BEYOND DISORDER: UNVEILING COOPERATIVENESS IN MULTIDIRECTIONAL ASSOCIATIVE MEMORIES

Andrea Alessandrelli *
Department of Computer Science
University of Pisa
Pisa, 56127, Italy
{andrea.alessandrelli}@phd.unipi.it

Adriano Barra †
Department of Basic and Applied Sciences for Engineering
Sapienza University
Rome, 00161, Italy
{adriano.barra}@uniroma1.it

Andrea Ladiana
Department of Mathematics and Physics
University of Salento
Lecce, 73100, Italy
{andrea.ladiana}@unisalento.it

Andrea Lepre
Department of Mathematics and Physics
University of Salento
Lecce, 73100, Italy
{andrea.lepre}@unisalento.it

Federico Ricci-Tersenghi ‡
Department of Physics
Sapienza University
Rome, 00185, Italy
{federico.ricci}@uniroma1.it

ABSTRACT

By leveraging tools from the statistical mechanics of complex systems, in these short notes we extend the architecture of a neural network for hetero-associative memory called *three-directional associative memories* (TAM) to explore supervised and unsupervised learning protocols. In particular, by providing entropic-heterogeneous datasets to its various layers, we predict and quantify a new emergent phenomenon - that we term layer's "cooperativeness"- where the interplay of dataset entropies across network's layers enhances their retrieval capabilities beyond those they would have without reciprocal influence. Naively we would expect layers trained with less informative datasets to develop smaller retrieval regions compared to those pertaining to layers that experienced more information: this does not happen and all the retrieval regions settle to the same amplitude, allowing for optimal retrieval performance globally. This cooperative dynamics marks a significant advancement in understanding emergent computational capabilities within disordered systems.

*Also at National Institute for Nuclear Physics (INFN), Lecce branch, Italy
†Also at National Institute for Nuclear Physics (INFN), Rome branch, Italy
‡Also at National Institute for Nuclear Physics (INFN), Rome branch, Italy

1 INTRODUCTION

John Hopfield's legacy, recently recognized with the Nobel Prize in Physics, continues to inspire the study of associative memories. His groundbreaking work established the foundation for understanding how networks of simple units can give rise to complex, emergent behaviors. Thanks to modern variations of Hopfield's networks, they are experiencing a resurgence of interest Krotov (2023); Krotov & Hopfield (2016; 2020); Ramsauer et al. (2020); Barra et al. (2023); Demircigil et al. (2017); Lucibello & Mézard (2024); Agliari et al. (2020b).

Building on the layered associative Hebbian network architecture introduced for pattern recognition and disentanglement tasks Agliari et al. (2025a;b), this paper extends the exploration to new learning paradigms, pushing the boundaries of what these associative networks may accomplish.
While earlier work demonstrated how such networks could autonomously extract fundamental components from composite inputs—like identifying individual notes from musical chords—we delve deeper into the dynamics that govern these emergent capabilities. This dual framework enables a comprehensive examination of how varying levels of data structure and supervision influence the network's performance, especially under noisy or corrupted conditions Agliari et al. (2023a;b).
The most striking finding of our study is the emergence of a phenomenon we term "cooperativeness". A detailed examination of the network phase diagrams, parameterized by dataset entropy values across distinct layers, reveals that the retrieval performance of each layer is not merely a reflection of its corresponding dataset's quality. Instead, it is governed by the collective interplay of datasets' entropy distributions across all layers (similar to the optimization tradeoff present in the Information Plane idea Shwartz-Ziv & Tishby (2017)). Interestingly, heterogeneous entropy levels diminish the retrieval performance of layers trained over more high-quality datasets while enhancing the performance of those associated with noisier ones. This dynamics results in a balanced retrieval capacity across the network, a synergistic interaction absent in classical associative networks where layers operate independently. Notably, our approach leverages tools from the statistical mechanics of complex systems Mézard et al. (1987); Talagrand (2003); Agliari et al. (2020a); Franz et al. (2001); Agliari et al. (2021); Albanese et al. (2022), allowing us to rigorously predict and quantify this cooperative phenomenon through mathematically robust frameworks. This cooperative dynamic represents a significant advancement in understanding emergent intelligence in disordered systems.

