PNAS RESEARCH ARTICLE | NEUROSCIENCE COMPUTER SCIENCES OPEN ACCESS

A general framework for interpretable neural learning based on local information-theoretic goal functions

Abdullah Makkeh[a,b,1,2], Marcel Graetz[a,c,1,2,3], Andreas C. Schneider[b,d,e,1,2], David A. Ehrlich[a,b], Viola Priesemann[b,d,e], and Michael Wibral[a,2]

Affiliations are included on p. 10.

Edited by Terrence Sejnowski, Salk Institute for Biological Studies, La Jolla, CA; received April 26, 2024; accepted December 19, 2024

Despite the impressive performance of biological and artificial networks, an intuitive understanding of how their local learning dynamics contribute to network-level task solutions remains a challenge to this date. Efforts to bring learning to a more local scale indeed lead to valuable insights, however, a general constructive approach to describe local learning goals that is both interpretable and adaptable across diverse tasks is still missing. We have previously formulated a local information processing goal that is highly adaptable and interpretable for a model neuron with compartmental structure. Building on recent advances in Partial Information Decomposition (PID), we here derive a corresponding parametric local learning rule, which allows us to introduce "infomorphic" neural networks. We demonstrate the versatility of these networks to perform tasks from supervised, unsupervised, and memory learning. By leveraging the interpretable nature of the PID framework, infomorphic networks represent a valuable tool to advance our understanding of the intricate structure of local learning.

information theory | partial information decomposition | neural networks | local learning

Significance

Which learning goals must individual computational elements pursue to contribute to a network-level task solution? This local understanding is missing in both biological, but also artificial neural networks, despite their impressive performance. We address this question by characterizing the information processing motifs of individual neurons as local goal functions, derived from first principles of information theory. A simple parameterization then enables the definition of an abstract goal function that spans a broad space of different learning rules and tasks. The resulting "infomorphic" networks offer a constructive approach to understanding local learning and information processing in neural networks, creating a bridge between theoretical neuroscience and artificial intelligence.

Both biological neural networks (BNNs) and artificial neural networks (ANNs) are capable of solving a variety of complex tasks, thanks to their interconnected structure comprising a large number of similar computational elements. The human neocortex employs a variety of neuron types organized into canonical, repeating microcircuits that show high functional flexibility (1–3), similar to how ANNs utilize relatively simple processing units arranged in repetitive structures (4). This structural repetition combined with functional flexibility enables both types of networks to scale drastically in size and complexity. Given the high intrinsic complexity of these networks, achieving an interpretable understanding of how local computational elements coordinate to address global tasks is challenging and remains an ongoing focus of intense research for both BNNs (5, 6) and ANNs (7, 8). Despite advances toward mechanistic interpretability of the inner local computational structures that emerge through learning (9, 10), the insights gained from post hoc approaches are specific to the data and network architecture, limiting their generality.

To foster a more general understanding of the local structures in neural networks, a data-independent description of the local algorithm is favorable. Such a description can be achieved through identifying a local optimization goal or learning rule, which prioritizes the learning process over the resulting representation. Traditionally, local learning has largely been formulated from two general perspectives: On one hand, the experimental study of BNNs has revealed activity-dependent changes of synaptic strengths. This has led researchers to propose a remarkable variety of local learning rules (11–15), most of which focus on biologically plausible mechanisms and require only locally available information. Despite these efforts, building large and powerful networks using only these mechanistic local learning rules has proven challenging (16). On the other hand, local learning in ANNs typically emerges implicitly by setting network-wide goal functions to satisfy global task requirements and then optimizing the network parameters via nonlocal gradient optimization. Such an approach hinders insights at the local scale, as the description of neuron function remains purely arithmetic. Nonetheless, efforts have recently intensified in developing learning rules that are both local, i.e., relying only on information that is available at the site, and show potential for scaling to larger, more capable networks (16–18). This includes learning rules based on concepts from contrastive learning (19, 20), predictive coding (21–23), local information maximization (15, 24) and many others (25–30).

Despite this large variety of fruitful efforts toward more local forms of learning, most existing approaches are limited to specific learning paradigms and implementations. What

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]A.M., M.G., and A.C.S. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: abdullah.alimakkeh@uni-goettingen.de, marcel.graetz@research.fchampalimaud.org, andreas.schneider@ds.mpg.de, or michael.wibral@uni-goettingen.de.

[3]Present address: Champalimaud Neuroscience Programme, Champalimaud Centre for the Unknown, Lisbon 1400-038, Portugal.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2408125122/-/DCSupplemental.

Published March 5, 2025.

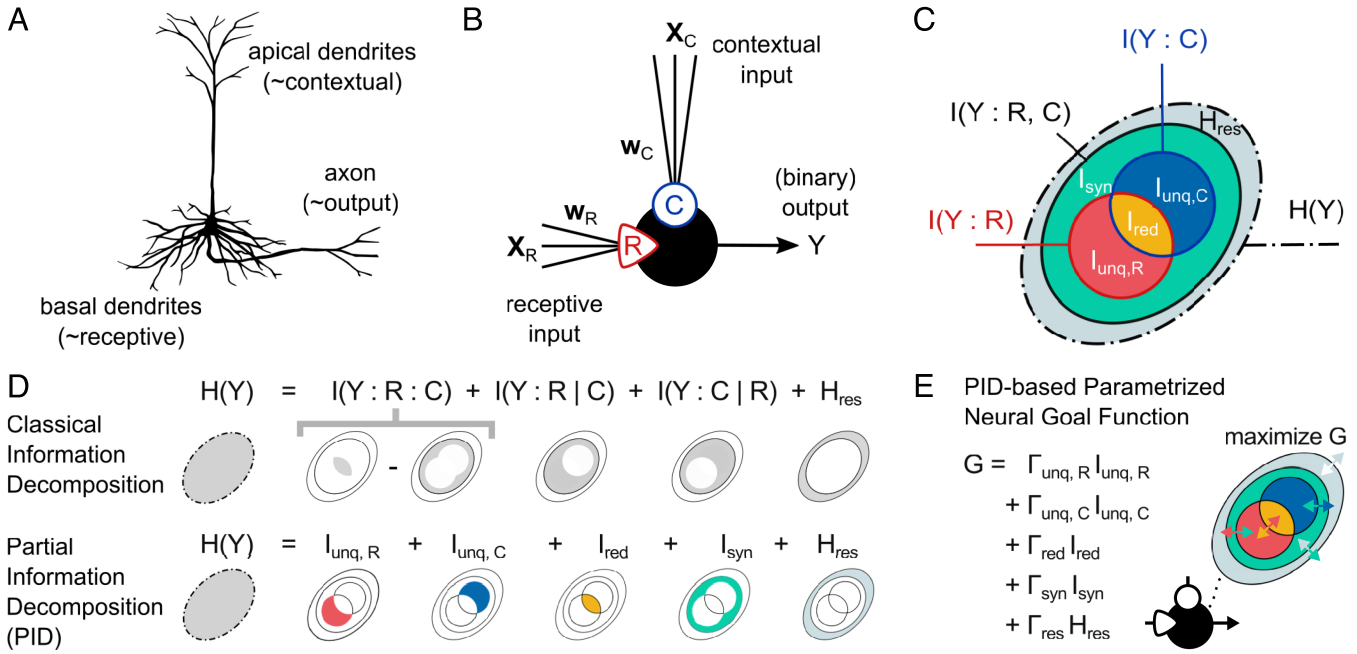

Fig. 1. The infomorphic neuron model, analogous to cortical pyramidal neurons, separately integrates two distinct classes of inputs. The neuron adjusts its synaptic weights to maximize the local goal function *G*, based on an information-theoretic decomposition of its own output information. (*A*) Cortical pyramidal neurons with separate synaptic integration sites for basal and apical dendrites, the former driving output and the latter providing contextual modulation. (*B*) The infomorphic neuron, modeled after these neurons, is characterized by two functionally distinct sets of inputs that are scaled by synaptic weights and added to obtain the integrated input signals *R* (receptive) and *C* (contextual). *R* and *C* contribute individually to the probabilities of the neuron's binary output, which are computed using an activation function $A(R, C)$ and a sigmoid transformation function. (*C*) The total Shannon output information $H(Y)$ of the neuron consists of the mutual information with the joint inputs $I(Y : R, C)$ and additional residual information $H_{\text{res}} = H(Y \mid R, C)$ that originates directly from the stochasticity of the neuron. Using Partial Information Decomposition (PID), the joint mutual information $I(Y : R, C)$ can be further subdivided into four information contributions: i) I_{red}, the redundant information that is provided by either *R* or *C* individually, ii) $I_{\text{unq},R}$, the unique information of *R* that is only provided by *R* but not by *C*, iii) $I_{\text{unq},C}$, the unique information of *C* that is only provided by *C* but not by *R*, and iv) I_{syn}, the synergistic information that is provided by *R* and *C* only when taken jointly but neither by *R* nor *C* taken individually. (*D*) Any classical mutual-information-based decomposition can only provide a linear combination of the underlying PID quantities. (*D*, *Upper*) Classical decomposition into four information contributions that formed the basis for prior work (32–35): the coinformation $I(Y : R : C)$, the two conditional mutual information values $I(Y : R \mid C)$ and $I(Y : C \mid R)$, and the stochasticity-caused residual entropy H_{res}. (*D*, *Lower*) The five contributions that are quantified using PID. (*E*) The neuron's synaptic weights **w** are optimized to maximize a goal function *G* that is based on the Partial Information Decomposition of the neuron's overall output information $H(Y)$ and parameterized by $\mathbf{\Gamma} = (\Gamma_{\text{unq},R}, \Gamma_{\text{unq},C}, \Gamma_{\text{red}}, \Gamma_{\text{syn}}, \Gamma_{\text{res}})$. Panel (*A*) is adapted from Fabian Mikulasch's original depiction of a pyramidal neuron (21) (CC0).

seems to be missing is a unifying framework to describe local learning goals—general enough to be applied across learning paradigms, datasets, and implementations, while still being interpretable. A promising starting point for developing such a framework from first principles is information theory (31–36). From an information-theoretic perspective, the local computational elements in a neural network can be interpreted as information channels that convert incoming signals into outgoing activity (37), with the conversion being specified by their synaptic weights. Previous research has demonstrated the feasibility and potential of a framework of local information-theoretic goal functions based on a decomposition of the output information of the individual computational elements (32, 35).

