Self-supervised predictive learning accounts for cortical layer-specificity

Kevin Kermani Nejad[1,2], **Paul Anastasiades**[4], **Loreen Hertäg**[3#], **Rui Ponte Costa**[1,2#*]

[1]Centre for Neural Circuits and Behaviour, Department of Physiology, Anatomy and Genetics, University of Oxford, Oxford, United Kingdom; [2]Bristol Computational Neuroscience Unit, Intelligent Systems Lab, Faculty of Engineering, University of Bristol, Bristol, BS8 1TH, United Kingdom; [3]Technische Universität Berlin & Bernstein Center for Computational Neuroscience Berlin, 10115 Berlin, Germany; [4]Department of Translational Health Sciences, University of Bristol, Whitson Street, Bristol, BS1 3NY, United Kingdom

Abstract The neocortex constructs an internal representation of the world, but the underlying circuitry and computational principles remain unclear. Inspired by self-supervised learning algorithms, we introduce a computational theory wherein layer 2/3 (L2/3) learns to predict incoming sensory stimuli by comparing previous sensory inputs, relayed via layer 4, with current thalamic inputs arriving at layer 5 (L5). We demonstrate that our model accurately predicts sensory information in context-dependent temporal tasks, and that its predictions are robust to noisy and occluded sensory input. Additionally, our model generates layer-specific sparsity and latent representations, consistent with experimental observations. Next, using a sensorimotor task, we show that the model's L2/3 and L5 prediction errors mirror mismatch responses observed in awake, behaving mice. Finally, through manipulations, we offer testable predictions to unveil the computational roles of various cortical features. In summary, our findings suggest that the multi-layered neocortex empowers the brain with self-supervised predictive learning.

Introduction

Internal models of the external world are thought to endow the brain with the ability to predict incoming sensory information and select appropriate action-outcome contingencies[1]. Internal models are widely believed to be encoded in the neocortex[2,3], whose hallmark feature is its laminar organization, comprising six distinct layers. Although much has been learned about the underlying cellular heterogeneity and connectivity of individual cortical layers, why the neocortex relies on a multi-layered structure remains unclear[4]. Unraveling its function could shed light on the neocortical algorithms responsible for building rich internal representations of the world.

Historically, it has been proposed that *unsupervised* learning in sensory cortices underpins the development of intricate sensory representations that are critical for driving behavior[5–7]. However, whether the laminar structure of neocortical microcircuits supports unsupervised learning remains unclear. Self-supervised learning is a form of unsupervised learning that leverages the inherent structure or patterns within the data as the target for learning. A common application of self-supervised learning is to predict the incoming input given past information[8–12]. Importantly, self-supervised learning algorithms learn representations that better capture experimentally observed latent representations while resulting in richer models of input statistics[12–16]. However, learning in these models is often treated as a black box, therefore it remains to be determined whether the brain is capable of employing such learning principles.

[#] These authors contributed equally.

*For correspondence: rui.costa@dpag.ox.ac.uk (RPC)

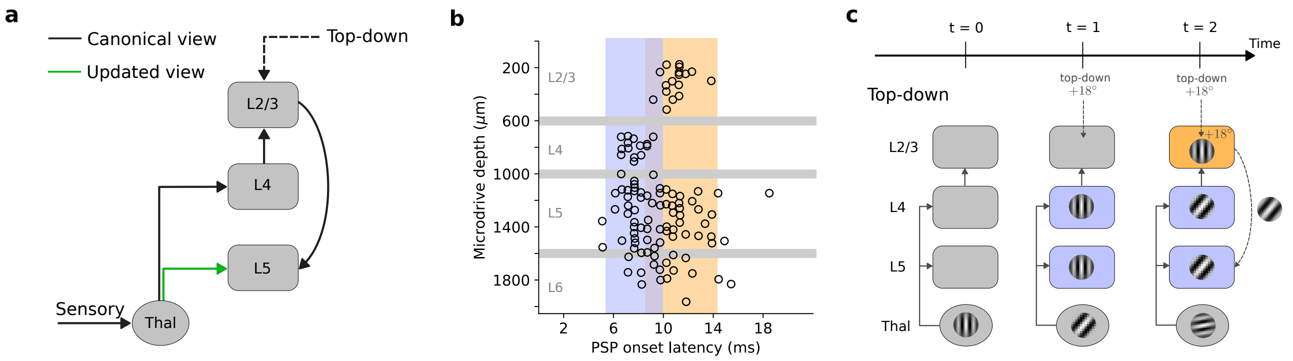


Figure 1. Information flow in neocortical circuits. **a,** The canonical and updated view of the neocortical microcircuit. Sensory input is initially processed by the thalamus, which, in the classical view, exclusively targets layer 4 (L4). L4 subsequently relays this information to layer 2/3 (L2/3). L2/3, in turn, combines L4 input with top-down contextual input that is fed forward to layer 5 (L5). However, recent studies have emphasized the need to update this view due to direct projections from sensory thalamic nuclei to L5 pyramidal cells[26] (green arrow). For the sake of clarity, we omitted feedback connections from the schematic which in our self-supervised model are responsible for carrying error signals that drive learning (see main text and Methods). **b,** Onset latencies of postsynaptic potentials (PSP) by cortical depth. These results demonstrate the simultaneous activation of L4 and L5 neurons by the thalamus, indicating a direct thalamic input to L5. Adapted from Constantinople and Bruno[26]. **c,** Proposed information flow of self-supervised temporal learning in the neocortical microcircuit. L2/3, informed by past sensory input from L4 and top-down contextual input, predicts the current sensory input arriving in L5. The direct thalamic inputs to L5 provide sensory input, which is used as a teaching signal to instruct the L2/3 predictive model.

The traditional view of the neocortical microcircuit postulates a sequential flow of sensory information. In this canonical view, sensory input is relayed via the thalamus to layer 4 (L4) of the neocortex[17,18]. L4 subsequently transmits this information to layer 2/3 (L2/3), which is thought to integrate ascending sensory information with top-down modulatory input from higher-order cortical areas[19–21]. L2/3 in turn projects to layer 5 (L5), with L5 outputting information to other brain areas (Fig. 1a). However, growing evidence suggests that this model does not capture the full diversity of connections in the neocortical microcircuit[18]. A body of experimental work suggests that L5 pyramidal cells receive direct thalamic input that can drive short-latency sensory evoked responses independently of activity within the cortical network (Fig. 1b)[22–26]. These observations imply two distinct sensory-driven pathways within the neocortex, one targeting L4 and the other L5 (Fig. 1a). However, why the cortex requires multiple inputs and the computations supported by such parallel pathways remain unknown.

Inspired by this refreshed view of the canonical microcircuit and the predictive capabilities of self-supervised machine learning algorithms[9,27], we propose a model in which L2/3, informed by past sensory input from L4 and top-down context, predicts incoming sensory input. In this model, the L4-to-L2/3 delay enables L2/3 to generate predictions based on previous sensory information. Direct thalamic input to L5 is responsible for providing sensory-based teaching signals required for L2/3 to adjust and refine its predictions (Fig. 1c). This perspective of the neocortical circuitry suggests that the L4-L2/3-L5 laminar structure with parallel thalamic innervation enables the brain to learn rich temporal representations.

