Trends in Neurosciences

CellPress OPEN ACCESS

Review

Where is the error? Hierarchical predictive coding through dendritic error computation

Fabian A. Mikulasch,[1,5,*] Lucas Rudelt,[1,5] Michael Wibral,[2] and Viola Priesemann[1,3,4]

Top-down feedback in cortex is critical for guiding sensory processing, which has prominently been formalized in the theory of hierarchical predictive coding (hPC). However, experimental evidence for error units, which are central to the theory, is inconclusive and it remains unclear how hPC can be implemented with spiking neurons. To address this, we connect hPC to existing work on efficient coding in balanced networks with lateral inhibition and predictive computation at apical dendrites. Together, this work points to an efficient implementation of hPC with spiking neurons, where prediction errors are computed not in separate units, but locally in dendritic compartments. We then discuss the correspondence of this model to experimentally observed connectivity patterns, plasticity, and dynamics in cortex.

Highlights

Hierarchical predictive coding has been considered one of the most promising unifying theories of cortical computation. Yet, in its classical form, it remains difficult to connect to single neuron physiology.

We review work that shows that hierarchical predictive coding can be implemented by neurons with dendritic compartments, where prediction errors are computed by the local voltage dynamics in the dendrites.

This connects the theories of predictive coding and efficient coding in balanced networks and provides a solution to the open problem of implementing predictive coding with spiking neurons.

This also links predictive coding to cortical physiology and voltage-dependent plasticity, which offers new ways to test for predictive coding in cortex.

Neural models of inference in cortex

A central feature of perception is that our internal expectations to a large degree shape how we perceive the world [1]. A long line of research aims to describe these expectation-guided computations in our brain by Bayesian **inference** (see Glossary) (i.e., statistically optimal perception) and, subsequently, could show that Bayesian inference often captures perception extraordinarily well [2,3] (for a critical discussion see also [4]). In light of these results, it has been proposed that the primary computation that is performed by the cortex is a hierarchically organized inference process, where cortical areas combine bottom-up sensory information and top-down expectations to find a consistent explanation of sensory data [5–8].

While the general idea of hierarchical inference in cortex found considerable experimental support [7,9,10], it is less clear how exactly this inference could be implemented by cortical neurons. A popular theory to describe the neural substrate of inference in cortex is classical **hierarchical predictive coding (hPC)** [6,11]. A central proposition of this theory is the existence of error units, which are thought to compare top-down predictions with bottom-up inputs, and guide neural computation and plasticity. However, classical hPC for the most part remains on the level of firing-rate dynamics of neural populations and it has proven difficult to connect the theory to the properties of single neurons with spiking dynamics [12,13].

Here we point towards a different, emerging theory of hierarchical inference in cortex, which relies on the local membrane dynamics in neural dendrites. The core idea of this theory, which we will refer to as dendritic hPC, is to shift error computation from separate neural populations into the dendritic compartments of **pyramidal neurons**. We will first discuss how this shift in perspective enables a biologically plausible implementation of hPC with spiking neurons, and how it connects hPC to theories of efficient coding in **balanced spiking networks** [14] and **neural sampling** [2]. In the second part, we will discuss the biological plausibility of dendritic hPC, explain how several experimental observations of hierarchical cortical computation fit into the picture, and highlight the experimental predictions that can be generated from the theory.

[1]Max-Planck-Institute for Dynamics and Self-Organization, Göttingen, Germany
[2]Göttingen Campus Institute for Dynamics of Biological Networks, Georg-August University, Göttingen, Germany
[3]Bernstein Center for Computational Neuroscience (BCCN), Göttingen, Germany
[4]Department of Physics, Georg-August University, Göttingen, Germany
[5]These authors contributed equally to this work

*Correspondence:
fabian.mikulasch@ds.mpg.de
(F.A. Mikulasch).

Dendritic predictive coding in balanced spiking neural networks

Classical models of predictive coding

Hierarchical predictive coding (hPC) describes the processing of sensory information as inference in a hierarchical model of sensory data (see Box 1 for mathematical details, which are not needed to understand the main text). The central idea of hPC is that activity of **prediction units** in one level of the hierarchy:

(i) should accurately predict sensory data or the prediction unit activity in a lower level, and
(ii) should be consistent with top-down predictions generated by higher levels in the hierarchy.

hPC tries to understand how these properties of neural activity can be ensured by neural dynamics on short timescales, and neural learning and plasticity on long timescales. The theory predicts that to this end, the prediction units in every level of the hierarchy need access to two types of errors:

Box 1. Mathematical details of classical predictive coding

The goal in hPC is to maximize the model log-likelihood [11] (for a detailed tutorial see [134])

$$\mathcal{L} = \sum_{i=1}^{N} \log p_\theta(\mathbf{r}^{i-1}|\mathbf{r}^i), \qquad [\text{I}]$$

where θ are the model parameters, $\mathbf{r}^i$ is neural activity of a neural network at level i, and inputs are provided by the previous level $\mathbf{r}^{i-1}$. This defines a hierarchy of processing stages that, for example, can be associated with different visual cortical areas (e.g., V1, V2, etc.), where $\mathbf{r}^0$ are visual inputs from LGN [11]. Typically, a linear model is assumed, where inputs are modeled according to

$$\mathbf{r}^{i-1} = \mathbf{D}^i\mathbf{r}^i + \mathbf{n}^{i-1}, \qquad [\text{II}]$$

with decoding matrix $\mathbf{D}^i$ and Gaussian white noise $\mathbf{n}^{i-1}$ with zero mean and variance σ_{i-1}^2. With this model, for a single level i, the relevant contributions of the negative log-likelihood $-\mathcal{L}^i$ take the intuitive form of the square sum of coding errors for bottom-up inputs and errors of top-down predictions:

$$\begin{aligned} &\text{bottom-up error}: \quad \mathbf{e}^{i-1} = \mathbf{r}^{i-1} - \mathbf{D}^i\mathbf{r}^i, \\ &\text{top-down error}: \quad \mathbf{e}^i = \mathbf{r}^i - \mathbf{D}^{i+1}\mathbf{r}^{i+1}, \end{aligned} \qquad [\text{III}]$$

$$-\mathcal{L}^i = \frac{1}{2\sigma_{i-1}^2}\mathbf{e}^{i-1\top}\mathbf{e}^{i-1} + \frac{1}{2\sigma_i^2}\mathbf{e}^{i\top}\mathbf{e}^i. \qquad [\text{IV}]$$

The goal is then to minimize the sum of coding errors on a fast timescale τ_r via neural dynamics $\frac{d}{dt}\mathbf{r}^i$, and with a slow learning rate η_D via neural plasticity on the weights $\mathbf{D}^i$, by performing gradient ascent on $\mathcal{L}$:

$$\text{dynamics}: \quad \tau_r\frac{d}{dt}\mathbf{r}^i = \frac{1}{\sigma_{i-1}^2}\mathbf{D}^{i\top}\mathbf{e}^{i-1} - \frac{1}{\sigma_i^2}\mathbf{e}^i \qquad [\text{V}]$$

$$\text{plasticity}: \quad \eta_D^{-1}\frac{d}{dt}\mathbf{D}^i = \frac{1}{\sigma_{i-1}^2}\mathbf{e}^{i-1}\mathbf{r}^{i\top}. \qquad [\text{VI}]$$

To yield a neural implementation, the key innovation in classical hPC was to represent prediction errors within a distinct neural population of error units. Error units integrate inputs of prediction units within the same level and subtract top-down predictions according to

$$\tau_e\frac{d}{dt}\mathbf{e}^i = -\mathbf{e}^i + \mathbf{r}^i - \mathbf{D}^{i+1}\mathbf{r}^{i+1}, \qquad [\text{VII}]$$

where decoding weights $\mathbf{D}^i$ now correspond directly to weights of neural connections [134]. Together with the dynamics of prediction units, this results in the hierarchical neural circuit shown in Figure 1A in the main text.

Glossary

Balanced spiking networks: recurrent networks of spiking neurons with E-I balance; these networks show asynchronous irregular spiking activity and can efficiently encode dynamic variables.
E-I balance: excitatory and inhibitory currents are 'balanced', when their magnitude approximately matches.
Hierarchical predictive coding (hPC): a theory of hierarchical inference in cortex.
Inference: in hPC, inference is the process of finding the underlying causes of sensory data; these underlying causes can be used to predict (or similarly, 'explain away') the sensory input or the activity in lower levels of the hierarchy.
Lateral inhibition: pyramidal neurons in a population compete via lateral inhibition through interneurons, which can be used to both increase the efficiency of the neural code and to distinguish between competing explanations of sensory data.
Neural sampling: instead of computing a single best explanation of sensory data, neural activity can sample possible explanations according to their likelihood.
Prediction neuron: pyramidal neuron that aims to predict the activity of other neurons, as proposed by dendritic hPC.
Prediction unit: abstract unit of neurons that aims to predict the activity of other units, as proposed by classical hPC.
Pyramidal neuron: the primary excitatory neuron in cortex, typically with a characteristic long 'apical' dendrite.
Tight balance: if the E-I balance is present not only on average, but also on short timescales, it is 'tight'.
Voltage-dependent plasticity (VDP): changes in synaptic strength that depend on the postsynaptic membrane potential in the vicinity of the synapse.

