Predictive coding with spiking neurons and feedforward gist signaling

Lee, K.; Dora, S.; Mejias, J.F.; Bohte, S.M.; Pennartz, C.M.A.

DOI
10.3389/fncom.2024.1338280

Publication date
2024

Document Version
Final published version

Published in
Frontiers in Computational Neuroscience

frontiers | Frontiers in Computational Neuroscience

TYPE Original Research
PUBLISHED 12 April 2024
DOI 10.3389/fncom.2024.1338280

OPEN ACCESS

EDITED BY
Guenther Palm,
University of Ulm, Germany

REVIEWED BY
Willem Wybo,
Helmholtz Association of German Research Centres (HZ), Germany
Jean-Philippe Thivierge,
University of Ottawa, Canada

*CORRESPONDENCE
Kwangjun Lee
✉ k.lee@uva.nl
Cyriel M. A. Pennartz
✉ c.m.a.Pennartz@uva.nl

RECEIVED 14 November 2023
ACCEPTED 14 March 2024
PUBLISHED 12 April 2024

CITATION
Lee K, Dora S, Mejias JF, Bohte SM and Pennartz CMA (2024) Predictive coding with spiking neurons and feedforward gist signaling.
Front. Comput. Neurosci. 18:1338280.
doi: 10.3389/fncom.2024.1338280

Predictive coding with spiking neurons and feedforward gist signaling

Kwangjun Lee[1]*, Shirin Dora[1,2], Jorge F. Mejias[1], Sander M. Bohte[1,3] and Cyriel M. A. Pennartz[1]*

[1]Cognitive and Systems Neuroscience Group, Swammerdam Institute for Life Sciences, Faculty of Science, University of Amsterdam, Amsterdam, Netherlands, [2]Department of Computer Science, School of Science, Loughborough University, Loughborough, United Kingdom, [3]Machine Learning Group, Centre of Mathematics and Computer Science, Amsterdam, Netherlands

Predictive coding (PC) is an influential theory in neuroscience, which suggests the existence of a cortical architecture that is constantly generating and updating predictive representations of sensory inputs. Owing to its hierarchical and generative nature, PC has inspired many computational models of perception in the literature. However, the biological plausibility of existing models has not been sufficiently explored due to their use of artificial neurons that approximate neural activity with firing rates in the continuous time domain and propagate signals synchronously. Therefore, we developed a spiking neural network for predictive coding (SNN-PC), in which neurons communicate using event-driven and asynchronous spikes. Adopting the hierarchical structure and Hebbian learning algorithms from previous PC neural network models, SNN-PC introduces two novel features: (1) a fast feedforward sweep from the input to higher areas, which generates a spatially reduced and abstract representation of input (i.e., a neural code for the gist of a scene) and provides a neurobiological alternative to an arbitrary choice of priors; and (2) a separation of positive and negative error-computing neurons, which counters the biological implausibility of a bi-directional error neuron with a very high baseline firing rate. After training with the MNIST handwritten digit dataset, SNN-PC developed hierarchical internal representations and was able to reconstruct samples it had not seen during training. SNN-PC suggests biologically plausible mechanisms by which the brain may perform perceptual inference and learning in an unsupervised manner. In addition, it may be used in neuromorphic applications that can utilize its energy-efficient, event-driven, local learning, and parallel information processing nature.

KEYWORDS

predictive processing, visual cortex, spiking neural network, Hebbian learning, unsupervised learning, representation learning, recurrent processing, sensory processing

1 Introduction

In the midst of chaotic barrages of sensory information, the brain achieves seamless perception of the world. Despite the apparent ease with which the brain achieves such a formidable feat, the problem of perception is computationally difficult, given that the brain has no direct access to the world. This renders perception into an inverse problem (Pizlo, 2001; Spratling, 2017): the brain has to infer a distal stimulus in the physical world (i.e., cause) from proximal sensations coded in the brain (i.e., effect) (Fechner, 1948). Moreover, given inherently noisy and ambiguous sensory information, the problem also becomes ill-posed. For example,

an exponentially growing number of object arrangements and viewing conditions in the three-dimensional world can form the same two-dimensional retinal image.

How does the brain overcome such ambiguity, find a unique and stable solution to the inverse problem, and facilitate seamless perception? A confluence of constructivist theories of perception (Helmholtz, 1867; Kant, 1908; MacKay, 1956; Neisser, 1967; Gregory, 1970; Pennartz, 2015) suggests that the brain imposes a priori constraints on possible solutions to the inverse problem based on an internal model of the world shaped by prior knowledge, experience, and context. In light of recent neurophysiological evidence that supports interaction of feedforward sensory inputs and feedback of a priori knowledge (Felleman and Van Essen, 1991; Bastos et al., 2012; Keller et al., 2012; Walsh et al., 2020), predictive coding (PC) has been proposed as a possible neural implementation of perception (Srinivasan et al., 1982; Mumford, 1992; Rao and Ballard, 1999; Friston, 2005; Pennartz et al., 2019). According to PC in its canonical version (Rao and Ballard, 1999), the brain employs hierarchical cortical circuits in which feedback connections carry predictions to lower areas, whereas feedforward connections carry the mismatch between actual and predicted neural activity (i.e., prediction error). The prediction error is used iteratively to correct the internal generative model, allowing to make more accurate inferences, but also to learn from errors. While its computational goal to explain away incoming sensory input resembles the ideas of redundancy reduction from information theory (Shannon, 1948) and the efficient coding hypothesis (Barlow, 1961), the probabilistic formalization of PC algorithms that approximate Bayesian inference (Friston, 2010) builds on the Bayesian brain hypothesis (Knill and Pouget, 2004) and Helmholtz machine (Dayan et al., 1995). In summary, PC offers a Bayes-inspired solution to the inverse and ill-posed problem of perception, and learning thereof, under the imperative of prediction error minimization. A primary goal of PC modeling is therefore to develop perceptual representations, from which inputs can be generatively reconstructed, in a biologically plausible manner, whereas the more "cognitive" goal of stimulus categorization or classification comes in second position.

Thanks to its potential to explain a multitude of cognitive and neural phenomena (Srinivasan et al., 1982; Rao and Ballard, 1999; Hosoya et al., 2005; Jehee et al., 2006; Spratling, 2010, 2016; Huang and Rao, 2011; Wacongne et al., 2012), PC has inspired many theoretical and computational models of perception. On the one hand, there are biologically motivated PC models in the literature to explain neural mechanisms of perception; on the other hand, machine learning inspired models seek missing ingredients that can place the perceptual capacity of artificial intelligence on par with nature's most intelligent machine. However, both approaches lack biological plausibility in their own respect. Biologically motivated PC models demonstrate how PC accounts for neuronal responses such as classical and extra-classical receptive field properties, but whether their efforts can be generalized across the cortical processing hierarchy remains an open question as their models were confined to specific components of the nervous system, such as the retina (Srinivasan et al., 1982; Hosoya et al., 2005), lateral geniculate nucleus (Huang and Rao, 2011), or V1 (Spratling, 2010), or had a limited depth of processing hierarchy (Rao and Ballard, 1999; Spratling, 2010; Wacongne et al., 2012). The machine learning inspired models show remarkable object recognition capabilities but lack the biological plausibility due to their reliance on supervised learning, convolutional filters, and backpropagation of errors (Whittington and Bogacz, 2017; Sacramento et al., 2018; Van den Oord et al., 2018; Wen et al., 2018; Han et al., 2019; Lotter et al., 2020). Meanwhile, there has been an effort to bridge the gap between the two approaches: a deep gated Hebbian PC (Dora et al., 2021) successfully learns internal representations of natural images across multiple layers of the visual processing hierarchy, while exhibiting neuronal response properties such as orientation selectivity, object selectivity, and sparseness. Yet, previous models relied on an artificial neural network, the basic computational unit of which mimics a real neuron with limited biological realism (Maass, 1997) and communicates using synchronous and continuous signals instead of spikes.

To advance the biological realism of computational models of PC and move toward a more biologically plausible model of perception, we developed a spiking neural network for predictive coding (SNN-PC) by introducing two novel features: (1) a spiking neuron model (Maass, 1997; Gerstner, 2002) that describes the behavior of neurons with more biological details than firing-rate based artificial neurons, such as using binary, asynchronous spikes for synaptic communication and replacing a simple non-linear activation function with synaptic and membrane potential dynamics; and (2) a feedforward gist (FFG) pathway that is added to a PC hierarchy, and mimics the gist of a scene or image (Oliva and Torralba, 2006), inspired by how the visual cortical system may rapidly recognize objects using a fast feedforward visual pathway (Thorpe et al., 1996; Serre et al., 2007; VanRullen, 2007). While our primary goal is to build a spiking neural network that learns a generative model of input image patterns (i.e., to perform image reconstructions) via predictive coding with biologically plausible mechanisms, we also investigate whether such a generative model can be used for a discriminative task (i.e., classification) despite having no explicit objective to optimize it. We hypothesize that having a coarse-level prior about incoming stimuli via the FFG pathway would help with forming classifiable latent representations. In the following sections, we describe non-trivial problems in implementing a spiking version of PC networks, such as encoding signed signals with binary spikes and finding error gradients for experience-dependent learning, and offer our biologically plausible solutions to make PC compatible with spikes. We show that, by putting together all the pieces, SNN-PC can learn hierarchical representations of MNIST hand-written digit images and infer unseen samples from spike signals of sensory inputs in an unsupervised manner.

2 Materials and methods

The following section is organized into four subsections, which address challenges of implementing a PC neural network with spiking neurons and propose biologically plausible mechanisms to facilitate perceptual inference and learning: (1) introduction of a spiking neuron model; (2) description of synaptic communication between spiking neurons for reliable signal transmission and

TABLE 1 Parameters for the adaptive exponential integrate-and-fire model.

Parameter	Value	Unit
C_m	281	pF
g_L	30	nS
E_L	−70.6	mV
V_θ	−50.4	mV
ΔT	2	mV
t_{ref}	2	ms
c	4	nS
b	0.0805	nA
τ_a	144	ms
τ_{rise}	5	ms
τ_{decay}	50	ms

Hebbian learning; (3) separation of error-computing neurons into two groups to encode signed signals with dynamic binary spikes; and (4) introduction of the FFG pathway to establish informed initial conditions for prediction-generating neurons as opposed to random initialization.

2.1 Single neuron model

The behavior of single neurons in SNN-PC was defined by the adaptive exponential integrate-and-fire model (Brette and Gerstner, 2005):

$$C_m \frac{dV}{dt} = -g_L * (V(t) - E_L) + g_L \Delta T \exp \frac{V(t) - V_\theta}{\Delta T} + I(t) - a(t) \tag{1}$$

$$\tau_a \frac{da}{dt} = c(V(t) - E_L) - a(t) \tag{2}$$

where C_m is the membrane capacitance, $V(t)$ the membrane potential, g_L the leak conductance, E_L the leak reversal potential, ΔT the slope factor, V_θ the action potential threshold, $a(t)$ the adaptation variable, $I(t)$ the incoming synaptic current, τ_a the adaptation time constant, and c the adaptation coupling parameter. The parameter values are taken from Brette and Gerstner (2005) and listed in Table 1.

The membrane potential dynamics (Equation 1) is described by a linear leak, a voltage-dependent exponential activation, which instantiates the fast activation of sodium channels (Badel et al., 2008), the incoming synaptic current, $I(t)$, and an abstract adaptation variable, a, which couples the membrane potential dynamics with voltage-dependent subthreshold and spike-triggered adaptation (Equation 2) (Gerstner et al., 2014).

At each time point of simulation, t, a neuron sums up all incoming current, $I(t)$, at its postsynaptic terminals to update its membrane potential, $V(t)$. Upon reaching the threshold, V_θ, a spike is generated [i.e., $s(t) = 1$]:

$$s(t) = \begin{cases} 1, & \text{if} \quad V(t) > V_\theta \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

A spike is followed by an instantaneous reset of the membrane potential, V, to the resting potential, ($V_r = -70.6$ mV), and an increase of adaptation variable (a) by an amount (b) to model membrane potential repolarization and spike-triggered current adaptation, respectively:

$$s(t) = 1 \quad \longrightarrow \quad V(t) = V_r \quad \& \quad a(t) = a(t) + b \tag{4}$$

2.2 Synaptic communication between spiking neurons

The behavior of the single neuron model described in the previous section (Equations 1–4) is a function that takes incoming current, $I(t)$, as input and a spike train, $s(t)$, as output. For synaptic communication between spiking neurons, there must be a way to convert the output of a source neuron to input of a target neuron. The current entering a postsynaptic cell j (postsynaptic current; PSC) through N synapses from presynaptic cells i can be formulated as a continuous variable, $I_j(t)$, by applying an exponential low-pass filter to each incoming binary spike train to compute a spike trace, $X_i(t)$, and weighting each spike trace with the corresponding synaptic strength, $W_{i,j}$, and summing the weighted spike traces over N synapses (Equation 5):

$$I_j(t) = \sum_i^N W_{i,j} X_i(t) \tag{5}$$

The spike trace, $X_i(t)$, is similar to the trace variable commonly used in spike timing dependent plasticity (STDP). With a proper choice of time constants, it approximates a generic excitatory postsynaptic current (EPSC) that reflects both the fast component driven by AMPA receptors ($\tau_{rise} = 5$ ms) and the slow component mediated by NMDA receptors ($\tau_{decay} = 50$ ms) (Forsythe and Westbrook, 1988; McBain and Dingledine, 1993) as follows:

$$\frac{dX}{dt} = \frac{Y(t)}{\tau_{rise}} - \frac{X(t)}{\tau_{decay}} \tag{6}$$

In particular, the NMDAR-mediated component of the EPSC ("NMDAR current") can be linked to the intracellular calcium concentration at the postsynaptic site, a high value of which leads to long term potentiation (LTP) (Barria and Malinow, 2005; Granger and Nicoll, 2014). Using the NMDAR currents of pre- and postsynaptic neurons, synaptic weights are adjusted by Hebbian learning. The modification of weights is described in a subsequent section.