2 THE SUPERVISED AND UNSUPERVISED HEBBIAN PROTOCOLS

We consider a neural network composed of three different families of binary neurons hereafter indicated by $\boldsymbol{\sigma} \equiv \{\sigma_i^A\}_{i=1,\dots,N_A}^{A=1,2,3}$ which interact in pairs via generalized Hebbian couplings (*vide infra*) whose goal lies in reconstructing the information encoded in a triplet of K binary archetypes $\{\boldsymbol{\xi}_\mu^A\}_{\mu=1,\dots,K}^{A=1,2,3}$ respectively of length N_1, N_2 and N_3. However, such archetypes are not provided directly to the network, hence the latter has to infer them by experiencing solely their noisy or corrupted versions. In particular, we assume that for each triplet of archetypes (μ, A), M examples $\eta_\mu^{a,A}$, with $a = 1, \dots, M$, are available, which are corrupted versions of the archetypes, such that for each $A = 1, 2, 3$ and $i = 1, \dots, N_A$ we have

$$\mathbb{P}(\eta_{i,\mu}^{a,A}|\xi_{i,\mu}^A) = \frac{1+r_A}{2}\delta(\eta_{i,\mu}^{a,A} - \xi_{i,\mu}^A) + \frac{1-r_A}{2}\delta(\eta_{i,\mu}^{a,A} + \xi_{i,\mu}^A) \tag{1}$$

where $r_A \in [0, 1]$ rules the quality of the dataset, i.e. for $r_A = 1$ the example matches perfectly the archetype, while for $r_A = 0$ it is totally random. To quantify the information content of the dataset it is useful to introduce the variables

$$\rho_A = \frac{1-r_A^2}{Mr_A^2}, \quad \rho_{AB} = \frac{1-r_A^2 r_B^2}{Mr_A^2 r_B^2} \quad \text{with} \;\; A, B \in \{1, 2, 3\} \tag{2}$$

that we shall refer to as the *dataset entropies* as deepened in Alemanno et al. (2023); Agliari et al. (2024). We observe that both ρ_A and ρ_{AB} approaches zero either when the examples perfectly match the archetypes (i.e., $r_A, r_B \to 1$), when the number of examples becomes infinite (i.e., $M \to \infty$), or under both conditions simultaneously.

The information related to the archetypes is encoded in the synaptic matrix, as outlined by the following cost function (or Hamiltonian):

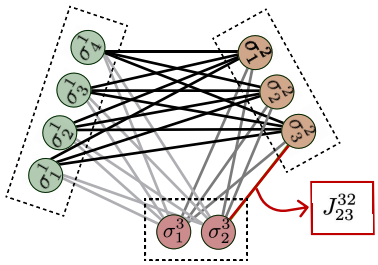

Figure 1: Schematic representation of the neural network described in equation 3 for the case $(N_1, N_2, N_3) = (4, 3, 2)$.

$$\mathcal{H}^{\boldsymbol{g}}_{\boldsymbol{N}}(\boldsymbol{\sigma}|\boldsymbol{J}) = -\frac{1}{2}\sum_{A\neq B}^{3} g_{AB} \sum_{i,j=1}^{N_A,N_B} J^{AB}_{ij}\sigma^A_i\sigma^B_j, \tag{3}$$

where $\boldsymbol{g} \in \mathbb{R}^{3\times 3}$ represents the strength of the interactions between different layers. The network architecture is sketched in Fig. 1. In order to let the network deal with examples rather than patterns, in this paper we inspect the following two variations of the above coupling matrix:

$$(J^{(unsup)})^{AB}_{ij} = \frac{1}{r_A r_B\sqrt{N_A N_B(1+\rho_A)(1+\rho_B)}}\sum_{\mu=1}^{K}\frac{1}{M}\sum_{a=1}^{M}\eta^{a,A}_{i,\mu}\eta^{a,B}_{j,\mu}, \tag{4}$$

$$(J^{(sup)})^{AB}_{ij} = \sqrt{\frac{(1+\rho_A)(1+\rho_B)}{N_A N_B}}\sum_{\mu=1}^{K}\left(\frac{1}{M}\sum_{a=1}^{M}\eta^{a,A}_{i,\mu}\right)\left(\frac{1}{M}\sum_{b=1}^{M}\eta^{b,B}_{j,\mu}\right). \tag{5}$$

