ASSOCIATIVE MEMORY LEARNING THROUGH REDUNDANCY MAXIMIZATION

Mark Blümel[1,2,*], **Andreas C. Schneider**[1,2,*], **David A. Ehrlich**[3,2], **Valentin Neuhaus**[1,2], **Marcel Graetz**[4], **Michael Wibral**[3,2], **Abdullah Makkeh**[3,2,†], **Viola Priesemann**[2,1,†]

[1]Faculty of Physics, Institute for the Dynamics of Complex Systems, University of Göttingen
[2]Max Planck Institute for Dynamics and Self-Organization, Göttingen
[3]Göttingen Campus Institute for Dynamics of Biological Networks, University of Göttingen
[4]Champalimaud Centre for the Unknown, Lisbon, Portugal
email: {mark.bluemel, andreas.schneider, abdullah.makkeh, viola.priesemann}@ds.mpg.de

ABSTRACT

Hopfield networks mark an important milestone in the development of modern artificial intelligence architectures. In this work, we argue that a foundational principle for solving such associative memory problems at the neuron scale is to promote redundancy between the input pattern and the network's internal state in the neurons' activity. We demonstrate how to quantify this redundancy in classical Hebbian Hopfield networks using Partial Information Decomposition (PID), and reveal that redundancy plays a dominant role compared to synergy or uniqueness when operating below capacity. Beyond analysis, we show that redundancy can be used as a learning goal for Hopfield networks by constructing associative memory networks from neurons that directly optimize PID-based goal functions. In experiments, we find that these "infomorphic" Hopfield networks greatly outperform the original Hebbian networks and achieve promising performance with the potential for further improvement. This work offers novel insights into how associative memory functions at an information-theoretic level of abstraction and opens pathways to designing new learning rules for different associative memory architectures based on redundancy maximization goals.

1 INTRODUCTION

Associative memory, a form of content-addressable memory network in which patterns are retrieved from noisy instances through recurrent dynamics, marks an important paradigm in neural learning (Hopfield, 1982; Ramsauer et al., 2021). Originally, such Hopfield networks have been trained using the biologically-inspired Hebbian learning rule, but more recently new learning rules have been introduced that enhance the capacity and noise resistance (Hillar et al., 2015; Tolmachev & Manton, 2020). Nevertheless, a key question remains: Is there an underlying principle that governs associative memory? And, if so, can it be exploited directly to improve performance?

Hopfield networks store patterns as attractors of their recurrent neural dynamics. To create these attractors, the weights of the network need to be trained by providing the patterns as a teaching signal. How the information of the recurrent dynamics and the teaching signal together predict the neuron's firing thus becomes pivotal to the network's performance. To describe this relation in an abstract, implementation-independent manner, we utilize information theory. Using mutual information, it is possible to quantify how much information about the neuron's output is contained either in the input from other neurons (the recurrent or lateral input) or the teaching signal (referred to as receptive input). However, using only mutual information it is not possible to differentiate *how* that information is carried in unique, redundant or synergistic ways. This differentiation can be done by a recent extension to information theory called Partial Information Decomposition (PID, Williams & Beer (2010)). In this work, we use PID both as a tool for analysis, revealing that classical

[*]Equally contributing first authors; [†]Equally contributing last authors

Hebbian Hopfield neurons are dominated by redundancy, and for constructing local goal functions which optimize for certain information processing goals directly.

The main contributions of this work are (i) information-theoretic analysis of a classical Hebbian Hopfield network which reveals that redundancy between the lateral input and the receptive input (teaching signal) dominates the neurons' output when operating below capacity, (ii) the construction of "infomorphic" associative memory neurons that can directly maximize a local PID goal function building on work by Makkeh et al. (2025), and (iii) experimental results showcasing that infomorphic neurons that maximize redundancy significantly outperform the original learning rule.

2 BACKGROUND

To find an underlying principle of associative memory, we investigate the information processing necessary at a single neuron. From an information-theoretic viewpoint, the neurons can be interpreted as channels that receive information as a weighted sum L of outputs of lateral neurons (i.e. the outputs of all other neurons in the network) to produce a binary output signal Y. The entropy of a neuron's output $H(Y)$ can therefore be quantified by the mutual information $I(Y : L)$ and the residual entropy as $H(Y) = I(Y : L) + H(Y \mid L)$.

For analysis or training, the receptive input (containing a single pattern element) is available as an additional random variable R that can help determine the task-relevance of the individual neuronal output activity through the mutual information $I(Y : R)$. To understand the local information *processing* necessary for associative memory function, however, it is important to understand how this information relates to the information from the lateral input $I(Y : L)$. While the information that the receptive input and the lateral input together hold about the output Y can be quantified by the joint mutual information $I(Y : R, L)$, classical information theory is unable to dissect *how* this information is provided by the two sources R and L (see Figure 1B): Some parts of the neuron's firing might be explainable *uniquely* by the lateral input (denoted by $\Pi_{\mathrm{unq},L}$) and thus be unrelated to the receptive input, or uniquely by the receptive input, i.e., relevant but not encoded in L ($\Pi_{\mathrm{unq},R}$). Other parts of the output information may be carried *redundantly* (Π_{red}) by both sources, meaning they are both relevant and encoded in L, while yet others may be carried *synergistically* (Π_{syn}), meaning that both information sources are necessary to uncover this piece of information. Enumerating and quantifying these information *atoms* is the subject of Partial Information Decomposition (Williams & Beer, 2010; Gutknecht et al., 2021).