The first term in Equation (6) governs the rising slope of the EPSC corresponding to the influx of cations into the postsynaptic neuron, whereas the second term describes its decay. The instantaneous reset of the variable, $Y(t)$, initializes glutamate release into the synaptic cleft, where it binds to AMPA and NMDA

receptors to open up ion channels in the postsynaptic membrane (Equation 7):

$$s(t) = 1 \quad \longrightarrow \quad Y(t) = 1 \tag{7}$$

The glutamate concentration in the synaptic cleft decays exponentially back to the resting state [i.e., $Y(t) \rightarrow 0$] (Equation 8):

$$\frac{dY}{dt} = -\frac{Y(t)}{\tau_{decay}} \tag{8}$$

In summary, the synaptic communication consists of three major steps (Equations 9–11), which can be seen as a serial adaptation of the spike emission and reception filters in the spike response model (Gerstner et al., 2014). For example, consider the following case where presynaptic neurons (indexed by i) project to a postsynaptic neuron j (Figure 1):

First, a spike train from neuron i, $s_i(t)$, is converted to an AMPA- and NMDA-receptor mediated postsynaptic current received by neuron j, $I_j(t)$:

$$h_1 : s_i(t) \mapsto I_j(t) \tag{9}$$

Second, the current arriving at the postsynaptic receptor site of neuron j, $I_j(t)$, influences the membrane potential of neuron j, $V_j(t)$:

$$h_2 : I_j(t) \mapsto V_j(t) \tag{10}$$

Third, the membrane potential, $V_j(t)$, generates a spike, $s_j(t)$, when it crosses the threshold, V_θ:

$$h_3 : V_j(t) \mapsto s_j(t) \tag{11}$$

2.3 Implementation of predictive coding

SNN-PC employed the same hierarchical structure as its two predecessors (Figure 2A) (Rao and Ballard, 1999; Dora et al., 2021). Each area (denoted by superscript ℓ) consists of two types of computational units (denoted by subscript i): (1) a representation unit, R_i^ℓ, which infers the causes (i.e., generates latent representations) of incoming sensory inputs in the area below ($R_i^{\ell-1}$) and makes predictions about the neural activity in the area below; and (2) an error unit, E_i^ℓ, which compares the prediction from the area above with inputs from the area below and propagates the difference (i.e., prediction error) to the representation units in the area above ($R_i^{\ell+1}$) to update the inferred causes and refine the prediction.

2.3.1 Error unit

An error unit computes prediction errors by taking the difference between the sensory input in the lowest area, or its latent representation in the case of higher areas, and the corresponding prediction from the area above. Depending on the relative strengths of the two signals, this difference can be positive (i.e., input > prediction) or negative (i.e., prediction < input). The signed nature of the prediction error poses no obstacle to PC models with artificial neurons, which can encode positive and negative signals. However, given the non-negative nature of spike signals, SNN-PC has to adopt a different solution that can encode both types of prediction error. As observed in the dopaminergic system, a neuron may encode both types of prediction error by expressing the magnitude of error in relation to its baseline firing rate: positive errors are encoded with activity above its baseline firing rate and negative errors with activity below (Schultz et al., 1997). However, such a neuron would require a very high spontaneous firing rate to encode the full range of negative responses, which contrasts with experimental evidence that suggests low baseline firing rates of layer 2/3 principal neurons (De Kock et al., 2007; Perrenoud et al., 2016). Moreover, in a system where single neurons encode bi-directional errors, a postsynaptic neuron that receives error signals must have a mechanism to subtract out the baseline firing rate of a presynaptic neuron. Given the discrete and non-linear dynamics of spiking neurons, which renders accurate approximation of synaptic transmission non-trivial, our attempts to implement bi-directional error coding with spiking neurons led to inaccurate propagation of prediction errors. Therefore, we separated the error unit into two subtypes, one coding positive and the other coding negative error (Figure 2B). The two units are complementary in propagating prediction errors to representation units in the next higher area. A positive error unit (Equation 12) integrates bottom-up excitatory inputs from representation units within the same area, $X_R^\ell(t)$, and top-down inhibitory inputs from representation units of the area above, $(W^{\ell,\ell+1})^T X_R^{\ell+1}(t)$, whereas a negative error unit (Equation 13) has the opposite arrangement of excitatory-inhibitory synapses:

$$I_{E+}^\ell(t) = X_R^\ell(t) - (W^{\ell,\ell+1})^T X_R^{\ell+1}(t) \tag{12}$$

$$I_{E-}^\ell(t) = (W^{\ell,\ell+1})^T X_R^{\ell+1}(t) - X_R^\ell(t) \tag{13}$$

Note that top-down predictions to both positive and negative units are the same ($(W^{\ell,\ell+1})^T X_R^{\ell+1}(t)$). Representation units and the two types of error unit within an area contain the same number of cells and connect to each other in a one-to-one fashion (i.e., $W_{i,j}^{\ell,\ell} = 1$ where $i = j$ and 0 elsewhere), so that error units receive the same bottom-up input, or its latent representation in the case of higher areas, and compare it to the top-down prediction.

2.3.2 Representation unit

Representation units infer the causes of sensory input via local interactions with error units in the area immediately below as well as those in the same area (Figure 2A). The interactions between two immediately adjacent areas are considered local, because they do not involve areas further down or up in the hierarchy (at least not directly) as would be commonly used in standard deep learning algorithms such as BP, which require global interactions from the top area to the lowest area. The inference process can be regarded as an iterative process of updating internal representations of sensory stimuli (or of neural activity of representation units, I_R^ℓ), and is mathematically formalized as performing a gradient descent on the

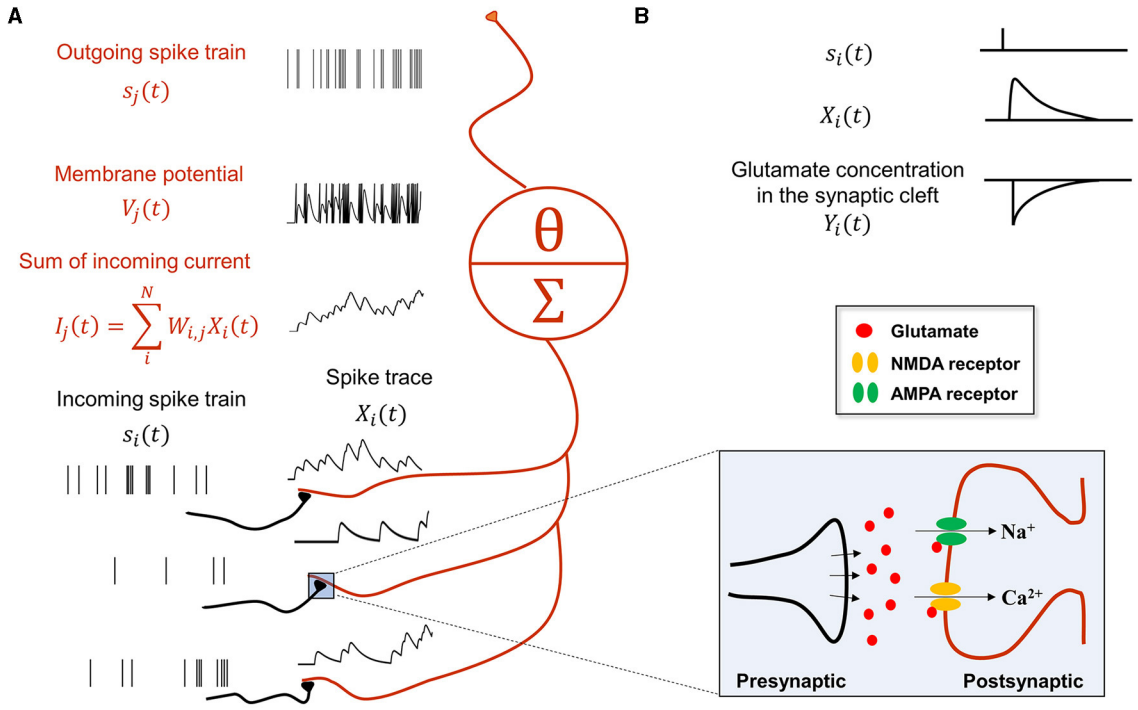

FIGURE 1
Synaptic transmission in a spiking neural network for predictive coding. **(A)** Schematic showing how spiking neurons in SNN-PC communicate with each other using spikes. Spikes from neurons *i* (black), $s_i(t)$, activating synapses that impinge on dendrites of neuron *j* (red), are converted to spike traces, $X_i(t)$. The sum of all spike traces, weighted by synaptic strength, $W_{i,j}$, make up the postsynaptic current, $I_j(t)$. In this particular scheme, we assume that all weights are 1 for simplicity. The postsynaptic membrane potential, $V_j(t)$, changes according to the incoming current and the cell emits spikes, $s_j(t)$, whenever it reaches threshold, V_θ. **(B)** A presynaptic spike ($s_i(t) = 1$) triggers glutamate release into the synaptic cleft [$Y_i(t) = 1$]. When glutamate binds to the postsynaptic AMPA and NMDA receptors, the inward current of cations (Na+ and Ca2+) depolarizes the postsynaptic cell (K+ efflux not shown here for brevity). Concentrations of glutamate in the synaptic cleft and cations (Na+ and Ca2+) in the postsynaptic terminal decrease exponentially [$Y_i(t) \to 0$ and $X_i(t) \to 0$, respectively] with time constants, τ_{rise} and τ_{decay}, respectively.

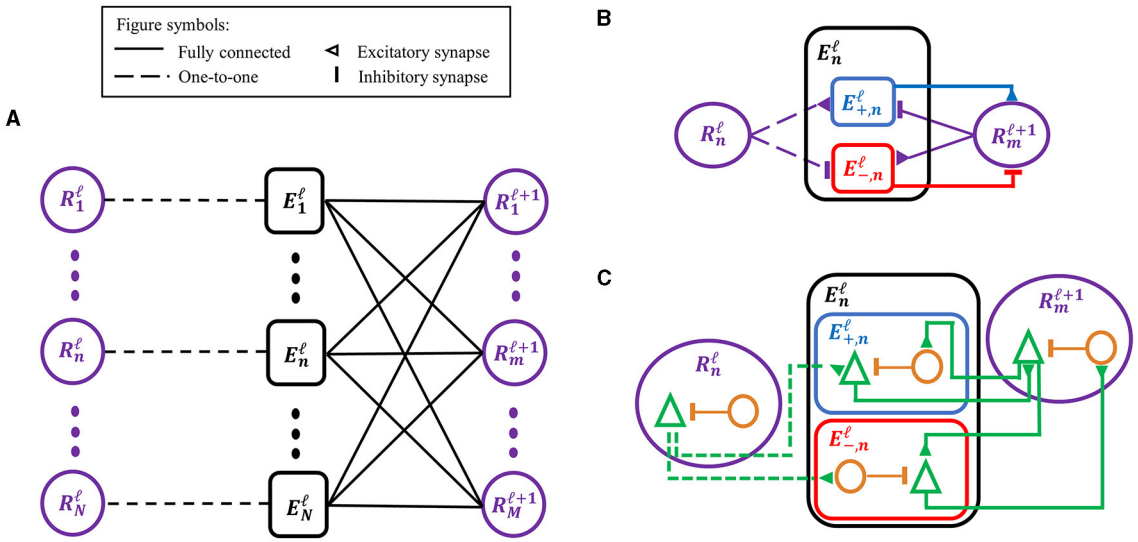

FIGURE 2
Adaptations of the classic predictive coding (PC) architecture to different levels of biological detail. **(A)** The classic PC architecture with representation (prediction) and error neurons as described by Rao and Ballard (1999). **(B)** To encode signed error signals with binary spikes, an error unit in A, E_n^ℓ, is separated into two units, $E_{+,n}^\ell$ and $E_{-,n}^\ell$, which compute a positive and negative error, respectively, via the opposite arrangement of excitatory-inhibitory synapses. The superscript (ℓ) denotes the cortical processing area, whereas the subscript (n or m) denotes an index for unit in an area. **(C)** A biological interpretation of a computational unit in SNN-PC, which consists of a pyramidal cell (green triangle) and an interneuron (orange circle). An excitatory synapse exists between pyramidal cells of two units, whereas a polysynaptic inhibitory connection is formed by a pyramidal cell in a computational unit to an interneuron in another unit, which inhibits the pyramidal cell within the same unit.

