\mu\right)\sigma_i\sigma_j. \tag{139}$$

We aim to calculate the quenched free energy of the model, defined as

$$A(\alpha, \beta) = \lim_{N\to\infty}\frac{1}{N}\mathbb{E}_\xi \ln \sum_{\sigma}^{2^N} e^{-\beta H_N(\boldsymbol{\sigma}|\boldsymbol{\xi})}, \tag{140}$$

in terms of the order parameters of the model that are the P Mattis magnetizations $m_\mu := \frac{1}{N}\sum_{i=1}^{N}\xi_i^\mu\sigma_i$ and the standard 2-replica overlaps.

[13] It is fair to state that we are using just two tools out of many produced by Francesco Guerra for spin glasses, see e.g. also [47, 75].

B.1 Hamilton-Jacobi Technique for the Hopfield Model in the Low Storage Regime

An elegant way to solve explicitly for the free energy (in the low storage sceanario, i.e. $\alpha = 0$), is by relying upon the *mechanical analogy* that consists in showing that the free energy obeys a Hamilton-Jacobi PDE (and thus it can be evaluated by relating upon tools typical of Analytical Mechanics rather than Statistical Mechanics): to this task we can introduce an interpolating free energy as the following Guerra action

$$f_N(t,\vec{x}) = \frac{1}{N}\mathbb{E}_\xi \ln \sum_{\sigma}^{2^N} \exp\left(-\frac{tN}{2}\sum_{\mu=1}^{P} m_\mu^2\right) + N\sum_{\mu=1}^{P} x_\mu m_\mu, \tag{141}$$

where $t \in \mathbb{R}^+$ and $\vec{x} = (x_1, ..., x_P) \in \mathbb{R}^P$.
Clearly, if we chose $t = -\beta$ and $\vec{x} = 0$, we get $\lim_{N\to\infty} f_N(t,\vec{x}) = A(\beta)$, hence the idea is to obtain $f_N(t,\vec{x})$ by solving the Hamilton-Jacobi PDE, perform the thermodynamic limit and then evaluate that function in this specific point of the spacetime (where the statistical mechanical scaffold is recovered).
We can easily evaluate the derivative w.r.t. the spacetime of the Guerra action to get

$$\frac{\partial f_N(t,\vec{x})}{\partial t} = -\frac{1}{2}\sum_{\mu=1}^{P}\mathbb{E}_\xi\omega(m_\mu^2) = -\frac{1}{2}\sum_{\mu=1}^{P}\langle m_\mu^2\rangle, \tag{142}$$

$$\nabla f_N(t,\vec{x}) = \sum_{\mu=1}^{P}\mathbb{E}_\xi\omega(m_\mu) = \sum_{\mu=1}^{P}\langle m_\mu\rangle. \tag{143}$$

As a consequence, by construction, at finite volume N it holds that

$$\frac{\partial f_N(t,\vec{x})}{\partial t} + \frac{1}{2}\left(\nabla f_N(t,\vec{x})\right)^2 + V_N(t,\vec{x}) = 0, \tag{144}$$

$$V_N(t,\vec{x}) = \frac{1}{2}\sum_{\mu=1}^{P}\left(\langle m_\mu^2\rangle - \langle m_\mu\rangle^2\right). \tag{145}$$

Note that, in this spacetime, we have an effective Hamiltonian $\mathcal{H}$ and an effective Lagrangian $\mathcal{L}$ that read, respectively,

$$\mathcal{H} = \frac{1}{2}\left(\nabla f_N(t,\vec{x})\right)^2 + V_N(t,\vec{x}), \quad \mathcal{L} = \frac{1}{2}\left(\nabla f_N(t,\vec{x})\right)^2 - V_N(t,\vec{x}).$$

Note further that the potential is the sum of the P variances of the related Mattis magnetizations around their means: as we expect that each Mattis magnetization self-averages as $|m_\mu - M_\mu| \sim O(\frac{c}{\sqrt{N}})$ for some constant c, as long as P does not scale linearly with N it is fair to assume that $\lim_{N\to\infty} V_N(t,\vec{x}) = 0$: in the thermodynamic limit and in the low storage regime, the Guerra action obeys a free-field evolution accordingly to the Hamilton-Jacobi PDE.
The solution $f(t,\vec{x}) = \lim_{N\to\infty} f_N(t,\vec{x})$ can be obtained as the Cauchy condition $f(t = 0,\vec{x}_0)$ plus the integral of the Lagrangian in $(0,t)$, namely

$$f(t,\vec{x}) = f(t=0,\vec{x}_0) + \int_0^t \mathcal{L}(t,\vec{x})dt. \tag{146}$$