In the first case, there is no external teacher who knows the labels and can organize the examples based on archetypes, as occurs in the second scenario. This distinction is why the two formulations are associated with unsupervised and supervised learning protocols, respectively Agliari et al. (2023a;b); Alemanno et al. (2023); Agliari et al. (2024).

As calculations will be performed in the thermodynamic limit, where $N_1, N_2, N_3 \to \infty$, it is important to highlight that the sizes of the three layers—and consequently the lengths of the corresponding examples—as well as the number of samples in each dataset, can differ from one another, meaning $N_1 \neq N_2 \neq N_3$ and $M_1 \neq M_2 \neq M_3$. Furthermore, despite these differences, the number of archetypes remains constant across all layers, denoted by K for each. Moreover, in order to ensure a meaningful asymptotic (thermodynamic) behavior, the ratio between the number of patterns and their respective lengths must remain finite. To achieve this, we impose the following conditions on K, N_1, N_2 and N_3:

$$\lim_{N_1,N_3\to\infty}\sqrt{\frac{N_1}{N_3}} = \alpha\,, \quad \lim_{N_1,N_2\to\infty}\sqrt{\frac{N_1}{N_2}} = \theta\,, \quad \lim_{N_1,K\to\infty}\frac{K}{N_1} = \gamma \tag{6}$$

where $\alpha, \theta, \gamma \in \mathbb{R}^+$ are *control parameters*. The parameter γ characterizes the storage capacity of the network, and our focus will be on the high-storage regime, where $\gamma > 0$.

Pivotal for a statistical mechanical analysis is the study of the quenched free energy in the thermodynamic limit, defined as

$$\mathcal{A}^{\boldsymbol{g}}_{\alpha,\theta,\gamma}(\beta) = \lim_{N_1,N_2,N_3\to\infty}\mathbb{E}\frac{1}{L}\ln\left[\sum_{\{\boldsymbol{\sigma}^1,\boldsymbol{\sigma}^2,\boldsymbol{\sigma}^3\}}\exp\left(-\beta\mathcal{H}^{\boldsymbol{g}}_{\boldsymbol{N}}(\boldsymbol{\sigma}|\boldsymbol{J})\right)\right], \tag{7}$$

where $\mathbb{E}$ averages over the $\boldsymbol{J}$ distributions, $L = \frac{1}{3}\left(\frac{1}{\sqrt{N_1N_2}} + \frac{1}{\sqrt{N_1N_3}} + \frac{1}{\sqrt{N_2N_3}}\right)$ and $\beta \in \mathbb{R}^+$ tunes the fast noise in the network such that for $\beta \to 0^+$ network's dynamics is a pure random walk in the neural configuration space (and any configuration is equally likely to occur), while for $\beta \to +\infty$ its