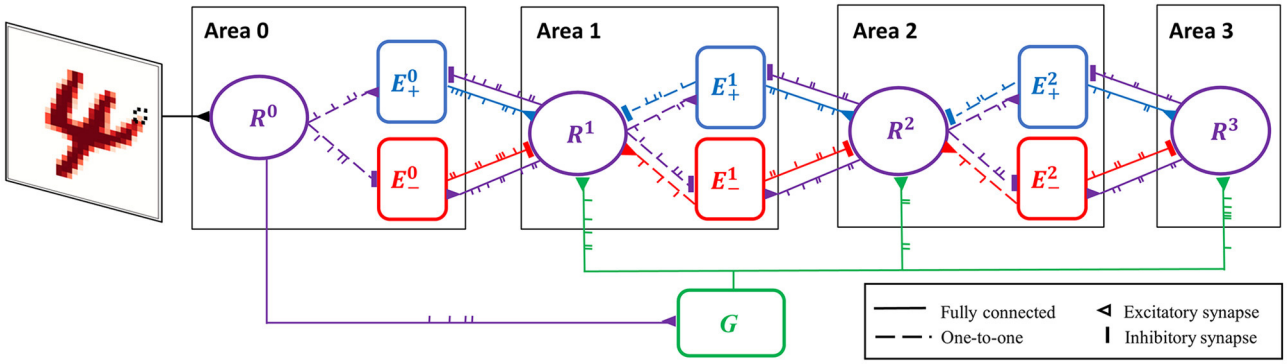

FIGURE 3
Spiking neural network for predictive coding (SNN-PC). SNN-PC modeling the first stages of the visual cortical processing hierarchy. The three areas roughly correspond to V1, V2, and V4, respectively (or LGN, V1, and V2). Each non-input area ($\ell > 0$) consists of a representation unit (purple circle; R^ℓ) and two units (blue and red squares; E^ℓ_+ and E^ℓ_-) that encode positive and negative prediction error, respectively. The representation unit in Area 0 acts as an input unit (R^0). Each pixel (e.g., the dotted box in the image of digit 4) is encoded by a single spiking neuron. Note that each unit consists of multiple spiking neurons. The feedforward gist pathway ($R^0 \rightarrow G \rightarrow R^\ell$) approximates the feedforward sweep of neuronal activity across the visual processing hierarchy. The solid lines between units indicate that they are fully connected, whereas the dotted lines indicate one-to-one connections. A triangle represents an excitatory synapse, whereas a thick vertical short ending represents an inhibitory synapse.

cost function of prediction error minimization with respect to the internal representation (Rao and Ballard, 1999). In SNN-PC, this first comes down to a sum of incoming synaptic current to each representation neuron:

$$I^\ell_R(t) = \beta^{\ell-1}_+ - \beta^{\ell-1}_- - \beta^\ell_+ + \beta^\ell_- \qquad (0 < \ell < L) \quad (14)$$

The first two terms in Equation (14) represent the bottom-up positive and negative prediction error ($\beta^{\ell-1}_+$ and $\beta^{\ell-1}_-$), computed as weighted sums of postsynaptic currents arising from the connections between the two error units and representation units, respectively (Equations 15, 16):

$$\beta^{\ell-1}_+ = W^{\ell-1,\ell} X^{\ell-1}_{E+}(t) \quad (15)$$

$$\beta^{\ell-1}_- = W^{\ell-1,\ell} X^{\ell-1}_{E-}(t) \quad (16)$$

The last two elements of Equation (14) are top-down positive and negative errors (β^ℓ_+ and β^ℓ_-), respectively, which are connected to representation units in a one-to-one fashion (Equations 17, 18):

$$\beta^\ell_+ = X^\ell_{E_+}(t) \quad (17)$$

$$\beta^\ell_- = X^\ell_{E_-}(t) \quad (18)$$

In the case of the highest area (e.g., Area 3 in Figure 3), these two terms are absent as it lacks top-down connections.

Sensory inputs are fed into the network via representation units in the lowest area (Area 0), each of which receives a constant current linearly proportional to the intensity of a pixel of the input image (Figure 3). The underlying assumption is that such a transduction is roughly comparable to a retinal image (with a resolution of a pixel). Given an MNIST digit sample (28 × 28 pixel image) as visual input, the number of units in Area 0 is 784.

2.4 Feedforward gist pathway

Visual cortical processing can be parsed into two distinct processes (Lamme and Roelfsema, 2000): (1) the fast feedforward sweep of neuronal activity across the visual processing hierarchy roughly within 150 ms of stimulus onset in primates, which is thought to generate coarse high-level representation of a visual scene and facilitates gist perception and rapid object recognition (Rousselet et al., 2005; Serre et al., 2007; VanRullen, 2007; Liu et al., 2009; Cauchoix et al., 2016); and (2) slow recurrent processing that iteratively refines the representation (a process henceforth referred to as inference). SNN-PC implements the former process with a FFG pathway and the latter process with the PC hierarchy.

The FFG pathway approximates the feedforward sweep across the visual hierarchy via sparse random projections from input to gist units (Figure 4):

$$I_G(t) = W_{I,G} X_G(t) \quad (19)$$

The weights between input and gist units ($W_{I,G}$ in Equation 19) were randomly sampled from a Gaussian distribution, the mean and standard deviation of which were defined as a ratio between the number of pre- and postsynaptic units. To induce sparsity, the connection probability between input and gist units was set to a low value ($P_c = 0.05$; Figure 4A).

To reflect the increasing receptive field size and complexity of tuning properties when ascending the visual processing hierarchy, the number of gist units (16) is set to be smaller than input units (784). The resulting neuronal activity patterns in gist units therefore correspond to a coarse-grained representation of incoming sensory input. As all input images are processed by the same set of non-plastic, sparse random weights, images that share more features (e.g., two samples belonging to the category of digit "1") have a higher chance to generate similar neural activity patterns in gist units than those that share less (e.g., a sample belonging to digit "0" and another belonging to "1"); in other words, by statistically sampling the same area in the visual field given different images,

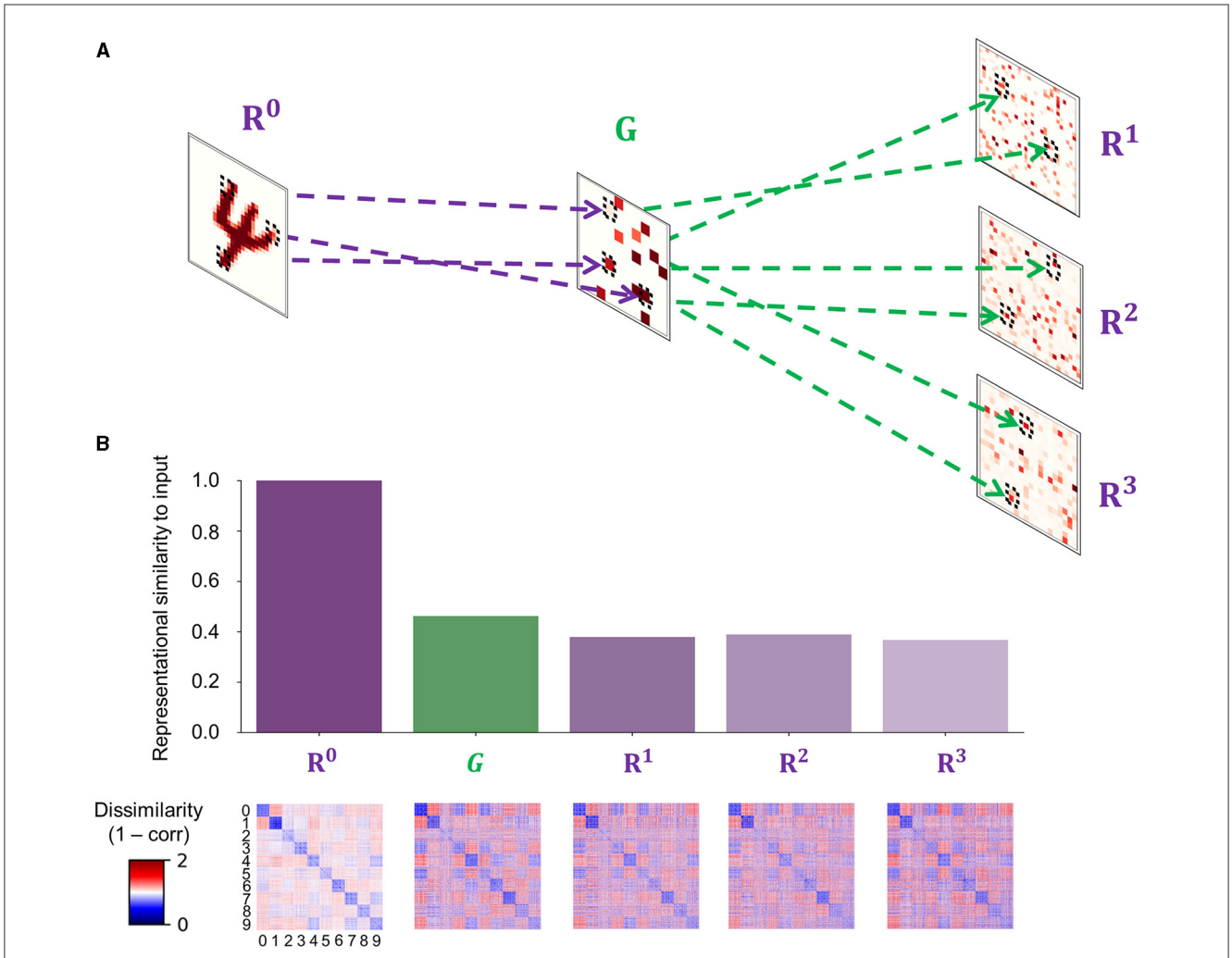

FIGURE 4
The fast feedforward gist pathway. **(A)** A fast feedforward sweep of neural activity from input (R^0) to representation units (R^ℓ where $\ell \in \{1, 2, 3\}$) via gist units (G) is instantiated by sparse random connections to generate a spatially reduced, abstract representation of an input image, which provides an informative baseline for the upcoming iterations of recurrent processing. Purple dashed lines show example projections from input to gist units, whereas green dashed lines show example projections from gist to representation units in areas 1–3 of the visual processing hierarchy. Each tilted square represents an example neural activity pattern activated by an input image in each area. The colored elements in each area correspond to mean incoming synaptic currents. **(B)** A comparison between internal representations formed in areas involved in the feedforward gist pathway (R^0, G, and R^ℓ where $\ell \in \{1, 2, 3\}$). Note that representation units only receive synaptic inputs from gist units and exclude other synaptic currents from the recurrent dynamics of the PC hierarchy ($R^\ell = W_{G,R^\ell} X_G$). The bar plot displays a measure of similarity between input images and internal representations. Strong correlations indicate that the representational geometry of input images is retained across the FFG pathway ($R^0 \rightarrow G$) and that representation units in each cortical area ($G \rightarrow R^\ell$ where $\ell \in \{1, 2, 3\}$) receive informative stimulus-based priors about the upcoming inference. Each element in a representational dissimilarity matrix (RDM) below each bar represents a measure of dissimilarity between 128 different samples per digit (0–9).

the latent representations of input images in gist units retain the representational geometry of input images (Figure 4B).

Gist units then project to representation units in each area of the PC hierarchy to modulate their activity. The synaptic input coming from gist units can be implemented into the inference step (Equation 14) by adding an extra term ($W_{G,R^\ell} X_G$):

$$I_R^\ell(t) = \beta_+^{\ell-1} - \beta_-^{\ell-1} - \beta_+^\ell + \beta_-^\ell + W_{G,R^\ell} X_G(t) \qquad (0 < \ell < L) \qquad (20)$$

The FFG pathway runs in parallel with the PC hierarchy and is active as long as the stimulus lasts to provide a high-level, coarse representation of the incoming sensory input to representation units in each PC area ($W_{G,R^\ell} X_G$; Equation 20) as a baseline activity.

In summary, the FFG pathway serves the role of initializing the neuronal activity of representation units in each area with a coarse representation of the incoming sensory input (e.g., the gist of a scene or object). Instead of starting the iterative process of prediction error minimization from zero or arbitrary activity in representation units, the gist-like latent representation of incoming sensory inputs operates in a biologically plausible manner.

2.5 Rate-based Hebbian learning

With non-differentiable binary spike signals, $s(t)$, the error gradient required to correct the internal model cannot be obtained.

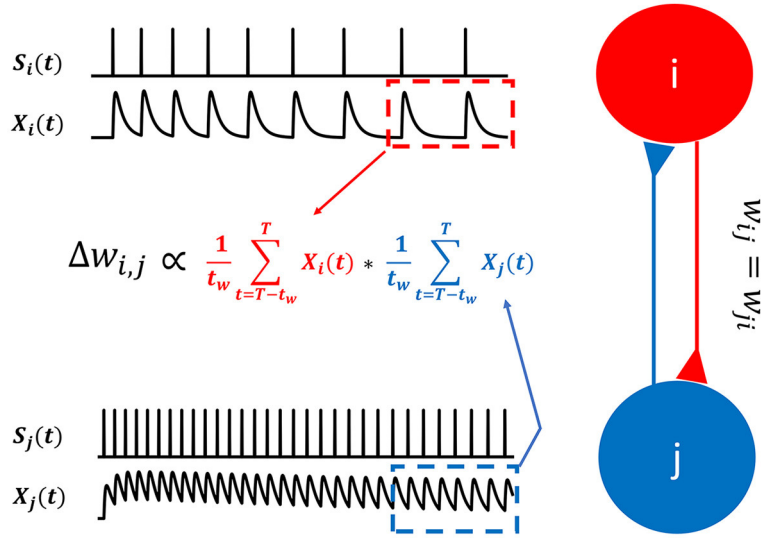

FIGURE 5
Hebbian learning for spiking neurons. In SNN-PC, synaptic plasticity is mediated by Hebbian learning. Instead of firing rates used in artificial neural networks, SNN-PC approximates NMDA receptor-mediated postsynaptic calcium dynamics, $X_i(t)$ and $X_j(t)$, in each cell, based on incoming spike trains, $s_i(t)$ and $s_j(t)$, and leverages them to compute a biologically plausible learning gradient, ΔW_{ij}. The red and blue dotted box indicate the time window, the last t_w ms (from T - t_w to T, where T is the total duration of stimulus presentation), during which the approximate postsynaptic calcium transient signals are averaged.