The integral of the Lagrangian in $(0,t)$ is simply the multiplication of the kinetic energy $T(t,\vec{x}) := \frac{1}{2}\sum_{\mu=1}^{P}\langle m_\mu^2\rangle$ by the time t as the potential term $V(t,\vec{x}) = 0$ and thus the kinetic

energy is constant along the (Galilean) trajectories of motion

$$\vec{x}(t) = \vec{x}_0 + \langle \vec{m} \rangle t. \tag{147}$$

Let us calculate the one body contribution, that is the Cauchy condition $f(t = 0, \vec{x}_0)$:

$$\begin{aligned} f(t = 0, \vec{x}_0) &= \frac{1}{N} \mathbb{E}_\xi \ln \sum_{\sigma}^{2^N} e^{N \sum_{\mu=1}^{P} x_0^\mu m_\mu} \\ &= \ln 2 + \mathbb{E}_\xi \ln \cosh \left(\vec{x}_0 \cdot \vec{\xi} \right) \qquad (148) \\ &= \ln 2 + \mathbb{E}_\xi \ln \cosh \left(\beta \sum_{\mu=1}^{P} m_\mu \xi^\mu \right), \qquad (149) \end{aligned}$$

where in the last passage we used (147) to make $\vec{x}_0$ explicit, then we imposed $\vec{x}(t) = 0$ and $t = -\beta$.
Overall, adding this term to the integral of the Lagrangian and evaluating the resulting expressions in the point $t = -\beta, \vec{x} = 0$, such that $f(t, \vec{x}) \to A(\beta)$ we get

$$A(\beta) = \ln 2 + \langle \ln \cosh \left(\beta \sum_{\mu=1}^{P} m_\mu \xi^\mu \right) \rangle - \frac{\beta}{2} \sum_{\mu=1}^{P} \langle m_\mu^2 \rangle, \tag{150}$$

$$\langle m_\mu \rangle = \langle \xi^\mu \tanh \left(\beta \vec{m} \cdot \vec{\xi} \right) \rangle, \tag{151}$$

that are the celebrated quenched free energy and related self-consistency equation for the Hopfield model in the low storage regime.[14]

B.2 Guerra Interpolation for the Hopfield Model in the High Storage Regime

Hereafter we assume $\alpha > 0$, namely that the network stores a linear amount of patterns $P = \alpha N$ (in the volume of the network). Via Guerra's interpolation we aim to recover the original Amit-Gutfreund-Sompolinsky prescription for the infinite volume limit of the quenched free energy and related self-consistency equations for the order parameters [50].
The strategy to follow is to keep considering the signal -say the one provided by the first pattern ξ^1 (with no loss of generality)- to be Boolean, that is $\xi_i^1 = \pm 1, \ \forall i \in (1, ..., N)$, while all the other patterns -that sum up to give the slow noise contribution- to be standard i.i.d. Gaussian: this can be done thanks to the universality of the quenched noise in spin glasses[15] [76, 77].
As a consequence, we also split the signal term in the Hamiltonian of the Hopfield model from the others, that is

$$H_N(\boldsymbol{\sigma}|\boldsymbol{\xi}) = -\frac{1}{2N} \sum_{i,j=1}^{N} \xi_i^1 \xi_j^1 \sigma_i \sigma_j - \frac{1}{2N} \sum_{i,j=1}^{N} \sum_{\mu>1}^{P} \left(\xi_i^\mu \xi_j^\mu \right) \sigma_i \sigma_j + \frac{1}{2N} \sum_{i,\mu}^{N,P} \left(\xi_i^\mu \right)^2 \tag{152}$$

[14] In the last expressions we used the brackets $\langle . \rangle$ to mean the quenched expectation $\mathbb{E}_\xi$ in order to reproduce the exact expressions reported in the Amit milestone [32].