(i) bottom-up errors (i.e., the mismatch between activity in lower levels and predictions generated within the level);
(ii) top-down errors (i.e., the mismatch between activity within the level and top-down predictions from higher levels).

In classical hPC [11], the key innovation was to represent these errors in distinct populations of error units that compare top-down predictions with the activity within a level (Figure 1A, Key figure). The elegance of this approach is that the same error units can mediate both, bottom-up errors to update prediction units in the next level, as well as top-down errors to neurons of the same level. Another central result of classical hPC is that the learning rules that improve the hierarchical model take the form [error × prediction], which turns out to be classic Hebbian plasticity (i.e., the multiplication of pre- and postsynaptic activity).

A functionally equivalent formulation of predictive coding with dendritic error computation

Although the idea of error units is undeniably elegant, it is not the only way to compute prediction errors in a neural circuit. More recent models showed that error computation can also be performed in the voltage dynamics of individual dendritic compartments [14–16] and, thus, without specialized error units. Combining these models allows for a reinterpretation of hPC, which we term dendritic hPC, where every **prediction neuron** will represent the two types of errors we discussed before in different sections of its dendritic tree (Figure 1B, see Box 2 for mathematical details):

Key figure

Implementation of predictive coding with dendritic error computation and spiking neurons

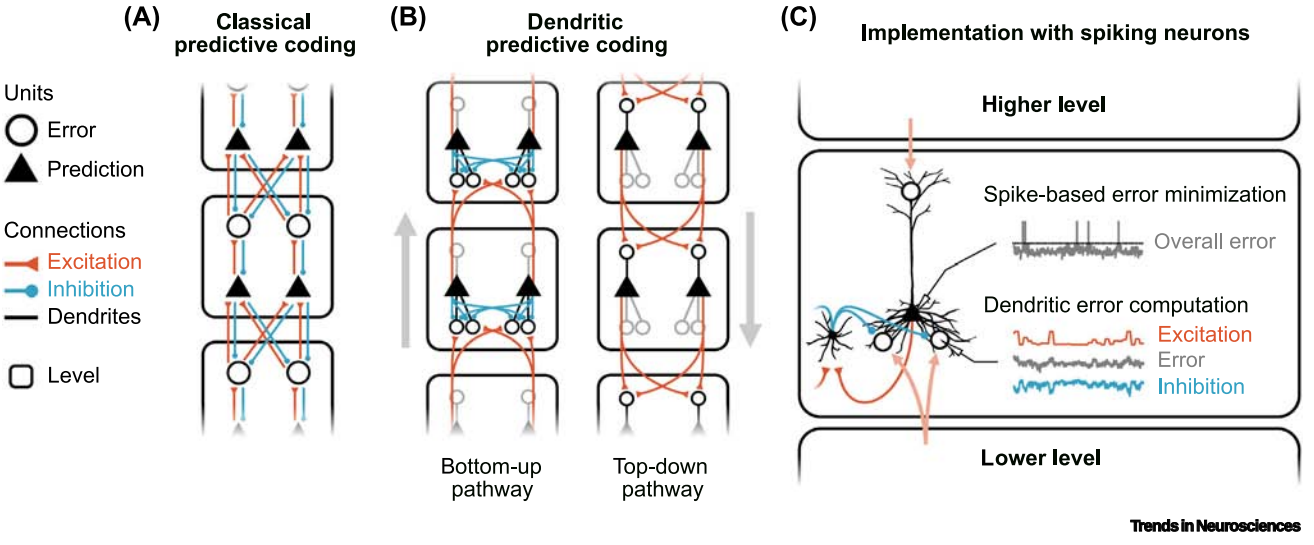

Figure 1. (A) Illustration of the classical model of hierarchical predictive coding (hPC). Errors and predictions are computed in different neural populations within one level of the hierarchy. Errors are sent up the hierarchy, while predictions are sent downwards. (B) In dendritic hPC, prediction neurons implement the same function, but errors are computed in neural dendrites. Predictions are sent up the hierarchy to basal dendrites, where they are balanced by lateral connections to compute bottom-up prediction errors (left). At the same time, predictions are sent down the hierarchy to apical dendrites, where they try to predict somatic spiking and guide the inference process (right). The pathways are shown separately for better visibility. (C) Dendritic hPC can be implemented with spiking neurons. The errors that are computed in the dendritic membrane potentials are integrated at the soma to form an overall error signal of the neuron's encoding. A spike is emitted when the somatic error potential grows too large and a spike would lead to a reduction in the overall error.

Box 2. Mathematical details of dendritic predictive coding

In dendritic hPC, the computation of errors in Equation VII is accomplished by the leaky voltage dynamics of dendritic compartments. Different models have explored this idea separately for basal dendrites [16,25,40] and apical dendrites (also with nonlinearities, which we here omit) [15,28], which we here combine to form a model that is equivalent to classical hPC. To this end, for each prediction neuron j, one introduces basal dendritic compartments $b^i_{jk} \approx D^i_{kj} e^{i-1}_k$, which are each innervated by a single synapse of a prediction neuron k of the previous level [16], as well as an apical compartment $a^i_j \approx -e^i_j$ that is innervated by prediction neurons of a higher level [15] (see Figure 1B in the main text). The error computation is then performed by voltage dynamics according to

$$\tau_b \frac{d}{dt} b^i_{jk} = -b^i_{jk} + D^i_{kj} r^{i-1}_k - \sum_l W^i_{jkl} r^i_l, \qquad [VIII]$$

$$\tau_a \frac{d}{dt} a^i_j = -a^i_j - r^i_j + \sum_l D^{i+1}_{jl} r^{i+1}_l, \qquad [IX]$$

where bottom-up inputs are balanced with lateral connections W^i_{jkl} (connection of neuron r^i_l to the kth dendritic compartment of neuron r^i_j), and top-down predictions are matched by the neurons own predictions r^i_j. The latter has been proposed to be implemented via the backpropagating action potential [15], solving the one-to-one connections problem of classical hPC [135]. To compute bottom-up errors, lateral weights have to be chosen as $W^i_{jkl} = D^i_{kj} D^i_{lj}$. Such weights can be found through a voltage-dependent plasticity rule, which enforces a tight balance in the kth dendritic compartment [16]

$$\eta_W^{-1} \frac{d}{dt} W^i_{jkl} = \frac{1}{\sigma^2_{i-1}} b^i_{jk} r^i_l. \qquad [X]$$

The dynamics of prediction neurons are then simply driven by the dendritic error potentials

$$\tau_r \frac{d}{dt} r^i_j = \frac{1}{\sigma^2_{i-1}} \sum_k b^i_{jk} + \frac{1}{\sigma^2_i} a^i_j, \qquad [XI]$$

and weights for bottom-up and top-down inputs can be learned with voltage-dependent rules (Equation XII proposed in [16], Equation XIII proposed in a generalized form in [15])

$$\eta_D^{-1} \frac{d}{dt} D^i_{kj} = \frac{1}{\sigma^2_{i-1}} \frac{1}{D^i_{kj}} b^i_{jk} r^i_j, \qquad [XII]$$

$$\eta_D^{-1} \frac{d}{dt} D^{i+1}_{jl} = -\frac{1}{\sigma^2_i} a^i_j r^{i+1}_l. \qquad [XIII]$$

Here, learning of bottom-up weights requires that lateral and bottom-up weights always align via $W^i_{jkl} = D^i_{kj} D^i_{lj}$, which in classical hPC is known as the weight transport problem [49,135]. For dendritic hPC a solution based on weight decay has been proposed in [16], which was demonstrated in a single-level model and is similar to a solution proposed for classical hPC [49]. Together, these equations yield an equivalent formulation of hPC for both learning and inference, where prediction errors are computed locally in dendritic compartments.

(i) bottom-up errors in basal dendritic compartments [16], where input from lower-level cortical areas is integrated [17];
(ii) the top-down prediction error (for the neuron's own activity) in an apical compartment [15], where higher-level cortical feedback arrives [17].

Besides the absence of error units, two additional central differences arise between the architectures of classical and dendritic hPC. First, in dendritic hPC both bottom-up and top-down signals are predictions, a possibility that has been discussed before [18]. Second, and more importantly, while prediction units in classical hPC inhibit each other through error units, prediction neurons in dendritic hPC directly compete through **lateral inhibition** on basal dendrites. Such networks with strong lateral inhibition (or similarly, winner-take-all-like dynamics [19]) have a long tradition in theoretical neuroscience, as models for the sparse and efficient encoding of sensory data [16,20–25] and as divisive normalization models of cortical computation [26,27]. Dendritic hPC

is closely related to these models, except that in these models it was not considered how exactly top-down connections could guide neural computations with predictions. In a more general context it has been proposed that top-down connections could provide these predictions by targeting apical dendrites [15,28–31]. Dendritic hPC combines these ideas of lateral competition and top-down predictions into a coherent theory of hierarchical inference in cortex.