However, since classical information theory is constructed from an information channel (simple input-to-output) perspective, it is fundamentally limited in its ability to describe all facets of information processing: Both the proposed biological learning in neurons and most proposed biologically plausible local learning in ANNs require at least two qualitatively different classes of inputs to the computational element, one carrying the information to be processed and the other one carrying contextual information on how to process it (e.g., feed-back, label, error, lateral, contrastive, or reward signals) (38–42). To be able to capture the general interactions that could arise between these two classes of inputs, Wibral et al. (31) proposed a generalization and unification of existing information-theoretic local goal functions by employing the more expressive and intuitive Partial Information Decomposition (PID). PID provides a comprehensive information-theoretic description of the complex interactions of multiple sources with respect to a target by dissecting the mutual information into unique, redundant, and synergistic contributions (43–46).*

For the case of two input classes, PID distinguishes four contributions to the overall information processing: Each class may contribute uniquely to the output, meaning they contribute information the other source does not have, they can provide redundant information or they could contribute synergistically, i.e., in a way that no input class can do alone. Taken by itself, each input class can only provide the redundant information and its unique contribution, while the synergy relies on access to both classes simultaneously (Fig. 1 *C* and *D*). Here, we argue that different learning paradigms require different processing of the local information from the two classes. For instance, in a supervised setting, where one class provides input data and the other provides the ground-truth labels, the intuitive goal becomes to encode in the output what is redundant between these two input classes, which enables the network to learn to extract the label information from the input signal. In general, the interpretability of these information atoms allows to intuitively

*Recently, PID has also been used to describe the function of cortical neurons (47) and the representation of information in artificial and biological neural networks (48–51).

identify which information processing is necessary at a local level to achieve a global task.

In this work, we derive a parametric local learning rule from a general PID-based goal function, leveraging the differentiable $i_\cap^{sx}$ PID measure (45). We provide a proof of principle that this local learning rule enables networks consisting of compartmental neurons to solve tasks across three classic learning paradigms—supervised, unsupervised, and associative memory learning. Our work additionally shows that PID-based goals can be flexibly applied to different datasets and architectures, while being intuitively interpretable. Note that the relatively simple networks studied in this work should be considered an initial step toward larger and more capable network structures and provide evidence for the promising potential of such a general framework of interpretable local learning goals.

Below, we first explain our view of neurons as information processors with multiclass input, efficiently characterized by PID. Based on these insights, we then introduce a compartmental neuron model and apply it to a collection of learning scenarios. We conclude with a discussion of strengths, limitations, and next steps. As a side note, the neurons and networks developed in this work are termed infomorphic—as a portmanteau of "information" and "morphous" to indicate that they are directly shaped by the information they process.

1. Using Information Theory to Describe the Information Processing of a Neuron

In general, a neuron can be regarded as a Shannon information channel receiving synaptic inputs $\mathbf{X}$ from its afferent synapses and producing its own activity Y as output. Here, both $\mathbf{X}$ and Y are modeled as random variables and their relationship can be, in the general case, stochastic. The mutual information (52, 53)

$$I(Y:\mathbf{X}) = \mathbb{E}_{y,\mathbf{x}} \log_2 p(y|\mathbf{x})/p(y)$$

then quantifies how much a neuron's output is influenced by its synaptic inputs, whereas the residual (or conditional) entropy

$$H(Y \mid \mathbf{X}) = -\mathbb{E}_{y,\mathbf{x}} \log_2 p(y|\mathbf{x})$$

quantifies the amount of stochasticity in the output of the neuron that is not predictable from its inputs. The sum of these quantities equals the total entropy or information content of the firing of the neuron

$$H(Y) = I(Y:\mathbf{X}) + H(Y \mid \mathbf{X}). \qquad [1]$$

1.1. Beyond Simple Channels: Differentiating Input Classes.

The picture of neurons as simple information channels has to be refined in light of the insight that different information streams into a neuron often play qualitatively different roles. In ANNs, forward-propagated signal and backpropagated gradients influence the neuron in very different ways. Similarly, biological neurons often have multiple classes of inputs with distinct information processing characteristics (54). An example of a biological neuron with two distinct input classes can be found in layer-5 pyramidal neurons (55). These neurons are ubiquitous in the cortex, involved in sensory, cognitive, and motor tasks, and have been hypothesized to play a role in conscious awareness (2, 56). They are typically embedded in a relatively stereotyped cortical microcircuit, at the junction of feed-forward and feed-back information streams in the cortical hierarchy (57). To process these two information streams, pyramidal neurons possess two distinct types of dendrites, the basal and apical dendrites (55). Basal dendrites receive input from hierarchically lower cortical areas and play a role in encoding the external features of the environment that are processed along the cortical hierarchy (38). Apical dendrites, in contrast, receive contextual input from higher cortical areas and have been shown to play an important role in modulating perception (58, 59). This connectivity is similar across a range of different brain areas and cognitive domains, motivating the assumption that the general function of pyramidal neurons is independent of the semantics of their input (Fig. 1*A* and ref. 60).

The two-compartment structure of layer-5 pyramidal neurons (61) is consistent with many biologically plausible local learning rules in ANNs that require at least two qualitatively different classes of inputs to the neuron, respectively carrying feedforward information to be processed and contextual information to guide this processing (e.g., feed-back, label, error, lateral, contrastive, or reward signals) (38–42). Once these two input classes are explicitly established, they motivate a local learning goal.

To prepare for the mathematical representation of a neuron's unique, redundant, and synergistic information, we will first reinterpret the source variable $\mathbf{X}$ from above as being a composite variable $\mathbf{X} = (\mathbf{X}_R, \mathbf{X}_C)$ of the receptive input $\mathbf{X}_R$, which is inspired by the input to the basal dendrites, and the contextual input $\mathbf{X}_C$, which is inspired by the input to the apical dendrites. Analogous to Eq. **1**, the total entropy of the neuron Y can now be written as

$$H(Y) = I(Y:\mathbf{X}_R,\mathbf{X}_C) + H(Y \mid \mathbf{X}_R,\mathbf{X}_C). \qquad [2]$$

The dissection of $\mathbf{X}$ additionally allows to consider the individual channels of the receptive or contextual inputs to the target, which are characterized by the mutual information terms $I(Y:\mathbf{X}_R)$ or $I(Y:\mathbf{X}_C)$, respectively. Note, however, that these two channels do not simply add up to the total mutual information $I(Y:\mathbf{X}_R,\mathbf{X}_C)$, because the sum $I(Y:\mathbf{X}_R) + I(Y:\mathbf{X}_C)$ contains information which is redundantly present in both input classes and will be double-counted, while synergistic information which only becomes apparent if one considers $\mathbf{X}_C$ and $\mathbf{X}_R$ simultaneously will be overlooked (see Fig. 1*C*, 53, 62). By introducing the coinformation

$$\begin{aligned} I(Y:\mathbf{X}_R:\mathbf{X}_C) = {} & I(Y:\mathbf{X}_R,\mathbf{X}_C) \\ & - I(Y:\mathbf{X}_R \mid \mathbf{X}_C) - I(Y:\mathbf{X}_C \mid \mathbf{X}_R), \end{aligned} \qquad [3]$$

the decomposition in Eq. **2** can be refined to

$$\begin{aligned} H(Y) = {} & I(Y:\mathbf{X}_R:\mathbf{X}_C) \\ & + I(Y:\mathbf{X}_R \mid \mathbf{X}_C) + I(Y:\mathbf{X}_C \mid \mathbf{X}_R) \\ & + H(Y \mid \mathbf{X}_R,\mathbf{X}_C), \end{aligned} \qquad [4]$$

where the conditional mutual information $I(Y:\mathbf{X}_R \mid \mathbf{X}_C)$ captures the remaining dependence of Y on $\mathbf{X}_R$ when $\mathbf{X}_C$ is known, and $I(Y:\mathbf{X}_C \mid \mathbf{X}_R)$ is defined analogously (53).

Kay used this decomposition as the starting point to construct models of learning neurons with information theoretic objective functions (32). In our work, we build on this concept by exploiting the superior expressiveness provided by the framework of Partial Information Decomposition to build infomorphic neurons.

1.2. Uncovering the Information Processing Between Different Input Classes Using Partial Information Decomposition. The above perspective of viewing a neuron as a collection of information channels still paints an incomplete picture of the information processing within a neuron because it cannot account for all the different ways in which the different information sources combine and determine the output information: While some of the information in the neuron's output activity Y might be provided uniquely by either the receptive input $\mathbf{X}_R$ or the contextual input $\mathbf{X}_C$, other parts might be redundantly supplied by both of them while yet others only become available synergistically when both sources are considered jointly (31). Classical information theory is insufficient for this distinction as it has no concept of "sameness" of information: While one can compute the total amount of information in the output that is coming from each source or from both sources together using mutual information, there is no way of quantifying how much of the information contributed to the output is the same, i.e., redundantly provided by the input variables about the output (43).