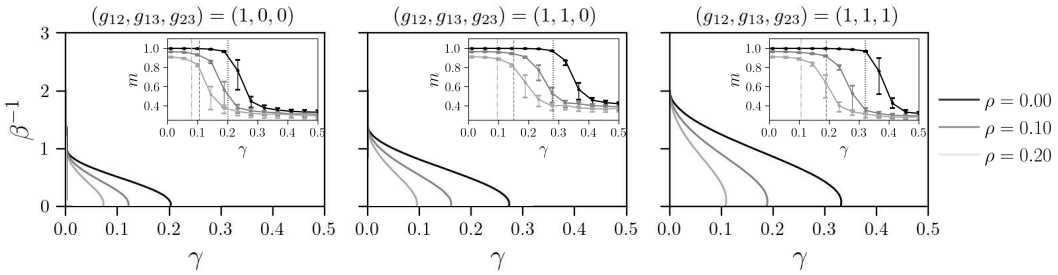

Figure 2: Phase diagrams of the TAM network in the supervised setting in the noise versus storage plane at $\alpha = \theta = 1$. The analysis includes different inter-layer interaction strengths and various values of the dataset entropy ρ, as indicated in the legend. Each solid line depicts the phase transition for the whole network splitting the working region (bottom left) -where archetypes are learned and thus their retrieval and generalization allowed- from the blackout region (up right) -where spin glass effects prevail- for a specific value of dataset entropy i.e., $\rho_1 = \rho_2 = \rho_3 = \rho$. The retrieval region is determined by the conditions $|m^1_{\xi_1}|, |m^2_{\xi_1}|, |m^3_{\xi_1}| > 0$: these inequalities are all satisfied simultaneously in the region below the solid line, while above it, all magnetizations vanish. The influence of ρ is clearly visible: as ρ increases, the retrieval region progressively shrinks in all diagrams. For $\rho = 0$, we recover the results of the standard Kosko's BAM case Kosko (1988) (first panel) and the novel ones pertaining to the TAM Agliari et al. (2025a) (second and third panels). In the insets of each plots: MC simulation at zero-fast noise ($\beta^{-1} = 0$) with a symmetric network ($N_1 = N_2 = N_3 = 1000$), showing the evolution of the Mattis magnetizations m_{ξ_1} across the layers as a function of network load (γ) for different ρ. The simulations agree with theoretical predictions, correctly depicting the maximum load beyond which the network stops functioning.

dynamics steepest descends toward the minima of the Hamiltonian equation 3.
More precisely, our aim is to find an expression of $\mathcal{A}^{\boldsymbol{g}}_{\alpha,\theta,\gamma}(\beta)$ in terms of a suitable set of macroscopic observables able to capture the global behavior of the system: these *order parameters* are the K archetype (ground truth) Mattis magnetizations that assess the quality of network's retrieval, defined as

$$m^A_{\xi^A_\mu} \quad = \quad \frac{1}{N_A} \sum_{i=1}^{N_A} \xi^A_{(i,\mu)} \sigma^A_i, \tag{8}$$

such that $m^A_{\xi^A_\mu} = 1$ accounts for a perfect retrieval of the archetype ξ^μ by layer A, its lacking being accounted by $m^A_{\xi^A_\mu} = 0$.

The application of Guerra's interpolation method Guerra (2001) allows us to derive an explicit expression for the quenched free energy in the thermodynamic limit, in terms of the control parameters ($\beta, \gamma, \alpha, \theta$ and $\boldsymbol{g}$) and the order ones, under the assumption of replica symmetry. This assumption implies that, in the thermodynamic limit, the observables defined in equation 8 exhibit negligible fluctuations around their means. Once the quenched free energy is expressed in terms of the control and order parameters, we can proceed to extremize it with respect to the order parameters. This process results in a set of self-consistent equations, whose solutions describe the behavior of the order parameters as functions of the control ones. By analyzing these solutions, we can construct the phase diagram, identifying regions in the control parameter space where the network successfully learns the archetypes from the examples, retrieves them, and is thus capable of generalization.

Focusing specifically on the retrieval of the pattern triplet labeled by $\mu = 1$ –without loss of generality– we can extract an explicit expression for the self-consistency equations governing the order parameters, under the large dataset limit assumption (i.e. $M \gg 1$) which allows us to construct the model self-diagrams both in the supervised and unsupervised scenarios.
For the sake of brevity, focusing on the supervised protocol (i.e. assuming the network has the coupling equation 5)[1], we first investigate the case of datasets sharing the same entropy (i.e.,

[1]For the unsupervised counterpart, outcomes are qualitatively similar, the only difference being in the definition of the dataset entropy (where ρ_A is replaced by ρ_{AB}).