Since in dendritic hPC error computation takes place in the voltage dynamics of basal and apical dendritic compartments, these local potentials play an important role for synaptic plasticity. For basal dendrites, dendritic hPC predicts that plastic lateral connections compute the errors for bottom-up inputs by establishing a **tight balance** locally in individual dendritic compartments (i.e., trying to closely match excitatory with inhibitory currents [32]). The intuitive explanation for this computation is that in a tightly balanced state, every input that can be predicted from other neurons is effectively canceled and the remaining unpredictable input constitutes the prediction error [14,16]. These errors can then be exploited by another voltage-dependent rule for bottom-up connections, in order to find an optimal encoding of inputs [16]. This learning rule is Hebbian-like (i.e., pairing postsynaptic firing with presynaptic input will induce potentiation of the synapse). At the same time, strong local inhibition during the postsynaptic spike would signal an over-prediction of the input and consequently should lead to long-term depression of the synapse. For apical dendrites, it has been proposed that error computation relies on the mismatch between apical prediction and somatic spiking [15]. In this theory of apical learning, plasticity of top-down connections is Hebbian-like as well, but synapses are depressed for a depolarization of the apical dendritic potential in the absence of somatic spiking. By employing these **voltage-dependent plasticity (VDP)** rules, dendritic hPC implements the same learning algorithm as classical hPC, but in prediction neurons with dendritic error computation (Box 2).

Dendritic error computation has also been used in a different context to implement the backpropagation algorithm in a cortical microcircuit [33–36]. Although this model of dendritic error backpropagation and dendritic hPC employ similar ideas, they ultimately pursue different goals and thus make distinct predictions for plasticity and **E-I balance** in basal and apical dendritic compartments (Figure 2).

Dendritic errors enable an efficient implementation of hPC with spiking neurons

Dendritic errors do not only yield an equivalent formulation of hPC, they also enable inference with spiking neurons. Here, the inferred variables have to be efficiently represented by spikes, which is

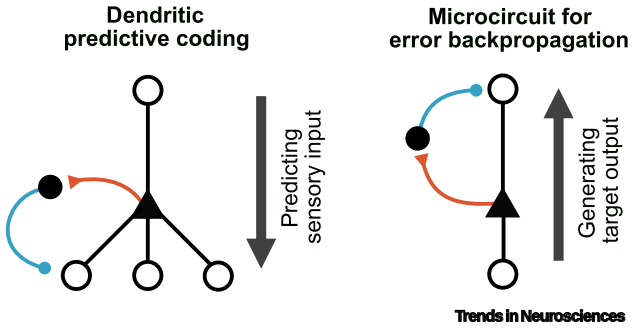

Figure 2. Relation of dendritic predictive coding to dendritic microcircuits for error backpropagation. (Left) In dendritic hierarchical predictive coding (hPC) the goal is to generate predictions of bottom-up sensory inputs. Here prediction errors are computed via balancing inhibition to basal dendrites and the mismatch of top-down predictions and somatic spiking at apical dendrites. (Right) In models that employ backpropagation the goal is to generate a target output at the highest level (e.g., a label) [33]. To this end an 'inverted' model of hPC is employed [35], where balancing inhibition at the apical dendrite is used to compute the backpropagated error of the output. While thorough testing of both theories remains to be conducted, a recent study indicates that pyramidal neurons learn predictive (and not balanced) apical activity [31], more consistent with dendritic hPC. However, this particular observation of course would not rule out that cortical networks could make use of both proposed mechanisms in different modes of operation or different neural populations.

Box 3. Mathematical details of dendritic predictive coding with spikes

Spike-based predictions of sensory data

A popular choice to mathematically formalize the prediction generated by a spike at time t_{sp} is via spike traces $\kappa(t, t_{sp}) = \exp(-(t - t_{sp})/\tau)$ that decay exponentially with some time constant τ [16,40]. Predictions of a neuron then change upon a spike according to $r(t) \rightarrow r(t) + \kappa(t, t_{sp})$, which approximately corresponds to the way spikes are read out in the membranes of postsynaptic neurons. With these predictions $r(t)$, the same formalism as before can be used to compute the instantaneous log-likelihood (see Box 1 in the main text):

$$\mathcal{L}(t) = \sum_{i=1}^{N} \log p_\theta\left(\mathbf{r}^{i-1}(t) | \mathbf{r}^i(t)\right). \quad \text{[XIV]}$$

However, due to the discontinuous nature of spikes, inference can no longer be implemented by simple gradient ascent.

Efficient spiking implementation of predictive coding with dendritic errors

One straightforward approach to implement inference with spikes is to deterministically fire a spike at time t if it instantly improves bottom-up and top-down errors, that is, the log-likelihood $\mathcal{L}(t)$ [40]:

$$\mathcal{L}(t|\text{neuron } j \text{ spikes at time } t) > \mathcal{L}(t|\text{no spike at time } t). \quad \text{[XV]}$$

This can be seen as a discrete implementation of gradient ascent to find the instantaneous maximum a posteriori (MAP) estimate for predictions r_j^i. From this principle it can be derived that a neuron should spike if its balanced membrane potential $u_j^i(t)$ surpasses a firing threshold T_j [40], that is, if

$$u_j^i(t) = \frac{1}{\sigma_{i-1}^2} \sum_k b_{jk}^i + \frac{1}{\sigma_i^2} a_j^i > T_j. \quad \text{[XVI]}$$

This equation is analogous to Equation XI, where $b_{jk}^i(t)$ are the balanced dendritic potentials of basal dendrites and $a_j^i(t)$ the potential of the apical dendrite.

Predictive coding with neural sampling

A more general approach to inference with spikes is to sample a (binary) spike train $\mathbf{S}_{0:T} = \{\mathbf{s}^i(t) | i \in \{1, \ldots, N\}, t \in \{0, \ldots, T\}\}$ from the posterior distribution of the generative model $\mathbf{S}_{0:T} \sim p_\theta(\mathbf{S}_{0:T} | \mathbf{r}_{0:T}^0)$ [16,136]. The posterior is implicitly defined via the model $p_\theta(\mathbf{r}^{i-1}(t) | \mathbf{r}^i(t))$, a prior on spiking $p_\theta(\mathbf{s}^N(t))$ and spike traces $\mathbf{r}^i(t) = \sum_{t'=0}^{t} \mathbf{s}^i(t') \kappa(t, t')$. While computing the posterior distribution exactly is intractable [16,136], approximate online sampling can be implemented with the same membrane potentials $u_j^i(t)$ and threshold T_j as before (up to a constant factor) and a soft spiking threshold mechanism

$$p(\text{neuron } j \text{ spikes at time } t) = \text{sig}\left(u_j^i(t) - T_j\right), \quad \text{[XVII]}$$

where $\text{sig}(x) = 1/(1 + \exp(-x))$ is the logistic function [16]. Note, that $u_j^i(t)$ and T_j are scaled by the precisions of errors $\frac{1}{\sigma_i^2}$ (Equation XVI) and thus the stochasticity of spiking will capture the uncertainty in inference. This model is a special case of the spike response model with escape noise [137] and can be implemented by a leaky-integrate-and-fire neuron with a noisy membrane potential. Equations XI, XVI, and XVII highlight the intimate relation that exists between the theories of hPC, efficient coding with spikes, and neural sampling.

possible if spikes are only fired if they reduce the overall prediction error [14,37,38] (see Box 3 for the mathematical details of dendritic hPC with spiking neurons). Since in dendritic hPC prediction errors are represented in the balanced membrane potentials, an efficient spike encoding can be found with a simple threshold mechanism that generates a spike when the error potential grows too large (Figure 1C), as demonstrated in single-level models [39,40]. Predictive coding thus serves a dual purpose in dendritic hPC, by enabling both inference in a hierarchical model and an efficient spike encoding of dynamical variables.

A central role in this inference scheme with spikes is played by noise in the neural dynamics, for two reasons. First, noise enables an efficient spike encoding in the face of transmission delays. With deterministic neurons, even a small delay of inhibition can lead to erratic network behavior, since inhibition will often arrive too late to prevent synchronous spiking of large parts of a

population [41]. Noise relaxes this constraint on the speed of feedback, since it effectively decouples and desynchronizes neural spiking [37,41,42]. Second, noise in spiking neural networks enables neural sampling [2,43–46]. Here, the idea is that neural activity samples possible predictions according to their likelihood, instead of computing a single best estimate as in classical hPC (Box 3). Neural sampling therefore is a principled way to represent uncertainty in inference via neural activity and has, for example, been used to explain variability in neural responses [47,139] and the origin of multistability in perception [48]. Recent models show that neural sampling and efficient spike coding with tight E-I balance can be combined in a single model with dendritic error computation [16,43], relating these concepts to the proposed model of dendritic hPC (Box 3).