However, we can use exponentially filtered spike trains, $X(t)$, to obtain the ingredients required for Hebbian learning (Figure 5).

A weight matrix, $W^{\ell,\ell+1}$, between error and representation units (Figure 3), is updated via:

$$\Delta W_{+}^{\ell,\ell+1} = \frac{1}{t_w}\sum_{t=T-t_w}^{T} X_{E+}^{\ell}(t) \times \frac{1}{t_W}\sum_{t=T-t_w}^{T} X_R^{\ell+1}(t) \quad (\ell < L) \quad (21)$$

$$\Delta W_{-}^{\ell,\ell+1} = \frac{1}{t_w}\sum_{t=T-t_w}^{T} X_{E-}^{\ell}(t) \times \frac{1}{t_w}\sum_{t=T-t_w}^{T} X_R^{\ell+1}(t) \quad (\ell < L) \quad (22)$$

The two terms in Equations (21, 22) are mean NMDAR current amplitudes entering the postsynaptic site of error and representation units from the last t_w milliseconds (ms) of stimulus presentation (total duration = T ms): Equation (22) specifies the use of positive error and Equation (22) of negative error. Apart from convergence and stability purposes to accommodate spiking dynamics, taking this mean reflects the calcium dynamics in dendritic spines, which induce NMDA receptor-dependent long term plasticity and depression (LTP and LTD) (Collingridge and Bliss, 1987; Malenka and Nicoll, 1993; Lüscher et al., 2000). The positive error units contact representation units with excitatory synapses to increase the calcium influx into representation units and can induce LTP, whereas the negative error units make inhibitory synaptic contact to decrease calcium influx and induce LTD (Mulkey and Malenka, 1992). Combining the two (Equations 21, 22) results in a weight update that is a linear combination of the Hebbian error gradients obtained between postsynaptic representation units and the two types of presynaptic error units (Equation 23):

$$\Delta W^{\ell,\ell+1} = \gamma_w\left(\Delta W_{+}^{\ell,\ell+1} - \Delta W_{-}^{\ell,\ell+1}\right) - \alpha_w\, g(W^{\ell,\ell+1}) \quad (\ell < L) \quad (23)$$

The weight change is controlled by the learning rate, γ_w. The last term, $\alpha_w g(W^{\ell,\ell+1})$, models the passive decay of weights by imposing a Laplacian prior on the weights (i.e., L1 regularization) (Dora et al., 2021) (Equation 24):

$$g : x \to \begin{cases} 1, & \text{if } x > 0 \\ 0, & \text{otherwise} \end{cases} \quad (24)$$

The resulting unsupervised learning algorithm can be regarded as a biologically plausible form of Hebbian learning. It uses the AMPA- and NMDA-receptor mediated postsynaptic currents between neurons located in adjacent cortical areas (i.e., area ℓ and $\ell + 1$), thereby requiring only locally available information. This is in contrast to backpropagation, which often requires explicit labels (supervised learning) and an end-to-end propagation of errors, from individual output units up to input units. Note that only inter-areal weights ($W^{\ell,\ell+1}$) are subject to synaptic plasticity. The intra-areal weights ($W^{\ell,\ell}$) are fixed.

2.6 Simulation details

2.6.1 Preprocessing of input image

The input images were normalized to unit vectors to limit the variance among pixel intensity distributions across different digits and samples and scaled to a range between 600 and 3,000 pA, within which the input and output synaptic currents approximated a linear relationship.

TABLE 2 Parameters for simulation.

Parameter	Meaning	Value
n_{R^0}	Number of units in R^0	784
$n_{E^0_+}$	Number of units in E^0_+	784
$n_{E^0_-}$	Number of units in E^0_-	784
n_{R^1}	Number of units in R^1	400
$n_{E^1_+}$	Number of units in E^1_+	400
$n_{E^1_-}$	Number of units in E^1_-	400
n_{R^2}	Number of units in R^2	225
$n_{E^2_+}$	Number of units in E^2_+	225
$n_{E^2_-}$	Number of units in E^2_-	225
n_{R^3}	Number of units in R^3	64
n_G	Number of units in G	16
dt	Simulation time step	1 ms
τ_w	Time window for synaptic plasticity	100 ms
T	Total simulation time per sample	350 ms
γ_w	Learning rate for synaptic plasticity	1e-7
α_w	Regularizer for synaptic plasticity	1e-5
n^{train}_{sample}	Number of samples in training set	5,120
n^{test}_{sample}	Number of samples in test set	1,280
n^{batch}_{sample}	Number of samples in a mini-batch	32
n^{epoch}_{batch}	Number of mini-batches per training epoch	160
n_{epoch}	Number of training epochs	50

2.6.2 Network size

Each area consisted of the same number of positive error units, negative error units, and representation units (Area 0 = 784 × 3 = 2,352; Area 1 = 400 × 3 = 1,200; Area 2 = 225 × 3 = 675), except the top area (Area 3) that only contained 64 representation units (Table 2). There were 16 gist units. In total, the number of units in SNN-PC was 4,307. Out of 431,842 total synapses in the network, 418,000 inter-areal synapses were subject to synaptic plasticity.

2.6.3 Training and testing

In order to test whether SNN-PC can learn statistical regularities of incoming sensory inputs and build latent representations thereof, we trained SNN-PC with a subset of the MNIST handwritten digit image dataset. The training set ($n^{sample}_{train} = 5,120$; Table 2) consisted of many different image samples per image class (digits 0–9; 512 samples / class). For efficient learning, we used mini-batch training ($n^{batch}_{sample} = 32$). During a single training epoch, the network goes through all mini-batches ($n^{epoch}_{batch} = 160$). After each mini-batch, the synaptic weights are updated. After 50 training epochs ($n_{epoch} = 50$), we tested the model to infer image samples it had not been exposed to during the training, taken from a test set ($n^{test}_{sample} = 1,280$). For statistical inference, the testing phase was repeated 100 times. Each test set was randomly sampled from 10,000 images that the network had not seen during the training.

For learning, we took the mean over the last 100 ms (300–400 ms) of synaptic current relative to the onset of the stimulus to ensure convergence. Weights were initialized randomly, but strictly positive, by sampling from a half-normal distribution around zero mean and a small standard deviation (0.3). They were updated after every batch with an initial learning rate (γ_w) of 1e-7 and a regularization parameter (α_w) of 1e-5. The learning rate for each pair of areas (e.g., Area 1 and 2) was adjusted by fitting the normalized root mean squared errors of input area (e.g., Area 1) to an exponential growth function.

2.7 Representational similarity analysis

A representational similarity analysis (RSA; Kriegeskorte et al. 2008) computes pairwise similarity between population responses for given inputs. The output of this analysis generates a representational dissimilarity matrix (RDM), each block of which represents the dissimilarity (1 − correlation) between responses to different images. To assess the consistency of information propagation through the processing hierarchy, we conducted RSA on the internal representations of input images (X^ℓ_R). We used the Spearman rank correlation coefficient as a measure of correlation distance between internal representations (see Kriegeskorte et al. 2008 for more choices of distance measures). A second-order RSA computes a similarity measure between RDMs and represents how similarly two areas of interest respond to a given set of input patterns, thereby revealing consistency in representational geometry (i.e., how well internal representations reflect the relationship between input images).

3 Results

3.1 Representational learning

As a generative model with an objective function of prediction error minimization, SNN-PC is expected to generate internal representations of input stimuli, which capture their underlying structures (i.e., probability distribution) in a high-dimensional latent space and, therefore, can be used to reconstruct them. To test this representational capacity, we trained it with a small subset of the MNIST handwritten image dataset ($n_{class} = 10$ and $n_{image} = 5120$; 8.5% of a full set, which has 60,000 images) and evaluated its internal representations of those images. Moreover, while a two-layer structure suffices for image reconstruction in principle, we investigated the impact of hierarchical dynamics on reconstruction performance and other cognitive functions such as image classification by introducing additional processing areas. Note that classification capacity of learned internal representations is only a serendipitous byproduct of our original goal: learning a generative model. We report that additional hierarchical constraints do not impair or improve reconstruction performance.

Learning performance can first be qualitatively assessed from the error units in the lowest area (Area 0), which receive bottom-up sensory inputs that directly correspond to pixel intensities of

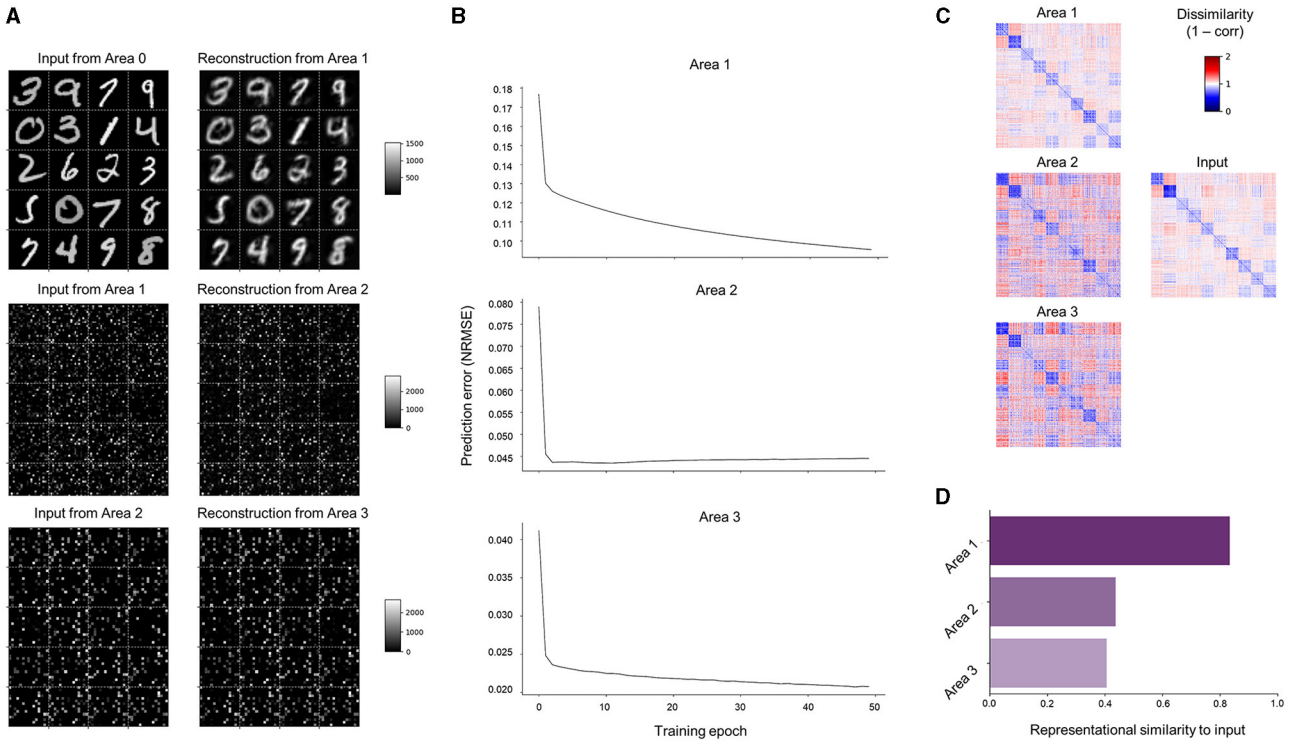

FIGURE 6
Learning of MNIST handwritten digits. **(A)** Reconstruction of sensory inputs. Input to each area (X_R^ℓ; left) is compared to the prediction ($W^{\ell,\ell+1}\ X_R^{\ell+1}$) made from the area immediately above (right). **(B)** Inference. Normalized root mean squared prediction errors (NRMSEs) in each area across training epochs computed by SNN-PC. **(C)** Representational dissimilarity matrices (RDMs) constructed with images in the input area (middle row, right column) and inferred causes of those images in each area from SNN-PC with (left column). Each element in an RDM represents a measure of dissimilarity between them (1 − correlation). **(D)** Second-order representational similarity analysis. The RDM of input neuron activities is compared against the RDMs of inferred causes of those input signals in each area of the SNN-PC model.

input images (X_R^0) and top-down predictions that reconstruct them ($W^{0,1}X_R^1$). By organizing neural activity patterns that correspond to the predictions (Area $1 \rightarrow 0$) into the input shape (28×28 pixels), the images reconstructed by SNN-PC can be visualized. SNN-PC was able to reconstruct digit images well (Figure 6A; top subpanel). While similar inspection in higher areas showed matching patterns between input and prediction (Figure 6A; Area $2 \rightarrow 1$ shown in the middle and Area $3 \rightarrow 2$ shown in the bottom subpanel) as well, predictions in higher areas ($W^{\ell+1,\ell}X_R^\ell$; Area $\ell + 1 \rightarrow \ell$ where $\ell > 1$) bear no immediately recognizable meaning to a human beholder; therefore, they are called latent representations. To show that learning takes place in all areas more evidently, we show the normalized root mean squared errors (NRMSE) as a scale-free measure of the discrepancy between incoming inputs (either sensory or from area 1 to 2 or area 2 to 3) and their predictions. The decreasing NRMSEs in all three areas across training epochs (Figure 6B) indicate that SNN-PC learned to hierarchically minimize prediction errors. Furthermore, a representational similarity analysis (RSA) (Kriegeskorte et al., 2008) on the internal representations of input images (X_R^ℓ) revealed that similar representations were formed across the hierarchy. Note that the colored square boxes shown in Figure 6C are representational dissimilarity matrices (RDMs), which illustrate all pairwise dissimilarities (1 − correlation) among the internal representations corresponding to the input images. To quantitatively evaluate how well those internal representations (X_R^0) reflect the relationship between input images (i.e., representational geometry), we conducted a second-order RSA. The results revealed that Area 1 exhibited a high correlation with inputs, whereas Area 2 and 3 displayed weak correlations with inputs (Figure 6D). Furthermore, despite the decreasing prediction errors (Figure 6B; middle and bottom subpanels), reconstructions of input images from Area 2 and 3 [i.e., $W^{0,1}(W^{1,2}X_R^2)$ and $W^{0,1}(W^{1,2}(W^{2,3}X_R^3))$] failed to produce the input images. Hence, for the subsequent analyses in this study, we will focus on the neuronal activities of Area 1 (X_R^1) and refer to them as SNN-PC's internal, latent representation of input images. Possible reasons for the underperformance in higher areas will be examined in the Discussion section. Overall, good reconstruction performance (Figure 6A; top subpanel), proper prediction error minimization (Figure 6B; top subpanel), and strong representational similarity (Figures 6C, D; Area 1) in Area 1 suggest that SNN-PC has successfully extracted statistical regularities from sensory inputs and updated its internal representations to infer their causes more accurately.