We first show that our learning rule for L2/3-to-L5 connections closely resembles experimentally observed long-term synaptic plasticity[28]. Using self-supervised learning, our model can learn and predict Gabor-like inputs in contextual sequential tasks, highlighting its effectiveness in capturing complex patterns. By ablating individual components of the model and evaluating their impact on performance, we reveal how the neocortical circuit components collaboratively enable self-supervised learning. Next, we demonstrate that self-supervised learning leads to predictions that are robust to sensory noise and occlusions. Moreover, the model predicts layer-specific sparsity that closely matches those found experimentally in sensory systems. Additionally, we demonstrate that our model produces L2/3 and L5 prediction errors in response to visuomotor mismatches, thereby providing an explanation to the mismatch responses observed in awake, behaving animals. Finally, we suggest a set of optogenetic experiments capable of testing the core predictions of our self-supervised learning model. Collectively, our findings support the notion that the L4→L2/3→L5 pathway is instrumental in enabling the brain to engage in temporal self-supervised learning, underscoring its potential significance in neural mechanisms of predictive learning.

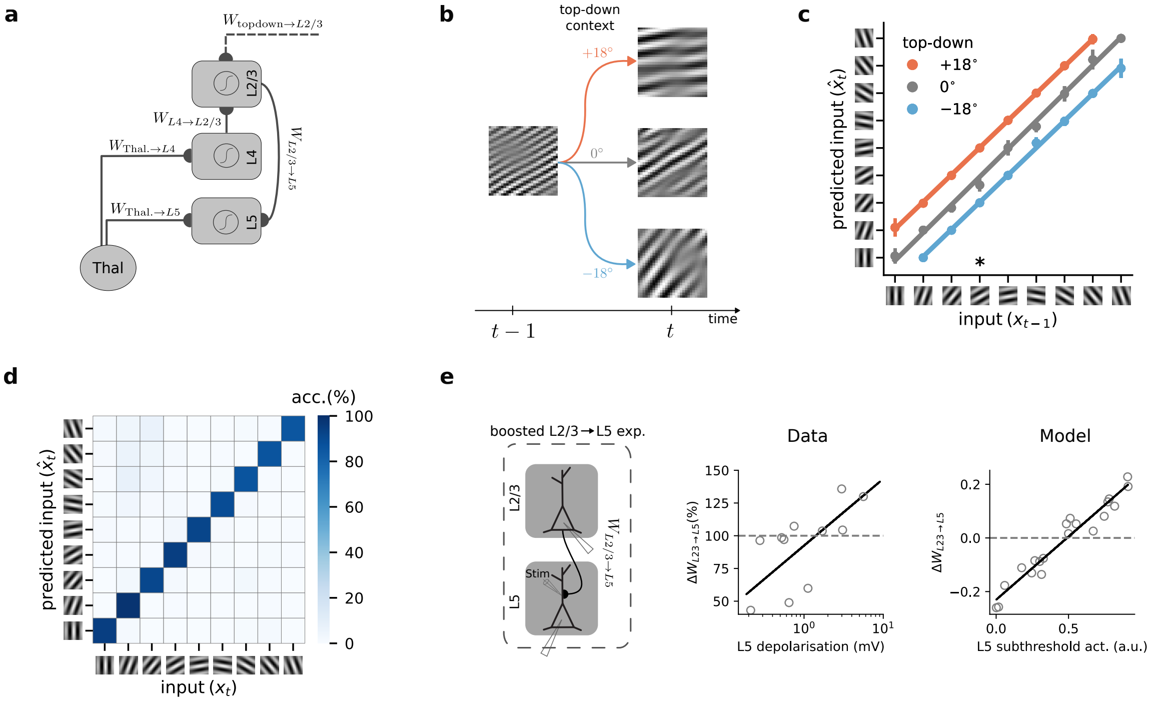


Figure 2. A model of temporal self-supervised learning in cortical circuits. a, Schematic of the cortical circuit model. **b,** Schematic of a sequential Gabor task. The generative factor that is provided to the model as top-down context at timestep t determines the orientation of the next Gabor patch at timestep $t+1$. **c,** Decoding accuracy of a linear model trained on the output of L2/3. For a given input, L2/3 predicts the incoming sensory input with high accuracy. Colors represent the three possible conditions ($-18°$, $0°$, and $+18°$). $*$ points to the example illustrated in panel b. **d,** Confusion matrix for classification accuracy of a linear model trained on the output of L5. The metrics in c and d are calculated over 5 different initial conditions. **e,** Left: schematic of the experimental setup in which an extracellular electrode was used to boost L5 activity while inducing long-term synaptic plasticity on L2/3-to-L5 connections[28]. Middle: observed changes in synaptic weights as a function of L5 depolarization (scatter plot: individual data points, solid line: linear fit to the data). Right: L2/3-to-L5 learning rule as predicted by our model as a function of L5 activity for multiple randomly drawn samples of L2/3 and L5 activity (circles), and linear fit to the data points (solid line). Error bars represent the standard error of the mean over 5 different initial conditions.

Results

Neocortical layers can implement self-supervised predictive learning

To understand how neocortical microcircuits process temporal information and learn latent representations in a self-supervised manner, we created a model that emulates the properties of cortical circuits. Our model contains three subnetworks with non-linear neurons separated into L2/3, L4, and L5 to reflect the laminar architecture of the neocortex (Fig. 2a). Within this framework, L4 receives ascending sensory information, x, at timestep t through input weights $W_{\mathrm{Thal.}\to L4}$. L2/3 receives delayed input, timestep $t-1$, from L4 via $W_{L4\to L2/3}$ synapses and top-down contextual input via $W_{\mathrm{top\text{-}down}\to L2/3}$ weights. We hypothesize that this combination of inputs enables L2/3 to make predictions about upcoming sensory information. L2/3 predictions are then sent down to L5 via $W_{L2/3\to L5}$ synapses and compared with the actual sensory input received by L5 at timestep t. This comparison enables an error function to be computed which is then fed back via L5-to-L2/3 connections to adjust the L2/3 predictive model of incoming input. The error function is defined as $\mathcal{C}_{L_{23}\to L_5} = \frac{1}{2}(\mathbf{z}_5^t - \underbrace{W_{L_{23}\to L_5}\mathbf{z}_{23}^t}_{\text{prediction, }\hat{\mathbf{z}}_5^t})^2$ where $\mathbf{z}_5^t$ is the activity of L5 neurons and $W_{L_{23}\to L_5}\mathbf{z}_{23}^t$ its prediction, $\hat{\mathbf{z}}_5^t$. In addition to this predictive error, L5 is also trained to reconstruct its own input (see Methods). During learning, we modify connections to minimize the error functions and facilitate the encoding of sensory input. Consequently, our model requires both feed-forward connections from L4 → L2/3, which relay sensory information, as well as feedback connections from L5 back to L2/3, to transmit error signals. All weights are optimized via gradient descent.

To demonstrate our model's ability to learn useful representations, we created two-step sequences of Gabor patches, which are commonly used to evoke responses in the primary visual cortex (see Methods)[29,30]. Starting with a random Gabor patch, at time point t, the orientation of the Gabor patch changes according to top-down contextual input that is randomly generated (see Methods). This higher-order contextual cue to L2/3 is provided at each time step and mimics signals such as those provided by the motor cortex[31]. Given this top-down input the Gabor patch either rotates anti-clockwise by -18°, remains the same, or rotates clockwise by $+18^\circ$ (Fig. 2b). For example, if the current Gabor patch has 0° orientation, the orientation of the subsequent input can be -18°, 0° or $+18^\circ$ for contextual cue values of -18°, 0°, and $+18^\circ$, respectively (Fig. 2c).