In addition to neural inference, dendritic errors also enable learning in populations of spiking neurons. This is not straightforward, since the switch from rate-based to spike-based models typically requires a modification of the learning algorithms. For example, when using spiking error units, as in classical hPC, it is not directly possible to represent both positive and negative errors by non-negative activity [49]. To resolve this, it was proposed that errors are represented by deviations relative to a baseline firing rate [49], but this would require high firing rates and therefore seems implausible considering the low firing rates in neocortex [50]. An alternative is to represent positive and negative errors in separate populations [11,50], but it is unclear how in this case biological plasticity can recombine the positive and negative parts, which are both required for the learning of single synapses. Due to these difficulties, to date, no complete implementation of hPC that uses spiking error units has been proposed [13]. By contrast, in dendritic hPC the same learning algorithm as for rate-based units can be straightforwardly applied to spiking neurons. The reason is that dendritic membrane potentials remain continuous quantities, despite the spiking nature of neural activity, and thus can easily represent the prediction errors that are required for the learning of bottom-up and top-down connections (Box 2), which has been successfully applied in [15,16].

Is dendritic predictive coding biologically plausible?

In the previous section we have introduced the two main assumptions of dendritic hPC, which are: (i) cortex implements inference in a hierarchical probabilistic model, and (ii) errors of the resulting predictions are computed in the local voltage dynamics of basal and apical dendrites. The implications of the first assumption have been discussed at length in the context of classical hPC and were found to align well with experimental observations [7,10,51]. In the following we will discuss the biological plausibility of the second assumption. Ultimately, we will argue that dendritic hPC can indeed be closely connected to many properties of pyramidal neurons and inhibitory connectivity in cortex.

Dendritic error computation and synaptic plasticity in pyramidal neurons

To compute errors in basal dendrites, a tight and local E-I balance is required. Indeed, it has been found in several instances that inhibitory and excitatory currents are tightly correlated, with inhibition trailing excitation by few milliseconds [14,52,53]. This tight balance leaves neurons only with a brief window of opportunity for spiking, which effectively decorrelates neural responses to inputs and thereby ensures an efficient neural code [25]. A tight E-I balance can therefore explain the origin of the irregular spiking patterns of neurons that have been observed throughout cortex [14,54]. Although models with a tight balance can reproduce irregular firing on the single neuron level, incorporating realistic synaptic transmission delays in these models can lead to oscillations on the population level [37]. Oscillations in cortical activity in the gamma frequency band have therefore been discussed as signatures of efficient coding in balanced networks [42] (and might also support efficient neural sampling [45,55]).

Consistent with dendritic hPC, this balance has also been found to extend to individual dendritic compartments [32,56,57]. Crucially, this local balance can be observed down to the scale of

(at least) single dendritic branches [56], since the attenuation of dendritic currents prevents that inhibitory postsynaptic potentials spread into other dendritic branches and influence the E-I balance there [58,59]. Experiments could also show that this local balance is maintained through localized synaptic plasticity, which re-establishes the balance after a perturbation and coordinates excitatory and inhibitory plasticity locally [56,60–65]. Overall, these findings are compatible with the idea that a local balance can compute prediction errors for specific synaptic contacts at basal dendrites.

Another prediction of dendritic hPC, which has been consistently observed in a range of experiments, is that the local membrane potential is a central determinant of synaptic plasticity [61,65–68]. This VDP is thought to be mainly mediated by the local calcium concentration, which follows the local membrane potential and modulates synaptic plasticity [59,69,70]. Based on these observations, VDP rules have been proposed that can reproduce several experiments of spike-timing-dependent plasticity in a unified picture [71–73]. An especially important consequence of locally organized VDP, which is also required by dendritic hPC, is that inhibition can strongly modulate synaptic plasticity in a very localized manner [32,65,74–76].

Are the VDP rules that can be derived from dendritic hPC consistent with these experimentally observed VDP rules? A distinction has to be made here between VDP rules in basal dendrites, which should enable the learning of neural representations [16], and VDP in apical dendrites, which should enable the prediction of somatic spiking [15]. For representation learning in basal dendrites, we have argued in [16] that previously proposed VDP rules [71,72] can be reconciled with the VDP rules derived from dendritic hPC. One prediction of these derived VDP rules is that strong local inhibition should promote the depression of excitatory synapses, an effect that has been observed in proximal dendrites of hippocampal pyramidal neurons [75] (similarly found in [77]). By contrast, for the learning of apical connections, an explicit correspondence to experimental VDP still has to be proposed. Experiments show that synaptic plasticity close to and far from the soma behaves vastly differently [31,78–80], which could support the different requirements for basal and apical synaptic plasticity in dendritic hPC. While more experimental and theoretical work is needed to clarify the connections between dendritic hPC and experimental VDP, these results suggest that cortical pyramidal neurons in principle are suited to implement the learning algorithm proposed by dendritic hPC.

A diversity of inhibitory interneurons is required for dendritic predictive coding

Since pyramidal neurons in general only excite other cells, additional inhibitory interneurons are required to implement the dendritic hPC model. The central inhibitory motif of dendritic hPC requires interneurons that balance bottom-up inputs to basal dendrites via lateral connections [16,25]. These interneurons show strong similarities to parvalbumin-expressing (PV) interneurons in cortex, which implement a precisely adjusted competition between pyramidal neurons [24,81–84]. PV positive, fast-spiking basket cells alone make up around 30–50% of all interneurons in the cortical microcircuit [85] and are especially adapted to tightly control pyramidal neuron spiking and the cortical E-I balance via very fast inhibition to somata and basal dendrites [86,87]. PV interneurons also seem to be responsible for the gamma oscillations that similarly arise through lateral inhibition in dendritic hPC [87–89]. Dendritic hPC is therefore closely linked to one of the defining inhibitory motifs of cortex.

Next to PV interneurons, most other interneurons in cortex can be classified as either somatostatin-expressing (SST) interneurons, which preferentially target the apical dendrites of pyramidal neurons, or vasoactive intestinal peptide-expressing (VIP) interneurons, which mainly inhibit other interneurons, especially SST [86,90]. SST and VIP interneurons, for example, have been observed to be responsible for top-down inhibitory control [91], which is also required in

dendritic hPC when top-down input predicts a decrease in activity. However, not all of the major connectivity patterns of SST and VIP cells can be straightforwardly explained by dendritic hPC: SST interneurons, for example, also mediate short-range lateral inhibition to apical dendrites, which allows them to contribute to surround suppression [92] and to gate top-down input [93,94]. The disinhibitory circuit of VIP contributes to this gating mechanism by specifically suppressing SST neurons during active behavior [93,95,96]. SST and VIP neurons have also been found to be crucial for gating apical plasticity, for example, during reward-based learning [97–99]. These connectivity motifs thus play a central role in how predictions are processed by apical dendrites, but precisely what functions they could implement, especially in the context of dendritic hPC, has yet to be understood [65].

Dendritic predictive coding in neocortical lamination

Neocortex employs multiple types of pyramidal neurons that reside on different cortical layers and exhibit specific connectivity [17]. We here propose that dendritic hPC in particular describes the computations of layer 2/3 neurons (Figure 3D). That layer 2/3 neurons are central in the hierarchical integration of information and the interpretation of sensory data has been proposed before, for example, based on cortical physiology [19] or in theories of classical hPC, where errors and predictions are first computed in layer 2/3 [6] (Figure 3C). There are several arguments for why dendritic hPC is particularly well suited to describe layer 2/3: first, like in dendritic hPC, layer 2/3 neurons combine bottom-up signals (sent from layer 4 to their basal dendrites) with top-down