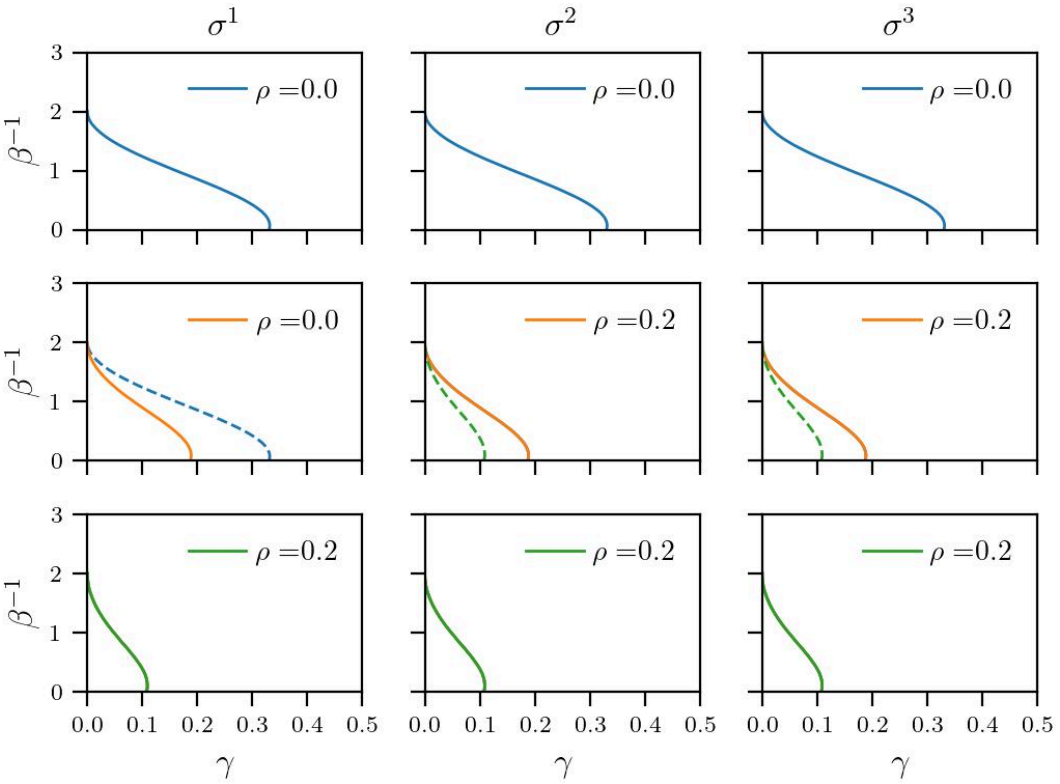

Figure 3: Phase diagrams in the symmetric case ($\alpha = \theta = 1$, $(g_{12}, g_{13}, g_{23}) = (1, 1, 1)$) highlighting the *cooperative behavior* among layers. Retrieval regions are shown for layers $\boldsymbol{\sigma}^1$, $\boldsymbol{\sigma}^2$, and $\boldsymbol{\sigma}^3$ under different dataset's entropy values: (top) zero-entropy datasets ($\rho_1 = \rho_2 = \rho_3 = 0$); (middle) heterogeneous-entropy datasets ($\rho_2 = \rho_3 = 0.2$, $\rho_1 = 0$); (bottom) homogeneous-entropy datasets ($\rho_1 = \rho_2 = \rho_3 = 0.2$). The comparison highlights that the amplitude of the retrieval regions depends on the interplay between dataset entropies across layers: in the middle row, the first layer's retrieval region shrinks -orange curve- despite the noiseless inputs would allow to reach the dashed blue line, allowing the noisy layers to expand theirs (from dashed green to orange boundary), an effect impossible without inter-layer interactions. For brevity, the case $\rho_1 = 0$, $\rho_2 = 0$, $\rho_3 = 0.2$ is not shown in the figure; however, it exhibits a slight deterioration in retrieval performance compared to the middle row.

$\rho_1 = \rho_2 = \rho_3 = \rho$ over all the layers): results are summarized by the phase diagrams in the inter-layer activation strength $\boldsymbol{g}$ vs noise presented in Fig. 2. As shown in Fig. 2, increasing ρ leads to a systematic reduction of the retrieval region—i.e., the domain in which the network can successfully reconstruct patterns from examples. As expected, for $\rho = 0$ results collapse to those of the standard Hebbian-like TAM scenario Agliari et al. (2025a). The main reward of this analysis is the determination of the thresholds for learning, namely the minimal values of ρ required to sustain a non-zero retrieval region, as it allows us to predict, a priori, through equation 2, the relationship between dataset quality (r_A) and dataset size (M) providing crucial insights for optimizing learning processes.