To evaluate the representations learned by our model, we use linear decoders[8,32,33]. We trained this decoder on L2/3's output and compared it with the Gabor patches received by L5 at a given timestep. L2/3 effectively learns to predict upcoming Gabor patches using the previous input, provided by L4, and the top-down context value (Fig. 2c). Next, we applied a linear classifier on L5's output. On average, L5 achieves a test accuracy of $\approx$ 89% (classification accuracy on a random model is $\approx$ 11%), indicating that L5 successfully identifies and encodes each Gabor patch's distinct features (Fig. 2d). In addition, we obtain similar results in a more complex task in which hand-written digits are used as input (Fig. S1). Our results indicate that L2/3 achieves near-perfect accuracy ($\approx$ 93%) in predicting the subsequent Gabor patch, outperforming L5. These results are in line with experimental findings showing that L2/3 can learn to predict image sequences[30].

Next, we compared the learning rule for the key predictive weights in our model, $W_{L2/3\rightarrow L5}$ (Methods Eq. 7), with observed long-term synaptic plasticity in primary sensory cortices (Fig. 2e)[28]. Our learning rule predicts a depression-to-potentiation switch as the activity of L5 neurons increases. This is in line with experimental observations showing a similar depression-to-potentiation switch of $W_{L2/3\rightarrow L5}$ connections with increasing depolarization of L5 pyramidal cells[28]. Hence, this experimental evidence corroborates our model's learning rule, showing that the model is consistent with the updated view of the neocortical circuit and known synaptic plasticity mechanisms in primary sensory cortices.

In summary, we have shown the model's ability to perform self-supervised learning in a temporal task and its consistency with synaptic plasticity observations. However, we have yet to explore the precise contribution of each circuit element to self-supervised learning.

Neocortical circuitry jointly underlies self-supervised learning

In our model, different cortical layers give rise to distinct computational roles. To demonstrate this, while generating experimentally testable predictions, we systematically ablated individual connections, allowing us to quantify their impact on both representational capability and performance (Fig. 3A).

When we knocked out the L2/3-to-L5 connection, L2/3 was unable to learn to make predictions about upcoming sensory information (Fig. 3b vs. Fig. 2c). This is due to the lack of communication between the source of the prediction in L2/3, and the source of error signals in L5. As a consequence, the L2/3 predictive model does not learn. However, because top-down contextual input is still present, L2/3 still shows contextual segregation.

Our model proposes a key function for the delay introduced by L4 as information propagates to L2/3. This delay creates a temporal discrepancy between the information available to L2/3 (past input) and L5 (present input) which enables L2/3 to learn predictive representations by attempting to anticipate the incoming sensory input. When this delay is removed, which would be equivalent to L2/3 receiving direct thalamic input, the entire network operates on incoming sensory input, i.e. at timestep t. Consequently, the network can no longer generate meaningful predictions of future inputs. This is reflected in Fig. 3c where L2/3 fails to reliably distinguish among potential future outcomes. This result highlights the key role that temporal delays may have in shaping predictive learning within the neocortical microcircuit. The L4 to L2/3 delay is essential for biasing L2/3 representations towards the future. Without it, both the Thal. $\rightarrow$ L4 $\rightarrow$ L2/3 $\rightarrow$ L5 and Thal. $\rightarrow$ L5 pathways end up representing the current sensory input, rendering the former pathway redundant.

Next, we investigated how the ablation of these different circuit elements affects the ability to decode *current* sensory information from L2/3 and L5 representations. For current input decoding (Fig. 3d), L5 demonstrated robust accuracy as long as it retained access to thalamic sensory input. This aligns with its role as a primary recipient of sensory data, together with L4[23,26,34]. L2/3 accuracy, however, was more dependent on the circuit properties. While top-down input to L2/3 provided useful context-dependent input (Fig. S2), any disruption to core pathways within the microcircuit, except the delay knockout, compromised L2/3's ability to represent the current sensory input.

Decoding the *previous* input (Fig. 3e) further differentiated L2/3 and L5. As anticipated, L5 exhibited limited infor-

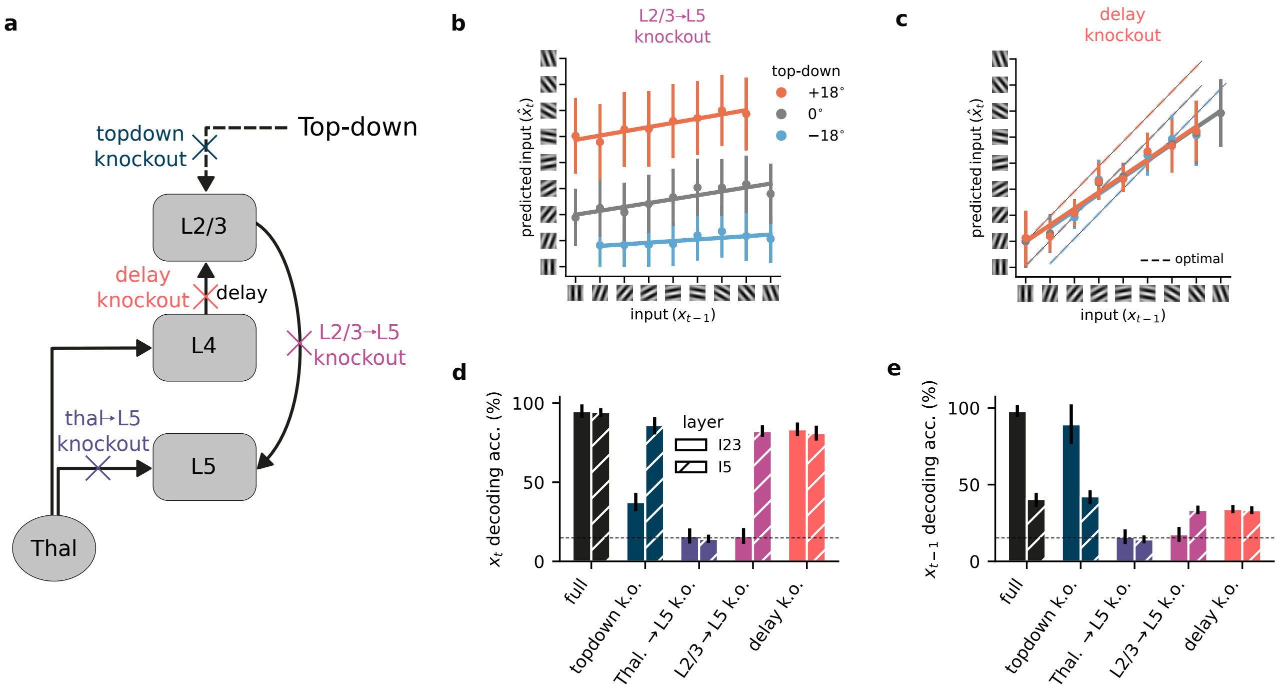


Figure 3. Neocortical circuitry jointly enables self-supervised learning. a, Schematic of the model with individual components knocked out (colored crosses) within the neocortical microcircuit architecture. **b,** Connections from L2/3 to L5 are necessary for L2/3 to learn a predictive representation of the input. **c,** Impact of L4-mediated delay in self-supervised learning (dashed lines represent the optimal prediction). **d,** Summary of decoding accuracy of the current input for L2/3 and L5 when specific connections are knocked out. The x-axis indicates the specific ablation, while the y-axis the decoding accuracy for the current input (x_t). **e,** Similar to d, but for the past input (x_{t-1}). Knockout components are color-coded as in panel a. Horizontal dashed lines in d and e represent chance decoding accuracy. Error bars represent the standard error of the mean over 5 different initial conditions.

mation about previous inputs due to its exclusive focus on current thalamic information. L2/3, however, encodes information about the past as a result of the delay introduced by L4. Complete loss of this past-input representation occurred only when critical learning pathways were ablated (Thal. → L5, L2/3 → L5), or when the delay was removed; thereby synchronizing L2/3 and L5 inputs.