Dissecting the mutual information between multiple source variables and a single target variable into nonoverlapping additive information atoms is the subject of Partial Information Decomposition (43, 46). Using PID, we can subdivide the entropy $H(Y)$ into five parts (Fig. 1*C*)

$$\begin{aligned} H(Y) = {} & I_{\text{unq}}(Y : \mathbf{X}_R) + I_{\text{unq}}(Y : \mathbf{X}_C) \\ & + I_{\text{red}}(Y : \mathbf{X}_R, \mathbf{X}_C) + I_{\text{syn}}(Y : \mathbf{X}_R, \mathbf{X}_C) \\ & + H(Y \mid \mathbf{X}_R, \mathbf{X}_C), \end{aligned} \quad [5]$$

where $I_{\text{unq}}(Y : \mathbf{X}_R)$ and $I_{\text{unq}}(Y : \mathbf{X}_C)$ are the unique information atoms of the receptive and contextual inputs, respectively, $I_{\text{red}}(Y : \mathbf{X}_R, \mathbf{X}_C)$ refers to the redundant (shared) information, and $I_{\text{syn}}(Y : \mathbf{X}_R, \mathbf{X}_C)$ refers to the synergistic (complementary) information. These four atoms can describe the information processing in Y of $\mathbf{X}_R$ and $\mathbf{X}_C$ in versatile ways, while also having meaningful interpretations: For example, if a neuron encodes the coherent parts of its inputs, this would be reflected in a high redundant information. Alternatively, a neuron might encode the information in its receptive input $\mathbf{X}_R$ that is specifically not present in the contextual input $\mathbf{X}_C$, which would translate to a high unique information contribution from $\mathbf{X}_R$. Finally, if the neuron's output contains information which cannot be obtained from any single source alone, for instance if the output Y reflected the logical "exclusive or" of its inputs, the synergy between the sources would be high. Overall, PID provides a decomposition framework with well-defined and intuitive interpretations for understanding a neuron's information processing.

Note that while the coinformation $I(Y : \mathbf{X}_R : \mathbf{X}_C)$ (in Eq. **3**) is equal to the difference between redundant and synergistic information

$$I(Y : \mathbf{X}_R : \mathbf{X}_C) = I_{\text{red}}(Y : \mathbf{X}_R, \mathbf{X}_C) - I_{\text{syn}}(Y : \mathbf{X}_R, \mathbf{X}_C), \quad [6]$$

classical information theory provides no tool to disentangle the two components.

To analyze the information processing of a neuron, the aforementioned PID atoms need to be quantified. Note that despite their strong relation to classical information-theoretic quantities through Eqs. **5** and **6**, the size of the PID atoms cannot be determined from classical information-theoretic quantities alone as there are four atoms with only three equations providing constraints (43). An additional quantity has to be defined for PID, which is typically, but not necessarily, the redundant information (43, 44, 46, and references therein). By now, a multitude of different measures for redundant information have been proposed, each fulfilling a number of partly mutually exclusive desiderata and drawing on concepts from different fields such as decision or game theory (44, 63–66). In this work, we use the PID measure $I_{\cap}^{\text{sx}}$ defined by Makkeh et al. (45) due to its analytical differentiability with respect to the underlying joint probability distribution $\mathbb{P}(Y, \mathbf{X}_R, \mathbf{X}_C)$, allowing for optimization of the PID quantities through gradient ascent.

2. Infomorphic Neurons

In a line of similar work, Kay (32) utilized the decomposition in Eq. **4** not as a post hoc analysis tool, but as a parameterizable optimization goal function, extending this idea in subsequent research (33–35). Even before the development of their differentiable PID measure, Wibral et al. (31) envisioned a similar, but more refined neural goal function derived from the decomposition in Eq. **5**. In the following paragraphs, we realize this idea in a neuron model closely aligned to prior work (32), which we refer to as the infomorphic neuron, and derive analytic gradients for the PID-based goal function.

2.1. Multicompartment Computation. Infomorphic neurons operate in discrete time and output values $Y \in \{-1, +1\}$ (referred to as "LOW" and "HIGH"), in analogy to time-binned spike trains of biological neurons. Akin to the basal and apical dendrites of layer-5 pyramidal neurons, an infomorphic neuron distinguishes between two classes of input synapses, namely "receptive" inputs $\mathbf{X}_R$ and "contextual" inputs $\mathbf{X}_C$ (Fig. 1 *A* and *B*). Inspired by how the inputs of different input classes are individually aggregated in separate compartments in these biological neurons (55), the inputs of the two classes of the infomorphic neuron are separately combined in a weighted sum to produce the aggregate inputs $R = \mathbf{w}_R^T \mathbf{X}_R - w_{0,R}$ and $C = \mathbf{w}_C^T \mathbf{X}_C - w_{0,C}$ (32). Here, $\mathbf{w}_R$ and $\mathbf{w}_C$ reflect the weights associated with the receptive and contextual inputs, respectively, while $w_{0,R}$ and $w_{0,C}$ denote constant bias values. At any time step, the probability θ of a neuron to be in the HIGH state depends only on the instantiation of its aggregate inputs r and c, as follows:

$$\theta(r, c) := \mathbb{P}(Y = 1 \mid R = r, C = c) := \sigma(A(r, c)),$$

where $\sigma(\xi) = 1/(1 + e^{-\xi})$ is a sigmoid nonlinearity, and A is an additional activation function. While the activation function can in principle be chosen arbitrarily, a biology-inspired choice of A may draw inspiration from layer-5 pyramidal neurons: By making the activation function be primarily dependent on the receptive inputs, one can imitate the privileged role that basal dendrites play in driving pyramidal neurons (33). In practice, we adapted the degree to which the contextual input influences the output, dependent on the requirements of each task. The choice of activation function will be individually motivated and discussed in the corresponding experimental sections.

2.2. Local Learning. Each infomorphic neuron optimizes its local information processing by changing the two sets of weights $\mathbf{w}_R$ and $\mathbf{w}_C$ of its incoming (afferent) synapses. This information processing can take on very different shapes: For some tasks, optimal information processing could mean coding for coherence between the receptive and contextual inputs, while for other tasks, optimal processing might entail extracting any piece of

information (e.g., a feature) exclusively provided by the receptive inputs that is not present in the contextual input.

Kay first derived a local goal function from an information-theoretic partition of the local mutual information of a neuron (32). Here, we argue for a similar local goal function involving a linear combination of the five components of the output entropy of a neuron as derived from PID and first established by (31):

$$\begin{aligned} G(Y:\tilde{R},\tilde{C}) &= \Gamma_{\mathrm{unq},R}\, I_{\mathrm{unq}}(Y:\tilde{R}) + \Gamma_{\mathrm{unq},C}\, I_{\mathrm{unq}}(Y:\tilde{C}) \\ &\quad + \Gamma_{\mathrm{red}}\, I_{\mathrm{red}}(Y:\tilde{R},\tilde{C}) + \Gamma_{\mathrm{syn}}\, I_{\mathrm{syn}}(Y:\tilde{R},\tilde{C}) \\ &\quad + \Gamma_{\mathrm{res}}\, H(Y \mid \tilde{R},\tilde{C}) \\ &=: (\Gamma_{\mathrm{unq},R}, \Gamma_{\mathrm{unq},C}, \Gamma_{\mathrm{red}}, \Gamma_{\mathrm{syn}}, \Gamma_{\mathrm{res}}) \\ &\quad \cdot (I_{\mathrm{unq},R}, I_{\mathrm{unq},C}, I_{\mathrm{red}}, I_{\mathrm{syn}}, H_{\mathrm{res}})^T. \end{aligned} \quad [7]$$

Here, the variables $\tilde{R}$ and $\tilde{C}$ are binned versions of the continuous-valued R and C inputs, necessary due to the lack of a differentiable PID measure for mixed discrete-continuous variables (67) and other conceptual difficulties of information theory in continuous networks (68, 69). Note that the binning procedure itself, while used in analogy to previous work (32, 35), is a nondifferentiable operation whose gradients we do not take into account here. Future work might circumvent this problem by using parametric (e.g., bivariate Gaussian) approximations of $p(R, C)$ (32, 35) combined with PID-estimators for mixed discrete-continuous variables. The neuron's local goal function G is a linear combination of the PID atoms that is defined a priori by choosing the parameters $\mathbf{\Gamma}$ (Fig. 1*E*). In the second equality, we introduced a short-hand vector notation of G.

2.3. Optimizing the Goal Function. The differentiability of the $I_\cap^{\mathrm{sx}}$ measure allows each neuron in an infomorphic network to optimize its own goal function G through gradient ascent.

The empirical gradients of G with respect to the weight vectors $\mathbf{w}_R$ and $\mathbf{w}_C$ can be analytically derived as

$$\frac{\partial \hat{G}}{\partial \mathbf{w}_R} = \frac{1}{N} \sum_{\mathbf{x}_R,\mathbf{x}_C} f^{\mathbf{\Gamma}}_{p(R,C)}(\tilde{r},\tilde{c}) \left.\frac{\partial A}{\partial r}\right|_{\tilde{r},\tilde{c}} \mathbf{x}_R \quad [8]$$

and

$$\frac{\partial \hat{G}}{\partial \mathbf{w}_C} = \frac{1}{N} \sum_{\mathbf{x}_R,\mathbf{x}_C} f^{\mathbf{\Gamma}}_{p(R,C)}(\tilde{r},\tilde{c}) \left.\frac{\partial A}{\partial c}\right|_{\tilde{r},\tilde{c}} \mathbf{x}_C, \quad [9]$$

where $\hat{G}$ indicates the estimator of G based on N input samples in the data, and f is implicitly dependent on the full probability distribution $p_{\tilde{R},\tilde{C}}$ of $\tilde{r}$ and $\tilde{c}$ over the dataset and the explicit current values of those variables, as well as the goal parameter vector $\mathbf{\Gamma}$. The full derivation of the gradients can be found in *SI Appendix*, section 1.