3.2 Robustness of latent representations

To examine the robustness of the representational capacity of SNN-PC, we tested it on a set of MNIST digit images, which it had not seen during the training (i.e., "Clean" set; $n_{class} = 10$ and

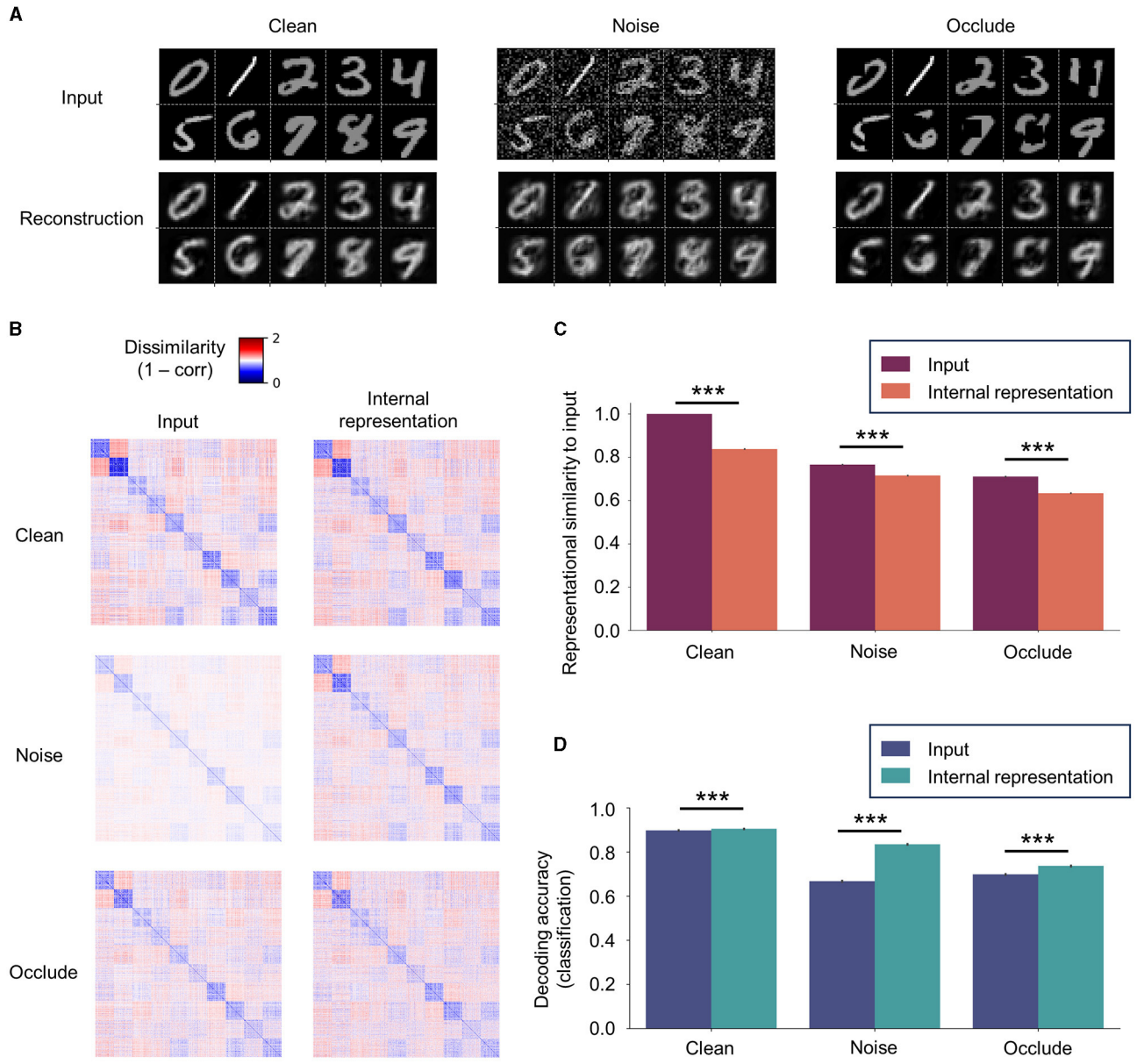

FIGURE 7
Robustness of internal representations. **(A)** SNN-PC's reconstruction of novel (previously unseen; "Clean"), noisy ("Noise"), and occluded ("Occlude") images. The network was able to generalize the internal model acquired during training to novel instances and denoise. However, it does not fully restore the occluded part of an image. **(B)** The robustness of internal representations against addition of noise or partial loss of input signals is shown by RDMs among novel input images and corresponding internal representations. **(C)** The internal representations were highly correlated to novel input images ($\rho > 0.6$ for all bars in **C**). Note that robustness refers to the network's representational capacity to preserve the relationship between input images in the high-dimensional, latent space (i.e., representational geometry) against input perturbations. **(D)** Digit class information can be decoded better from internal representations than input images with or without input perturbations. For statistical comparisons in **(C, D)**, we computed RDMs and classification accuracy on 100 test sets (n_{img} = 1,280) randomly sampled from the "Clean" test set (n_{img} = 10,000). Differences between the means were then assessed by Mann-Whitney *U*-test. Statistical significance is indicated by asterisks (***$p \leq 0.001$). Error bars indicating 95% confidence intervals. Note that a linear classifier was trained only once on Clean images and internal representations inferred from those images and tested on Clean, Noise, and Occlude dataset images and corresponding internal representations.

$n_{image\ per\ class} = 128$). Subsequently, we challenged the robustness by testing the network on two additional datasets created by modifying the input statistics of the Clean set: (1) the first set ("Noise") was given additive Gaussian noise, $\epsilon \sim \mathcal{N}(0, 300)$ (with pA as unit), that spanned across the whole image; and (2) the second set ("Occlude") was masked by occlusion patches at random locations on images (patch size = 9 × 9 pixel; 10.3% of an image). Note that the network had been trained only on Clean images but tested on all three input variants.

We found that SNN-PC was able to generalize the internal model it had acquired during the training onto novel instances as shown by both faithful reconstruction (Figure 7A; Clean) and

RDMs that highly correlated with input images ($\rho > 0.8$, where ρ is Spearman's rank correlation coefficient; Figures 7B, C, Clean). It was also able to denoise (Figure 7A; Noise) and retain representational geometry of input images ($\rho > 0.6$; Figures 7B, C, Noise). Meanwhile, pixels behind random occlusion patches were not fully restored (Figure 7A; Occlude); thus, the Clean version of the input was not pattern-completed, whereas the occluded input as offered was in fact faithfully reconstructed. However, internal representations generated from partially occluded images still showed strong correlation with input images ($\rho > 0.6$; Figures 7B, C, Occlude).

The contrast between dissimilarities within and between classes in an RDM reflects the network's capability of encoding sensory inputs into meaningful latent representations; a greater contrast indicates more generalizable representations and often leads to a better classification performance. The reduced contrast in the RDMs of Noise or Occlude input neuron activities compared to those of Clean input neuron activities (Figure 7B; "Clean" vs. Noise and Clean vs. Occlude at input level) reflects the effect of additive Gaussian noise or random patch occlusion on input images. This loss of representational similarity from input perturbations persisted in the RDMs of latent representations (Figure 7C; "Clean" vs. Noise and Clean vs. Occlude at the internal representation level). However, the corresponding internal representations in both input variants were still strongly correlated with original images ($\rho > 0.6$).

To assess the robustness of the discriminative capacity of SNN-PC against input perturbations, we trained a linear classifier on the internal representations of images from the training set without any perturbations (the same set as in Figure 6; $n_{class} = 10$ and $n_{image\ per\ class} = 512$) and tested it on the images from test sets with (Clean) and without input perturbations (Noise and Occlude). For statistical comparison, Mann-Whitney's U-test was used on the classification results of 100 test sets ($n_{class} = 10$ and $n_{image\ per\ class} = 128$) randomly sampled from 10,000 images. Our results revealed that digit class could be decoded better from internal representations than from input images themselves in all three input variants (Figure 7D; $p < 0.001$).

3.3 The effect of the feedforward gist pathway

While the faithful input reconstruction, consistent representational geometry, and improved decoding observed with novel (Clean) and corrupted (Noise and Occlude) datasets could indicate meaningful representation learning of input statistics, the effect of the FFG pathway remained elusive thus far. For instance, the network could also have learned to build an internal model entirely based on feedforward gist inputs, thereby rendering the recurrent dynamics of PC redundant, or vice versa. To investigate the contribution of the FFG pathway to perceptual inference and learning, we trained another network on the same training set without the FFG pathway ("PC-only"), which learns the underlying structures of input images purely from the recurrent dynamics of PC. We compared reconstruction, representational geometry, and classification results of the PC-only model against those of the original model ("PC + FFG"), which employs both the PC hierarchy and the FFG pathway, and the FFG pathway model ("FFG-only"), in which input signals were processed only by the FFG pathway.

Without PC, the FFG pathway could not faithfully reconstruct input images (Figure 8A; FFG-only), formed internal representations that were only weakly correlated to input images ($\rho < 0.4$; Figures 8B, C; FFG-only), and performed poorly on a digit classification task [decoding accuracy of FFG-only model or DA (FFG-only) $\approx$0.6; Figure 8D; FFG-only]. These results were consistent with our modeling objectives for the FFG pathway: it was designed to generate only a coarse representation of input images. Note that, to visualize the gist-like internal representations, we reconstructed input images using synaptic weights from a model trained with both FFG and PC (PC + FFG).

On the other hand, PC could reconstruct input images in the absence of FFG just as well as in its presence (Figure 8A; PC-only). In fact, internal representations correlated significantly more with input images without than with FFG inputs to the network [ρ (PC-only) $> \rho$(PC + FFG) with $p < 0.001$; Figure 8B]. However, a model classified the digit class of images better when it had learned statistical regularities of those images with both PC and FFG pathway than PC alone [DA (PC + FFG) > DA (PC-only) with $p < 0.001$; Figure 8D].

The reconstruction of sensory inputs was achieved with or without the FFG pathway, because internal representations are inferred via the PC hierarchy which is configured to minimize the difference between sensory inputs and a linear transformation of internal representations (i.e., prediction error). However, the difference in representational similarity to input images between the two cases (Figure 8C; PC + FFG vs. PC-only; $p < 0.001$) suggests that they did not converge on the same internal representations, despite both preserving the representational geometry of input images ($\rho > 0.8$ in both cases). The only difference between the two cases in how they inferred internal representations of input images was the prior: with the FFG pathway, representation units are provided with informative priors about input images through sparse connections ($\rho > 0.3$; Figure 8C; FFG-only); whereas a purely PC network assumes uniform priors about input images. In sum, our results suggest that the FFG pathway aids representational learning via PC by providing informative priors about sensory inputs for the upcoming inference process, helping to infer internal representations that are more discriminative of class information than the uniform priors of PC alone. Effectively, the FFG can be said to install a prior in the network based on which the PC machinery refines the representation.