Our ablation and decoding analyses suggest that predictive learning within the neocortical microcircuit depends on a complex interplay between L2/3 and L5. We find that the L2/3-to-L5 connection is essential for the model to learn predictive representations, suggesting that L2/3 prediction errors drive learning. Furthermore, the temporal delay between L4 and L2/3 is crucial for generating future-oriented predictions, but not current representations. In terms of decoding past and present sensory input, our results demonstrate that L2/3 specializes in representing temporal context, while L5 primarily encodes immediate sensory information. This result aligns with the experimental observations showing that L2/3 effectively encodes temporal information with high precision[35].

Self-supervised learning leads to robustness to noise and occlusion in cortical networks

As a consequence of learning a robust predictive model of the sensory input, the network should disregard unpredictable aspects like noise. Therefore, we hypothesized that the self-supervised L2/3 → L5 predictive component would help L5 filter out sensory noise. To test this, we ablated the L2/3 → L5 projection in our model. This simulated ablation shows that removing the predictive component (i.e. $L2/3 \rightarrow L5$) dramatically reduces robustness to different noise levels (Fig. 4a; similar results are obtained when ablating the L4-to-L2/3 delay, Fig. S3). This denoising capability emerges naturally from the model's design, even though it was not explicitly designed for this purpose, which are comparable to a near-optimal denoising autoencoder network.

To further investigate the model's robustness to input perturbations we tested its ability to reconstitute input patterns during partial input occlusions. We modified our sequential Gabor task, by randomly occluding parts of the input (Fig. 4b). After training the model without occlusions, we assessed the robustness of the learned representa-

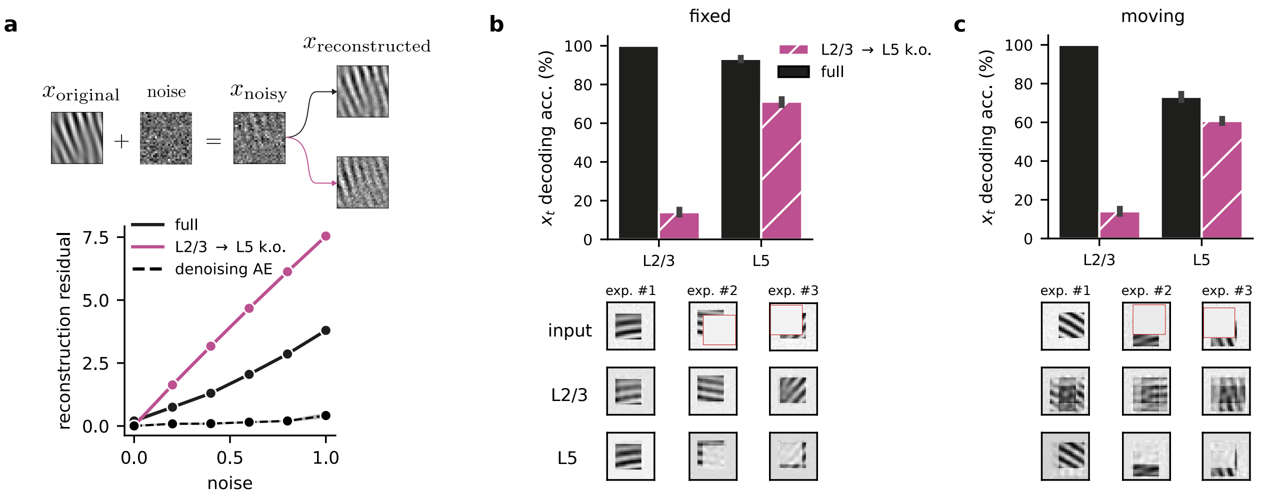


Figure 4. L2/3-to-L5 predictions are crucial for denoising and resolving occluded stimuli. a, L2/3-to-L5 connections promote noise suppression in L5 representations. Top: Schematic of noise added to the original inputs. Bottom: Noise-corrupted input samples lead to higher L5 reconstruction residuals ($\hat{x}_t - x_t$) when the L2/3→L5 pathway is ablated (purple) compared to the full model (solid black). We also provide the reconstruction residual for an autoencoder that was explicitly trained to denoise the input (dashed black line). **b,** Top: Decoding accuracy with and without L2/3→L5 for a Gabor task with occlusion. Bottom: Three examples depicting L2/3's ability to recover occluded information, compared to L5's incomplete reconstructions (top row: original occluded input; middle row: L2/3 prediction; bottom row: L5 reconstruction). **c,** Top: Accuracy with and without L2/3→L5 connections for a task in which Gabor patches move randomly. Bottom: Examples illustrating the robustness with moving Gabor patches (top row: original input with motion (cf. panel b); middle row: L2/3 prediction; bottom row: L5 reconstruction). L2/3 reconstruction encodes uncertainty about the future possible input location. Error bars represent the standard error of the mean over 5 different initial conditions.

tions in each layer by evaluating their ability to classify occluded sensory input. We observed that L2/3 achieves higher decoding accuracy compared to L5 and that L5 decoding is not affected by L2/3 → L5 knockout (Fig. 4b, top). Further analysis shows that L2/3 can fully reconstruct the input while L5 is only able to reconstruct the observable parts of the input (Fig. 4b, bottom). These results support the idea that a strong predictive model leads to representations that are robust to several perturbations.

Finally, we explored whether L2/3 can also encode the uncertainty about the possible input locations. To test this, we introduced random shifts in the position of the Gabor patch on the blank canvas during the task. Decoding performance remained similar to the fixed-position task, but reconstructions were different (Fig. 4c, top). L2/3 representations reflect the input's positional uncertainty (blurred reconstructions across possible locations), while L5 again encodes only the visible parts (Fig. 4c, bottom). This suggests that self-supervised learning also leads to useful L2/3 representations when in the presence of sensory uncertainty.

Collectively, these results underscore the robustness exhibited by the proposed neocortical predictive learning model across diverse input conditions. Consequently, our model offers valuable insights into the mechanisms through which cortical circuits deal with the considerable variability inherent in naturalistic environments.

Layer-specific sparseness emerges from self-supervised learning

Sparse coding, in which only a small subset of neurons are strongly active for a given stimulus, is a widespread phenomenon across the neocortex[36–40]. This sparsity is particularly pronounced in superficial layers (L2/3) compared to deeper layers (e.g. L5)[41]. However, it is unclear why the degree of sparsity varies across cortical layers and how it relates to their computational role.