The four PID atoms are related to the three classical mutual information quantities with the target variable Y through the so-called consistency equations

$$I(Y : R, L) = \Pi_{\mathrm{red}} + \Pi_{\mathrm{unq},R} + \Pi_{\mathrm{unq},L} + \Pi_{\mathrm{syn}} \tag{1}$$

$$I(Y : R) = \Pi_{\mathrm{red}} + \Pi_{\mathrm{unq},R} \tag{2}$$

$$I(Y : L) = \Pi_{\mathrm{red}} + \Pi_{\mathrm{unq},L}. \tag{3}$$

Note that this system is underdetermined with four unknown quantities Π, but only three classical mutual information quantities providing constraints. To resolve this underdetermination, an additional concept needs to be defined, which is usually a measure for redundancy. Throughout the literature, a plethora of different redundancy measures have been devised, which fulfill different requirements and have different operational interpretations (e.g. Lizier et al., 2018, and references therein). Throughout this work, we use the $I_\cap^{\mathrm{sx}}$ measure introduced by Makkeh et al. (2021), due to its differentiability, which is essential for being able to optimize it (see Appendix B).

3 REDUNDANCY IN HEBBIAN LEARNING

As a first step we analyze the information dynamics in classical Hopfield networks to examine the role of redundant information for pattern memorization. We train Hopfield networks with one-shot Hebbian learning (described in Algorithm 1), then initialize them in one of the memorized patterns and conduct a single update step. We then quantify which information processing each neuron performs by computing the PID of $I(Y : R, L)$ between the output of each neuron Y as the PID target, and the receptive input R (task-relevant information) and the lateral input L as the PID sources. The analysis reveals that when the Hopfield network memorizes patterns below its capacity

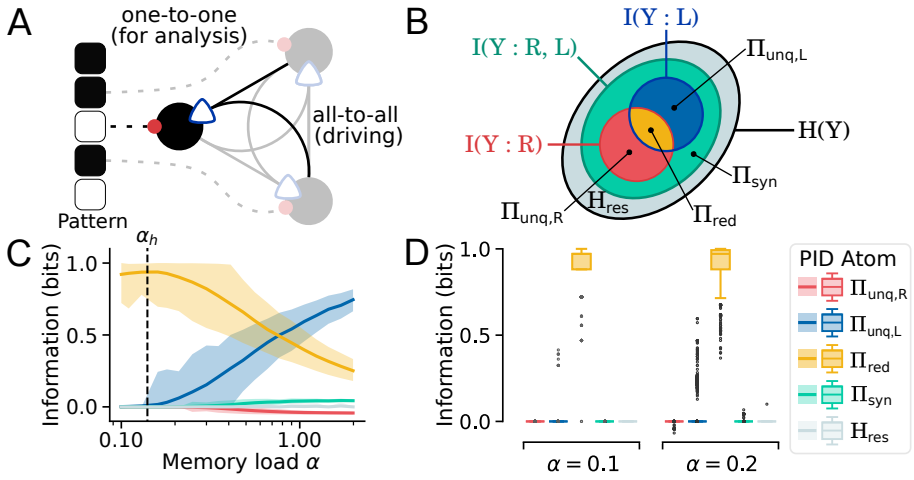

Figure 1: **Hebbian learning exhibits high redundant information of the lateral input and the receptive input about the neurons' output, which decreases significantly when Hebbian learning fails. A:** The setup for measuring PID: in addition to the lateral input L, each neuron receives its corresponding element of the pattern via a second source variable R. **B:** PID diagram for two source variables, decomposing the mutual information $I(Y : R, L)$ into unique, redundant, synergistic information atoms. **C:** PID profile for increasing memory load. The memory capacity $\alpha_h \approx 0.14$ is indicated in black. **D:** PID profiles for the network at a load of $\alpha = 0.1$ and $\alpha = 0.2$. Below the memory capacity, redundancy is high with a small number of outlier neurons with lower redundancy. Just above the memory capacity, the number of outliers increases and outliers with higher unique information appear. The median PID terms do not change significantly, however.

$\alpha_h \approx 0.14$ (Hopfield, 1982), redundant information constitutes most of $I(Y : R, L)$ (see Figure 1C). However, redundant information begins decreasing when this capacity is exceeded, while unique lateral information gradually increases (see Figure 1C). These information dynamics suggest that Hebbian Hopfield neurons encode information that is redundant between their receptive input and lateral input to enable associative memory learning, and that lateral information unrelated to the receptive input leads to memorization errors above capacity.

Moreover, beyond the memory capacity, the trend of decreasing redundancy and increasing unique lateral information is not uniform across the neurons in the network (see Figure 1D). While most neurons still retain high redundancy, a fraction of neurons increase their lateral unique information. This suggests that, above (and starting even already below) the memory capacity, islands of "dysfunctional" neurons emerge, while most of the network behaves correctly. Forcing these dysfunctional neurons to also encode redundancy by directly optimizing for redundancy across the whole network thus appears as a likely path towards improving network performance.