In practice, we only update the network parameters after a fixed number of discrete network time steps, referred to as a minibatch. For each minibatch, we estimate the full binned probability distribution $p_{\tilde{R},\tilde{C}}$ from the histogram of inputs and finally conduct a single weight update. We report the number of minibatches as the training time t. Instead of using minibatches, it would also be possible due to the pointwise nature of the $i_\cap^{\mathrm{sx}}$ PID measure to keep running estimates with exponential forgetting of past samples, which would allow for weight updates after each network time step.

3. Infomorphic Networks Encompass Various Learning Paradigms

The parameterized information-theoretic goal function enables groups of infomorphic neurons, i.e., "infomorphic networks," to serve as a very general and versatile approach to learning. In the following, we demonstrate their broad applicability by providing three example applications of infomorphic networks, on supervised learning, unsupervised learning, and online learning of associative memories.

In correspondence to classical ANNs, infomorphic networks require choices on network topology, activation functions, and goal functions, where the latter are chosen by setting the goal function hyperparameters for each neuron. The ability to arbitrate between different local goals by setting these hyperparameters is a major strength of our framework, and we will motivate and discuss our specific hyperparameter choices in all three presented applications.

3.1. Supervised Learning by Encoding Coherence Between Input and Label Information. We construct a single-layer infomorphic network for supervised classification of MNIST digits (70).

3.1.1. Topology and inputs. Each of $k = 10$ neurons receives the full MNIST image via a set of $28 \cdot 28 = 784$ receptive input synapses, with $\mathbf{X}_R \in \{0, \frac{1}{255}, \ldots, 1\}^{28\times 28}$, and a single element of a one-hot label vector as contextual input, with $X_C \in \{-1, 1\}$ (Fig. 2*A*). In this setup, each neuron becomes a one-vs.-all classifier of its assigned digit.

3.1.2. Goal functions. Viewed through the lens of PID, supervised learning requires extracting from input data the same information that is contained in the label. This is achieved if each neuron's output information is redundantly determined by its two input classes, motivating the goal function $G = I_{\mathrm{red}}$. In practice, weak incentives for the other PID quantities $\mathbf{\Gamma}_{\mathrm{supervised}} = (\Gamma_{\mathrm{unq},R}, \Gamma_{\mathrm{unq},C}, \Gamma_{\mathrm{red}}, \Gamma_{\mathrm{syn}}, \Gamma_{\mathrm{res}}) = (0.1, 0.1, 1, 0.1, 0)$ improve performance and stability of learning by preventing neurons from going silent, i.e., always outputting the same value.

3.1.3. Activation functions. To ensure that the receptive inputs are strong enough to drive the neurons during the test phase when the label is missing (all label input $x_C = 0$), we set the activation function to $A_\sigma(r, c) := r(0.5 + \sigma(2rc))$. This makes the binary output probabilities mostly dependent on the receptive input and only weakly modulated by the contextual information, rendering the label input a teacher signal that strongly influences learning but hardly the dynamics.

3.1.4. Protocol. In the training phase, we present the MNIST images and labels sequentially in random order, with a weight update after each minibatch (See *SI Appendix*, section 2.B for all chosen training parameters). In the test phase, we present previously unseen MNIST images and set the contextual input to $x_C = 0$, calculating winner-take-all classification accuracy. Note that instead of calculating $\mathbb{P}(Y \mid r, c = 0)$, an optimal predictor would marginalize over $\mathbb{P}(c_{\mathrm{train}})$ to estimate $\mathbb{P}(Y \mid r) \approx \sum_{c_{\mathrm{train}}} \mathbb{P}(Y \mid r, c_{\mathrm{train}})\, \mathbb{P}(c_{\mathrm{train}})$. However, this would require running each test input twice for the two different labels, and performing computations that infomorphic neurons do not implement. Fortunately, due to the merely modulating role of the contextual input, the test accuracy of both approaches is virtually identical (see *SI Appendix*, section 3.A and Fig. S3).

3.1.5. Performance and outcome. The infomorphic networks reach an average test accuracy of 89.7% (Fig. 2*D*), slightly lower than the 91.9% we find for multinomial logistic regression. Indeed, logistic regression upper-bounds the network performance,

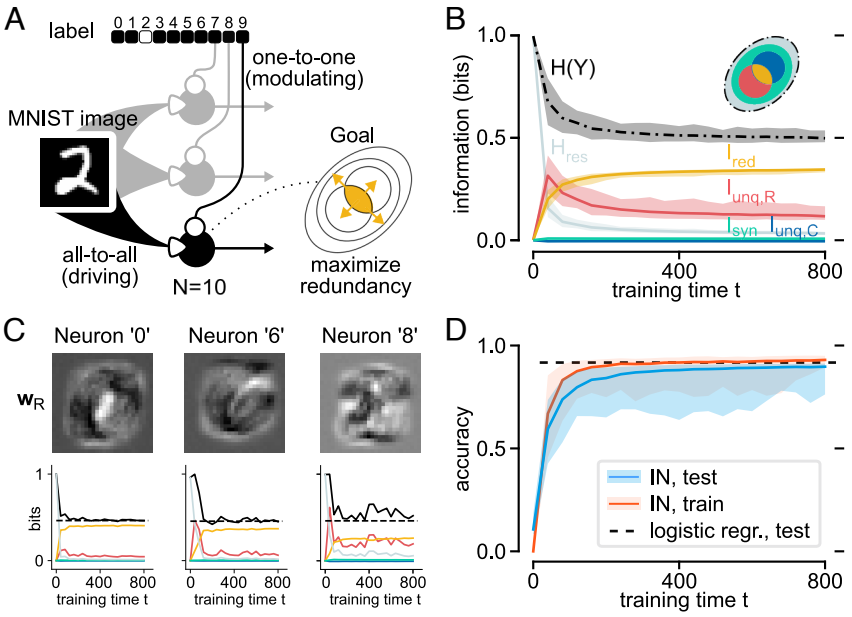

Fig. 2. Supervised learning in single-layer infomorphic networks. By maximizing redundant information between image and label, the neurons learn to identify MNIST digits with a test accuracy comparable to logistic regression. (*A*) Network architecture with one-hot encoded label and 10 neurons, each receiving all 28 × 28 image pixels as $\mathbf{X}_R$ and one element of the label vector as X_C. Activation function $A(r, c) := r(0.5 + \sigma(2rc))$ is chosen such that c has only modulating effect on the binary output probabilities, in line with the label's role as context for learning. The goal of each neuron is to transmit maximal information I_{red} that is redundant between R (image) and C (label element), thereby learning to act as a detector of its respective digit. The learning shows best stability if the goal function sets weak incentives for additionally maximizing the unique and synergistic information: $\mathbf{\Gamma}_{\text{supervised}} = (0.1, 0.1, 1, 0.1, 0)$. (*B*) Information quantities averaged over all neurons, shown for 100 independent training runs. (*C*) Receptive fields ($\mathbf{w}_R$) and information quantities for three sample neurons for a single training run, the dotted line indicating the expected $H(Y)$ in case of perfect classification (one-vs.-all entropy of label in test dataset). (*D*) The average training and test accuracies across 100 training runs, with test accuracy approaching that of logistic regression (reaching on average 89.7% vs. 91.9% for log. regr.). Note that in (*B*) and (*D*) the 95-percentile is being displayed.

because by setting $x_C = 0$ in the test phase, the activation function simplifies to $A(r, c = 0) = r$ and the firing probability becomes $\theta = \sigma(\mathbf{w}_R \cdot \mathbf{x}_R - w_{0,R})$, identical to logistic regression. The receptive weights $\mathbf{w}_R$ of individual neurons after training, plotted as receptive fields, visually reveal their assigned digit and qualitatively match the corresponding receptive fields found in logistic regression (Fig. 2*C* and *SI Appendix*, section 3.A and Fig. S2).

3.1.6. Information dynamics. Analyzing the information atoms of individual neurons over the course of training, we find an expected increase in redundant information I_{red} (Fig. 2 *B* and *C*). This increase is less pronounced in neurons corresponding to digits that are more likely to be confused (*SI Appendix*, section 3.A and Fig. S1). For these neurons we also find higher unique information from the receptive input $I_{\text{unq},R}$, indicating that they are still encoding image information that is not present in their label. Additionally, the average output entropy $H(Y)$ of the neurons decreases and approaches the average entropy of the one-hot label encoding (label k present vs. absent) of 0.47 bits, while the residual entropy $H(Y|R, C)$ decreases fast, reflecting a decrease in the neurons' stochasticity.

3.2. Unsupervised Learning of Independent Features by Maximizing Each Neuron's Unique Information About the Stimulus.

We construct a very simple data compression task that requires recurrent communication between neurons.

3.2.1. Topology and inputs. Each of the $k = 8$ neurons receives 8×8-pixel binary images as receptive inputs, $\mathbf{X}_R \in \{-1, 1\}^{8\times 8}$, with each image containing 8 horizontal bars appearing independently with probability $P = 0.5$ (Fig. 3*A*). As contextual input, each neuron receives the activity of all other neurons in the previous time step, with $\mathbf{X}_C \in \{-1, 1\}^7$.