4 Discussion

While inspired by biological neurons, artificial neural networks make many assumptions for the sake of functional and computational efficiency. For instance, the assumption underlying the use of firing rates as a measure of neural activity is that information is rate coded. However, the brain is also thought to encode information using both the timing of spikes (e.g., phase coding and neural synchrony) (Gray et al., 1989; O'Keefe and

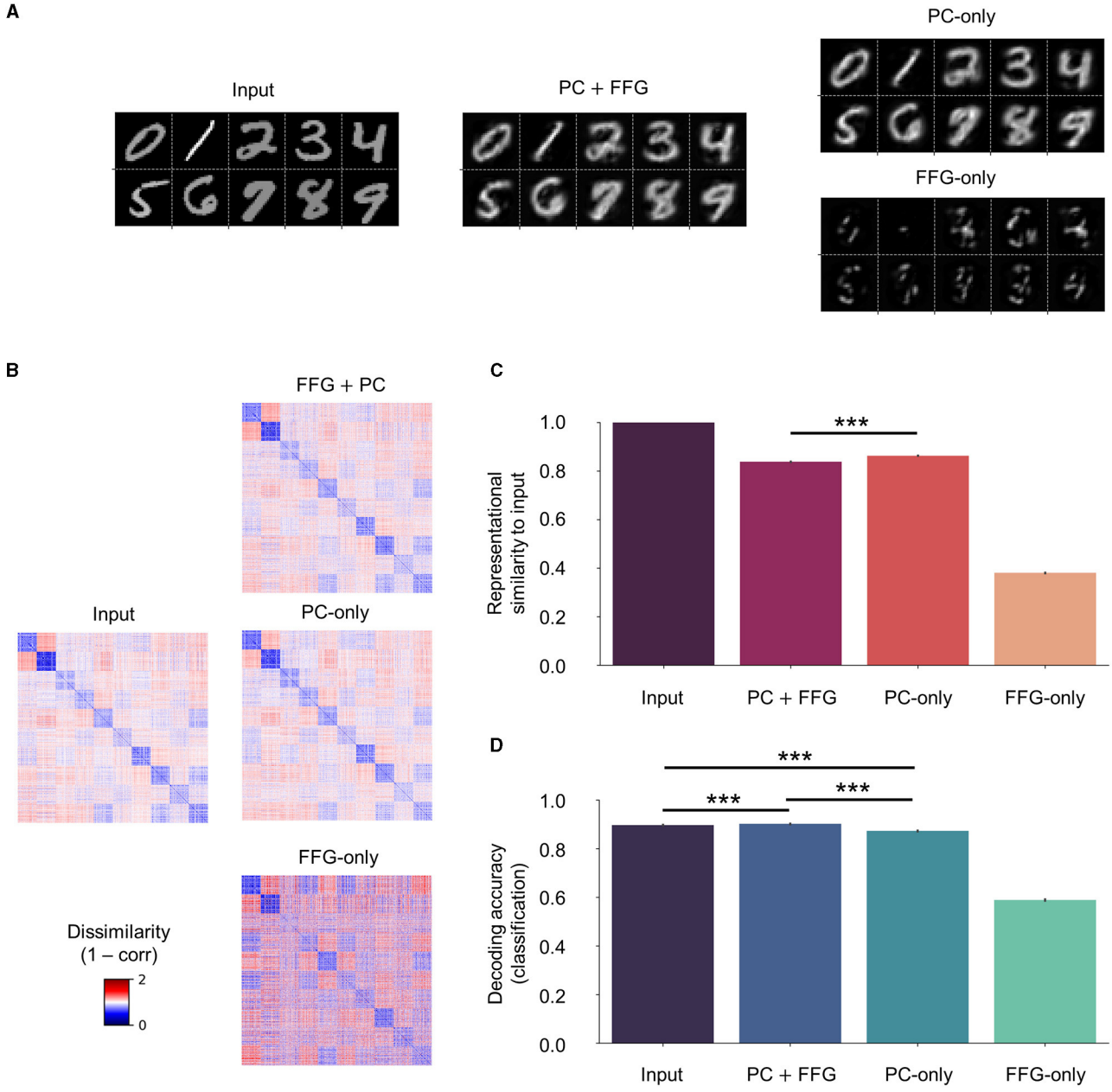

FIGURE 8
Effect of the feedforward gist pathway on representational learning. Three models were compared against each other on input reconstruction, representational geometry, and classification: (1) a model that processes input stimuli via both the PC hierarchy and the FFG pathway ("PC + FFG"); (2) a model that utilizes only the PC hierarchy ("PC-only"); and (3) the FFG pathway model ("FFG-only"). **(A)** Input reconstruction is a result of the recurrent dynamics of PC (PC + FFG and PC-only). The FFG pathway alone cannot account for the generative capacity (FFG-only). **(B)** PC preserves the metric relationship (1 − correlation) among input images in the representational geometry. **(C)** In the presence of PC, internal representations show a strong correlation with inputs ($\rho > 0.8$). The FFG pathway alone generates coarse, gist-like representations that only weakly correlate to input images ($\rho < 0.4$). **(D)** The FFG pathway helps slightly but significantly in inferring internal representations from input images that are more discriminative of digit class than PC-only. For statistical comparisons in **(C, D)**, we computed RDMs and classification accuracy on 100 test sets (n_{img} = 1,280) randomly sampled from the "Clean" test set (n_{img} = 10,000). Differences between the means were then assessed by Mann-Whitney *U*-test. Statistical significance is indicated by asterisks (***$p \leq 0.001$). Error bars indicating 95% confidence intervals. Note that four separate linear classifiers were trained on input images and internal representations inferred from those images by the three model variants.

Recce, 1993; Singer, 1999; Knoblauch and Palm, 2001; Van Rullen and Thorpe, 2001; Brette, 2012; Ono and Oliver, 2014) and aggregate responses of neuronal ensembles (i.e., population coding) (Georgopoulos et al., 1986; Lee et al., 1988; Pouget et al., 2000; Averbeck et al., 2006). By collapsing the temporal dimension into an arbitrary iteration step, artificial neurons cannot leverage asynchronous, event-based, and sparse information processing for energy efficiency (Pfeiffer and Pfeil, 2018; Tavanaei et al., 2019; Deng et al., 2020). Moreover, the non-local, end-to-end error propagation of BP, often used in ANNs, poses a serious challenge

to biological plausibility (Rumelhart et al., 1986; Bengio et al., 2015; Sacramento et al., 2018; Whittington and Bogacz, 2019; Lillicrap et al., 2020; Song et al., 2020). To create a realistic system that performs complex cognitive behaviors on par with a human agent, we need to study and incorporate principles of neural computation and architectures from the biological agent we want to mimic, namely the mammalian brain. A straightforward choice in pursuing such endeavors therefore was to introduce spiking neurons, as they provide a biophysically realistic level of detail to simulate basic computations in the brain.

To this end, we developed a biologically grounded neural network for generative visual modeling (SNN-PC), based on the following four components: (1) a predictive coding model that provides computational algorithms and a neural architecture for generative perceptual inference via recurrent sensory processing; (2) a FFG pathway that accounts for rapid feedforward processing in the visual cortical system; (3) spiking neurons that reflect the time-varying pulsatile behavior of neurons better than rate-based neurons; and (4) Hebbian learning enabled by NMDAR-mediated synaptic plasticity. The model learned to reconstruct and develop latent representations of the MNIST handwritten digit images using only a small subset of the image dataset (8.5% of the full dataset) and an unsupervised learning method that requires only locally available information at each level of the hierarchy (as opposed to end-to-end backpropagation). Furthermore, our implementation of PC is based on biologically grounded mechanisms such as an adaptive spike generation mechanism, synaptic transmission modeling the effect of presynaptic spikes on the postsynaptic membrane potential, and synaptic plasticity based on calcium transients in postsynaptic dendritic spines.

4.1 Robustness against noise and occlusion

While previous studies have explored robustness of PC network's generative capacity against noise and partial occlusion, both denoising and pattern completion require structural modifications such as lateral connections and auxiliary connections between non-local areas (Ororbia, 2023) or algorithmic adjustments such as adding a memory vector and unclamping input units of the missing pixels from an input image and allowing it to vary from top-down prediction (Salvatori et al., 2021) or conditional inference on pre-trained labels (Salvatori et al., 2022). Our results exhibit a contrary case, where the missing part is not filled in by the top-down prediction; rather, the SNN-PC infers faithfully from the actual sensory inputs (i.e., with occlusion). This is because, without structural or algorithmic modifications, the local prediction error minimization loop always aligns predictions to inputs. Also, assuming that the reconstruction of sensory inputs from Area 1 of SNN-PC is a direct prediction of the retinal image, the pattern completion should not occur as a subject never actually sees the occluded part. For example, if you walk on a street and find an uncovered manhole, it is not in your interest to fill it in based on priors built upon previous encounters with covered manholes. However, the digit class information was better decoded from internal representations of occluded images in the latent space than from actual images themselves (Figure 7D; Occlude). This implies that SNN-PC was able to capture underlying structures of digit images during training, which were indeed robust against a partial loss of pixels. A recent study (Papale et al., 2023) showed comparable experimental evidence to our results: multi-unit spiking activity of monkey V1 neurons exhibited a significantly weaker response to occluded regions than non-occluded regions; nevertheless, the scene information could be decoded from a cross-decoding experiment (i.e., training on non-occluded images and testing on occluded images).

Additionally, SNN-PC was able to denoise. While structural similarity index measure (SSIM; Wang et al., 2004) values decrease with an increasing level of noise, the network was able to denoise up to a high level of Gaussian noise, and the reconstruction performance did not acutely break down within a range of 0–200 % (figure not shown here). Given the nature of an inference model that builds upon statistical regularities of input signals, SNN-PC simply cannot make predictions about noise, which by definition is unpredictable. Hence, it leveraged on those input signals that could be predicted and showed denoised reconstructions, robust internal representations, and better decodability from internal representations than noisy images themselves.

4.2 Novel features of the predictive coding model with spiking neurons

To our knowledge, no predictive coding model has been proposed before that is operating purely with spiking neurons, except for the spiking neural coding network (SpNCN) proposed in Ororbia (2023). While SNN-PC and SpNCN similarly implement synaptic transmission (i.e., low-pass filtering of spike trains) and weight updating (i.e., Hebbian learning), a few key differences arise from the additional biological constraints we placed on our model. For example, spiking neurons in SNN-PC are based on the AdEx model, which offers a biophysically more accurate description of a neuron's behavior than the leaky-and-integrate fire (LIF) model used in SpNCN (Brette and Gerstner, 2005). However, the most noteworthy difference is that error units in SNN-PC are explicitly modeled as spiking neurons and separated into two groups to encode both positive and negative error with binary spikes, as opposed to being an arbitrary unit that signals a signed difference between two exponentially filtered spike trains, as in Ororbia (2023). While such an arbitrary error unit might be biologically implemented in a dendritic compartment model (e.g., Urbanczik and Senn 2014; Mikulasch et al. 2023), our implementation of error neurons follows experimentally observed mismatch between feedforward and feedback signals in visual cortical neurons (e.g., of the somatostatin-positive type) (Keller et al., 2012; Zmarz and Keller, 2016; Attinger et al., 2017). Moreover, having two types of error unit to compute positive and negative error separately also circumvents the need to employ unrealistic negative synaptic weights for inhibitory connections (Figure 2B). A growing body of recent studies suggests that layer 2/3 of cortex indeed contains neurons that express positive or negative errors (Keller et al., 2012; Jordan and Keller, 2020; O'Toole et al., 2023).

Despite having solved the implausible negative weight problem, all units in SNN-PC (when viewed as single neurons) do not adhere to another biological property (i.e., Dale's principle) as they can form both excitatory (in case of the $R^{\ell} \rightarrow E_{+}^{\ell-1}$ projection) and inhibitory ($R^{\ell} \rightarrow E_{-}^{\ell-1}$) synapses onto other units. However, such a violation can be mitigated by replacing computational units in SNN-PC (R_j^{ℓ}, $E_{+,i}^{\ell}$, and $E_{-,i}^{\ell}$; Figure 2B) by cortical microcircuits that consist of pyramidal cells and interneurons (e.g., green triangle and orange circle wrapped inside purple, blue, and red contours, respectively; Figure 2C). Note that we used one pyramidal cell and one interneuron in a microcircuit for visual presentation purposes only. Using this microcircuit, we predict that an excitatory synapse between two microcircuits (e.g., R_m^{ℓ} and $E_{+,n}^{\ell}$) is formed between their pyramidal cells, whereas an inhibitory synapse between the two microcircuits consists of an excitatory connection from pyramidal cells in a microcircuit (e.g., R_m^{ℓ}) to interneurons in another microcircuit (e.g., $E_{-,n}^{\ell}$), which then inhibits pyramidal cells in the same microcircuit (e.g., $E_{-,n}^{\ell}$). Future research will have to show how PC can be implemented using known anatomical connections (Douglas and Martin, 1991), laminar organization (Bastos et al., 2012; Pennartz et al., 2019), and different cell types (e.g., pyramidal, SST, VIP, and PV) (Keller and Mrsic-Flogel, 2018; O'Toole et al., 2023).

Besides offering a biologically plausible solution to accommodate binary spiking dynamics, the explicit separation of error units into positive and negative elements may also be beneficial for the algorithmic and computational efficiency of neuromorphic hardware. It only requires a straightforward subtraction between input and prediction, whereas bi-directional error units (such as postulated in reinforcement learning models based on prediction error coding by mesencephalic dopamine neurons) (Schultz et al., 1997) must first compute the difference and then compare it against a baseline firing rate to determine the sign of the error. With a certain baseline firing rate, positive and negative errors can be encoded by the range above and below it, respectively. However, for a full coverage of prediction error ranges, bidirectional error units must maintain a high baseline firing rate, thereby leading to a higher energy cost. The neuron targeted by a bidirectional error unit would also have to be equipped with a mechanism to discount for the baseline firing rate to obtain the true prediction error.

4.3 Feedforward gist

Another important novel feature of SNN-PC is the FFG pathway, which combines the fast feedforward sweep and the slow recurrent PC to account for a more comprehensive picture of visual processing. PC reconciles bottom-up and top-down accounts of perception by casting it as an inferential process that involves hierarchical recurrent interactions. However, the inference process requires multiple loops of recurrent processing to converge on an accurate representation of incoming sensory input. The expected latency of visual responses arising from recurrent predictive processing is not in accordance with the rapid forward spread of object- and context-sensitive neuronal activity across the visual cortical hierarchy within 100 ms of stimulus onset (Lamme and Roelfsema, 2000). While the precise contributions of feedforward and recurrent processes to perception are yet to be determined (Kreiman and Serre, 2020), we aimed to combine these two temporally distinct processes by integrating the FFG pathway in a PC architecture and asked whether the FFG pathway can improve network performance.

Reflecting on the temporal dichotomy of the two visual processes, the FFG pathway quickly establishes a high-level, coarse representation of input signals (e.g., the gist of a scene or object) and feeds it to each area of the PC hierarchy to aid the recurrent processing that slowly refines the representation. Instead of starting the iterative process of prediction error minimization from zero or arbitrary activity in representation units, the gist-like latent representation of incoming sensory inputs offers a biologically plausible starting point for predictive coding. This suggests a novel function of the feedforward activity relative to the classic hypothesis of rapid image recognition (Thorpe et al., 1996).