4 Infomorphic Networks

Building on the empirical evidence of high redundancy in classical Hebbian Hopfield networks, we now demonstrate that Hopfield networks can be trained by maximizing this redundancy directly. To do so, we take a constructive approach by devising 'infomorphic Hopfield neurons" for associative memory that directly optimize a PID-based goal function (Makkeh et al., 2025).

During memory retrieval, an infomorphic Hopfield neuron operates identically to a classical Hopfield neuron. The neuron integrates outputs from all other neurons in a weighted sum, which is then fed into the binary activation function $\mathrm{sgn}(L)$. After initializing in a pattern of choice, the network is run until a fixed point is reached or for up to $N_{\text{iter}} = 100$ time steps.

During the training phase, each neuron is provided with a receptive input (a single element of the pattern multiplied with a weight W_r) as well as with the lateral input L (the standard weighted sum over the output of other neurons) by augmenting the activation function to $\sigma(R + L)$. The resulting activation is then interpreted as a firing probability, according to which the binary output is drawn.

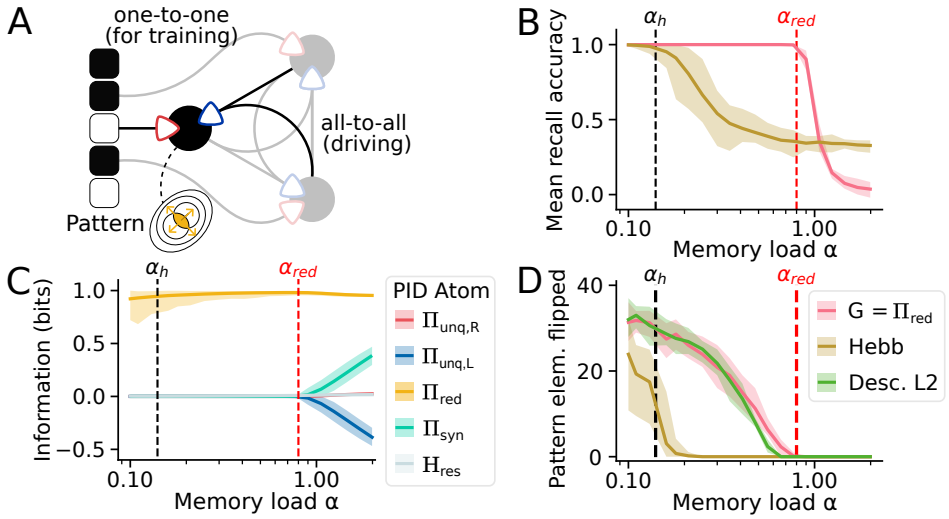

Figure 2: **Explicit maximization of redundant information in infomorphic networks leads to increased capacity and noise resistance. A:** The extended setup for infomorphic learning, where the receptive input is used as a second input to the neurons. During training, a stochastic activation function is used. **B:** The performance of the infomorphic rule as a function of memory load α, measured in terms of the overlap between patterns and stable states of the system. For comparison, the performance of the original Hebbian rule is shown. **C:** The PID profile for the infomorphic rule maximizing redundancy. Above the capacity redundancy only slightly declines, but other atoms become more prominent. **D:** Noise resistance for the redundancy rule as a function of memory load. For comparison, the original Hebbian rule as well as the rule 'Descent L2' (with λ=0.5) from Tolmachev & Manton (2020) is included.

Given this second information source R, a general PID-based learning goal can be formulated as a weighted sum of PID atoms as

$$G = \gamma_{\text{red}}\Pi_{\text{red}} + \gamma_{\text{unq},R}\Pi_{\text{unq},R} + \gamma_{\text{unq},L}\Pi_{\text{unq},L} + \gamma_{\text{syn}}\Pi_{\text{syn}} + \gamma_H H_{\text{res}}, \tag{4}$$

where the $\boldsymbol{\gamma}$ are fixed coefficients and $H_{\text{res}} = H(Y \mid R, L)$ is the residual entropy of the output not explained by either R or L. To train, we initialize the network in each of the memory patterns, present the elements of the same pattern in the external input R, and run the network for one time step to obtain firing probabilities $p(Y = y \mid R = r, L = l)$. In addition, we construct the empirical probability mass function $p(R = r, L = l)$ from the inputs R and L, binned to a $(2, 60)$ grid. From this, the PID terms and the loss function are computed according to Equation 4. Finally, the weights are updated using gradient descent with a learning rate of 0.1. Pseudocode of the described training procedure can be found in algorithm 2.

5 Maximizing Redundant Information in Infomorphic Hopfield Networks

Our experiments reveal that using the simple goal function $G = \Pi_{\text{red}}$, infomorphic Hopfield networks are able to significantly increase their capacity to a memory load of $\alpha_{\text{red}} \approx 0.8$ compared to $\alpha_{\text{h}} \approx 0.14$ of Hebbian learning (see Figure 2B). Furthermore, their noise resistance is promising and on par with the associative memory learning rules presented in Tolmachev & Manton (2020) (see Figure 2D). The increased memory capacity by maximizing redundant information shows the sufficiency of redundancy as an underlying information processing principle in associative memory learning.