We wondered whether our network, equipped for temporal self-supervised learning, could reproduce experimentally observed sparsity distributions. Moreover, we wanted to investigate how different network features may control sparsity across different neocortical layers. To this end, we trained our model on the sequential Gabor task. After training, we measured population sparseness across layers using established metrics[42] (see Methods). Interestingly, our results closely mirror experimental findings[41]: L2/3 presents the highest sparseness, followed by L4 and then L5 (Fig. 5a). This alignment suggests that self-supervised learning, focused on input prediction, could be a key factor

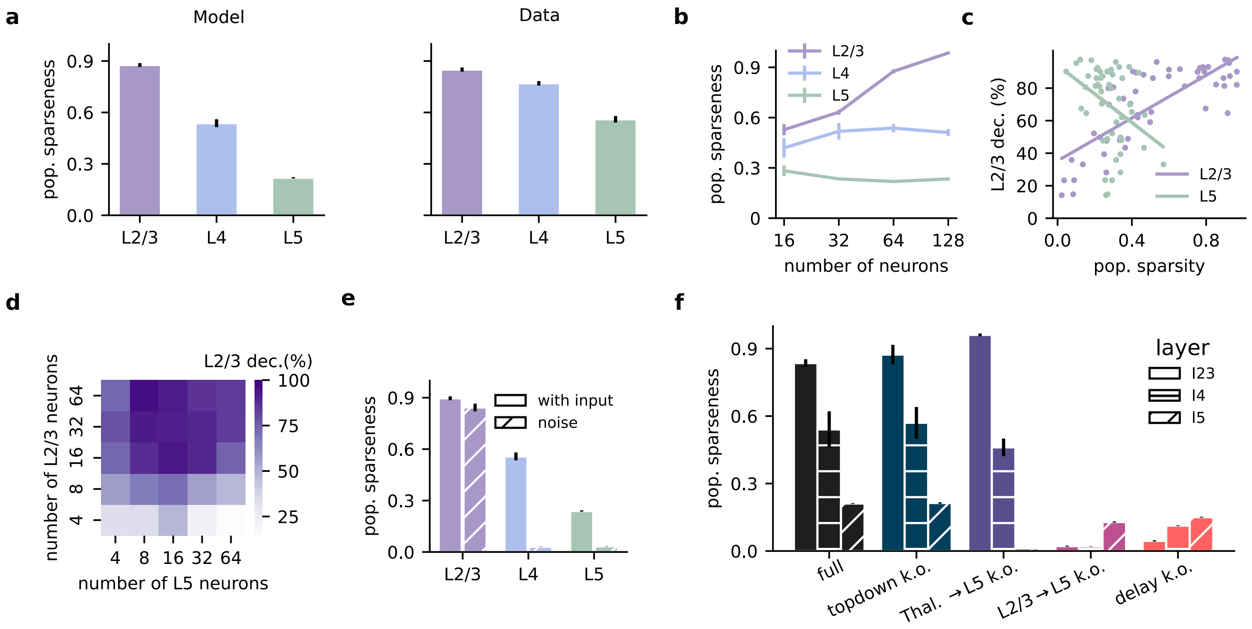


Figure 5. Population sparseness depends on neocortical layer. a, Population sparseness across layers in the model (left) and experimental data[41] (right). **b,** Population sparseness as a function of the number of neurons. The qualitative relationship between layers is preserved, but L2/3 sparseness increases with network size. **c,** Decoding accuracy of current input as a function of the population sparsity of L2/3 and L5 (L2/3: r=0.78, p=2.1e-11; L5: r=-0.35 p=0.01). **d,** L2/3 decoding accuracy of the current input as a function of the number of neurons in L2/3 and L5. **e,** Population sparseness with or without sensory input (noise condition) after learning. L2/3 remains sparse, while L4 and L5 show a strong reduction in response sparsity. **f,** Population sparseness following ablation during learning of different model components. Top-down input ablation slightly increases L2/3 sparseness. Thalamic input ablation to L5 decreases L2/3 sparseness while increasing L5 sparseness. Ablation of L2/3-to-L5 connections abolishes sparseness across all layers. Error bars represent the standard error of the mean over 5 different initial conditions.

driving sparsity in biological neural networks.

Layer 2/3 has undergone rapid expansion relative to other layers within the human evolutionary lineage[43,44]. Could this expansion support greater predictive learning capabilities? We found that L2/3 sparseness increased with network size, while sparsity in L4 and L5 remained relatively stable (Fig. 5b). Consistent with the increased sparsity in L2/3, we find that an increase in the number of L2/3 neurons also results in improved L2/3 decodability of upcoming sensory inputs (Fig. 5c,d), in line with previous work[45]. In contrast, the relationship between the number of neurons, sparsity, and decoding accuracy was not present in L5 neurons (Fig. 5c,d; Fig. S4).

To determine whether sparsity is due to the encoding of sensory input, or is simply an underlying feature of the circuitry, we replaced sensory input with Gaussian background noise. When sensory input was replaced with background noise, L2/3 retained high sparsity, whereas L4 and L5 responses showed a strong decrease in response sparsity (Fig. 5e). Importantly, this is an effect that emerges over learning (Fig. S4). These results suggest that L2/3 sparsity is primarily a consequence of learning to predict sensory input, whereas deeper layers rely more heavily on ongoing input dynamics.

Next, we ablated different model elements during learning to test their contribution to the emergence of response sparsity (Fig. 5f). Removing top-down input had a minimal effect on sparseness. However, ablating thalamic input to L5 during learning selectively decreases L5 sparseness, which is likely due to the resulting random L5 responses. Finally, ablating L2/3-to-L5 connections or the delay component completely abolished sparsity across all layers, demonstrating their crucial role in encouraging sparseness over learning.

Together, these results show that sparsity emerges as a function of input-driven predictive learning as postulated by our model, thus providing an explanation for layer-specific sparsity as observed experimentally.

L5 → L2/3 feedback is required for self-supervised learning

A cardinal feature of self-supervised learning models is that they require an error, or teaching signal to instruct plasticity across the network. This error signal prompts adjustments in the synaptic weights, thereby refining the network's