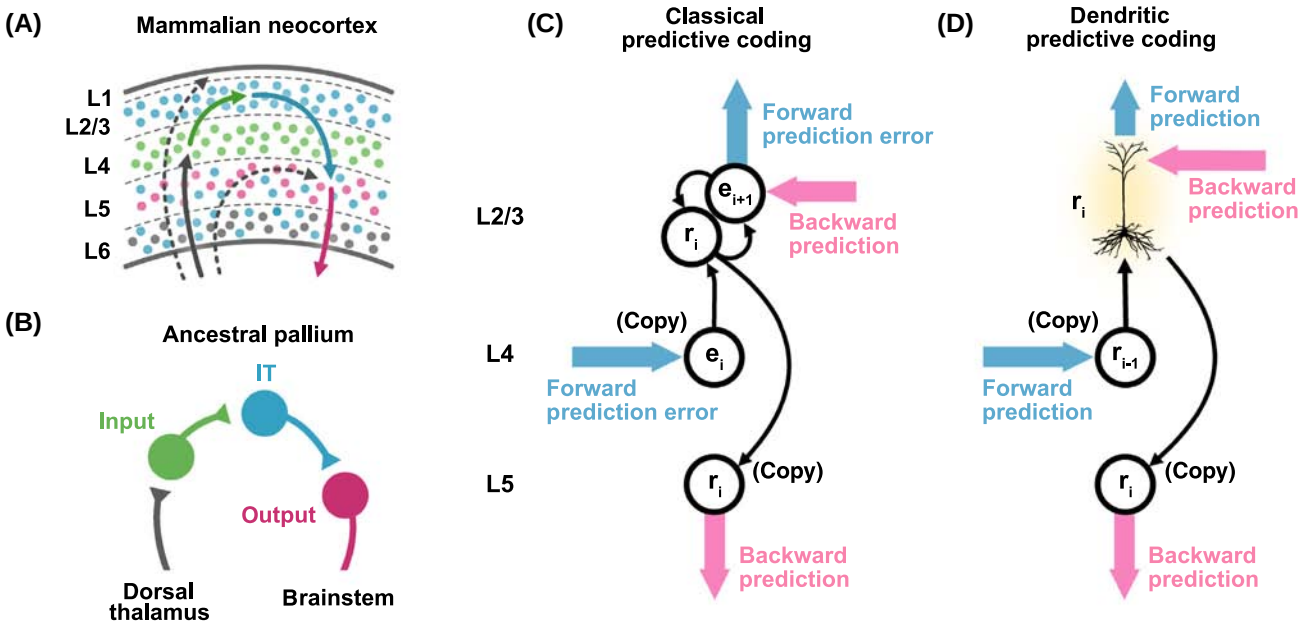

Figure 3. How could dendritic predictive coding be embedded into neocortical microcircuits and lamination? (A) Core circuitry of mammalian neocortex, as shown in [102,104]. Input neurons in layer 4 (green) receive sensory information from the dorsal thalamus, layer 2/3 intratelencephalic (IT) neurons (blue) further process this information, and output neurons in layer 5 (red) project to the brainstem and other areas. Additional connections, for example, from thalamus to layer 1 (mostly relayed from other cortical areas [94]) or layer 5 (broken lines), or within layer 2/3 between areas exist [17,138], but will be omitted in the following for simplicity. (B) Theories of cortical evolution hypothesize that these input, IT, and output cells are homologous to cells that existed in the ancestral amniote pallium [104]. Also, in birds and non-avian reptiles, homologous cell types exist, but are organized in architectures that differ from the laminar organization of mammalian neocortex. (C) The predictive coding microcircuit as proposed by [6] (here presented in a simplified form) follows the organization of the neocortical microcircuit. Predictions (r_i) and prediction errors (e_i) are computed in layer 2/3. Deeper layers mainly act as communication hubs by copying signals from layer 2/3. (D) Speculative microcircuit for dendritic predictive coding. Here, deeper layers fulfill the same role as communication hubs (and possibly complementary functions [19]), but layer 2/3 only computes predictions.

signals (sent from layer 5 or layer 2/3 to their apical dendrites) [6,17,19]. Second, layer 2/3 neurons exhibit sparse activity, which is mainly enforced by lateral inhibition via PV interneurons [83,84,100], a motif that is present in dendritic hPC but not in other theories of hPC [6,35]. Last, superficial cortical layers show pronounced gamma oscillations [6,88,89] that are expected to arise through lateral inhibition in dendritic hPC [37,42].

Importantly, these properties implied by dendritic hPC are not general features of pyramidal neurons, which in other layers likely implement different functions. Layer 5 neurons, for example, employ a dense and not a sparse code [100] and show less gamma oscillations [6,89]. These properties, together with the position of layer 5 neurons as downstream elements in the microcircuit [17], have led to the suggestion that layer 5 might be employed in long-range communication [100] and output selection [19]. Layer 4 in turn shows an abundance of PV interneurons [86] and could implement a preprocessing of bottom-up inputs [17]. These different roles of deeper layers are also in line with theories of cortical evolution, which hypothesize that deeper layers have migrated from previously separate 'input' and 'output' neural populations to neocortex in order to integrate cortical neurons more deeply with the rest of the brain and other cortical areas [101–104] (Figure 3A,B). Hence, the different functions of deeper layers could complement the computations of dendritic hPC in important ways, but how exactly such an interaction could look has yet to be formulated.

Another aspect of cortical lamination that could support the computations of dendritic hPC are neuromodulators. Neuromodulators act on a wide range of scales [105] and can target specific cortical layers, where they might modulate computations in specific dendritic domains of pyramidal cells [94,106–108]. For example, acetylcholine (ACh), which is associated with attention and learning, has been found to promote (dis-)inhibition of apical or basal dendrites through distinct mechanisms, possibly in a very targeted manner [96,99,107–109]. In the context of hPC, ACh and other neuromodulators have been proposed to set the precisions of the internal model and thereby determine the influence of sensory and top-down information [110–112]. The separation of top-down and bottom-up inputs across cortical layers, as in dendritic hPC, could therefore be a central factor to enable the targeted modulation of these pathways. This might not only apply to the effects of ACh on neural gain, but also to the various other effects ACh and other neuromodulators have on cortical dynamics and plasticity [105].

How can error responses arise in prediction neurons?

One of the central features of classical hPC is its ability to explain a variety of experimental observations through the concept of error neurons. Error neurons have, for example, been used to explain extra-classical receptive field effects in visual cortex [11], as well as mismatch responses in cortex, which are neural responses that appear to signal the mismatch between an internal model and sensory data [10]. Thus, an important question for dendritic hPC is if and how these experimental observations can arise in a model without error neurons.

The first experimental observation that has been explained with error neurons in hPC is the extra-classical receptive field effect of endstopping [11]. In endstopping it is found that, first, the response of a neuron in V1 to a bar stimulus decreases when the bar extends over its receptive field, and second, this effect is reduced when feedback from higher-level areas is disabled [113,114]. Recent theoretical work showed that endstopping behavior, as well as other extra-classical receptive field effects, also occur in prediction neurons, where top-down connections strengthen these effects [7,115,116]. Here, endstopping is mainly mediated by lateral inhibition between neurons with overlapping receptive fields [116]. Top-down connections from higher-level areas predict the activity patterns that arise from these lateral interactions and enhance

them, which strengthens endstopping behavior [115]. This cooperation of lateral and top-down interactions could be important to help the network to cope with noise in the inputs and improve visual processing [115,117] and has been widely observed in visual cortex [114,117–119].

Mismatch responses have been observed in different forms, such as responses to the omission of expected stimuli [10], responses to a mismatch between information in different modalities (e.g., visual and motor information) [120–122], strong responses to unexpected stimuli [7,123], or suppressed responses to expected stimuli [1,124]. Omission responses can already occur in straightforward prediction neuron responses, as prediction neurons can be active even without the expected input [10]. Recent work from our group has also shown that multimodal mismatch responses can naturally arise in prediction neurons, when different cortical areas jointly infer a consistent explanation of sensory data [125]. This joint inference aims to find single causes that underlie stimuli in multiple modalities, meaning cortical areas should suppress predictable activity in other areas (as in [122,126]), but might also drive activity in case of a prediction mismatch (as in [120,121,127]). Strong/suppressed responses to unexpected/expected stimuli in turn have so far not been explained with pure prediction responses, but it has been argued that they might be mediated by other mechanisms, such as attention to interesting stimuli, the variance in neural sampling, or adaptation mechanisms [7,124,128]. In conclusion, the observed mismatch responses can be explained by a variety of plausible mechanisms in prediction neurons, which, however, in some cases might not be directly relatable to the computations of dendritic hPC.

Testable predictions

To better assess the potential as well as the limitations of dendritic hPC to describe inference in cortex, we here propose experiments that: (i) test predictions for specific neural mechanisms, and (ii) aim to distinguish between the different implementations of hPC with and without error neurons.

Predictions for specific neural mechanisms

- Bottom-up excitation to basal dendrites of layer 2/3 pyramidal cells should be locally matched and balanced with lateral inhibition, likely via PV interneurons (an indication that such a precise matching is possible, e.g., in dendritic spines, has been found in [129]). This could be tested in detail, for example, using large-scale connectomics datasets [130].
- Plasticity for excitatory bottom-up connections is predicted to be modulated by local inhibitory input, which is expected to turn long-term potentiation into depression. While such modulation of plasticity has been found (e.g., in hippocampal neurons in a spike-timing-dependent plasticity experiment [56]), it would be interesting to test this more specifically in layer 2/3 basal dendrites, with a particular focus on the predicted impact of the strength and timing of inhibition on plasticity [16].
- Similar experiments could be conducted for top-down connections to apical dendrites, where plasticity should be Hebbian, but switch to depression when presynpaptic spikes depolarize the dendrite while the neuron remains silent. Also, here it would be interesting to explicitly test for the predicted dependence of plasticity on the dendritic membrane potential [15].
- As a consequence of these plasticity mechanisms, activity in basal dendrites is expected to be decreased (‘explained away’) in the course of learning, whereas activity in apical dendrites should increase and become predictive of somatic spiking (similar to what was found in [31]). An important experiment would be to test explicitly if apical activity indeed becomes predictive on a single neuron level, which would also distinguish dendritic hPC from theories of dendritic error backpropagation that predict a clear decrease of apical activity (Figure 2).