Examining the PID profile for $\alpha \leq 0.8$, the redundant information is successfully maximized and constituted most of $I(Y : R, L)$ as desired (see Figure 2C). At the onset of failure of infomorphic Hopfield networks ($\alpha > 0.8$ in Figure 2), however, the redundancy is still maximized, but at the expense of driving the unique information of L to be negative indicating that each neuron's activity is not fully predictable by their lateral input (for more details, see Appendix B).

6 DISCUSSION

This work provides a step towards understanding the underlying principle that governs associative memories at the individual neuron scale. Such an underlying principle can be revealed by studying the information processing of the neurons in associative networks. In particular, by regarding the neuron as an information processor, Partial Information Decomposition quantifies the information contribution of the receptive input and the lateral input to the neuron's output—revealing in which way the task-relevant information is encoded by the internal state in order to store the patterns (see section 2). We argue that for an associative memory network to store patterns, the lateral input and the receptive input need to redundantly determine the firing of each neuron.

To check the validity of this claim, we first analyze the information processing of Hebbian Hopfield networks (see section 3). While the information dynamics of the Hebbian Hopfield neurons are indeed dominated by redundancy up to their memory capacity, the redundancy starts degrading and gets replaced by unique lateral information when memory capacity is exceeded, implying that a fraction of neurons becomes unable to encode information relevant to the patterns (see Figure 1C).

Subsequently, we show that maximizing local redundant information is sufficient for associative memory learning (see section 4). To this end, we construct infomorphic Hopfield networks such that every neuron strives to maximize the redundant information that receptive input and its lateral input provide about its output (see section 5). The simulations show that infomorphic Hopfield networks strongly outperform Hebbian Hopfield networks, their noise resistance being on par with state-of-the-art associative memory learning rules (see Figure 2B,D). These results assert that promoting redundancy can be regarded as an underlying principle to associative memory learning.

Related Works. Information theory has been employed to understand various properties of Hopfield networks. In particular, Dominguez et al. (2004; 2009) found that the mutual information of the pattern and the internal state $I(R:L)$ was optimal for a sparsely connected network which ultimately enhanced the memory capacity. Montazeri & Schmidt (2024) concluded using $I(R:L)$ that the robustness of Hopfield networks depends on the sparsity of the patterns. Knoblauch et al. (2010) quantified the information storage of Hopfield networks by measuring the mutual information between the patterns and the synaptic weights. Beyond binary Hopfield networks, Bollé et al. (2000) showed that $I(R:L)$ during retrieval is related to the size of the basins of attraction in ternary Hopfield networks. Dominguez & Korutcheva (2000) quantified mutual information in mean-field ternary Hopfield networks to construct an energy function which optimized their retrieval properties. Makkeh et al. (2025) constructed infomorphic associative memories that learn by optimizing the goal function $G = 0.9\Pi_{\mathrm{red}} + 0.1I(Y:R,L)$, employing identical neuron properties for training and testing, but a more complex activation function. In addition to analyzing Hebbian Hopfield networks, here we implement infomorphic Hopfield networks with a simpler activation function $A(R,L) = \sigma(R+L)$, which gets replaced by a Heaviside activation during testing. These changes establish a more clear-cut connection to traditional associative memories, while also leading to a significantly higher capacity than the one reported by Makkeh et al. (2025).

Limitations and Outlook. Our simulations show that while maximizing redundancy is capable of significantly improving the memory capacity, promoting redundancy alone proves unable to reach the theoretically optimal memory capacity of twice the number of neurons (Cover, 1965). To achieve higher capacity, it might be beneficial to suppress other quantities like unique lateral or synergistic information. Future work should conduct hyperparameter optimization to determine the best goal function parameters. Beyond classical associative memories, Krotov & Hopfield (2016); Krotov (2023) showed that the aforementioned optimal capacity can be exponentially improved by opting for a hierarchical architecture known as Dense Associative Memory. In addition, Schneider et al. (2025) constructed hierarchical infomorphic network architectures with up to three layers, significantly improving performance on supervised learning. This opens an intriguing avenue for future research: Utilizing the framework of infomorphic networks to understand which information processing mechanisms are sufficient to implement Dense Associative Memories.

In summary, PID and infomorphic networks constitute a powerful pair of analytic and constructive tools, providing a novel level of interpretability that can in the future be used to understand the strengths and shortcomings of other associative memory learning rules and help in designing novel network architectures and learning rules.

Acknowledgements We would like to thank Giordano De Marzo as well as the members of the Wibral and Priesemann groups for valuable feedback and discussions. M.B. was supported by the Max Planck Society. A.S., V.N. and V.P. were funded via the MBExC by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy-EXC 2067/1-390729940. A.S., V.N., M.W., and V.P. were supported and funded by the DFG – GRK2906 – project number 502807174 and acknowledge support from the Max Planck Society. V.N. was partly supported by the Else Kröner Fresenius Foundation via the Else Kröner Fresenius Center for Optogenetic Therapies. D.A.E. and M.W. were supported by a funding from the Ministry for Science and Education of Lower Saxony and the Volkswagen Foundation through the "Niedersächsisches Vorab" under the program "Big Data in den Lebenswissenschaften" – project "Deep learning techniques for association studies of transcriptome and systems dynamics in tissue morphogenesis". M.W. and A.M. are employed at the Campus Institute for Dynamics of Biological Networks (CIDBN) funded by the Volkswagenstiftung. M.W., V.P. and A.M. received funding from the DFG via the SFB 1528 "Cognition of Interaction" - project-ID 454648639. M.W. was supported by the flagship science initiative of the European Commission's Future and Emerging Technologies program under the Human Brain project, HBP-SP3.1-SGA1-T3.6.1. M.G. received a PhD scholarship by the Champalimaud Foundation.