3.2.2. Goal functions. The network-level goal is to encode all 8 bits of the information provided by the image distribution, distributed over the neurons. This can be achieved if each neuron encodes one full bit of image information that is not already encoded by the other neurons, and motivates a goal function maximizing for the conditional mutual information $G = I(Y : R|C) = I_{\text{unq},R} + I_{\text{syn}}$. In order to encourage the network to explicitly disentangle the contributions of each neuron, we chose to only encourage the unique information of the receptive input: $\mathbf{\Gamma}_{\text{unsupervised}} = (\Gamma_{\text{unq},R}, \Gamma_{\text{unq},C}, \Gamma_{\text{red}}, \Gamma_{\text{syn}}, \Gamma_{\text{res}}) = (1, 0, 0, 0, 0)$.

3.2.3. Activation functions. To avoid temporal oscillations in the network, we chose the same activation function $A_\sigma(r, c) := r(0.5 + \sigma(2rc))$ as in the supervised context, making the recurrent connections relevant for learning, but less so for the dynamics.

3.2.4. Protocol. We sequentially present randomly sampled images containing between 0 and 8 bars. Due to the time delay in recurrent connections, presentation of a new image introduces a mismatch between the receptive and contextual inputs of all neurons. We compensate for the resulting noise by presenting each image for 8 successive time steps.

Early in training, neurons compete for which information to encode. Two neurons choosing to encode for the same bar leads to a local optimum where both neurons try to increase their receptive weights to reduce stochasticity, however cannot obtain high unique information. To avoid this local maximum and prolong the critical initial phase of high stochasticity and competition, we introduce a strong weight decay (linear downscaling of all receptive weights after each time step) during the first half of training (see *SI Appendix*, section 2.C for all chosen training parameters).

3.2.5. Performance and outcome. Over the course of training almost all neurons learn to encode mutually different individual bars (Fig. 3*C*). Rare encoding errors occur exclusively when two neurons encode the same bar (*SI Appendix*, Fig. S8). Correspondingly, the total mutual information of the layer $I(\mathbf{X}_R : Y_1, \ldots, Y_8)$ approaches the entropy of the dataset, indicating successful compression of the receptive input information (Fig. 3*D*).

3.2.6. Information dynamics. The average unique receptive information of the neurons $I_{\text{unq},R}$ converges to 0.77 bits, with the average conditional mutual information $I(Y : R|C) = I_{\text{unq},R} + I_{\text{syn}}$ reaching close to 1.0 bits (Fig. 3 *B* and *C*). This suboptimal result is mostly attributable to individual neurons not having fully converged onto their chosen bar and to the weak contextual cross-talk between neurons (*SI Appendix*, section 3.B and Fig. S4). However, the above-mentioned rare

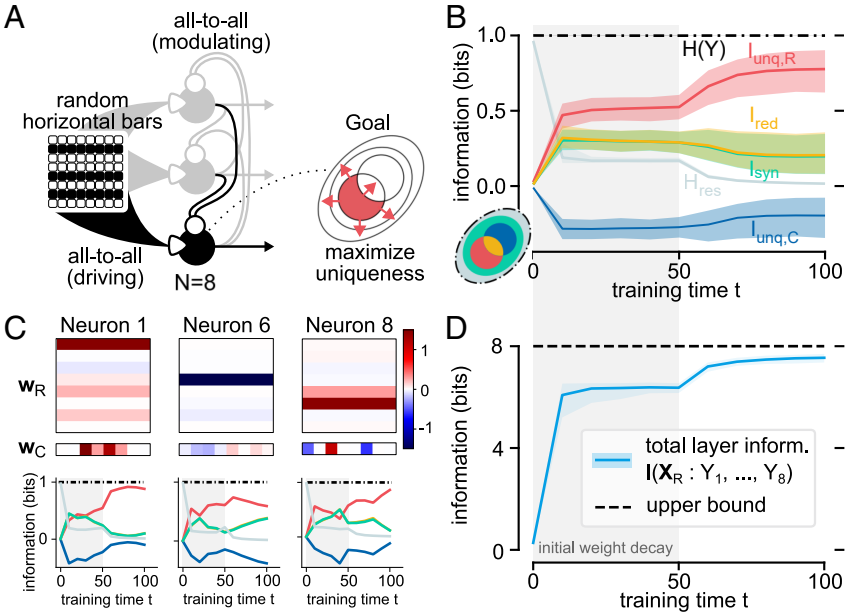

Fig. 3. Unsupervised feature learning in recurrent infomorphic networks. By maximizing unique information with respect to all other neurons, the neurons self-organize to create a highly informative representation of the input. (*A*) Network architecture for unsupervised learning with 8 neurons, each receiving 8×8 binary pixel inputs as $\mathbf{X}_R$ and output of all other neurons as $\mathbf{X}_C$. The image consists of 8 independent horizontal bars, randomly appearing with $P = 0.5$. Activation function $A(r, c) := r(0.5 + \sigma(2rc))$ is chosen such that c has only modulating effect on the binary output probabilities, which leads to recurrent connections mainly acting as context for learning. The goal of each neuron is set to maximize unique information $I_{\text{unq},R}$ of its own receptive input R with respect to the output of all other neurons, received as contextual input C: $\mathbf{\Gamma}_{\text{unsupervised}} = (1, 0, 0, 0, 0)$. (*B*) Information quantities averaged over all neurons for 300 independent training runs, showing two-phase training for feature competition and stabilization. (*C*) Receptive and contextual fields ($\mathbf{w}_R$, $\mathbf{w}_C$) and information quantities of three sample neurons for a single training run. (*D*) Mutual information $I(\mathbf{X}_R : Y_1, \ldots, Y_8)$ between all neurons' outputs and input image, approaching full encoding capacity and input information content of 8 bits. Note that in (*B*) and (*D*) the 95-percentile is being displayed.

encoding errors, i.e., two neurons encoding the same bar, show a strikingly different signature of low unique information $I_{\text{unq},R}$ and high redundancy I_{red} (*SI Appendix*, section 3.B and Fig. S7).

3.3. Online Associative Memory Learning by Maximizing the Local Coherence Between Network Firing and External Input.

We construct an (auto)associative memory network, similar to the Hopfield network (71), with an infomorphic online learning rule.

3.3.1. Topology and inputs. Each of $k = 100$ neurons receives a single element of a ($P = 0.5$)-sparse memory vector as receptive input, with $X_R \in \{-1, 1\}$, and the activity of all other neurons in the previous time step as contextual input, with $\mathbf{X}_C \in \{-1, 1\}^{99}$ (Fig. 4*A*).

3.3.2. Goal functions. The network-level goal is to align with any external input, when present, and over time inscribe it as an attractor, i.e., a memory, into the recurrent dynamics. The external input thus functions as both the memory cue and the teaching signal, depending on the duration of presentation. Learning a memory pattern upon repeated presentation implies that the recurrent contextual inputs learn to align with the external receptive inputs by providing the same information about the firing of a neuron as the external inputs, motivating the goal function $G = I_{\text{red}}$. As in the supervised experiment, weak incentives for the other PID quantities $\mathbf{\Gamma}_{\text{memory}} = (\Gamma_{\text{unq},R}, \Gamma_{\text{unq},C}, \Gamma_{\text{red}}, \Gamma_{\text{syn}}, \Gamma_{\text{res}}) = (0.1, 0.1, 1, 0.1, 0)$ improve performance by preventing neurons from going silent.

3.3.3. Activation functions. In the absence of a receptive input the neurons should be driven by the contextual synapses, while receptive input, if present, should overrule this recurrent drive and force the neurons into a new firing pattern. As a consequence, each input class individually needs to be able to drive the neuron. To make both inputs driving and encourage high weights, we choose the symmetric activation function

$$A(r, c) = \frac{r^8 \text{sign}(r) + c^8 \text{sign}(c)}{r^8 + c^8} \cdot (r^8 + c^8)^{1/8}$$

that is monotonic in both r and c and aligns with the positive and negative 8-norm for $r, c > 0$ and $r, c < 0$, respectively. (See *SI Appendix*, section 2.D for all chosen training parameters).

3.3.4. Protocol. In the training phase, we sequentially presented a set of memory patterns to the network in random order. As in the unsupervised experiments, to compensate for the time delay of the recurrent connections, we presented each memory pattern for 8 consecutive time steps. Note that structured sequences of patterns presented for only one time step lead to heteroassociative memory formation (results not reported here, but see ref. 72).

In the test phase, we present noise-corrupted memory patterns for a single time step, with the noise level $0 \leq \alpha \leq 1$ indicating the fraction of pattern elements set at random. After presentation, we set the receptive input of all neurons to $x_R = 0$ and assess retrieval accuracy by computing the cosine similarity between the network state and the noiseless memory pattern after 20 time steps (Fig. 4*D*). We define the noise-dependent capacity of the network as the highest number of trained memory patterns such that the average retrieval accuracy exceeds 95%.

3.3.5. Performance and outcome. For noiseless initialization, infomorphic networks attain a capacity of 35 patterns per 100 neurons, which significantly exceeds the limit of 14 patterns in classical Hopfield networks (71). Note that unlike Hopfield networks our readout includes no binarization but remains stochastic, thus likely even limiting the measured capacity in the direct comparison. Furthermore, infomorphic networks outperform Hopfield networks up to a noise level of $\alpha = 0.4$, making them far more robust to random pattern distortion, even though training is conducted on noise-free patterns (Fig. 4*C*). Interestingly, infomorphic networks cannot reliably encode very few patterns, as in this case, many neurons receive the exact same input for every pattern, resulting in 0 bits of information in the receptive inputs.

3.3.6. Information dynamics. We find an expected increase in redundant information I_{red} over the course of training (Fig. 4*B*). Surprisingly, this increase is also present in networks that are seemingly above capacity, but then coincides with misinformative unique contextual information $I_{\text{unq},C} < 0$, indicating that each neuron's activity is not fully predictable by the other neurons in this case (*SI Appendix*, section 3.C and Fig. S9).