Distinguishing between hPC with and without error neurons

The central challenge in distinguishing between different implementations of hPC is that their underlying mathematical framework is the same, hence they predict the same computations in prediction neurons. Thus, since classical hPC as yet does not make clear predictions on the single neuron level, the main distinguishing characteristic between classical and dendritic hPC is the presence or absence of error units. For specific computations, this might be used to rule out one of the models:

- As we discussed, mismatch responses are explained via distinct mechanisms in models with or without error neurons, which could be tested on a case-by-case basis. For example, mismatch responses in multimodal mismatch experiments are transient [131], where classical hPC predicts this decrease to be caused by top-down inhibition, while in dendritic hPC one would expect the origin in adaptation or other bottom-up mechanisms [125] (for other experiments, see also discussion in [7,124]).
- Another, more direct approach would be to map out the functional circuits in cortex, where classical hPC expects a clear separation between error and prediction units (i.e., error units only receive predictions and vice versa), but dendritic hPC expects no such separation. For example, in several experiments reporting 'error' and 'prediction' neurons, their populations appear intermixed [123,132] and it would be important to clarify whether or not there exists a clear feedforward–feedback circuit motif between these populations (e.g., if bottom-up excitation and inhibition always arrives first in one of the populations).

For these experiments it is important to note that dedicated error neurons (or even classical hPC) might coexist with dendritic hPC for complementary computations. For example, it is well known that dopaminergic neurons code for reward prediction errors to guide behavioral learning [133]. However, it is unclear whether there exists an advantage to implement the same computation, such as inference in sensory cortex, simultaneously with two different implementations of hPC.

Concluding remarks

Since its conception over 20 years ago, hPC has been considered one of the most promising unifying theories of cortical computation, but – in its classical form – it is still facing substantial questions regarding its biological plausibility. Here, we outlined an emerging hPC scheme based on dendritic error computation, which is functionally equivalent, but provides solutions to the most pressing open problems of the established theory of classical hPC: first, it can explain the lack of clear empirical evidence for the coexistence of error and prediction neurons [10,51], and second, it overcomes the unresolved question of how learning can be efficiently implemented with spiking error neurons [13]. Moreover, we explained how dendritic hPC could connect the microscopic properties of neural dendrites, such as the local E-I balance [14,32,57] and VDP [72,75], to neural dynamics [14] and learning [15,16,115] in the cortical hierarchy.

These advances open up several interesting paths for future research. Next to experimentally testing for the predicted mechanisms of inference and learning in cortex (see section 'Testable predictions'), there are a number of open theoretical challenges, especially concerning the details of the biological implementation (see Outstanding questions). Going forward, it will also be important to understand how the learning of a hierarchical model of sensory data interacts with complementary mechanisms, such as attention and behavioral learning, not only for dendritic hPC, but also for hPC and other theories of inference in cortex more generally.

Outstanding questions

Dendritic hPC has been derived under the assumption of linear dendrites for a linear encoding of sensory data, but dendrites often show nonlinear behavior. How can the ideas of dendritic hPC be transported to a model with nonlinear dendrites and could this allow for a nonlinear and thus more versatile encoding?

Pyramidal cells show extensive lateral excitatory connectivity, which could be used to learn and predict temporal sequences within a single level. Can these mechanisms interact purposefully with the learning of predictions in a hierarchical model?

When cortical areas communicate there might be substantial challenges, such as long transmission delays or sparse activity in both areas. Are there additional mechanisms that could improve neural communication under these conditions, such as communication through coherence, and how could they be integrated into dendritic hPC?

Pyramidal cells are not a uniform class of cells, for example, the different physiology of layer 2/3 and layer 5 apical dendrites leads to different integration of top-down inputs, but also layers 2 and 3 contain slightly different subtypes of pyramidal cells. What are the functional reasons for these properties and how are they related to dendritic hPC?

We have suggested that dendritic hPC describes the computations of layer 2/3 pyramidal neurons. Under this assumption, what are the roles of deeper cortical layers and how can they be integrated into the framework?

Inference has not only been used to model sensory processing, but also computations in hippocampus, and some of the core predictions of dendritic hPC also seem to apply to hippocampal pyramidal cells. Are principles of dendritic hPC also employed by different brain regions, or different neuron types?

Often indirect measures of neural activity (e.g., electroencephalography, fMRI) have been used to search for evidence of classical hPC. How would

CellPress
OPEN ACCESS

error computation in specialized neurons and in dendrites differ in these measures?

Acknowledgments

We would like to thank Abdullah Makkeh, Beatriz Belbut, Caspar Schwiedrzik, David Ehrlich, Georg Keller, and members of the Priesemann Lab, especially Andreas Schneider, Kjartan van Driel, and Matthias Loidolt, for helpful discussions and comments on the manuscript. F.A.M. and L.R. were funded by the German Research Foundation (DFG), SFB 1286. V.P. and M.W. received support from the German Research Foundation (DFG), SFB 1528, Cognition of Interaction. F.A.M., L.R., and V.P. acknowledge support by the Max Planck Society.

Declaration of interests

The authors declare no competing interests in relation to this work.

References

1. De Lange, F.P. *et al.* (2018) How do expectations shape perception? *Trends Cogn. Sci.* 22, 764–779
2. Fiser, J. *et al.* (2010) Statistically optimal perception and learning: from behavior to neural representations. *Trends Cogn. Sci.* 14, 119–130
3. Knill, D.C. and Pouget, A. (2004) The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends Neurosci.* 27, 712–719
4. Rahnev, D. and Denison, R.N. (2018) Suboptimality in perceptual decision making. *Behav. Brain Sci.* 41
5. Lee, T.S. and Mumford, D. (2003) Hierarchical Bayesian inference in the visual cortex. *JOSA A* 20, 1434–1448
6. Bastos, A.M. *et al.* (2012) Canonical microcircuits for predictive coding. *Neuron* 76, 695–711.1
7. Aitchison, L. and Lengyel, M. (2017) With or without you: predictive coding and Bayesian inference in the brain. *Curr. Opin. Neurobiol.* 46, 219–227.1
8. Heeger, D.J. (2017) Theory of cortical function. *Proc. Natl. Acad. Sci.* 114, 1773–1782
9. Cao, Y. *et al.* (2019) Causal inference in the multisensory brain. *Neuron* 102, 1076–1087
10. Walsh, K.S. *et al.* (2020) Evaluating the neurophysiological evidence for predictive processing as a model of perception. *Ann. N. Y. Acad. Sci.* 1464, 242
11. Rao, R.P.N. and Ballard, D.H. (1999) Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* 2, 79–87.1
12. Kogo, N. and Trengove, C. (2015) Is predictive coding theory articulated enough to be testable? *Front. Comput. Neurosci.* 9, 111
13. Millidge, B. *et al.* (2021) Predictive coding: a theoretical and experimental review. *arXiv* https://doi.org/10.48550/arXiv.2107.12979
14. Denève, S. and Machens, C.K. (2016) Efficient codes and balanced networks. *Nat. Neurosci.* 19, 375–382.X
15. Urbanczik, R. and Senn, W. (2014) Learning by the dendritic prediction of somatic spiking. *Neuron* 81, 521–528
16. Mikulasch, F.A. *et al.* (2021) Local dendritic balance enables learning of efficient representations in networks of spiking neurons. *Proc. Natl. Acad. Sci.* 118, e2021925118
17. Harris, K.D. and Shepherd, G.M.G. (2015) The neocortical circuit: themes and variations. *Nat. Neurosci.* 18, 170–181
18. Spratling, M.W. (2008) Predictive coding as a model of biased competition in visual attention. *Vis. Res.* 48, 1391–1408.1
19. Douglas, R.J. and Martin, K.A.C. (2004) Neuronal circuits of the neocortex. *Annu. Rev. Neurosci.* 27, 419–451.X
20. Földiak, P. (1990) Forming sparse representations by local anti-Hebbian learning. *Biol. Cybern.* 64, 165–170
21. Olshausen, B.A. and Field, D.J. (1996) Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 381, 607–609.1
22. Boerlin, M. *et al.* (2013) Predictive coding of dynamical variables in balanced spiking networks. *PLoS Comput. Biol.* 9, e1003258.1
23. Bill, J. *et al.* (2015) Distributed Bayesian computation and self-organized learning in sheets of spiking neurons with local lateral inhibition. *PLoS One* 10, e0134356
24. Chettih, S.N. and Harvey, C.D. (2019) Single-neuron perturbations reveal feature-specific competition in V1. *Nature* 567, 334–340
25. Brendel, W. *et al.* (2020) Learning to represent signals spike by spike. *PLoS Comput. Biol.* 16, e1007692
26. Carandini, M. and Heeger, D.J. (2012) Normalization as a canonical neural computation. *Nat. Rev. Neurosci.* 13, 51–62
27. Burg, M.F. *et al.* (2021) Learning divisive normalization in primary visual cortex. *PLoS Comput. Biol.* 17, e1009028
28. Brea, J. *et al.* (2016) Prospective coding by spiking neurons. *PLoS Comput. Biol.* 12, e1005003
29. Larkum, M. (2013) A cellular mechanism for cortical associations: an organizing principle for the cerebral cortex. *Trends Neurosci.* 36, 141–151
30. Aru, J. *et al.* (2020) Cellular mechanisms of conscious processing. *Trends Cogn. Sci.* 24, 814–825
31. Gillon, C.J. *et al.* (2021) Learning from unexpected events in the neocortical microcircuit. *bioRxiv* https://doi.org/10.1101/2021.01.15.426915
32. Hennequin, G. *et al.* (2017) Inhibitory plasticity: balance, control, and codependence. *Annu. Rev. Neurosci.* 40, 557–579
33. Sacramento, J. *et al.* (2018) Dendritic cortical microcircuits approximate the backpropagation algorithm. In *Advances in Neural Information Processing Systems*, pp. 8721–8732
34. Richards, B.A. and Lillicrap, T.P. (2019) Dendritic solutions to the credit assignment problem. *Curr. Opin. Neurobiol.* 54, 28–36
35. Whittington, J.C.R. and Bogacz, R. (2019) Theories of error back-propagation in the brain. *Trends Cogn. Sci.* 23, 235–250
36. Haider, P. *et al.* (2021) Latent equilibrium: a unified learning theory for arbitrarily fast computation with arbitrarily slow neurons. *Adv. Neural Inf. Proces. Syst.* 34, 17839–17851
37. Kadmon, J. *et al.* (2020) Predictive coding in balanced neural networks with noise, chaos and delays. *Adv. Neural Inf. Proces. Syst.* 33
38. Yoon, Y.C. (2016) LIF and simplified SRM neurons encode signals into spikes via a form of asynchronous pulse sigma–delta modulation. *IEEE Trans. Neural Netw. Learn. Syst.* 28, 1192–1205
39. Mancoo, A. *et al.* (2020) Understanding spiking networks through convex optimization. *Adv. Neural Inf. Proces. Syst.* 33, 8824–8835
40. Boerlin, M. and Denève, S. (2011) Spike-based population coding and working memory. *PLoS Comput. Biol.* 7, e1001080
41. Rullán Buxó, C.E. and Pillow, J.W. (2020) Poisson balanced spiking networks. *PLoS Comput. Biol.* 16, e1008261
42. Chalk, M. *et al.* (2016) Neural oscillations as a signature of efficient coding in the presence of synaptic delays. *Elife* 5, e13824
43. Savin, C. and Deneve, S. (2014) Spatio-temporal representations of uncertainty in spiking neural networks. *Adv. Neural Inf. Proces. Syst.* 27, 2024–2032
44. Buesing, L. *et al.* (2011) Neural dynamics as sampling: a model for stochastic computation in recurrent networks of spiking neurons. *PLoS Comput. Biol.* 7, e1002211
45. Aitchison, L. and Lengyel, M. (2016) The Hamiltonian brain: efficient probabilistic inference with excitatory-inhibitory neural circuit dynamics. *PLoS Comput. Biol.* 12, e1005186
46. Petrovici, M.A. *et al.* (2016) Stochastic inference with spiking neurons in the high-conductance state. *Phys. Rev. E* 94, 042312
47. Orbán, G. *et al.* (2016) Neural variability and sampling-based probabilistic representations in the visual cortex. *Neuron* 92, 530–543
48. Gershman, S.J. *et al.* (2012) Multistability and perceptual inference. *Neural Comput.* 24, 1–24