[15] In a nutshell, this prescription states that, solely for the quenched noise, it does not matter if we use Rademacher ± 1 or Gaussian $\mathcal{N}[0, 1]$ variables, results are the same and, as the latter are by far more useful within the Guerra's route (as they allow to use Wick theorem/Stein lemma [47]) we assume all the not-retrieved patterns to be Gaussian distributed.

$$= -\frac{N}{2}m_1^2 - \frac{1}{2N}\sum_{i,j=1}^{N}\sum_{\mu>1}^{P}\left(\xi_i^\mu \xi_j^\mu\right)\sigma_i\sigma_j + \frac{1}{2N}\sum_{i,\mu}^{N,P}\left(\xi_i^\mu\right)^2, \tag{153}$$

where the last term, negligible in the low storage regime, contributes in the $\alpha > 0$ scenario but solely to add the numerical value $-\alpha\beta/2$ to the free energy in the thermodynamic limit: it can be forgotten by now and added at the end of the calculations.
The partition function reads as

$$Z_N(\beta) = \sum_{\{\sigma\}}^{2^N} e^{\beta\frac{N}{2}m_1^2} e^{\frac{\beta}{2N}\sum_{i,j=1}^{N}\sum_{\mu>1}^{P}\left(\xi_i^\mu\xi_j^\mu\right)\sigma_i\sigma_j} \tag{154}$$

$$= \sum_{\{\sigma\}}^{2^N} e^{\beta\frac{N}{2}m_1^2}\int_{\mathcal{R}^{P-1}}\prod_{\mu>1} d\mu(z_\mu) e^{\sqrt{\frac{\beta}{N}}\sum_{i,\mu}\xi_i^\mu\sigma_i z_\mu}, \tag{155}$$

where in the last line, via Gaussian integration, we achieved the dual representation of the Hopfield model in terms of a bipartite network: by using the above integral representation, we aim to evaluate explicitly the infinite volume limit of

$$A(\alpha,\beta) = \lim_{N\to\infty}\frac{1}{N}\mathbb{E}_\xi \ln\sum_{\sigma}^{2^N} e^{\beta\frac{N}{2}m_1^2}\int_{\mathcal{R}^{P-1}}\prod_{\mu>1} d\mu(z_\mu) e^{\sqrt{\frac{\beta}{N}}\sum_{i,\mu}\xi_i^\mu\sigma_i z_\mu} - \frac{\alpha\beta}{2}. \tag{156}$$

We now introduce the Guerra's interpolation strategy: the idea is to interpolate between the original model and an effective one-body model that is mathematically solvable and to connect these extrema via the fundamental theorem of calculus (see Eq. (160)).
The one-body model has to mimic the field (or post-synaptic potential in neural network jargon) experienced by each neuron within the two layers: the N binary neurons σ_i perceive a field $\sim \sum_\mu \xi_i^\mu z_\mu$ while the P Gaussian neurons z_μ perceive a field $\sim \sum_i \xi_i^\mu\sigma_i$. These must be substituted by external fields whose values must resemble the original ones but these new effective fields are no longer functions of the neurons of the other layer (making the model effectively one-body).
To this task, once introduced an interpolating parameter $t \in (0, 1)$, we define the interpolating free energy as

$$A(\alpha,\beta|t) = \lim_{N\to\infty}\frac{1}{N}\mathbb{E}_\xi \ln\sum_{\sigma}^{2^N} e^{t\beta\frac{N}{2}m_1^2}\int_{\mathcal{R}^{P-1}}\prod_{\mu>1} d\mu(z_\mu) e^{\sqrt{t\frac{\beta}{N}}\sum_{i,\mu}\xi_i^\mu\sigma_i z_\mu} - \frac{\alpha\beta}{2} \cdot \tag{157}$$

$$\cdot\, e^{t\beta M\sum_i \xi_i^1\sigma_i} e^{\sqrt{1-t}\left(A\sum_i J_i\sigma_i + B\sum_\mu J_\mu z_\mu\right)} e^{(1-t)\frac{C}{2}\sum_\mu z_\mu^2}, \tag{158}$$

where the N couplings J_i and the P couplings J_μ are all standard Gaussian distributions $Z[0, 1]$ and the correct values of the constants A, B, C can be determined by requesting that the order parameters self-average around their means, that is

$$\lim_{N\to\infty}\mathcal{P}(m_1) = \delta(m_1 - M_1), \quad \lim_{N\to\infty}\mathcal{P}(q_{ab}) = \delta(q_{ab} - \bar{q}), \quad \lim_{N\to\infty}\mathcal{P}(p_{ab}) = \delta(p_{ab} - \bar{p}). \tag{159}$$