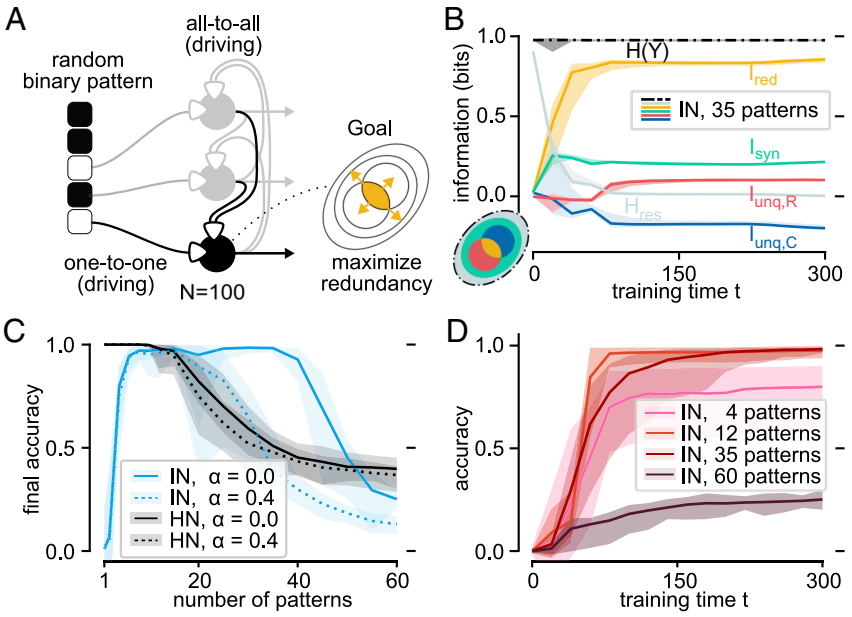

Fig. 4. Associative memory learning in recurrent infomorphic networks. By maximizing redundant information between external input and output of all other neurons, the neurons learn to memorize a maximum number of patterns exceeding that of classical Hopfield networks. (*A*) Network architecture for memorizing binary patterns with 100 recurrently connected neurons, each receiving single element of target patterns as X_R and output of all other neurons as $\mathbf{X}_C$. Activation function $A(r, c)$ is chosen such that both r and c can directly drive the neuron's output probabilities, to enable an active recovery of patterns based only on recurrent connections at test time (Section 3.3). The goal of each neuron is to learn to predict their respective element of the pattern based on activity of other neurons, which can be done by maximizing the redundant information in the output. Similar to the supervised example, the goal function used here includes additional, weak incentives for maximizing unique and synergistic PID contributions: $\mathbf{\Gamma}_{\text{memory}} = (0.1, 0.1, 1, 0.1, 0)$. (*B*) Information quantities averaged over all neurons, for 25 independent runs trained on 35 memory patterns. (*C*) Final accuracy of infomorphic networks and classical Hopfield networks over different numbers of patterns on the horizontal axis. Shown are the results for two noise levels used for testing the recovery of memorized patterns. (*D*) Accuracy over training for all 25 runs for different numbers of patterns. Note that in (*B*–*D*) the 95-percentile is being displayed.

4. Discussion

In this work, we defined the infomorphic neuron, an artificial neuron with two input classes and a flexible, parameterized local goal function derived from Partial Information Decomposition (43). Like classical information theory, PID provides an abstract, high-level description of neural functioning, yet enriches this description with the additional structure of redundancy, uniqueness, and synergy. This structure is inherited by the infomorphic neuron and leads to a highly interpretable and flexible description of local goals, independent of task, substrate, type of signals, and encoding of information therein (31, 37).

The experiments conducted provide a proof of principle that the level of abstraction gained by an information-based approach does not compromise the ability of model neural networks to learn and solve diverse tasks. Concretely, we find that maximizing the encoding of redundant information between the input and the label enables a single-layer network of infomorphic neurons to do supervised learning. Furthermore, we show that input information can be distributed between multiple infomorphic neurons in a recurrent network in an unsupervised learning task, by making each neuron maximize its encoding of unique input information with respect to the activity of other neurons. Finally, we find that maximizing the encoding of redundant information between an external input and the activity of other neurons in a recurrent infomorphic network leads to the formation of robust associative memories that exceed the memory capacity of classical Hopfield networks.

Notably, these experiments only explore a fraction of the available space of parameterized goal functions, and other terms of the goal function might become relevant in other learning scenarios. To demonstrate this, we construct primitive error neurons in the spirit of predictive coding (23, 73) (*SI Appendix*, section 4.B and Fig. S10), which receive simulated observations as receptive input and simulated model predictions as contextual input. They are trained to maximize synergistic information and both unique information atoms, while simultaneously minimizing redundant information. In a similar spirit, goals can be linearly combined to create more complex goal functions in other scenarios. One example of this is the weak incentives for unique and synergistic atoms in the supervised and associative memory experiments. We expect the residual entropy H_{res} to be another useful incentive to either reduce or artificially inject noise into infomorphic networks to improve learning.

Note that changing the hyperparameters $\mathbf{\Gamma}$ and the activation functions allows us to arbitrate between three very different learning tasks with little effort, providing practical evidence that our goal function G is highly interpretable and provides an intuitively accessible understanding of the local goal of each neuron in solving various tasks. Such interpretability is hard to establish in conventional ANNs, where a global error minimization goal is automatically backpropagated to the local level to adjust neuron parameters (e.g., refs. 74 and 75). Furthermore, a similar understanding of local goal functions might be an insightful target in our description of biological neural networks, and ultimately help to bridge the gap between artificial and biological intelligent systems. To this end, the synthetic methodology of infomorphic networks can easily be combined with post hoc analyses of trained BNNs and ANNs. In particular, it remains an open question which local information quantities these existing networks are effectively maximizing (37, 47, 49).

4.1. Future Work. Both the supervised and unsupervised experiments reported here focused on small single-layer neural networks, yet the ultimate strength of neural networks lies in their scalability to multilayered networks with a large number of neurons (4). Currently, this scalability is not present in infomorphic networks due to a conundrum that is implicitly solved in backpropagation: In most architectures, a neuron does not only need to get a feedforward input and a context signal that conveys target information (like a reward, supervision, or self-supervision signal), but additionally it requires knowledge about what other neurons in the same layer are coding for, such that it can choose to provide a complementary contribution with respect to the target (22). In backpropagation, this information flows to the neuron implicitly through the gradient signal from higher layers (22, 76), yet in infomorphic networks, it needs to be provided explicitly, because it fulfills a different role from the feed-back information: While the neuron typically needs to

follow the feed-back signal (redundancy), it simultaneously needs to be different from the lateral signal (uniqueness). In follow-up work, we define an infomorphic neuron with three input classes that combines these ideas, leading to greatly improved supervised learning performance (77).

Additionally, it has been shown that encoding for synergy plays a role in integration of information from multiple sources or sensory streams, both in the brain (78, 79) and in ANNs (80). This could be tested constructively in infomorphic neurons with four input classes, which could be trained to extract information that is synergistic between two different receptive inputs, redundant with a contextual input and unique with respect to other lateral neurons.

Infomorphic networks also offer a natural connection to other information-based learning algorithms. Prominently, under the infomax principle (81) it has been shown that in the low-noise regime, neurons maximize global information encoding by finding unique, independent features, which is in line with uniqueness maximization in our unsupervised learning experiments. However, under high noise, as might be prevalent in biological networks (82), more cooperative, redundant representations emerge (81). It remains an intriguing open question whether increasing admixtures of redundancy to the individual neural goals of infomorphic neurons may lead them to develop similar noise-robust representations. To test this hypothesis, our current framework allows increasing the noise by introducing an additional H_{res} term or stronger weight decay. However, learning under constant noise is a constrained optimization problem and we leave the introduction of the required Lagrange multipliers to future work.

Despite their fully local computation, infomorphic neurons currently lack biological plausibility due to the complexity of the gradient equations (Eqs. **8** and **9**) as well as the memory-expensive histogram estimation method in Section 2. Whether the gradients can be effectively approximated by simpler equations remains an open question for future work. However, a parameterized estimation of $p(R, C)$ combined with PID-estimators for mixed discrete-continuous variables would significantly reduce memory cost, e.g., to only 5 parameters for a two-dimensional Gaussian (32, 35), while for some tasks like supervised classification, more expressive multimodal distributions might be necessary. In addition, infomorphic neurons are functional, not anatomical, units and might thus be modeled by small microcircuits instead of individual spiking neurons.

Finally, notice that the high degree of interpretability of PID and our simple task setups allowed for a very intuitive reasoning about hyperparameters with only minor fine-tuning. However, more complex tasks with bigger infomorphic networks might require more systematic hyperparameter optimization techniques, a variety of which are easily accessible in modern-day machine learning tools (83). Fortunately, the resulting hyperparameter sets will then still be formulated in the language of Partial Information Decomposition and thus potentially provide crucial insights into the optimal local goals to enable the self-organization and collaboration of local units to solve a variety of global tasks (77).

4.2. Conclusion. Leveraging Partial Information Decomposition, this work establishes the infomorphic neuron, a neuron model permitting the flexible and direct optimization of interpretable information-theoretic goals. Through several lines of experimentation, the versatility of these neurons to solve a variety of machine learning tasks has been demonstrated. We propose infomorphic neurons as abstract neuron models that can provide a foundation for studying information processing in neural networks in the language of local goals, opening up many exciting avenues for future research.