49. Alonso, N. and Neftci, E. (2021) Tightening the biological constraints on gradient-based predictive coding. In *International Conference on Neuromor- phic Systems 2021*, pp. 1–9
50. Keller, G.B. and Mrsic-Flogel, T.D. (2018) Predictive processing: a canonical cortical computation. *Neuron* 100, 424–435.X
51. Heilbron, M. and Chait, M. (2018) Great expectations: is there evidence for predictive coding in auditory cortex? *Neuroscience* 389, 54–73.X
52. Wehr, M. and Zador, A.M. (2003) Balanced inhibition underlies tuning and sharpens spike timing in auditory cortex. *Nature* 426, 442–446
53. Okun, M. and Lampl, I. (2008) Instantaneous correlation of excitation and inhibition during ongoing and sensory-evoked activities. *Nat. Neurosci.* 11, 535–537
54. Vogels, T.P. *et al.* (2011) Inhibitory plasticity balances excitation and inhibition in sensory pathways and memory networks. *Science* 334, 1569–1573.1
55. Korcsak-Gorzo, A. *et al.* (2022) Cortical oscillations support sampling-based computations in spiking neural networks. *PLoS Comput. Biol.* 18, e1009753
56. Liu, G. (2004) Local structural balance and functional interaction of excitatory and inhibitory synapses in hippocampal dendrites. *Nat. Neurosci.* 7, 373–379.1
57. Iascone, D.M. *et al.* (2020) Whole-neuron synaptic mapping reveals spatially precise excitatory/inhibitory balance limiting dendritic and somatic spiking. *Neuron* 106, 566–578
58. Spruston, N. *et al.* (2016) Principles of dendritic integration. *Dendrites* 351, 361–364
59. Müllner, F.E. *et al.* (2015) Precision of inhibition: dendritic inhibition by individual GABAergic synapses on hippocampal pyramidal cells is confined in space and time. *Neuron* 87, 576–589
60. Field, R.E. *et al.* (2020) Heterosynaptic plasticity determines the set point for cortical excitatory-inhibitory balance. *Neuron* 106, 842–854
61. Chen, S.X. *et al.* (2015) Subtype-specific plasticity of inhibitory circuits in motor cortex during motor learning. *Nat. Neurosci.* 18, 1109–1115
62. Hu, H.Y. *et al.* (2019) Endocannabinoid signaling mediates local dendritic coordination between excitatory and inhibitory synapses. *Cell Rep.* 27, 666–675
63. Bourne, J.N. and Harris, K.M. (2011) Coordination of size and number of excitatory and inhibitory synapses results in a balanced structural plasticity along mature hippocampal CA1 dendrites during LTP. *Hippocampus* 21, 354–373.1
64. D'amour, J.A. and Froemke, R.C. (2015) Inhibitory and excitatory spiketiming-dependent plasticity in the auditory cortex. *Neuron* 86, 514–528
65. Herstel, L.J. and Wierenga, C.J. (2021) Network control through coordinated inhibition. *Curr. Opin. Neurobiol.* 67, 34–41
66. Artola, A. *et al.* (1990) Different voltage-dependent thresholds for inducing long-term depression and long-term potentiation in slices of rat visual cortex. *Nature* 347, 69–72
67. Lisman, J. and Spruston, N. (2005) Postsynaptic depolarization requirements for LTP and LTD: a critique of spike timing-dependent plasticity. *Nat. Neurosci.* 8, 839–841
68. Lisman, J. and Spruston, N. (2010) Questions about STDP as a general model of synaptic plasticity. *Front. Synaptic Neurosci.* 2, 140
69. Higley, M.J. (2014) Localized GABAergic inhibition of dendritic Ca 2+ signalling. *Nat. Rev. Neurosci.* 15, 567–572
70. Augustine, G.J. *et al.* (2003) Local calcium signaling in neurons. *Neuron* 40, 331–346
71. Shouval, H.Z. *et al.* (2002) A unified model of NMDA receptor-dependent bidirectional synaptic plasticity. *Proc. Natl. Acad. Sci.* 99, 10831–10836
72. Clopath, C. and Gerstner, W. (2010) Voltage and spike timing interact in STDP–a unified model. *Front. Synaptic Neurosci.* 2, 25
73. Clopath, C. *et al.* (2010) Connectivity reflects coding: a model of voltage-based STDP with homeostasis. *Nat. Neurosci.* 13, 344
74. Meredith, R.M. *et al.* (2003) Maturation of long-term potentiation induction rules in rodent hippocampus: role of GABAergic inhibition. *J. Neurosci.* 23, 11142–11146
75. Hayama, T. *et al.* (2013) GABA promotes the competitive selection of dendritic spines by controlling local Ca 2+ signaling. *Nat. Neurosci.* 16, 1409–1416
76. Wang, L. and Maffei, A. (2014) Inhibitory plasticity dictates the sign of plasticity at excitatory synapses. *J. Neurosci.* 34, 1083–1093
77. Steele, P.M. and Mauk, M.D. (1999) Inhibitory control of LTP and LTD: stability of synapse strength. *J. Neurophysiol.* 81, 1559–1566
78. Sjöström, P.J. and Hausser, M (2006) A cooperative switch determines the sign of synaptic plasticity in distal dendrites of neocortical pyramidal neurons. *Neuron* 51, 227–238
79. Letzkus, J.J. *et al.* (2006) Learning rules for spike timing-dependent plasticity depend on dendritic synapse location. *J. Neurosci.* 26, 10420–10429
80. Froemke, R.C. *et al.* (2010) Dendritic synapse location and neocortical spiketiming-dependent plasticity. *Front. Synaptic Neurosci.* 2, 29
81. Yoshimura, Y. and Callaway, E.M. (2005) Fine-scale specificity of cortical networks depends on inhibitory cell type and connectivity. *Nat. Neurosci.* 8, 1552–1559
82. Znamenskiy, P. *et al.* (2018) Functional selectivity and specific connectivity of inhibitory neurons in primary visual cortex. *bioRxiv* https://doi.org/10.1101/294835
83. Petersen, C.C.H. and Crochet, S. (2013) Synaptic computation and sensory processing in neocortical layer 2/3. *Neuron* 78, 28–48
84. Avermann, M. *et al.* (2012) Microcircuits of excitatory and inhibitory neurons in layer 2/3 of mouse barrel cortex. *J. Neurophysiol.* 107, 3116–3134
85. Kubota, Y. (2014) Untangling GABAergic wiring in the cortical microcircuit. *Curr. Opin. Neurobiol.* 26, 7–14
86. Tremblay, R. *et al.* (2016) GABAergic interneurons in the neocortex: from cellular properties to circuits. *Neuron* 91, 260–292
87. Ferguson, B.R. and Gao, W.-J. (2018) PV interneurons: critical regulators of E/I balance for prefrontal cortex-dependent behavior and psychiatric disorders. *Front. Neural Circ.* 12, 37
88. Cardin, J.A. *et al.* (2009) Driving fast-spiking cells induces gamma rhythm and controls sensory responses. *Nature* 459, 663–667
89. David, F. *et al.* (2022) Layer-specific stimulations of parvalbumin-positive cortical interneurons in mice entrain brain rhythms to different frequencies. *bioRxiv* https://doi.org/10.1101/2021.03.31.437894
90. Markram, H. *et al.* (2004) Interneurons of the neocortical inhibitory system. *Nat. Rev. Neurosci.* 5, 793–807
91. Zhang, S. *et al.* (2014) Long-range and local circuits for top-down modulation of visual cortex processing. *Science* 345, 660–665
92. Adesnik, H. *et al.* (2012) A neural circuit for spatial summation in visual cortex. *Nature* 490, 226–231
93. Gentet, L.J. *et al.* (2012) Unique functional properties of somatostatin-expressing GABAergic neurons in mouse barrel cortex. *Nat. Neurosci.* 15, 607–612
94. Schuman, B. *et al.* (2021) Neocortical layer 1: an elegant solution to top-down and bottom-up integration. *Annu. Rev. Neurosci.* 44, 221–252
95. Lee, S. *et al.* (2013) A disinhibitory circuit mediates motor integration in the somatosensory cortex. *Nat. Neurosci.* 16, 1662–1670
96. Yu, F. *et al.* (2014) A cortical circuit for gain control by behavioral state. *Cell* 156, 1139–1152
97. Pi, H.-J. *et al.* (2013) Cortical interneurons that specialize in disinhibitory control. *Nature* 503, 521–524
98. Garrett, M. *et al.* (2020) Experience shapes activity dynamics and stimulus coding of VIP inhibitory cells. *Elife* 9, e50340
99. Letzkus, J.J. *et al.* (2015) Disinhibition, a circuit mechanism for associative learning and memory. *Neuron* 88, 264–276
100. Harris, K.D. and Mrsic-Flogel, T.D. (2013) Cortical connectivity and sensory coding. *Nature* 503, 51–58
101. Shepherd, G.M. and Rowe, T.B. (2017) Neocortical lamination: insights from neuron types and evolutionary precursors. *Front. Neuroanat.* 11, 100