Then, by keeping the network symmetric in terms of both sizes and activation strengths [2] (i.e., $\alpha = \theta$, $g_{12} = g_{13} = g_{23} = 1$), we deepened the analysis for datasets characterized by different entropies: results of this investigation are shown in Fig. 3. A detailed examination of the resulting phase diagrams, parameterized by the entropy values across the distinct layers, reveals an intriguing phenomenon: the retrieval region pertaining to each layer does not merely reflect the entropy of its corresponding dataset but is instead governed by the collective interplay of entropy distributions

[2]The general case in which the activation strengths differ from 1, is being studied, and the results will be presented in a forthcoming publication.

across all the layers. Indeed, when the network has to handle entropic-heterogeneous datasets, a redistribution effect spontaneously appears: the retrieval region pertaining to the layer associated with the most informative dataset shrinks, while those of the layers at work with messy datasets enlarge, benefiting from their mutual interaction. This results in a more balanced retrieval performance across the network's layers. This emergent effect highlights their intrinsic cooperative nature, wherein the presence of a low entropy training set can enhance the performance of noisier ones, fostering a form of mutual reinforcement. We propose the term "*cooperativeness*" to describe this emergent property of the network, a feature that inherently arises from the reciprocal influence among the layers. This cooperative behavior is not merely a byproduct of parameter tuning but an intrinsic characteristic of the multipartite network structure, which only becomes evident through a comprehensive analytical treatment of the system's self-consistency equations and whose presence can be crucial when dealing with dirty or small datasets.

3 Conclusion

Our work focuses on Hebbian information processing by a hetero-associative model able to cope with three sources of information simultaneously (i.e. the TAM network). In our setting, rather than directly providing the network with the original archetypes (i.e. the patterns), we expose it to examples—corrupted versions of them—thereby assessing its ability to learn and generalize from incomplete or noisy data.

Through a statistical mechanics analysis, we obtained the phase diagrams of the network: the latter highlights how the amplitude of the retrieval region is affected by the entropy of the experienced datasets. Our results emphasize that successful pattern retrieval depends critically on both the quality and the quantity of examples provided, much like how human learning benefits from both clear instruction and repeated exposure.

The most noteworthy finding of this study is the *cooperative behavior* emerging among the layers of the network. A detailed examination of the phase diagrams, further corroborated by extensive numerical simulations, has revealed that layers associated with higher-informative datasets actively assist those provided with lower-informative ones, enhancing the amplitude of their retrieval regions by sacrificing their own. This effect arises because the lower the entropy of a dataset, the larger the retrieval region of the corresponding layer; thus, the stronger layer, benefiting from a higher-quality dataset, can partially reduce its own retrieval region for the overall advantage of the system. This trade-off results in an optimal redistribution of learning and retrieval capacity across the network, fostering a form of mutual reinforcement that is absent in classical associative memory models.

This phenomenon is particularly striking because it has no direct counterpart in the existing literature: our findings suggest that cooperativity may play a crucial and previously overlooked role in multidirectional associative models, potentially offering new insights into both artificial and biological memory systems and their applications.

References

Elena Agliari, Francesco Alemanno, Adriano Barra, Martino Centonze, and Alberto Fachechi. Neural networks with a redundant representation: Detecting the undetectable. *Physical review letters*, 124(2):028301, 2020a.

Elena Agliari, Adriano Barra, Peter Sollich, and Lenka Zdeborová. Machine learning and statistical physics: theory, inspiration, application. *J. Phys. A: special issue*, 2020b.

Elena Agliari, Linda Albanese, Francesco Alemanno, and Alberto Fachechi. A transport equation approach for deep neural networks with quenched random weights. *Journal of Physics A: Mathematical and Theoretical*, 54(50):505004, 2021.