REFERENCES

Désiré Bollé, DRC Dominguez, and Shun-ichi Amari. Mutual information of sparsely coded associative memory with self-control and ternary neurons. *Neural Networks*, 13(4-5):455–462, 2000.

Thomas M Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE transactions on electronic computers*, (3):326–334, 1965.

David Dominguez, Kostadin Koroutchev, Eduardo Serrano, and Francisco B Rodríguez. Mutual information and topology 1: Asymmetric neural network. In *Advances in Neural Networks–ISNN 2004: International Symposium on Neural Networks, Dalian, China, August 2004, Proceedings, Part I 1*, pp. 14–19. Springer, 2004.

David Dominguez, Mario González, Eduardo Serrano, and Francisco B Rodríguez. Structured information in small-world neural networks. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 79(2):021909, 2009.

David R Carreta Dominguez and Elka Korutcheva. Three-state neural network: From mutual information to the hamiltonian. *Physical Review E*, 62(2):2620, 2000.

Aaron J Gutknecht, Michael Wibral, and Abdullah Makkeh. Bits and pieces: Understanding information decomposition from part-whole relationships and formal logic. *Proceedings of the Royal Society A*, 477(2251):20210110, 2021.

Christopher Hillar, Jascha Sohl-Dickstein, and Kilian Koepsell. Efficient and optimal binary hopfield associative memory storage using minimum probability flow, 2015. URL `https://arxiv.org/abs/1204.2916`.

John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.

Andreas Knoblauch, Günther Palm, and Friedrich T Sommer. Memory capacities for synaptic and structural plasticity. *Neural Computation*, 22(2):289–341, 2010.

Dmitry Krotov. A new frontier for Hopfield Networks. *Nature Reviews Physics*, pp. 1–2, 2023.

Dmitry Krotov and John J. Hopfield. Dense associative memory for pattern recognition. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL `https://proceedings.neurips.cc/paper/2016/file/eaae339c4d89fc102edd9dbdb6a28915-Paper.pdf`.

Joseph T Lizier, Nils Bertschinger, Jürgen Jost, and Michael Wibral. Information decomposition of target effects from multi-source interactions: Perspectives on previous, current and future work. *Entropy*, 20(4):307, 2018.

Abdullah Makkeh, Aaron J Gutknecht, and Michael Wibral. Introducing a differentiable measure of pointwise shared information. *Physical Review E*, 103(3):032149, 2021.

Abdullah Makkeh, Marcel Graetz, Andreas C. Schneider, David A. Ehrlich, Viola Priesemann, and Michael Wibral. A general framework for interpretable neural learning based on local information-theoretic goal functions. *Proceedings of the National Academy of Sciences*, 122 (10):e2408125122, 2025. doi: 10.1073/pnas.2408125122. URL https://www.pnas.org/doi/abs/10.1073/pnas.2408125122.

Ali Montazeri and Robert Schmidt. Robustness in hopfield neural networks with biased memory patterns. *bioRxiv*, pp. 2024–10, 2024.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Lukas Gruber, Markus Holzleitner, Thomas Adler, David Kreil, Michael K Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. Hopfield networks is all you need. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=tL89RnzIiCd.

Andreas C. Schneider, Valentin Neuhaus, David A. Ehrlich, Abdullah Makkeh, Alexander S. Ecker, Viola Priesemann, and Michael Wibral. What should a neuron aim for? designing local objective functions based on information theory. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=CLE09ESvul.

Pavel Tolmachev and Jonathan H. Manton. New insights on learning rules for hopfield networks: Memory and objective function minimisation. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, July 2020. doi: 10.1109/ijcnn48605.2020.9207405. URL http://dx.doi.org/10.1109/IJCNN48605.2020.9207405.

Paul L Williams and Randall D Beer. Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*, 2010.

APPENDIX

A MODEL PARAMETERS

In Table 1, we explain the parameters of infomorphic networks and their values that were used during training.

Table 1: The model parameters.

Parameter	Definition/Meaning	Value
N_{Neurons}	number of neurons	100
$\mathbf{W}_r$	initialization of the receptive weights	diag(1)
$\mathbf{W}_l$	initialization of the lateral weights	Kaiming Uniform
optimizer	algorithm to maximize G	SGD
η	learning rate	0.1
epochs	number of complete passes of the entire input patterns	1000
reps	number of consecutive times the pattern is presented	1
n_r	number of receptive bins	2
n_l	number of lateral bins	60
sequential	True if states are updated sequentially	False
N_{iter}	maximum iterations during testing	100

B SXPID MEASURE

In the simulations, we used infomorphic networks that optimize PID-based goal functions. This PID is computed using the shared-exclusion redundancy measure $I_\cap^{\text{sx}}$ introduced by Makkeh et al. (2021). In what follows, we first motivate and then briefly explain the definition of this measure.