When we tested the impact of the FFG pathway on the learning of two-dimensional visual images, our results showed that its presence during training leads to reduced consistency between internal representations in higher areas with input image statistics (Figure 8). Despite the faithful reconstruction of novel and perturbed sensory inputs, which is largely driven by the recurrent dynamics of the prediction error minimization loop, the diminished classification accuracy in the absence of FFG suggests that the latent representations formed without gist inputs have extracted less information from the image statistics (Figure 8D). These findings suggest that the FFG plays a modest, but statistically significant role, in achieving classifiable latent representations by placing an a priori constraint on the inference process. Effectively, the FFG can be said to install a prior in the network based on which the PC machinery refines the representation, and which improves classifiability.

Meanwhile, it is not clear whether the same populations of cortical neurons may be involved in both the feedforward sweep and recurrent PC. Instead of performing a series of feedforward feature extraction and integration steps to elicit object-sensitive responses in high visual areas, feedforward connections in the PC network convey prediction errors. Therefore, the two processes might take separate routes. For instance, recurrent processes governed by PC may occur via cortico-cortical pathways, whereas the feedforward sweep may be mediated by the pulvino-cortical pathway or by bottom-up projections from low-level visual areas like lateral occipital cortex (Vinberg and Grill-Spector, 2008; Jaramillo et al., 2019). Such involvement of subcortical pathways is in line with the brain not being strictly hierarchical (Suzuki et al., 2023). Alternatively, the FFG pathway can be explained as a combination of the fast feedforward sweep and subsequent top-down modulation: an instantaneous feedforward sweep may activate gist units, conceptualized as IT, PFC, or other high-level cells; this may then be followed by top-down projection of activity from gist to representation units in lower visual areas (e.g., V1, V2, V3 and V4).?

In both scenarios, we assume innate and non-plastic feedforward connections in the FFG pathway. A

recent study (Tschantz et al., 2023) examined how such feedforward connections can be trained via amortized inference and also showed robust perceptual capacity with shorter error convergence time and fewer training samples. However, we note that SNN-PC is spike-based and unsupervised, whereas the model in Tschantz et al. (2023) is rate-based and uses a mix of supervised and unsupervised learning.

4.4 Limitations on scalability

While our results show that SNN-PC can generalize what it had learned from a training set to novel instances of the test set (Figure 7; Clean), higher areas (area 2 and 3) were excluded from analyses beyond Figure 6 due to their weak correlations to input images and subpar reconstruction performance, implying that their representations in the latent space do not capture the underlying statistics well. Despite the decreasing prediction error in higher areas in the PC hierarchy during training (Figure 6B; areas 2 and 3), the progressive decrease in representational similarity across the hierarchy (Figure 6C; areas 2 and 3) suggests a loss of information about the inputs. A possible source of information leakage is the spiking mechanism of a neuron, which fires upon reaching a threshold membrane potential (V_θ), that renders the input-output curve of a current-based spiking neuron non-linear. By consequence, a signal (e.g., an input current of 1,000 pA) loses its strength as it propagates through a series of neurons. Given that each cortical processing area in SNN-PC can reliably reconstruct inputs from the lower network area (Figure 6A) and that Area 1 generates internal representations of input images with a high decodability (Figures 7D, 8D), the most likely location for the information leak would be between representation and error neurons within the same area ($R^\ell \rightarrow E^\ell_{+/-}$). Addressing this leakage will require further work in future studies.

Again, we want to stress that the classification capacity of learned representations is only a byproduct of our original goal: learning a generative model. We also emphasize that, despite the low number of training samples, and the adjustments made to the PC algorithms that facilitate spike communication, we demonstrate generative capacity via reconstruction of previously unseen images (Figure 7A; Clean) and robustness against noise and occlusion (Figure 7A; Noise and Occlude). However, as for classification results, the decoding accuracy did not increase when ascending hierarchical processing areas. This is of no surprise, given the sole objective function of local prediction error minimization: there is no constraint during the training phase to learn to categorize inputs; the network is only instructed to learn internal representations that can reconstruct inputs from the area immediately below. In fact, SNN-PC fulfills this objective (Figures 6A, B). Meanwhile, we think that SNN-PC can be converted to a competitive discriminative model, if the topmost area would be clamped to class labels corresponding to inputs to the lowest area during training to learn class-specific representations via supervised learning (as in Whittington and Bogacz, 2017). This approach, however, would also obviously compromise the pursuit of biological realism.

4.5 Future directions

There are various other ways in which future studies may extend the biological details and/or perceptual capacities of SNN-PC. To name a few, first, a point spiking neuron can be replaced by a cortical microcircuit with multiple interneurons (e.g., Figure 2C) and with a columnar organization to replicate experimental findings and make predictions for new experiments. Second, spiking neuron behavior or connectivity can be altered to implement receptive fields and response properties to construct an invariant object representation (c.f. Brucklacher et al., 2023). Third, self-recurrent loops or online learning rules such as STDP may be employed to deal with a continuous stream of sensory inputs. Fourth, a population coding regime can be implemented to improve the reliability of signal transmission (Boerlin and Denève, 2011). Fifth, a different sensory modality can be added to perform multi-sensory integration, following the rate-based predecessor of our current SNN-PC model (Dora et al., 2021), which has been implemented in a rodent robot that performs bimodal integration of vision and touch to navigate in a maze (Pearson et al., 2021). Sixth, the network can be scaled up to better reflect the areas involved in the visual processing hierarchy. Seventh, while requiring novel learning rules, different coding schemes known to exist in the brain, such as temporal coding (Konishi, 2000; Van Rullen and Thorpe, 2001; Ono and Oliver, 2014), phase coding (O'Keefe and Recce, 1993), and the use of neural synchrony (Gray et al., 1989; Singer, 1999; Knoblauch and Palm, 2001; Brette, 2012), can be explored to make use of computational advantages offered by spiking signals.

5 Conclusion

We have described how to build a PC model of visual perception using biologically plausible components such as spiking neurons, Hebbian learning, and a FFG pathway. As one of the first purely spike-based and completely unsupervised PC models of visual perception, SNN-PC successfully performs perceptual inference and learning as shown by reconstruction of MNIST digit images. Also, it can denoise and show robust decodability of class information from noisy and partially occluded images. Our findings may inspire machine learning, neuromorphics, neuroscience and cognitive science communities to seek avenues moving closer to mimic the nature's most intelligent and efficient system, the brain.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

KL: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Conceptualization. SD: Writing – review & editing, Methodology, Investigation, Conceptualization. JM: Writing – review & editing, Methodology, Conceptualization. SB: Writing – review & editing, Supervision, Methodology, Investigation, Conceptualization. CP: Writing – review & editing, Supervision, Resources, Methodology, Investigation, Funding acquisition, Conceptualization.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This project has received funding from the European Union's Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreement No. 945539 (Human Brain Project SGA3 to CP). We acknowledge the use of Fenix Infrastructure resources, which are partially funded from the European Union's Horizon 2020 Research and Innovation Programme through the ICEI project under the grant agreement No. 800858.

Acknowledgments

The authors would like to thank Matthias Brucklacher, Giulia Moreni, Giovanni Pezzulo, and Walter Senn for constructive discussions.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

References

Attinger, A., Wang, B., and Keller, G. B. (2017). Visuomotor coupling shapes the functional development of mouse visual cortex. *Cell* 169, 1291–1302. doi: 10.1016/j.cell.2017.05.023

Averbeck, B. B., Latham, P. E., and Pouget, A. (2006). Neural correlations, population coding and computation. *Nat. Rev. Neurosci.* 7, 358–366. doi: 10.1038/nrn1888

Badel, L., Lefort, S., Brette, R., Petersen, C. C., Gerstner, W., and Richardson, M. J. (2008). Dynamic iv curves are reliable predictors of naturalistic pyramidal-neuron voltage traces. *J. Neurophysiol.* 99, 656–666. doi: 10.1152/jn.01107.2007

Barlow, H. (1961). Possible principles underlying the transformations of sensory messages. *Sens. Commun.* 1, 217–234.

Barria, A., and Malinow, R. (2005). Nmda receptor subunit composition controls synaptic plasticity by regulating binding to camkii. *Neuron* 48, 289–301. doi: 10.1016/j.neuron.2005.08.034

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., and Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron* 76, 695–711. doi: 10.1016/j.neuron.2012.10.038

Bengio, Y., Lee, D.-H., Bornschein, J., Mesnard, T., and Lin, Z. (2015). Towards biologically plausible deep learning. *arXiv* [preprint]. doi: 10.48550/arXiv.1502.04156

Boerlin, M., and Denève, S. (2011). Spike-based population coding and working memory. *PLoS Comput. Biol.* 7:e1001080. doi: 10.1371/journal.pcbi.1001080

Brette, R. (2012). Computing with neural synchrony. *PLoS Comput. Biol.* 8:e1002561. doi: 10.1371/journal.pcbi.1002561

Brette, R., and Gerstner, W. (2005). Adaptive exponential integrate-and-fire model as an effective description of neuronal activity. *J. Neurophysiol.* 94, 3637–3642. doi: 10.1152/jn.00686.2005

Brucklacher, M., Bohte, S. M., Mejias, J. F., and Pennartz, C. M. (2023). Local minimization of prediction errors drives learning of invariant object representations in a generative network model of visual perception. *Front. Comput. Neurosci.* 17:1207361. doi: 10.3389/fncom.2023.1207361

Cauchoix, M., Crouzet, S. M., Fize, D., and Serre, T. (2016). Fast ventral stream neural activity enables rapid visual categorization. *Neuroimage* 125, 280–290. doi: 10.1016/j.neuroimage.2015.10.012

Collingridge, G. L., and Bliss, T. (1987). Nmda receptors-their role in long-term potentiation. *Trends Neurosci.* 10, 288–293. doi: 10.1016/0166-2236(87)90175-5

Dayan, P., Hinton, G. E., Neal, R. M., and Zemel, R. S. (1995). The helmholtz machine. *Neural Comput.* 7, 889–904. doi: 10.1162/neco.1995.7.5.889

De Kock, C. P. J., Bruno, R. M., Spors, H., and Sakmann, B. (2007). Layer- and cell-type-specific suprathreshold stimulus representation in rat primary somatosensory cortex. *J. Physiol.* 581, 139–154. doi: 10.1113/jphysiol.2006.124321

Deng, L., Wu, Y., Hu, X., Liang, L., Ding, Y., Li, G., et al. (2020). Rethinking the performance comparison between snns and anns. *Neural Netw.* 121, 294–307. doi: 10.1016/j.neunet.2019.09.005

Dora, S., Bohte, S. M., and Pennartz, C. M. A. (2021). Deep gated Hebbian predictive coding accounts for emergence of complex neural response properties along the visual cortical hierarchy. *Front. Comput. Neurosci.* 15:65. doi: 10.3389/fncom.2021.666131

Douglas, R. J., and Martin, K. (1991). A functional microcircuit for cat visual cortex. *J. Physiol.* 440, 735–769. doi: 10.1113/jphysiol.1991.sp018733

Fechner, G. T. (1948). *Elements of Psychophysics, 1860*. East Norwalk, CT: Appleton-Century-Crofts.

Felleman, D. J., and Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex* 1, 1–47. doi: 10.1093/cercor/1.1.1

Forsythe, I. D., and Westbrook, G. L. (1988). Slow excitatory postsynaptic currents mediated by n-methyl-d-aspartate receptors on cultured mouse central neurones. *J. Physiol.* 396, 515–533. doi: 10.1113/jphysiol.1988.sp016975

Friston, K. (2005). A theory of cortical responses. *Philos. Transact. R. Soc. B* 360, 815–836. doi: 10.1098/rstb.2005.1622

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138. doi: 10.1038/nrn2787

Georgopoulos, A. P., Schwartz, A. B., and Kettner, R. E. (1986). Neuronal population coding of movement direction. *Science* 233, 1416–1419. doi: 10.1126/science.3749885

Gerstner, W. (2002). *Spiking Neuron Models: Single Neurons, Populations, Plasticity*. Cambridge; New York, NY: Cambridge University Press.

Gerstner, W., Kistler, W. M., Naud, R., and Paninski, L. (2014). *Neuronal Dynamics*. Cambridge: Cambridge University Press.

Granger, A. J., and Nicoll, R. A. (2014). Expression mechanisms underlying long-term potentiation: a postsynaptic view, 10 years on. *Philos. Transact. R. Soc. B Biol. Sci.* 369:20130136. doi: 10.1098/rstb.2013.0136

Gray, C. M., König, P., Engel, A. K., and Singer, W. (1989). Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. *Nature* 338, 334–337. doi: 10.1038/338334a0

Gregory, R. L. (1970). *The Intelligent Eye*. New York, NY: Mcgraw-Hill.

Han, T., Xie, W., and Zisserman, A. (2019). "Video representation learning by dense predictive coding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*.

Helmholtz, H. (1867). *Handbuch der physiologischen Optiks, Allgemeine Encyclopädie der Physik, IX. Band*. Leipzig: Leopold Voss.