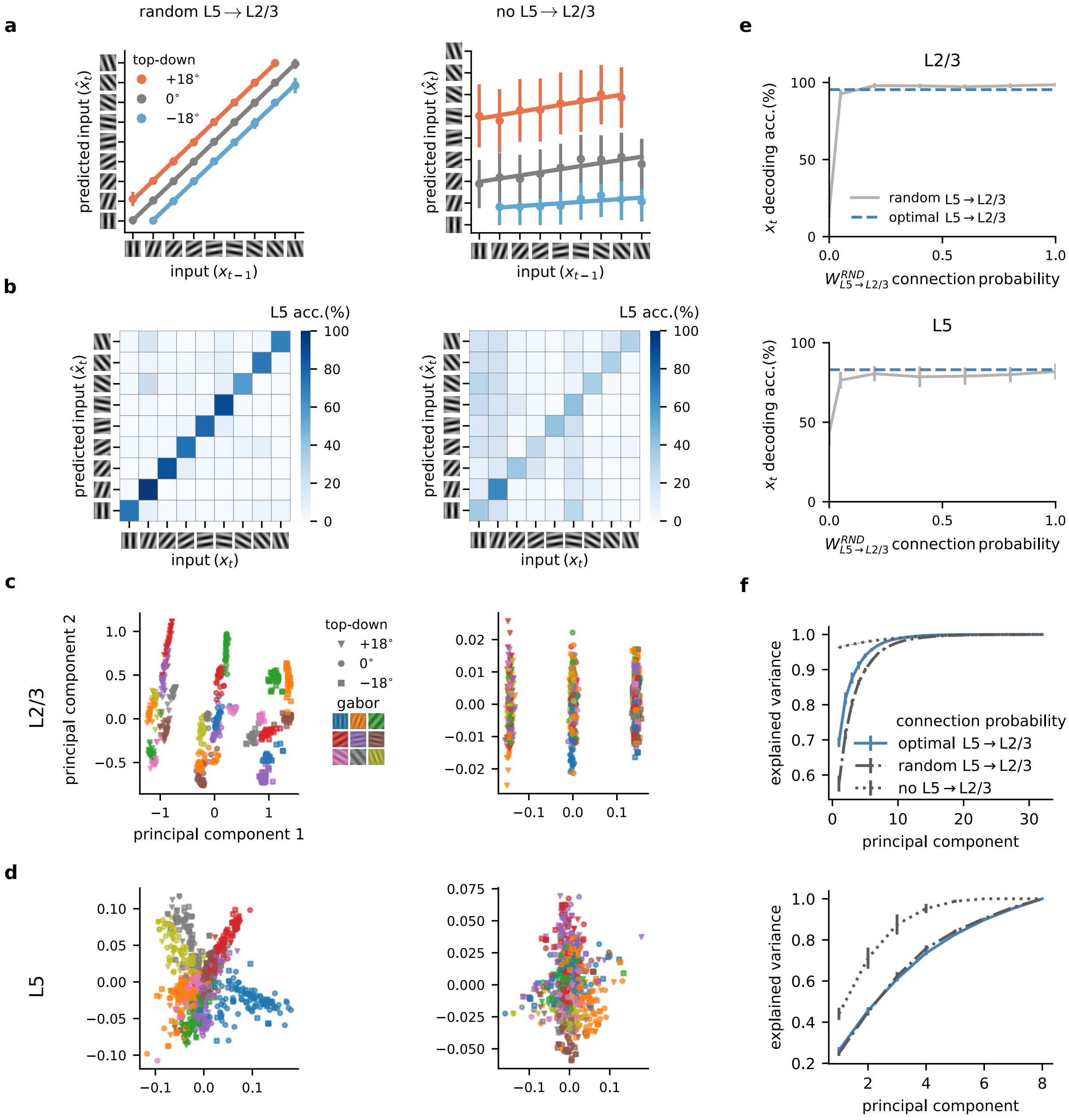


Figure 6. Role of L5-to-L2/3 feedback connections in self-supervised predictive learning. a, L2/3 learns to predict the input in the presence of random feedback (left) but fails to do so without L5-to-L2/3 feedback (right). **b,** L5 learns to represent the inputs accurately with random feedback (left) but shows lower decodability without feedback (right). **c,** Two main principal components of layer 2/3 representations for random (left) and no feedback (right) across different top-down contexts (symbols) and input Gabor orientations (colors). **d,** Two main principal components of layer 5 representations for random (left) and no feedback (right) across different top-down contexts (symbols) and input Gabor orientations (colors). **e,** L2/3 (top) and L5 (bottom) decodability for different degrees of L5-to-L23 feedback. **f,** Explained variance of L2/3 (top) and L5 (bottom) learnt representations. Error bars represent the standard error of the mean over 5 different initial conditions.

activity to enhance its predictive model. Since the error signal that drives learning in our model originates in L5, the resulting error signals should, in principle, be transmitted back to L2/3 pyramidal neurons to improve the L2/3 predictive model. Therefore, this suggests the need for a feedback connection that propagates this information from L5 to L2/3. Although the vast majority of work on neocortical circuits has disregarded feedback connections from L5 to L2/3 pyramidal cells[17,18], growing evidence shows that they are more abundant than previously assumed[46,47].

Here we explore the importance of the L5 to L2/3 feedback connection for learning in the model. In particular, we contrast *optimal* feedback, as used in previous figures, with *random* and *no feedback*. The optimal feedback condition corresponds to a setting in which the feedback weights mirror the feedforward weights (i.e. $W_{L_5 \to L_{23}} = W^T_{L_{23} \to L_5}$), whereas in the random feedback condition the feedback weights are set to a random weight matrix[48].

Inspired by work showing that random feedback weights are sufficient for credit assignment[48], we tested whether this form of unstructured feedback was sufficient for L2/3 to learn. We observed that L2/3 was indeed able to learn to predict the input with random feedback weights (Fig. 6a), in line with the optimal feedback (cf. Fig. 2c). These results suggest that unstructured feedback may be sufficient in enabling L2/3 to develop useful predictive representations. Furthermore, our findings demonstrate a significant drop in the decoding accuracy of L5 when feedback connections from L5-to-L2/3 (Fig. 6b) were removed. This decline is due to L2/3's inability to learn, causing L5 to adopt erroneous representations influenced by L2/3's unlearned state. These results demonstrate the need for L5 $\to$ L2/3 feedback.

Next, to study the neuronal representations learned by the different layers, we analyzed the two main principal components of L2/3. This revealed a notable difference in the structural organization across feedback conditions. With feedback, L2/3 representations were differentiated based on the identity of the Gabor patch as well as the top-down context (Fig. 6c, left). Without feedback, the L2/3 representations were only distinguished based on top-down inputs, indicating a limitation in the network's learning capability (Fig. 6C, right).

Similar analysis of L5 showed that with random feedback, representations are grouped as a function of Gabor patches, suggesting a structured learning process (Fig. 6d, left; similar to the optimal feedback condition, Fig. S5). These observations are in line with the increased sparsity we observed in L2/3 compared to L5 (Fig. 5). In contrast, when feedback was absent, L5 representations were less organized (Fig. 6d, right).

Classically, feedback connections within the neocortex occur at lower probabilities than the corresponding feedforward pathway[46,47]. To test how connection density influenced the properties of the network, we next explored how the linear decoding accuracy of both L5 and L2/3 varies with the probability of feedback connections from L5 to L2/3. An increase in connection probability corresponded to enhanced decoding accuracy (Fig. 6e). However, while a very low feedback connection probability was sufficient for learning the task considered here, for more complex tasks higher connection probabilities may be required for optimal performance (Fig. S6).

Finally, to determine how distributed information was, we examined how the explained variance, as assessed by the number of principal components (PCs), changed with varying feedback probabilities. The absence of feedback required a greater number of PCs to explain the data effectively, while random feedback closely mirrored the efficiency of optimal feedback connections (Fig. 6f). This increase in the number of PCs to capture the same variance is consistent with our findings above, showing the importance of feedback in organizing sensory information in superficial layers (cf. Fig. 6c,d).

This analysis underscores the critical role of feedback connections in neural networks, particularly in enhancing predictive capabilities and structuring neural representations. The nuanced differences observed across varying feedback types and intensities offer insights into the role of observed L5-to-L2/3 feedback connections[46,47] in learning and information processing in neural networks.

Model generates sensorimotor prediction error signals consistent with experimental observations

Our study thus far evidences the capability of our cortical model to make predictions about upcoming sequences. We next sought to determine how the model responds when those predictions are violated and if these responses differ for superficial and deep cortical layers. We also wanted to test whether our network generates prediction error signals that resemble those observed in cortical networks of behaving animals[19,31,49]. For example, recent *in vivo* awake experiments have observed error-like mismatch responses in a visuomotor task[19,49].