This request enters in the following sum rule

$$A(\alpha,\beta) = A(\alpha,\beta|t=1) = A(\alpha,\beta|t=0) + \int_0^1 \frac{dA(\alpha,\beta|t)}{dt}, \tag{160}$$

to be evaluated in the thermodynamic limit as we explain hereafter.
First note that, in this way, we are left to evaluate $A(\alpha, \beta|t=0)$ and the integral of $\dot{A}(\alpha, \beta|t)$. The former is a one-body calculation achievable by brute force that returns

$$A(\alpha, \beta|t=0) = \ln 2 + \mathbb{E}_\xi \left(\ln\cosh\left(\beta M_1 \xi^1 + AJ\right)\right) - \frac{\alpha}{2}\ln(1-C) + \frac{\alpha\beta^2}{2(1-C)} - \frac{\alpha\beta}{2}. \tag{161}$$

The latter, instead, gives

$$\frac{dA(\alpha, \beta|t)}{dt} = \frac{\beta}{2}\langle\left(m_1^2 - 2M_1 m_1\right)\rangle + \frac{1}{2N}\left(\beta - B^2 - C\right)\sum_\mu \langle z_\mu^2\rangle \tag{162}$$

$$- \frac{\alpha\beta}{2}\langle q_{12}p_{12}\rangle - \frac{A^2}{2}(1 - \langle q_{12}\rangle) + \frac{\alpha B^2}{2}\langle p_{12}\rangle. \tag{163}$$

Thus, by choosing $A = \sqrt{\alpha\beta\bar{p}}$ and $B = \sqrt{\beta\bar{q}}$ and $C = \beta(1-\bar{q})$, we can write each term at the r.h.s. of the above expression as a mean value plus a variance (that vanishes as $N \to \infty$), e.g. $\langle\left(m_1^2 - 2M_1 m_1\right)\rangle = \langle(m_1 - M_1)^2\rangle - M_1^2$ such that the derivative of the free energy (162) can be rewritten as

$$\frac{dA(\alpha, \beta|t)}{dt} = -\frac{\beta}{2}M_1^2 - \frac{\alpha\beta}{2}\bar{p}(1-\bar{q}) - \frac{\beta}{2}\langle(m_1 - M_1)^2\rangle\frac{\alpha\beta}{2}\langle(q_{12} - \bar{q})(p_{12} - \bar{p})\rangle. \tag{164}$$

Now, crucially, note that the first two terms at the r.h.s. of the above equation contain solely mean values of the order parameters while the last two terms at the r.h.s. contain solely fluctuations around these means that, in the self-averaging regime induced by the replica symmetry (requested in (159)), go to zero (in the thermodynamic limit), hence

$$\int_0^1 \frac{dA(\alpha, \beta|t)}{dt} = -\frac{\beta}{2}M_1^2 - \frac{\alpha\beta}{2}\bar{p}(1-\bar{q}) \tag{165}$$

as the integration of constant terms (in $t \in (0, 1)$) reduces to the multiplication by one.
As a consequence, the sum rule (160) reads as

$$A(\alpha, \beta) = \ln 2 + \mathbb{E}_\xi \left(\ln\cosh\left(\beta M_1 \xi^1 + \sqrt{\alpha\beta\bar{p}}\right)\right) - \frac{\beta}{2}M_1^2 \tag{166}$$

$$- \frac{\alpha}{2}\ln(1 - \beta(1-\bar{q})) + \frac{\alpha\beta\bar{q}}{2(1-\beta(1-\bar{q}))} - \frac{\alpha\beta}{2}\bar{p}(1-\bar{q}) - \frac{\alpha\beta}{2}, \tag{167}$$

where the mean values of the order parameters must respect $\nabla_{M_1,\bar{q},\bar{p}} A(\alpha, \beta) = 0$: once this extremization is performed the celebrated AGS picture of the Hopfield model, originally achieved vie the replica trick, is completely recovered.
All the detailed calculations can be deepened in the related paper, see e.g. [45, 46, 58, 59].