5. Materials and Methods

Here, we provide the material and methods that apply to all our simulation experiments, while specific variations are discussed in the respective subsections of Section 3. In all simulation experiments, we run discrete-time neural networks. After choosing the network topology, the goal function, and activation function of each neuron, we initialize all weights at small values. Stimuli vary by experiment and are always presented to the networks sequentially, without reinitialization of neural activities. For each experiment, we segment the execution time into minibatches of fixed length. After each minibatch, we construct the full binned probability distribution $p_{Y,\tilde{R},\tilde{C}}$ of each neuron from the histogram of its inputs and outputs, which are obtained by quantizing R and C in uniform bins. Assuming the histogram as constant, we compute the gradients of the information-theoretic neural goal function G with respect to the weights according to Eqs. **8** and **9**, and conduct a weight update by applying gradient descent with a fixed learning rate.

To evaluate each experiment, we calculate performance metrics over the course of training by interleaving the training data stream with test data at regular intervals, while suppressing weight updates. These metrics include both the "information dynamics," i.e., the size of all information atoms of each neuron, and more traditional performance metrics like task accuracy, which our algorithm does not explicitly optimize for. For details on the parameter choices and results of each experiment, we refer the reader to the respective experimental Sections 3.1, 3.2, and 3.3 and to *SI Appendix*, sections 2 and 3. The experiments have been implemented in Python and have been made available on GitLab at https://gitlab.gwdg.de/wibral/infomorphic_networks (84).

Data, Materials, and Software Availability. The full derivation of the learning rules, parameters for all experiments, as well as supplementary figures providing additional information on the experiments are provided in *SI Appendix*. Code for reproducing all experiments is available on GitLab at https://gitlab.gwdg.de/wibral/infomorphic_networks (84). The raw data of the experiments are accessible via Göttingen Research Online Data at https://doi.org/10.25625/0M1PTJ (85). All other data are included in the manuscript and/or *SI Appendix*.

ACKNOWLEDGMENTS. We would like to thank William Phillips and Jim Kay for opening this line of research and fruitful discussions on Partial Information Decomposition and neural networks in general. We would like to thank Fabian Mikulasch, Matthias Loidolt, and Lucas Rudelt for extensive discussions on this topic. We would also like to thank Valentin Neuhaus, Kjartan van Driel, Paul Spitzner, Johannes Zierenberg, Aaron Gutknecht, Jonas Dehning, and the rest of the Wibral and Priesemann groups for their valuable comments and feedback. Finally, we would like to thank the reviewers for their insightful input. M.G. received a scholarship by the Champalimaud Foundation, the German Academic Scholarship Foundation, as well as the FIH-D Scholarship by the Department of Chemistry and Applied Biosciences at Eidgenössische Technische Hochschule Zürich. A.M. and M.W. are employed at the Göttingen Campus Institute for Dynamics of Biological Networks funded by the Volkswagenstiftung. 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.A.C.S. and V.P. acknowledge support from the Max Planck Society and the Deutsche Forschungsgemeinschaft (German Research Foundation) under Germany's Excellence Strategy-EXC 2067/1-390729940 as well as the RTG 2906 Curiosity. V.P. and M.W. received funding from the Deutsche Forschungsgemeinschaft (German Research Foundation) via the SFB 1528 "Cognition of Interaction" - project-ID 454648639. 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." Open access funding provided by the Max Planck Society.

Author affiliations: [a]Department of Data-driven Analysis of Biological Networks, Göttingen Campus Institute for Dynamics of Biological Networks, University of Göttingen, Göttingen 37077, Germany; [b]Complex Systems Theory, Max Planck Institute for Dynamics and Self-Organization, Göttingen 37077, Germany; [c]Department of Chemistry and Applied Biosciences, ETH Zurich, Zurich 8092, Switzerland; [d]University of Göttingen, Göttingen 37073, Germany; and [e]Cluster of Excellence "Multiscale Bioimaging: from Molecular Machines to Networks of Excitable Cells" (MBExC), University of Göttingen, Göttingen 37073, Germany

Author contributions: A.M., M.G., A.C.S., and M.W. designed research; A.M., M.G., and A.C.S. performed research; A.M., M.G., A.C.S., and D.A.E. analyzed data; A.M. and A.C.S. analyzed, interpreted results, and visualized results; M.G. and D.A.E. analyzed and interpreted results; and A.M., M.G., A.C.S., D.A.E., V.P., and M.W. wrote the paper.