We aimed to test if our model could reproduce the mismatch responses in both L2/3 and L5 recently observed experimentally[19]. In the study by Jordan and Keller[19] the authors explored the prediction error responses in a setting where animals learn to couple the speed of the visual flow (that is, sliding vertical gratings) to speed of locomotion (Fig. 7a). This paradigm allows for the systematic investigation of how neural responses in the primary visual cortex are shaped by the interplay of external sensory stimuli (visual flow) and internal contextual expectations (running speed). Using whole-cell recordings, Jordan and Keller showed that when the visuomotor coupling is temporarily

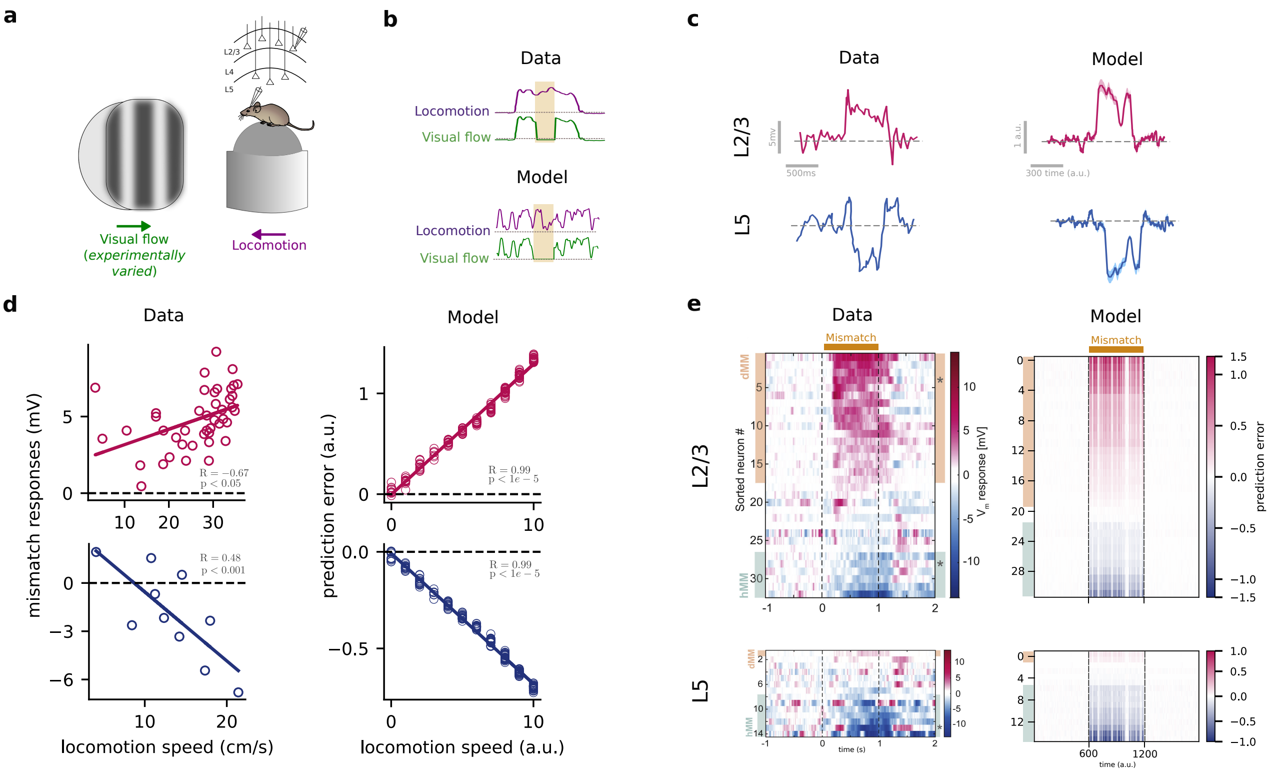


Figure 7. Model generates sensorimotor mismatch prediction errors in line with experimental observations. a, Illustration of visuomotor task used by Jordan and Keller [19] in which mice learned to associate visual flow with locomotion. **b,** Sample of training data from experiments (top) and our synthetic dataset (bottom). As in the experimental setup, we randomly halt the visual flow (flat green line) to generate visuomotor mismatches. **c,** When the visual flow is halted, a sample neuron in L2/3 of the mouse visual cortex shows depolarisation, while a sample neuron in L5 shows hyperpolarization (left). In our model, in line with the data, L2/3 shows a positive mismatch error, while L5 shows a negative mismatch error (right). Shaded areas represent standard deviation over 5 runs. **d,** The mismatch error signals in the model are correlated with the modelled locomotion speed when the visual flow is halted (right), in line with experimental observations [19,49] (left). **e,** Model generates a distribution of mismatch prediction errors which are biased towards positive errors in L2/3 and negative errors in L5 (right), in line with mismatch responses observed in primary visual cortex [19] (left). Schematics in (a) and data in (b), (d), and (e) were adapted from Jordan and Keller [19], Padamsey and Rochefort [50].

broken (i.e. 'visuomotor mismatch'), the majority of L2/3 pyramidal neurons depolarize, whereas the majority of L5 excitatory neurons hyperpolarize. We propose that the opposing mismatch responses between L2/3 and L5 observed experimentally [19] can be explained by visuomotor prediction errors in a network implementing self-supervised predictive learning.

To this end, we produced a synthetic dataset that captures the correlation between visual flow and running speed. Specifically, running speed is dictated by a random walk process whereby the speed at any given instance is determined based on the preceding speed plus a random variation (see Methods). Under normal conditions, changes in visual flow were linked to changes in speed by linearly scaling the visual flow vector according to the running speed. Occasionally, we uncoupled visual flow and locomotion by setting the visual flow to zero, thus generating visuomotor mismatches (Fig. 7b). With this setup, we can investigate how the network responds to both intact and uncoupled visuomotor integration.

We first explored the sign and magnitude of prediction errors in our model. To this end, we calculated L2/3 and L5 prediction errors using the gradients of the self-supervised error function with respect to L2/3 and L5 neurons, respectively (see Methods). When visual flow was randomly halted during locomotion, we observed a positive error signal in L2/3 neuronal activity and a negative error signal in L5 (Fig. 7c). This discrepancy arises from their roles, as postulated by our self-supervised learning model: L2/3 predicts the upcoming visual flow while L5 encodes the actual visual flow. Hence, when visual flow halts and given top-down running speed, L2/3 predicts a non-zero flow, thereby resulting in a positive prediction error. L5, in contrast, encodes the actual zero flow, and, hence, generates a negative prediction error due to the (positive) prediction provided by L2/3 input. Both L2/3 and L5 mismatch errors in our

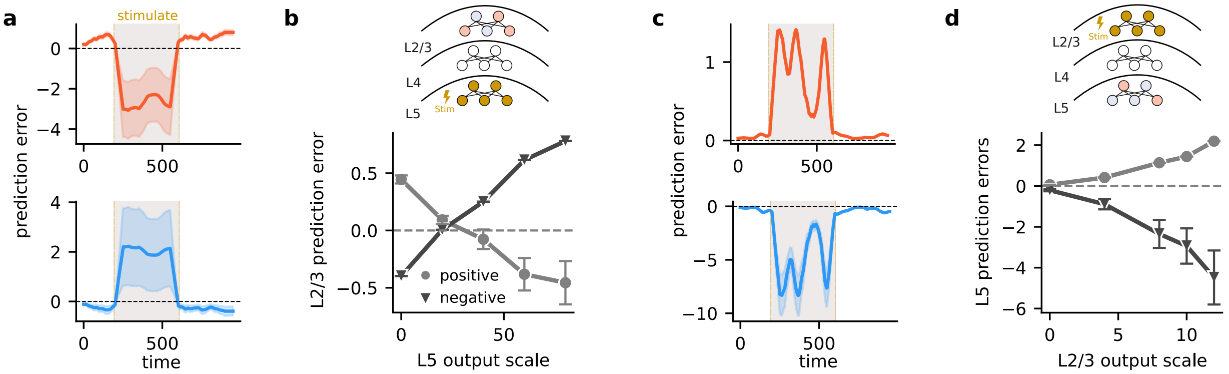


Figure 8. Differential role of L5 and L2/3 activation in generating mismatch prediction errors. a, Stimulating L5 during the mismatch interval causes prediction errors in L2/3 to switch signs. Neurons with positive errors switch their error signal with scaling of L5 activity (top) and vice versa (bottom). **b,** Positive errors gradually shift towards negative and vice versa, demonstrating a direct relationship between L5 stimulation and L2/3 prediction error modulation. **c,** Stimulating L2/3 during the mismatch interval amplifies existing prediction errors within L5. **d,** Plot demonstrates a proportional increase or decrease in prediction error magnitude as the output of L2/3 is scaled.

model are consistent with mismatch responses observed experimentally (Fig. 7c). Furthermore, we found that the mismatch responses in our model scale linearly with the running speed (Fig. 7d), in line with experiments [19,49].