References

1. O'Keefe, J., Dostrovsky, J.: The hippocampus as a spatial map: preliminary evidence from unit activity in the freely-moving rat. Brain Res. **34**, 171–175 (1971)
2. Moser, E.I., Kropff, E., Moser, M.B.: Place cells, grid cells, and the brain's spatial representation system. Ann. Rev. Neurosci. **31**, 69–89 (2008)

3. Best, P.J., White, A.M., Minai, A.: Spatial processing in the brain: the activity of hippocampal place cells. Ann. Rev. Neurosci. **24**, 459–486 (2001)
4. Coolen, A.C.C., Kuehn, R., Sollich, P.: Theory of neural information processing systems. OUP, Oxford (2005)
5. Samsonovich, A., McNaughton, B.: Path integration and cognitive mapping in a continuous attractor neural network model. J. Neurosci. **17**(15), 5900–5920 (1997)
6. Monasson, R., Rosay, S.: Transitions between spatial attractors in place-cell models. Phys. Rev. Lett. **115**(9), 098101 (2015)
7. Rosay, S., Monasson, R.: Cross-talk and transitions between multiple environments in an attractor neural network model of the hippocampus. BMC Neurosci. **14**, O15 (2013)
8. Tsodyks, M.V., Sejnowski, T.: Rapid state switching in balanced cortical network models. Comp. Neural Syst. **6**(02), 111 (1995)
9. Schönsberg, F., Monasson, R., Treves, A.: Continuous quasi-attractors dissolve with too much-or too little-variability. bioRxiv (2023). https://doi.org/10.1093/pnasnexus/pgae525
10. Battaglia, F.P., Treves, A.: Attractor neural networks storing multiple space representations: a model for hippocampal place fields. Phys. Rev. E **58**(6), 7738 (1998)
11. Ryom, K., Treves, A.: Speed inversion in a potts glass model of cortical dynamics. PRX Life **1**(3), 013005 (2023)
12. Agliari, E., et al.: Replica symmetry breaking in neural networks: a few steps toward rigorous results. J. Phys. A **53**(41), 415005 (2020)
13. Alberici, D., et al.: Annealing and replica-symmetry in deep Boltzmann machines. J. Stat. Phys. **180**, 665–677 (2020)
14. Alberici, D., et al.: Deep Boltzmann machines: rigorous results at arbitrary depth. Annal. Hen. Poin. **22**, 2619–2642 (2021)
15. Barbier, J., Macris, N.: The adaptive interpolation method: a simple scheme to prove replica formulas in Bayesian inference. Probab. Theory Relat. Fields **174**, 1133–1185 (2017)
16. Barbier, J., Macris, N.: Exact mutual information of sparse superposition codes from the adaptive path interpolation method. In: Int. Zurich Sem. Inform. & Commu. (2018)
17. Agoritsas, E., et al.: Explaining the effects of non-convergent MCMC in the training of energy-based models. In: Int. Conf. Mach. Learn. PMLR (2023)
18. Catania, G., et al., The Copycat Perceptron: Smashing Barriers Through Collective Learning (2023). arXiv:2308.03743
19. Decelle, A., Ricci-Tersenghi, F., Zhang, P.: Data quality for the inverse Ising problem. J. Phys. A **49**(38), 384001 (2016)
20. Marino, R., Ricci-Ternseghi, F.: Phase transitions in the mini-batch size for sparse and dense neural networks (2023). arXiv:2305.06435
21. Sollich, P., et al.: Extensive parallel processing on scale-free networks. Phys. Rev. Lett. **113**(23), 238106 (2014)
22. Albanese, L., et al.: About the de Almeida-Thouless line in neural networks. Physica A **633**, 129372 (2024)
23. Decelle, A., Hwang, S., Tantari, D.: Inverse problems for structured datasets using parallel TAP equations and restricted Boltzmann machines. Sci. Rep. **11**(1), 19990 (2021)
24. Alemanno, F., et al.: Hopfield model with planted patterns: a teacher-student self-supervised learning model. Appl. Math. Comput. **458**, 128253 (2023)
25. Lucibello, C., Mezard, M.: Exponential capacity of dense associative memories. Phys. Rev. Lett. **132**(7), 077301 (2024)
26. Krotov, D., Hopfield, J.J.: Dense associative memory for pattern recognition. Adv. Neural Inf. Proc. Syst. **29** (2016)
27. Vardi, R., Tugendhaft, Y., Kanter, I.: Neuronal plasticity features are independent of neuronal holding membrane potential. Physica A **632**, 129351 (2023)
28. Agliari, E., et al.: Machine learning and statistical physics: theory, inspiration, application. J. Phys. A (2020). https://doi.org/10.1088/1751-8121/abca75
29. Huang, H.: Statistical Mechanics of Neural Networks. Springer, Singapore (2021)
30. Kang, L., Toyoizumi, T.: Hopfield-like network with complementary encodings of memories. Phys. Rev. E **108**, 054410 (2023)
31. Kang, L., Toyoizumi, T.: Distinguishing examples while building concepts in hippocampal and artificial networks. Nat. Commun. **15**, 647 (2024)
32. Amit, D.J.: Modeling Brain Function: The World of Attractor Neural Networks. Cambridge University Press, Cambridge (1989)
33. Rolls, E.T., Treves, A.: Neural Networks and Brain Function. Oxford University Press, Oxford (1998)