The $I_\cap^{\text{sx}}$ is built on the Bayesian intuition that the mutual information $I(T : S_1, S_2)$ between a target variable T and two source variables S_1 and S_2 can be interpreted as an average measure for how the prior belief of the target event $T = t$ is updated in light of the event of observing *both* source events $S_1 = s_1$ *and* $S_2 = s_2$ simultaneously:

$$I(T : S_1, S_2) = \sum_{t,s_1,s_2} p(T = t, S_1 = s_1, S_2 = s_2) \log_2 \frac{p(T = t \mid S_1 = s_1 \wedge S_2 = s_2)}{p(T = t)}.$$

If $I(T : S_1, S_2) > 0$ then the posterior belief of event $T = t$ is higher than the prior on average and S_1 and S_2 hold information about T, otherwise the prior belief is equal to the posterior on average and S_1 and S_2 hold no information about T.

Following the same logic, Makkeh et al. (2021) define redundancy as an average measure for how the prior belief about the target event $T = t$ is updated if instead it is only known that $S_1 = s_1$ *or* $S_2 = s_2$ have occurred:

$$I_\cap^{\text{sx}}(T : S_1, S_2) = \sum_{t,s_1,s_2} p(T = t, S_1 = s_1, S_2 = s_2) \log_2 \frac{p(T = t \mid S_1 = s_1 \vee S_2 = s_2)}{p(T = t)}.$$

The definition is symmetric with respect to permutation of the sources, fulfills a target chain rule and is differentiable with respect to the underlying probability distribution (Makkeh et al., 2021), which makes it a suitable definition for optimizing objective functions.

One counter-intuitive property of $I_\cap^{\text{sx}}$ is that $I_\cap^{\text{sx}}(T : S_1, S_2)$ can be larger than $I(T : S_1)$ or $I(T : S_2)$. In these networks, the source variables are given by the neuron inputs R and L while the target variable is given by the neuron output Y (thus $T = Y$, $S_1 = R$ and $S_2 = L$). The additional information in $I_\cap^{\text{sx}}(Y : R, L)$ beyond $I(Y : R)$ or $I(Y : L)$ cannot be attributed to the sources themselves and thus does not contribute to the learning of infomorphic neurons, but rather indicates misinformative behaviour. This misinformation is reflected in obtaining $\Pi_{\text{unq,i}} < 0$ whenever $I_\cap^{\text{sx}}(Y : R, L) > I(Y : X_i)$ for $X_i \in \{R, L\}$, which indicates that one of the variables X_i

provides misinformative unique information about Y. A case in point is that for infomorphic Hopfield networks with loading $\alpha > \alpha_{\mathrm{red}}$ in section 5, $I_{\cap}^{\mathrm{sx}}(Y : R, L) > I(Y : R)$ does not help the neurons to learn their correct output. Instead the negative $\Pi_{\mathrm{unq,L}}$ reflects that the lateral input is misinforming the output indicating that the neurons did not learn properly. This counter-intuitive property becomes a signature for failure in learning.

C PSEUDOCODE FOR TRAINING

Function 1: TrainHebbianHopfieldModel

Input: data
Output: trained model

```
INITIALIZE model;
INITIALIZE model.neuron_weights ← zero_matrix;
foreach pattern in data do
    model.neuron_weights ← neuron_weights + outer(pattern, pattern);
return model
```

Function 2: TrainInfomorphicHopfieldModel

Input: data, num_epochs, goal_params
Output: trained model

```
INITIALIZE model;
foreach epoch in range(num_epochs) do
    INITIALIZE model_outputs;
    foreach pattern in data do
        INITIALIZE network_state ← pattern;
        network_state ← model.forward_network(r=pattern, l=network_state);
        model_outputs.append(network_state);
    foreach neuron in model do
        TrainInfomorphicNeuron(neuron, goal_params, y=output_state[neuron], r=data,
         l=model_outputs);
return model
```

Function 3: TrainInfomorphicNeuron

Input: neuron, goal_params, y, r, l
Output: None

```
BIN continuous values r in 2 and l in 60 equally sized bins;
COUNT occurrences of tuples (r, l);
COMPUTE empirical probability masses p(r, l);
EVALUATE conditional probabilities p(y | r, l) from the neurons;
CONSTRUCT full joint probability mass function p(y, r, l) = p(r, l)p(y | r, l);
isx_redundancies ← ComputeIsxRedundancies(p(y, r, l));
pid_atoms ← ComputePIDAtoms(isx_redundancies);
goal ← scalar_product(goal_params, pid_atoms);
PERFORM autograd to maximize goal;
UPDATE neuron.weights;
```

Function 4: ComputeIsxRedundancies

Input: Joint probability mass function $p(y, r, l)$
Output: Isx Redundancy Measure Values
foreach antichain $\beta \in \big\{\{\{1\},\{2\}\},\{\{1\}\},\{\{2\}\},\{\{1,2\}\}\big\}$ **do**
  COMPUTE conditional probability mass functions $p(Y = y \mid \bigvee_{\boldsymbol{b}\in\beta} \bigwedge_{i\in\boldsymbol{b}} S_i = s_i)$;
  COMPUTE marginal probability mass function $p(Y = y)$;
  $I_\cap^{\mathrm{sx}}(Y : S_\beta) \leftarrow \sum_{y,r,l} p(Y = y, R = r, L = l) \log_2 \frac{p(Y=y|\bigvee_{\boldsymbol{b}\in\beta} \bigwedge_{i\in\boldsymbol{b}} S_i=s_i)}{p(Y=y)}$;
return $I_\cap^{\mathrm{sx}}(Y : S_\beta)$ for all antichains β