Elena Agliari, Linda Albanese, Francesco Alemanno, Andrea Alessandrelli, Adriano Barra, Fosca Giannotti, Daniele Lotito, and Dino Pedreschi. Dense hebbian neural networks: a replica symmetric picture of supervised learning. *Physica A: Statistical Mechanics and its Applications*, 626: 129076, 2023a.

Elena Agliari, Linda Albanese, Francesco Alemanno, Andrea Alessandrelli, Adriano Barra, Fosca Giannotti, Daniele Lotito, and Dino Pedreschi. Dense hebbian neural networks: A replica sym-

metric picture of unsupervised learning. *Physica A: Statistical Mechanics and its Applications*, 627:129143, 2023b.

Elena Agliari, Francesco Alemanno, Miriam Aquaro, Adriano Barra, Fabrizio Durante, and Ido Kanter. Hebbian dreaming for small datasets. *Neural Networks*, 173:106174, 2024.

Elena Agliari, Andrea Alessandrelli, Adriano Barra, Martino Salomone Centonze, and Federico Ricci-Tersenghi. Generalized hetero-associative neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2025(1):013302, 2025a.

Elena Agliari, Andrea Alessandrelli, Adriano Barra, Martino Salomone Centonze, and Federico Ricci-Tersenghi. Networks of neural networks: more is different, 2025b. URL https://arxiv.org/abs/2501.16789.

Linda Albanese, Francesco Alemanno, Andrea Alessandrelli, and Adriano Barra. Replica symmetry breaking in dense hebbian neural networks. *Journal of Statistical Physics*, 189(2):24, 2022.

Francesco Alemanno, Miriam Aquaro, Ido Kanter, Adriano Barra, and Elena Agliari. Supervised hebbian learning. *Europhysics Letters*, 141(1):11001, 2023.

Adriano Barra, Giovanni Catania, Aurélien Decelle, and Beatriz Seoane. Thermodynamics of bidirectional associative memories. *Journal of Physics A: Mathematical and Theoretical*, 56(20): 205005, 2023.

Mete Demircigil, Judith Heusel, Matthias Löwe, Sven Upgang, and Franck Vermet. On a model of associative memory with huge storage capacity. *Journal of Statistical Physics*, 168:288–299, 2017.

Silvio Franz, Michele Leone, Federico Ricci-Tersenghi, and Riccardo Zecchina. Exact solutions for diluted spin glasses and optimization problems. *Physical review letters*, 87.12:127209, 2001.

Francesco Guerra. Sum rules for the free energy in the mean field spin glass model. *Fields Institute Communications*, 30(11), 2001.

Bart Kosko. Bidirectional associative memories. *IEEE Transactions on Systems, man, and Cybernetics*, 18(1):49–60, 1988.

Dmitry Krotov. A new frontier for hopfield networks. *Nature Reviews Physics*, 5(7):366–367, 2023.

Dmitry Krotov and John Hopfield. Large associative memory problem in neurobiology and machine learning. *arXiv preprint arXiv:2008.06996*, 2020.

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

Carlo Lucibello and Marc Mézard. Exponential capacity of dense associative memories. *Physical Review Letters*, 132(7):077301, 2024.

Marc Mézard, Giorgio Parisi, and Miguel Angel Virasoro. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific Publishing Company, 1987.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, et al. Hopfield networks is all you need. *arXiv preprint arXiv:2008.02217*, 2020.

Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

Michel Talagrand. *Spin glasses: a challenge for mathematicians*. Springer Press, 2003.

ACKNOWLEDGMENTS

A.B. acknowledges PRIN 2022 grant *Statistical Mechanics of Learning Machines: from algorithmic and information theoretical limits to new biologically inspired paradigms* n. 20229T9EAT funded by European Union - Next Generation EU.
A.B., A.A. are members of GNFM-INdAM which is acknowledged.
The research has received financial support from the 'National Centre for HPC, Big Data and Quantum Computing–HPC', Projects CN-00000013, CUP B83C22002940006, NRP Mission 4 Component 2 Investment 1.5, Funded by the European Union–NextGenerationEU.