34. Barra, A., et al.: On the equivalence of Hopfield networks and Boltzmann machines. Neural Netw. **34**, 1–9 (2012)
35. Barra, A., Genovese, G., Sollich, P., Tantari, D.: Phase diagram of restricted Boltzmann machines and generalized Hopfield networks with arbitrary priors. Phys. Rev. E **97**(2), 022310 (2018)
36. Mezard, M.: Mean-field message-passing equations in the Hopfield model and its generalizations. Phys. Rev. E **95**(2), 022117 (2017)
37. Tubiana, J., Monasson, R.: Emergence of compositional representations in restricted Boltzmann machines. Phys. Rev. Lett. **118**(13), 138301 (2017)
38. Marullo, C., Agliari, E.: Boltzmann machines as generalized Hopfield networks: a review of recent results and outlooks. Entropy **23**(1), 34 (2020)
39. Barra, A., Di Biasio, A., Guerra, F.: Replica symmetry breaking in mean-field spin glasses through the Hamilton-Jacobi technique. J. Stat. **09**, P09006 (2010)
40. Barra, A., Del Ferraro, G., Tantari, D.: Mean field spin glasses treated with PDE techniques. Eur. Phys. J. B **86**, 1–10 (2013)
41. Chen, H.B., Xia, J.: Hamilton-Jacobi equations from mean-field spin glasses (2022). arXiv:2201.12732
42. Mourrat, J.C.: Hamilton-Jacobi equations for mean-field disordered systems. Ann. Henri Lebesgue **4**, 453–484 (2021)
43. Mourrat, J.C.: The Parisi formula is a Hamilton-Jacobi equation in Wasserstein space. Can. J. Math. **74**(3), 607–629 (2022)
44. Dominguez, T., Mourrat, J.C.: Statistical mechanics of mean-field disordered systems: a Hamilton-Jacobi approach (2023). arXiv:2311.08976
45. Barra, A., Genovese, G., Guerra, F., Tantari, D.: How glassy are neural networks? J. Stat. **2012**(07), P07009 (2012)
46. Barra, A., Genovese, G., Guerra, F.: The replica symmetric approximation of the analogical neural network. J. Stat. Phys. **140**(4), 784–796 (2010)
47. Guerra, F.: Broken replica symmetry bounds in the mean field spin glass model. Commun. Math. Phys. **233**, 1–12 (2003)
48. Eliav, T., et al.: Multiscale representation of very large environments in the hippocampus of flying bats. Science **372**(6545), eabg4020 (2021)
49. Hopfield, J.J.: Neural networks and physical systems with emergent collective computational abilities. Proc. Natl. Acad. Sci. USA **79**, 2554–2558 (1982)
50. Amit, D.J., Gutfreund, H., Sompolinsky, H.: Storing infinite numbers of patterns in a spin-glass model of neural networks. Phys. Rev. Lett. **55**(14), 1530–1533 (1985)
51. Bovier, A., Picco, P.: Mathematical Aspects of Spin Glasses and Neural Networks. Springer, Boston (2012)
52. Alme, C.B., et al.: Place cells in the hippocampus: eleven maps for eleven rooms. Proc. Natl. Acad. Sci. USA **111**(52), 18428–18435 (2014)
53. Spalla, D., et al.: Can grid cell ensembles represent multiple spaces? Neural Comput. **31**(12), 2324–2347 (2019)
54. Barra, A.: The mean field Ising model trough interpolating techniques. J. Stat. Phys. **132**, 787–809 (2008)
55. Talagrand, M.: Spin Glasses: A Challenge for Mathematicians: Cavity and Mean Field Models. Springer, Berlin (2006)
56. Treves, A.: Are spin-glass effects relevant to understanding realistic auto-associative networks? J. Phys. A **24**, 2645 (1991)
57. Baskerville, N.P., et al.: The loss surfaces of neural networks with general activation functions. J. Stat. **6**, 064001 (2021)
58. Agliari, E., et al.: Generalized Guerra's interpolation schemes for dense associative neural networks. Neural Netw. **128**, 254–267 (2020)
59. Agliari, E., Barra, A., Longo, C., Tantari, D.: Neural networks retrieving Boolean patterns in a sea of Gaussian ones. J. Stat. Phys. **168**, 1085–1104 (2017)
60. Fentone, A., et al.: Unmasking the CA1 ensemble place code by exposures to small and large environments: more place cells and multiple, irregularly arranged, and expanded place fields in the larger space. J. Neurosci. **28**(44), 11250–11262 (2008)
61. Rich, P.D., Liaw, H.P., Lee, A.K.: Large environments reveal the statistical structure governing hippocampal representations. Science **345**(6198), 814–817 (2014)
62. Amit, D.J.: Neural networks counting chimes. Proc. Natl. Acad. Sci. USA **85**(7), 2141–2145 (1988)
63. Kosko, B.: Bidirectional associative memories. IEEE Trans. Syst Man Cybern. **18**(1), 49–60 (1988)
64. Barra, A., Catania, G., Decelle, A., Seoane, B.: Thermodynamics of bidirectional associative memories. J. Phys. A **56**(20), 205005 (2023)