D LEARNING IN THE REGIME OF LOW NUMBER OF PATTERNS

Interestingly, infomorphic networks cannot encode a single pattern, as in this case all neurons receive only a constant receptive input, resulting in zero bits of input entropy $H(R) = 0$ and redundancy $\Pi_{\mathrm{red}} = 0$ for the respective neuron, such that the goal function G

$$G = \gamma_{\mathrm{red}} \Pi_{\mathrm{red}}$$

becomes constant and equal to zero for any γ_{red}. Without a gradient, the neurons are not able to adapt their weights, failing to memorize the pattern. The problem persists even when the number of patterns m is increased but remains very small (see Figure 3A and Figure 4A). Since the patterns are drawn from a binomial distribution with $p = 0.5$, a significant fraction of neurons will receive all $+1$ or all -1 across all patterns. As a result, these neurons have constant receptive inputs R with $H(R) = 0$ (see Figure 3A and Figure 4A with $m \leq 4$). and fire arbitrarily (based on their initial weights). They additionally negatively affect those neurons with $H(R) > 0$ by transmitting an unreliable signal to other neurons, hindering overall performance. For example, when learning $m = 4$ patterns, about 16% of neurons fire arbitrarily, affecting recall accuracy (see Figure 3C and Figure 5A).

The arbitrary firing of neurons due to their constant receptive inputs can be avoided by using $G = \Pi_{\mathrm{red}} - H_{\mathrm{res}}$ instead of $G = \Pi_{\mathrm{red}}$. Additionally minimizing H_{res} promotes neurons with constant receptive input to optimize their lateral weights to become less stochastic and reliably align with their receptive input (see Figure 4 and Figure 5).

In detail, this is possible due to $A(R, L) = \sigma(R + L)$ through which the neuron can trivially reduce its H_{res} by adjusting $\mathbf{W}_l$ in such a way that aligns the sign of L with that of R. This alignment pushes $|A(R, L)|$ to a high value resulting in $p(y = \pm 1 \mid r, l) \approx 1$ (see Figure 4). For instance, if the neuron receives $+1$ across all patterns and has $\mathbf{W}_r = 1$, $R = 1$ and to minimize H_{res} the neuron will adapt its weights $\mathbf{W}_l$ to also produce a high lateral input $L > 0$. This results in saturation of the activation function, leading to $p(Y = 1 \mid R, L) \approx 1$ and thus making the neuron fire constantly and aligning with the pattern element $+1$. In the same manner, if the pattern element is a constant -1 the neuron will output a -1 most of the time, encoding the pattern element correctly. In essence, using $G = \Pi_{\mathrm{red}} - H_{\mathrm{res}}$, neurons receiving R with $H(R) = 0$ minimize H_{res} by encoding their receptive input in order to be deterministic, whereas those receiving R with $H(R) > 0$ will consistently maximize Π_{red}, learning as intended (see Figure 4 and Figure 5). Additionally minimizing residual entropy in the goal therefore leads to an overall better associative memory learning (see Figure 4A).

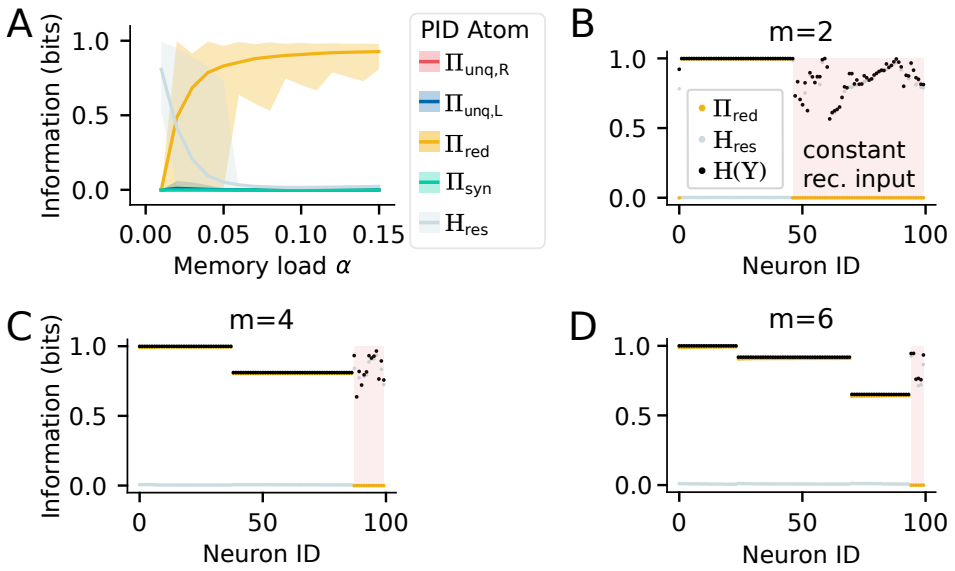

Figure 3: **A fraction of neurons receive no information via their receptive inputs at very low memory loads, leading to a failure of learning when relying on maximizing redundancy only. A:** PID profile for trained infomorphic Hopfield networks at low memory loads α. For α close to zero the residual entropy H_{res} dominates over Π_{red}. **B-D:** Information measures across neurons of single networks after training with $G = \Pi_{\text{red}}$ at $m = 2, 4, 6$ patterns. A fraction of neurons (background shaded in red) that by chance receive a constant receptive input across all patterns during training have zero redundant information Π_{red} and fail to learn. Residual entropy H_{res} and total entropy $H(Y)$ remains high for these neurons, indicating arbitrary firing.