Hosoya, T., Baccus, S. A., and Meister, M. (2005). Dynamic predictive coding by the retina. *Nature* 436, 71–77. doi: 10.1038/nature03689

Huang, Y., and Rao, R. P. (2011). Predictive coding. *Wiley Interdiscip. Rev. Cogn. Sci.* 2, 580–593. doi: 10.1002/wcs.142

Jaramillo, J., Mejias, J. F., and Wang, X.-J. (2019). Engagement of pulvino-cortical feedforward and feedback pathways in cognitive computations. *Neuron* 101, 321–336. doi: 10.1016/j.neuron.2018.11.023

Jehee, J. F., Rothkopf, C., Beck, J. M., and Ballard, D. H. (2006). Learning receptive fields using predictive feedback. *J. Physiol.* 100, 125–132. doi: 10.1016/j.jphysparis.2006.09.011

Jordan, R., and Keller, G. B. (2020). Opposing influence of top-down and bottom-up input on excitatory layer 2/3 neurons in mouse primary visual cortex. *Neuron* 108, 1194–1206. doi: 10.1016/j.neuron.2020.09.024

Kant, I. (1908). *Critique of Pure Reason. 1781*. Cambridge, MA: Houghton Mifflin, Modern Classical Philosophers.

Keller, G. B., Bonhoeffer, T., and Hübener, M. (2012). Sensorimotor mismatch signals in primary visual cortex of the behaving mouse. *Neuron* 74, 809–815. doi: 10.1016/j.neuron.2012.03.040

Keller, G. B., and Mrsic-Flogel, T. D. (2018). Predictive processing: a canonical cortical computation. *Neuron* 100, 424–435. doi: 10.1016/j.neuron.2018.10.003

Knill, D. C., and Pouget, A. (2004). The bayesian brain: the role of uncertainty in neural coding and computation. *Trends Neurosci.* 27, 712–719. doi: 10.1016/j.tins.2004.10.007

Knoblauch, A., and Palm, G. (2001). Pattern separation and synchronization in spiking associative memories and visual areas. *Neural Netw.* 14, 763–780. doi: 10.1016/S0893-6080(01)00084-3

Konishi, M. (2000). Study of sound localization by owls and its relevance to humans. *Comp. Biochem. Physiol. Part A Mol. Integr. Physiol.* 126, 459–469. doi: 10.1016/S1095-6433(00)00232-4

Kreiman, G., and Serre, T. (2020). Beyond the feedforward sweep: feedback computations in the visual cortex. *Ann. N. Y. Acad. Sci.* 1464, 222–241. doi: 10.1111/nyas.14320

Kriegeskorte, N., Mur, M., and Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* 2:4. doi: 10.3389/neuro.06.004.2008

Lamme, V. A., and Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends Neurosci.* 23, 571–579. doi: 10.1016/S0166-2236(00)01657-X

Lee, C., Rohrer, W. H., and Sparks, D. L. (1988). Population coding of saccadic eye movements by neurons in the superior colliculus. *Nature* 332, 357–360. doi: 10.1038/332357a0

Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J., and Hinton, G. (2020). Backpropagation and the brain. *Nat. Rev. Neurosci.* 21, 335–346. doi: 10.1038/s41583-020-0277-3

Liu, H., Agam, Y., Madsen, J. R., and Kreiman, G. (2009). Timing, timing, timing: fast decoding of object information from intracranial field potentials in human visual cortex. *Neuron* 62, 281–290. doi: 10.1016/j.neuron.2009.02.025

Lotter, W., Kreiman, G., and Cox, D. (2020). A neural network trained for prediction mimics diverse features of biological neurons and perception. *Nat. Mach. Intellig.* 2, 210–219. doi: 10.1038/s42256-020-0170-9

Lüscher, C., Nicoll, R. A., Malenka, R. C., and Muller, D. (2000). Synaptic plasticity and dynamic modulation of the postsynaptic membrane. *Nat. Neurosci.* 3, 545–550. doi: 10.1038/75714

Maass, W. (1997). Networks of spiking neurons: the third generation of neural network models. *Neural Netw.* 10, 1659–1671. doi: 10.1016/S0893-6080(97)00011-7

MacKay, D. M. (1956). The epistemological problem for automata. *Autom. Stud.* 34, 235–251.

Malenka, R. C., and Nicoll, R. A. (1993). Nmda-receptor-dependent synaptic plasticity: multiple forms and mechanisms. *Trends Neurosci.* 16, 521–527. doi: 10.1016/0166-2236(93)90197-T

McBain, C. J., and Dingledine, R. (1993). Heterogeneity of synaptic glutamate receptors on ca3 stratum radiatum interneurones of rat hippocampus. *J. Physiol.* 462, 373–392. doi: 10.1113/jphysiol.1993.sp019560

Mikulasch, F. A., Rudelt, L., Wibral, M., and Priesemann, V. (2023). Where is the error? Hierarchical predictive coding through dendritic error computation. *Trends Neurosci.* 46, 45–59. doi: 10.1016/j.tins.2022.09.007

Mulkey, R. M., and Malenka, R. C. (1992). Mechanisms underlying induction of homosynaptic long-term depression in area ca1 of the hippocampus. *Neuron* 9, 967–975. doi: 10.1016/0896-6273(92)90248-C

Mumford, D. (1992). On the computational architecture of the neocortex. *Biol. Cybern.* 66, 241–251. doi: 10.1007/BF00198477

Neisser, U. (1967). *Cognitive Psychology*. New York, NY: Appleton-Century-Crofts.

O'Keefe, J., and Recce, M. L. (1993). Phase relationship between hippocampal place units and the eeg theta rhythm. *Hippocampus* 3, 317–330. doi: 10.1002/hipo.450030307

Oliva, A., and Torralba, A. (2006). Building the gist of a scene: the role of global image features in recognition. *Prog. Brain Res.* 155:23–36. doi: 10.1016/S0079-6123(06)55002-2

Ono, M., and Oliver, D. L. (2014). The balance of excitatory and inhibitory synaptic inputs for coding sound location. *J. Neurosci.* 34, 3779–3792. doi: 10.1523/JNEUROSCI.2954-13.2014

Ororbia, A. (2023). Spiking neural predictive coding for continually learning from data streams. *Neurocomputing* 544:126292. doi: 10.1016/j.neucom.2023.126292

O'Toole, S. M., Oyibo, H. K., and Keller, G. B. (2023). Molecularly targetable cell types in mouse visual cortex have distinguishable prediction error responses. *Neuron* 111, 2918–2928. doi: 10.1016/j.neuron.2023.08.015

Papale, P., Wang, F., Morgan, A. T., Chen, X., Gilhuis, A., Petro, L. S., et al. (2023). The representation of occluded image regions in area v1 of monkeys and humans. *Curr. Biol.* 33, 3865–3871. doi: 10.1016/j.cub.2023.08.010

Pearson, M. J., Dora, S., Struckmeier, O., Knowles, T. C., Mitchinson, B., Tiwari, K., et al. (2021). Multimodal representation learning for place recognition using deep hebbian predictive coding. *Front. Robot. AI* 8:732023. doi: 10.3389/frobt.2021.732023

Pennartz, C. (2015). *The Brain's Representational Power: On Consciousness and the Integration of Modalities*. Cambridge, MA: The MIT Press.

Pennartz, C. M., Dora, S., Muckli, L., and Lorteije, J. A. (2019). Towards a unified view on pathways and functions of neural recurrent processing. *Trends Neurosci.* 42, 589–603. doi: 10.1016/j.tins.2019.07.005

Perrenoud, Q., Pennartz, C. M., and Gentet, L. J. (2016). Membrane potential dynamics of spontaneous and visually evoked gamma activity in v1 of awake mice. *PLoS Biol.* 14:21. doi: 10.1371/journal.pbio.1002383

Pfeiffer, M., and Pfeil, T. (2018). Deep learning with spiking neurons: opportunities and challenges. *Front. Neurosci. 12*:774. doi: 10.3389/fnins.2018.00774

Pizlo, Z. (2001). Perception viewed as an inverse problem. *Vision Res.* 41, 3145–3161. doi: 10.1016/S0042-6989(01)00173-0

Pouget, A., Dayan, P., and Zemel, R. (2000). Information processing with population codes. *Nat. Rev. Neurosci.* 1, 125–132. doi: 10.1038/35039062

Rao, R. P., and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* 2, 79–87. doi: 10.1038/4580

Rousselet, G., Joubert, O., and Fabre-Thorpe, M. (2005). How long to get to the "gist" of real-world natural scenes? *Vis. Cogn.* 12, 852–877. doi: 10.1080/13506280444000553

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature* 323, 533–536. doi: 10.1038/323533a0

Sacramento, J., Ponte Costa, R., Bengio, Y., and Senn, W. (2018). Dendritic cortical microcircuits approximate the backpropagation algorithm. *Adv. Neural Inform. Process. Syst.* 31, 8721–8732.

Salvatori, T., Song, Y., Hong, Y., Sha, L., Frieder, S., Xu, Z., et al. (2021). Associative memories via predictive coding. *Adv. Neural Inf. Process. Syst.* 34, 3874–3886.

Salvatori, T., Pinchetti, L., Millidge, B., Song, Y., Bao, T., Bogacz, R., et al. (2022). Learning on arbitrary graph topologies via predictive coding. *Adv. Neural Inf. Process. Syst.* 35, 38232–38244.

Schultz, W., Dayan, P., and Montague, P. R. (1997). A neural substrate of prediction and reward. *Science.* 275, 1593–1599. doi: 10.1126/science.275.5306.1593

Serre, T., Oliva, A., and Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proc. Nat. Acad. Sci. U. S. A.* 104, 6424–6429. doi: 10.1073/pnas.0700622104

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.* 27, 379–423. doi: 10.1002/j.1538-7305.1948.tb01338.x

Singer, W. (1999). Neuronal synchrony: a versatile code for the definition of relations? *Neuron* 24, 49–65. doi: 10.1016/S0896-6273(00)80821-1

Song, Y., Lukasiewicz, T., Xu, Z., and Bogacz, R. (2020). Can the brain do backpropagation?–exact implementation of backpropagation in predictive coding networks. *Adv. Neural Inf. Process. Syst.* 33, 22566–22579.

Spratling, M. W. (2010). Predictive coding as a model of response properties in cortical area v1. *J. Neurosci.* 30, 3531–3543. doi: 10.1523/JNEUROSCI.4911-09.2010

Spratling, M. W. (2016). A neural implementation of bayesian inference based on predictive coding. *Conn. Sci.* 28, 346–383. doi: 10.1080/09540091.2016.1243655

Spratling, M. W. (2017). A review of predictive coding algorithms. *Brain Cogn.* 112, 92–97. doi: 10.1016/j.bandc.2015.11.003

Srinivasan, M. V., Laughlin, S. B., and Dubs, A. (1982). Predictive coding: a fresh view of inhibition in the retina. *Proc. R. Soc. London Ser. B Biol. Sci.* 216, 427–459. doi: 10.1098/rspb.1982.0085

Suzuki, M., Pennartz, C. M., and Aru, J. (2023). How deep is the brain? The shallow brain hypothesis. *Nat. Rev. Neurosci.* 24, 778–791. doi: 10.1038/s41583-023-00756-z

Tavanaei, A., Ghodrati, M., Kheradpisheh, S. R., Masquelier, T., and Maida, A. (2019). Deep learning in spiking neural networks. *Neural Netw.* 111, 47–63. doi: 10.1016/j.neunet.2018.12.002

Thorpe, S., Fize, D., and Marlot, C. (1996). Speed of processing in the human visual system. *Nature* 381, 520–522. doi: 10.1038/381520a0

Tscshantz, A., Millidge, B., Seth, A. K., and Buckley, C. L. (2023). Hybrid predictive coding: inferring, fast and slow. *PLoS Comput. Biol.* 19:e1011280. doi: 10.1371/journal.pcbi.1011280

Urbanczik, R., and Senn, W. (2014). Learning by the dendritic prediction of somatic spiking. *Neuron* 81, 521–528. doi: 10.1016/j.neuron.2013.11.030

Van den Oord, A., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv* [preprint]. doi: 10.48550/arXiv.1807.03748

Van Rullen, R., and Thorpe, S. J. (2001). Rate coding versus temporal order coding: what the retinal ganglion cells tell the visual cortex. *Neural Comput.* 13, 1255–1283. doi: 10.1162/08997660152002852

VanRullen, R. (2007). The power of the feed-forward sweep. *Adv. Cogn. Psychol.* 3:167. doi: 10.2478/v10053-008-0022-3

Vinberg, J., and Grill-Spector, K. (2008). Representation of shapes, edges, and surfaces across multiple cues in the human visual cortex. *J. Neurophysiol.* 99, 1380–1393. doi: 10.1152/jn.01223.2007

Wacongne, C., Changeux, J.-P., and Dehaene, S. (2012). A neuronal model of predictive coding accounting for the mismatch negativity. *J. Neurosci.* 32, 3665–3678. doi: 10.1523/JNEUROSCI.5003-11.2012

Walsh, K. S., McGovern, D. P., Clark, A., and O'Connell, R. G. (2020). Evaluating the neurophysiological evidence for predictive processing as a model of perception. *Ann. N. Y. Acad. Sci.* 1464, 242–268. doi: 10.1111/nyas.14321

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Transact. Image Process.* 13, 600–612. doi: 10.1109/TIP.2003.819861

Wen, H., Han, K., Shi, J., Zhang, Y., Culurciello, E., and Liu, Z. (2018). "Deep predictive coding network for object recognition," in *International Conference on Machine Learning* (PMLR), 5266–5275.

Whittington, J. C., and Bogacz, R. (2017). An approximation of the error backpropagation algorithm in a predictive coding network with local hebbian synaptic plasticity. *Neural Comput.* 29, 1229–1262. doi: 10.1162/NECO_a_00949

Whittington, J. C., and Bogacz, R. (2019). Theories of error back-propagation in the brain. *Trends Cogn. Sci.* 23, 235–250. doi: 10.1016/j.tics.2018.12.005

Zmarz, P., and Keller, G. B. (2016). Mismatch receptive fields in mouse visual cortex. *Neuron* 92, 766–772. doi: 10.1016/j.neuron.2016.09.057