1. O. D. Creutzfeldt, Generality of the functional structure of the neocortex. *Naturwissenschaften* **64**, 507–517 (1977).
2. S. Lodato, P. Arlotta, Generating neuronal diversity in the mammalian cerebral cortex. *Annu. Rev. Cell Dev. Biol.* **31**, 699–720 (2015).
3. K. D. Harris, G. M. Shepherd, The neocortical circuit: Themes and variations. *Nat. Neurosci.* **18**, 170–181 (2015).
4. O. Montesinos-López, A. Montesinos, J. Crossa, "Fundamentals of artificial neural networks and deep learning" in *Multivariate Statistical Machine Learning Methods for Genomic Prediction* (Springer, 2022), pp. 379–425.
5. N. S. Y. Dumont *et al*., Biologically-based computation: How neural details and dynamics are suited for implementing a variety of algorithms. *Brain Sci.* **13**, 245 (2023).
6. R. Q. Quiroga, Concept cells: The building blocks of declarative memory functions. *Nat. Rev. Neurosci.* **13**, 587–597 (2012).
7. T. Räuker, A. Ho, S. Casper, D. Hadfield-Menell, "Toward transparent AI: A survey on interpreting the inner structures of deep neural networks" in *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)* (2023), pp. 464–483.
8. P. P. Angelov, E. A. Soares, R. Jiang, N. I. Arnold, P. M. Atkinson, Explainable artificial intelligence: An analytical review. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **11**, e1424 (2021).
9. G. Goh *et al*., Multimodal neurons in artificial neural networks. *Distill* **6**, e30 (2021).
10. R. Tan, L. Gao, N. Khan, L. Guan, Interpretable artificial intelligence through locality guided neural networks. *Neural Netw.* **155**, 58–73 (2022).
11. J. Konorski, *Conditioned Reflexes and Neuron Organization* (CUP Archive, 1948).
12. E. Oja, Simplified neuron model as a principal component analyzer. *J. Math. Biol.* **15**, 267–273 (1982).
13. E. L. Bienenstock, L. N. Cooper, P. W. Munro, Theory for the development of neuron selectivity: Orientation specificity and binocular interaction in visual cortex. *J. Neurosci.* **2**, 32–48 (1982).
14. G. Bi, M. Poo, Synaptic modifications in cultured hippocampal neurons: Dependence on spike timing, synaptic strength, and postsynaptic cell type. *J. Neurosci.* **18**, 10464–10472 (1998).
15. T. Isomura, T. Toyoizumi, A local learning rule for independent component analysis. *Sci. Rep.* **6**, 28073 (2016).
16. I. Jeon, T. Kim, Distinctive properties of biological neural networks and recent advances in bottom-up approaches toward a better biologically plausible neural network. *Front. Comput. Neurosci.* **17**, 1092185 (2023).
17. T. P. Lillicrap, A. Santoro, L. Marris, C. J. Akerman, G. Hinton, Backpropagation and the brain. *Nat. Rev. Neurosci.* **21**, 335–346 (2020).
18. B. A. Richards *et al*., A deep learning framework for neuroscience. *Nat. Neurosci.* **22**, 1761–1770 (2019).
19. B. Illing, J. Ventura, G. Bellec, W. Gerstner, Local plasticity rules can learn deep representations using self-supervised contrastive predictions. *Adv. Neural Inf. Proces. Syst.* **34**, 30365–30379 (2021).
20. M. A. Ahamed, J. Chen, A. A. Z. Imran, Forward-forward contrastive learning. arXiv [Preprint] (2023). http://arxiv.org/abs/2305.02927 (Accessed 4 May 2023).
21. F. A. Mikulasch, L. Rudelt, M. Wibral, V. Priesemann, Where is the error? Hierarchical predictive coding through dendritic error computation. *Trends Neurosci.* **46**, 45–59 (2023).
22. J. Sacramento, R. Ponte Costa, Y. Bengio, W. Senn, Dendritic cortical microcircuits approximate the backpropagation algorithm. *Adv. Neural Inf. Proces. Syst.* **31**, 8721–8732 (2018).
23. B. Millidge, A. Tschantz, C. L. Buckley, Predictive coding approximates backprop along arbitrary computation graphs. *Neural Comput.* **34**, 1329–1368 (2022).
24. S. Löwe, P. O'Connor, B. Veeling, Putting an end to end-to-end: Gradient-isolated learning of representations. *Adv. Neural Inf. Proces. Syst.* **32**, 3039–3051 (2019).
25. J. Launay, I. Poli, F. Boniface, F. Krzakala, Direct feedback alignment scales to modern deep learning tasks and architectures. *Adv. Neural Inf. Proces. Syst.* **33**, 9346–9360 (2020).
26. G. Hinton, The forward-forward algorithm: Some preliminary investigations. arXiv [Preprint] (2022). http://arxiv.org/abs/2212.13345 (Accessed 27 December 2022).
27. R. Høier, D. Staudt, C. Zach, "Dual propagation: Accelerating contrastive hebbian learning with dyadic neurons" in *International Conference on Machine Learning*, A. Krause *et al.*, Eds. (PMLR, 2023), pp. 13141–13156.
28. D. H. Lee, S. Zhang, A. Fischer, Y. Bengio, "Difference target propagation" in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2015, Porto, Portugal, September 7-11, 2015, Proceedings, Part I 15*, A. Appice *et al.*, Eds. (Springer, 2015), pp. 498–515.
29. A. Nøkland, L. H. Eidnes, "Training Neural Networks with local error signals" in *International Conference on Machine Learning*, K. Chaudhuri, R. Salakhutdinov, Eds. (PMLR, 2019), pp. 4839–4850.
30. F. Lässig, P. V. Aceituno, M. Sorbaro, B. F. Grewe, Bio-inspired, task-free continual learning through activity regularization. *Biol. Cybern.* **117**, 345–361 (2023).
31. M. Wibral, V. Priesemann, J. W. Kay, J. T. Lizier, W. A. Phillips, Partial information decomposition as a unified approach to the specification of neural goal functions. *Brain Cogn.* **112**, 25–38 (2017).
32. J. Kay, "Information-theoretic neural networks for unsupervised learning: Mathematical and statistical considerations" (Tech. Rep. 1573387449478062080, Scottish Agricultural Statistics, 1994).
33. J. Kay, W. A. Phillips, Activation functions, computational goals, and learning rules for local processors with contextual guidance. *Neural Comput.* **9**, 895–910 (1997).
34. J. Kay, *Neural Networks for Unsupervised Learning Based on Information Theory* (Oxford University Press Inc, United States, 2000), pp. 25–63.
35. J. W. Kay, W. Phillips, Coherent infomax as a computational goal for neural systems. *Bull. Math. Biol.* **73**, 344–372 (2011).
36. V. Koren, G. Bondanelli, S. Panzeri, Computational methods to study information processing in neural circuits. *Comput. Struct. Biotechnol. J.* **21**, 910–922 (2023).
37. M. Wibral, J. T. Lizier, V. Priesemann, Bits from brains for biologically inspired computing. *Front. Robot. AI* **2**, 5 (2015).
38. R. Chéreau, L. E. Williams, T. Bawa, A. Holtmaat, "Circuit mechanisms for cortical plasticity and learning" in *Seminars in Cell & Developmental Biology*, T. Van Raay, P. Opazo, V. Anggono, Eds. (Elsevier, 2022), vol. 125, pp. 68–75.
39. W. Schultz, Predictive reward signal of dopamine neurons. *J. Neurophysiol.* **80**, 1–27 (1998).
40. E. Rolls, A. Treves, *Neural Networks and Brain Function* (Oxford University Press, 1997).
41. Y. Shu, A. Hasenstaub, D. A. McCormick, Turning on and off recurrent balanced cortical activity. *Nature* **423**, 288–293 (2003).
42. S. Manita *et al*., A top-down cortical circuit for accurate sensory perception. *Neuron* **86**, 1304–1316 (2015).
43. P. L. Williams, R. D. Beer, Nonnegative decomposition of multivariate information. arXiv [Preprint] (2010). http://arxiv.org/abs/1004.2515 (Accessed 14 April 2010).
44. J. T. Lizier, N. Bertschinger, J. Jost, M. Wibral, Information decomposition of target effects from multi-source interactions: Perspectives on previous, current and future work. *Entropy* **20**, 307 (2018).
45. A. Makkeh, A. J. Gutknecht, M. Wibral, Introducing a differentiable measure of pointwise shared information. *Phys. Rev. E* **103**, 032149 (2021).
46. A. J. Gutknecht, M. Wibral, A. Makkeh, Bits and pieces: Understanding information decomposition from part-whole relationships and formal logic. *Proc. R. Soc. A* **477**, 20210110 (2021).
47. J. M. Schulz, J. W. Kay, J. Bischofberger, M. E. Larkum, Gaba b receptor-mediated regulation of dendro-somatic synergy in layer 5 pyramidal neurons. *Front. Cell. Neurosci.* **15**, 718413 (2021).
48. T. Tax, P. A. Mediano, M. Shanahan, The partial information decomposition of generative neural network models. *Entropy* **19**, 474 (2017).
49. D. A. Ehrlich, A. C. Schneider, V. Priesemann, M. Wibral, A. Makkeh, A measure of the complexity of neural representations based on partial information decomposition. *Trans. Mach. Learn. Res.* **2023** (2023).
50. T. F. Varley, O. Sporns, S. Schaffelhofer, H. Scherberger, B. Dann, Information-processing dynamics in neural networks of macaque cerebral cortex reflect cognitive state and behavior. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2207677120 (2023).
51. A. Ingel, A. Makkeh, O. Corcoll, R. Vicente, Quantifying reinforcement-learning agent's autonomy, reliance on memory and internalisation of the environment. *Entropy* **24**, 401 (2022).
52. C. E. Shannon, A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948).
53. T. M. Cover, J. A. Thomas, *Elements of Information Theory* (John Wiley & Sons, Inc., ed. 2, 2006).
54. M. E. Larkum, T. Nevian, Synaptic clustering by dendritic signalling mechanisms. *Curr. Opin. Neurobiol.* **18**, 321–331 (2008).
55. R. Y. Cajal *et al*., Nuevo concepto de la histología de los centros nerviosos. *Ind. Med. Gaz.* **2**, 565–566 (1894).
56. J. Aru, M. Suzuki, M. E. Larkum, Cellular mechanisms of conscious processing. *Trends Cogn. Sci.* **24**, 814–825 (2020).
57. A. M. Bastos *et al*., Canonical microcircuits for predictive coding. *Neuron* **76**, 695–711 (2012).
58. N. Takahashi, T. G. Oertner, P. Hegemann, M. E. Larkum, Active cortical dendrites modulate perception. *Science* **354**, 1587–1590 (2016).
59. N. Takahashi *et al*., Active dendritic currents gate descending cortical outputs in perception. *Nat. Neurosci.* **23**, 1277–1285 (2020).
60. A. Rockel, R. W. Hiorns, T. Powell, The basic uniformity in structure of the neocortex. *Brain J. Neurol.* **103**, 221–244 (1980).
61. K. P. Körding, P. König, Supervised and unsupervised learning with two sites of synaptic integration. *J. Comput. Neurosci.* **11**, 207–215 (2001).
62. W. McGill, Multivariate information transmission. *Trans. IRE Prof. Group Inf. Theory* **4**, 93–111 (1954).
63. M. Harder, C. Salge, D. Polani, Bivariate measure of redundant information. *Phys. Rev. E* **87**, 012130 (2013).
64. N. Bertschinger, J. Rauh, E. Olbrich, J. Jost, N. Ay, Quantifying unique information. *Entropy* **16**, 2161–2183 (2014).
65. R. A. Ince, Measuring multivariate redundant information with pointwise common change in surprisal. *Entropy* **19**, 318 (2017).
66. C. Finn, J. T. Lizier, Pointwise partial information decompositionusing the specificity and ambiguity lattices. *Entropy* **20**, 297 (2018).
67. D. A. Ehrlich *et al*., Partial information decomposition for continuous variables based on shared exclusions: Analytical formulation and estimation. *Phys. Rev. E* **110**, 014115 (2024).
68. A. M. Saxe *et al*., On the information bottleneck theory of deep learning. *J. Stat. Mech: Theory Exp.* **2019**, 124020 (2019).
69. Z. Goldfeld *et al*., Estimating information flow in deep neural networks. *Int. Conf. Mach. Learn.* **97**, 2299–2308 (2019).
70. Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition. *Proc. IEEE* **86**, 2278–2324 (1998).
71. J. J. Hopfield, Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. U.S.A.* **79**, 2554–2558 (1982).
72. M. Graetz, "Infomorphic networks: Locally learning neural networks derived from partial information decomposition," M.Sc. thesis, ETH, Zürich, Switzerland (2021).
73. R. P. Rao, D. H. Ballard, Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* **2**, 79–87 (1999).
74. Y. LeCun, Y. Bengio, G. Hinton, Deep learning. *Nature* **521**, 436–444 (2015).
75. W. Samek, T. Wiegand, K. R. Müller, Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. arXiv [Preprint] (2017). http://arxiv.org/abs/1708.08296 (Accessed 28 August 2017).

76. J. C. Whittington, R. Bogacz, Theories of error back-propagation in the brain. *Trends Cogn. Sci.* **23**, 235–250 (2019).
77. A. C. Schneider *et al.*, "What should a neuron aim for? Designing local objective functions based on information theory" in *The Thirteenth International Conference on Learning Representations* (2025). https://openreview.net/forum?id=CLE09ESvul. Accessed 11 February 2025.
78. A. I. Luppi *et al.*, A synergistic workspace for human consciousness revealed by integrated information decomposition. *Elife* **12**, RP88173 (2024).
79. A. I. Luppi *et al.*, A synergistic core for human brain evolution and cognition. *Nat. Neurosci.* **25**, 771–782 (2022).
80. A. M. Proca *et al.*, Synergistic information supports modality integration and flexible learning in neural networks solving multiple tasks. *PLoS Comput. Biol.* **20**, e1012178 (2024).
81. R. Linsker, Self-organization in a perceptual network. *Computer* **21**, 105–117 (1988).
82. M. London, A. Roth, L. Beeren, M. Häusser, P. E. Latham, Sensitivity to perturbations in vivo implies high noise and suggests rate coding in cortex. *Nature* **466**, 123–127 (2010).
83. B. Bischl *et al.*, Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **13**, e1484 (2023).
84. M. Graetz, A. Makkeh, A. C. Schneider, Code for "A general framework for interpretable neural learning based on local information-theoretic goal functions." GitLab. https://gitlab.gwdg.de/wibral/infomorphic_networks. Deposited 22 November 2024.
85. A. A. Makkeh *et al.*, A general framework for interpretable neural learning based on local information-theoretic goal functions, V1. Göttingen Research Online Data. https://doi.org/10.25625/0M1PTJ. Accessed 2 July 2024.