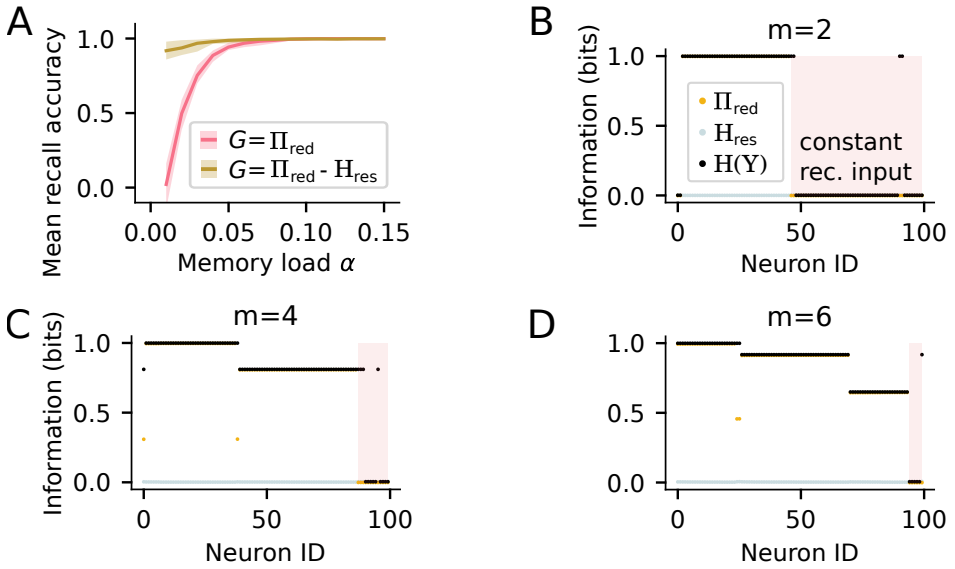

Figure 4: **Additionally minimizing residual entropy during training leads to improved performance at very low memory loads. A:** Performance of networks trained using $G = \Pi_{\text{red}}$ and $G = \Pi_{\text{red}} - H_{\text{res}}$ at low memory loads. **B-D:** Information measures across neurons of single networks after training with $G = \Pi_{\text{red}}$ at $m = 2, 4, 6$ patterns. The neurons with constant receptive input (background shaded in red) are able to learn, resulting in low residual entropy H_{res} and total entropy $H(Y)$.

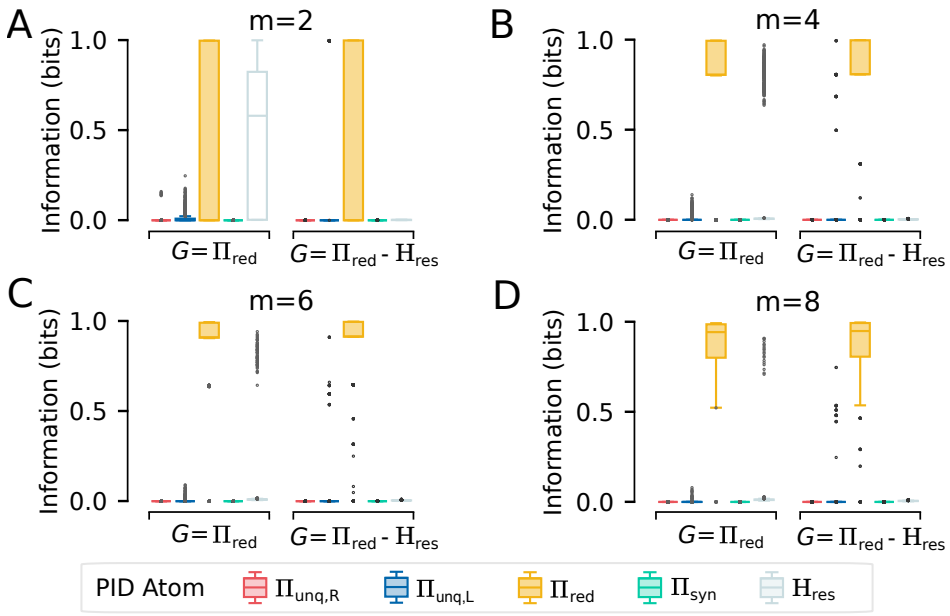

Figure 5: **Neurons with constant receptive input fire arbitrarily when only maximizing redundancy, which can be avoided by additionally minimizing residual entropy.** PID distribution for all neurons in comparison between $G = \Pi_{\text{red}}$ and $G = \Pi_{\text{red}} - H_{\text{res}}$ after training across 20 network initializations for very low numbers of patterns $m = 2, 4, 6, 8$.