102. Dugas-Ford, J. *et al.* (2012) Cell-type homologies and the origins of the neocortex. *Proc. Natl. Acad. Sci.* 109, 16974–16979
103. Karten, H.J. (2013) Neocortical evolution: neuronal circuits arise independently of lamination. *Curr. Biol.* 23, R12–R15
104. Briscoe, S.D. and Ragsdale, C.W. (2018) Homology, neocortex, and the evolution of developmental mechanisms. *Science* 362, 190–193
105. Mei, J. *et al.* (2022) Informing deep neural networks by multiscale principles of neuromodulatory systems. *Trends Neurosci.* 45, 237–250
106. Allaway, K.C. *et al.* (2020) Cellular birthdate predicts laminar and regional cholinergic projection topography in the forebrain. *Elife* 9, e63249
107. Urban-Ciecko, J. *et al.* (2018) Precisely timed nicotinic activation drives SST inhibition in neocortical circuits. *Neuron* 97, 611–625
108. Brombas, A. *et al.* (2014) Activity-dependent modulation of layer 1 inhibitory neocortical circuits by acetylcholine. *J. Neurosci.* 34, 1932–1941
109. Kruglikov, I. and Rudy, B. (2008) Perisomatic GABA release and thalamocortical integration onto neocortical excitatory cells are regulated by neuromodulators. *Neuron* 58, 911–924
110. Moran, R.J. *et al.* (2013) Free energy, precision and learning: the role of cholinergic neuromodulation. *J. Neurosci.* 33, 8227–8236
111. Iglesias, S. *et al.* (2021) Cholinergic and dopaminergic effects on prediction error and uncertainty responses during sensory associative learning. *Neuroimage* 226, 117590
112. Barron, H.C. *et al.* (2020) Prediction and memory: a predictive coding account. *Prog. Neurobiol.* 192, 101821
113. Bolz, J. and Gilbert, C.D. (1986) Generation of end-inhibition in the visual cortex via interlaminar connections. *Nature* 320, 362–365
114. Nassi, J.J. *et al.* (2013) Corticocortical feedback contributes to surround suppression in V1 of the alert primate. *J. Neurosci.* 33, 8504–8517
115. Boutin, V. *et al.* (2021) Sparse deep predictive coding captures contour integration capabilities of the early visual system. *PLoS Comput. Biol.* 17, e1008629
116. Spratling, M.W. (2010) Predictive coding as a model of response properties in cortical area V1. *J. Neurosci.* 30, 3531–3543
117. Liang, H. *et al.* (2017) Interactions between feedback and lateral connections in the primary visual cortex. *Proc. Natl. Acad. Sci.* 114, 8637–8642
118. Marques, T. *et al.* (2018) The functional organization of cortical feedback inputs to primary visual cortex. *Nat. Neurosci.* 21, 757–764
119. Nurminen, L. *et al.* (2018) Top-down feedback controls spatial summation and response amplitude in primate visual cortex. *Nat. Commun.* 9, 1–13
120. Zmarz, P. and Keller, G.B. (2016) Mismatch receptive fields in mouse visual cortex. *Neuron* 92, 766–772
121. Jordan, R. and Keller, .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
122. Eliades, S.J. and Wang, X. (2008) Neural substrates of vocalization feedback monitoring in primate auditory cortex. *Nature* 453, 1102–1106
123. Fiser, A. *et al.* (2016) Experience-dependent spatial expectations in mouse visual cortex. *Nat. Neurosci.* 19, 1658–1664
124. Feuerriegel, D. *et al.* (2021) Evaluating the evidence for expectation suppression in the visual system. *Neurosci. Biobehav. Rev.* 126, 368–381
125. Mikulasch, F.A. *et al.* (2022) Visuomotor mismatch responses as a hallmark of explaining away in causal inference. *bioRxiv* https://doi.org/10.1101/2022.04.07.486697
126. Garner, A.R. and Keller, G.B. (2022) A cortical circuit for audio-visual predictions. *Nat. Neurosci.* 25, 98–105
127. Keller, G.B. and Hahnloser, R.H.R. (2009) Neural processing of auditory feedback during vocal practice in a songbird. *Nature* 457, 187–190
128. Mlynarski, W.F. and Hermundstad, A.M. (2018) Adaptive coding for dynamic sensory inference. *Elife* 7, e32055
129. Kubota, Y. *et al.* (2007) Neocortical inhibitory terminals innervate dendritic spines targeted by thalamocortical afferents. *J. Neurosci.* 27, 1139–1150
130. Alexander Bae, J. *et al.* (2021) Functional connectomics spanning multiple areas of mouse visual cortex. *bioRxiv* https://doi.org/10.1101/2021.07.28.454025
131. Keller, G.B. *et al.* (2012) Sensorimotor mismatch signals in primary visual cortex of the behaving mouse. *Neuron* 74, 809–815
132. Hamm, J.P. *et al.* (2021) Cortical ensembles selective for context. *Proc. Natl. Acad. Sci.* 118, e2026179118
133. Schultz, W. (2016) Dopamine reward prediction-error signalling: a two-component response. *Nat. Rev. Neurosci.* 17, 183–195
134. Bogacz, R. (2017) A tutorial on the free-energy framework for modelling perception and learning. *J. Math. Psychol.* 76, 198–211
135. Millidge, B. *et al.* (2020) Relaxing the constraints on predictive coding models. *arXiv* https://doi.org/10.48550/arXiv.2010.01047
136. Kappel, D. *et al.* (2014) STDP installs in winner-take-all circuits an online approximation to hidden Markov model learning. *PLoS Comput. Biol.* 10, e1003511.X
137. Gerstner, W. *et al.* (2014) *Neuronal Dynamics: From Single Neurons to Networks and Models of Cognition*, Cambridge University Press
138. Constantinople, C.M. and Bruno, R.M. (2013) Deep cortical layers are activated directly by thalamus. *Science* 340, 1591–1594
139. Echeveste, R. *et al.* (2020) Cortical-like dynamics in recurrent circuits optimized for sampling-based probabilistic inference. *Nat. Neurosci.* 23, 1138–1149