65. Centonze, M.S., Kanter, I., Barra, A.: Statistical mechanics of learning via reverberation in bidirectional associative memories. Physica A **637**, 129512 (2024)
66. Thompson, L.T., Best, P.J.: Place cells and silent cells in the hippocampus of freely-behaving rats. J. Neurosci. **9**(7), 2382 (1989)
67. Solstad, T., Moser, E., Einevoll, G.: From grid cells to place cells: a mathematical model. Hippocampus **16**(12), 1026–1031 (2006)
68. Caswell, B., et al.: The boundary vector cell model of place cell firing and spatial memory. Rev. Neurosci. **17**(1), 71–98 (2006)
69. Lesieur, M.T., Lelarge, M., Krzakala, F., Zdeborova, L.: Statistical and computational phase transitions in spiked tensor estimation. In: IEEE International Symposium on Information Theory (ISIT), pp. 511-515 (2017)
70. Bovier, A.: Mathematical Aspects of Spin Glasses and Neural Networks. Springer Press, New York (2012)
71. Talagrand, M.: Mean Field Models for Spin Glasses: Voloume I: Basic Examples. Springer, Berlin (2010)
72. Mezard, M., Parisi, G., Virasoro, M.A.: Spin Glass Theory and Beyond: An Introduction to the Replica Method and Its Applications. World Sci. Publ., Singapore (1987)
73. Talagrand, M.: The Parisi formula. Ann. Math. **163**, 221–263 (2006)
74. Barra, A., Guerra, F., Mingione, E.: Interpolating the Sherrington-Kirkpatrick replica trick. Philos. Mag. **92**(1), 78 (2012)
75. Ghirlanda, S., Guerra, F.: General properties of overlap probability distributions in disordered spin systems. Towards Parisi ultrametricity. J. Phys. A **31**(46), 9149 (1998)
76. Genovese, G.: Universality in bipartite mean field spin glasses. J. Math. Phys. **53**, 123304 (2012)
77. Carmona, P., Hu, Y.: Universality in Sherrington-Kirkpatrick's spin glass model. Annal. Henri Poincare **42**, 215–222 (2006)