To further test whether the mismatch errors across neurons resemble those that have been found experimentally [19], we analyzed the distribution of prediction errors predicted by our model. Our results show that the majority of L2/3 neurons exhibit positive mismatch responses while the majority of L5 neurons exhibit negative mismatch responses (Fig. 7e; see also Fig. S7 for prediction errors across different model conditions), again in line with Jordan and Keller [19].

Differential role of L2/3 and L5 during sensorimotor mismatches

We have shown that the prediction-error responses observed in L2/3 and L5 of awake behaving animals [19] can be explained by the different roles played by cortical layers equipped with self-supervised predictive learning. This suggests that layer-specific manipulations, for example through simulated optogenetic stimulation, of L2/3 and L5 neurons may reveal their unique functions. To test this in our model, we selectively increase the neuronal response of either L2/3 or L5 neurons during sensorimotor mismatches to determine their contributions to the prediction errors in the other layer, L5 and L2/3, respectively.

To this end, we first scale the output of L5 neurons. Upon moderate scaling, the the model predicts that both positive and negative mismatch errors in L2/3 should decrease in magnitude (Fig. 8b). By increasing the L5 activity further, the neurons that displayed a positive mismatch error before scaling, now switch their sign and become negative mismatch error neurons. Similarly, the neurons that displayed a negative mismatch error before scaling, now signal a positive mismatch error (Fig. 8a,b). These results are a consequence of L5 neurons representing zero input during the sensorimotor mismatch period while L2/3 neurons predict a non-zero visual flow as a result of non-zero topdown. By increasing the L5 output activity, eventually, the feedback signal from L5 becomes larger than the prediction in L2/3, which leads to a change in the signs of prediction errors in L2/3.

Thus far, we have scaled the L5 output altogether, without distinguishing between neurons exhibiting a positive or negative mismatch error. When we scaled only the outputs of neurons exhibiting a positive mismatch in L5, the changes in L2/3 mismatch errors were heterogeneous (Fig. S8). This is likely due to the low number of neurons with positive mismatch errors in L5, limiting their impact on L2/3. In contrast, scaling only the output of the L5 neurons exhibiting negative mismatch errors reverses the mismatch errors in L2/3 (Fig. S8), similar to the impact of scaling the L5 output altogether (Fig. 8b).

Next, we studied how changes in L2/3 affect the mismatch errors in L5. Scaling the output of L2/3 neurons altogether enhances the mismatch errors in L5. That is, neurons exhibiting a negative mismatch error before scaling now show even stronger negative mismatch errors after L2/3 activity scaling. Likewise, neurons exhibiting a positive mismatch error before scaling, now exhibit even stronger positive mismatch errors after scaling (Fig. 8c,d). This remains true even when we only scale the neurons with a positive mismatch error in L2/3 (Fig. S8). However, scaling only the L2/3 neurons with a negative mismatch error reverses the mismatch errors in L5 (Fig. S8). Hence, our model predicts a dynamic interplay between L2/3 and L5, with L2/3 stimulation generally amplifying L5 mismatch errors. Interestingly, this effect is primarily driven by L2/3 neurons signaling positive mismatch errors, while stimulating the negative mismatch neurons in L2/3 reverses the direction of prediction errors in L5.

These results can be best explained by the asymmetric contribution of L5 and L2/3 in generating mismatch errors. The strong effect of neurons with a positive mismatch error in L2/3 aligns with their role in predicting sensory input; increasing their activity strengthens this prediction signal within L5. Conversely, neurons with a negative mismatch error in L2/3 likely represent neurons suppressed by a greater-than-expected input. Enhancing their activity during a mismatch further emphasizes this "less than expected" signal, ultimately reversing the sign of the prediction error in L5.

Overall, these targeted manipulations of layer-specific mismatch signals offer a means by which to dissect the distinct functional roles of L5 and L2/3 populations. This approach could be further explored experimentally to advance our understanding of how neocortical layers contribute to predictive learning of sensory streams.

Discussion

Inspired by a refreshed view of the canonical neocortical circuitry and modern self-supervised learning algorithms, we introduce a computational theory wherein L2/3 learns to anticipate incoming sensory input. We demonstrated L2/3's capacity to predict incoming sensory information using temporal-contextual tasks. As a result, L2/3 develops latent sensory representations that are resilient to sensory noise and occlusions, improving the ability of cortical networks to encode partially observable information. Additionally, the proposed optimization leads to layer-specific sparsity, in line with experimental findings. Subsequently, by employing a sensorimotor task, we reveal that the model's prediction errors align with L2/3 and L5 mismatch responses observed in awake, behaving mice. Finally, using manipulations we generated predictions for the role of specific circuit elements in self-supervised predictive learning.

Our study focuses on the canonical L4-L2/3-L5 three-layered motif[17,18]. This classical view of the neocortical microcircuit emphasizes the feedforward flow of information across layers. However feedback projections are also evident[46,47] and both anatomical and electrophysiological data suggest the existence of direct thalamic input onto L5 pyramidal cells that effectively bypasses this feedforward circuity[22,23,25,26,51]. Our model explores the computational significance of these pathways by mapping them onto a self-supervised learning framework. Our results suggests that the two parallel thalamic pathways play critical, yet distinct roles: the L4-L2/3 pathway makes a temporal prediction, while the thalamic-L5 pathway provides the self-supervised target (i.e. incoming sensory input) with which to test this prediction. Feedback from L5-to-L2/3 connects the two parallel systems to guide learning.

Our model proposes that a critical feature of L2/3's integrative capacity is to use past information to predict incoming sensory input. This aligns with experimental observations showing that superficial layers display stronger temporal integration of sensory information compared to deep layers[52,53]. Our work indicates that the delay introduced by the thalamic-L4-L2/3 pathway is critical for the emergence of these properties. It also predicts that the delay introduced by these neurons and synapses sets the predictive time scale achievable by L2/3 neurons. Since this delay is approximately 10-20 milliseconds[26,54], this suggests that the temporal horizon for prediction in L2/3 of primary sensory cortices is limited. However, the cortex is highly hierarchical. Our framework, combined with higher-order top-down inputs, may provide a circuit-based explanation for why higher-order brain areas show longer timescales[55,56]. For example, the secondary visual cortex (V2) also gets primary thalamic projections, indicating that V2 could introduce further delays in the sensory input via cortico-cortical projections onto the L4→L2/3 pathway. These delays could then be compared with the incoming sensory input in L5, allowing the brain to learn predictive representations over hundreds of milliseconds or even longer.

It has been well documented that superficial layers respond more sparsely to sensory stimuli than deep layers[41]. However, it was not known how this feature emerges. In our model, layer-specific sparsity occurs naturally due to the proposed predictive function of L2/3 (Fig. 5). Our results also imply that simple measures like sparsity can help infer the optimization/learning processes of various brain structures. Indeed, we show how selective ablations of individual components of the network alter sparsity in a layer- and pathway-specific manner. While these findings provide important insights into the nature of sparsity in cortical networks, fully understanding the interplay between input, connectivity, and the emergence of sparsity during learning requires further investigation.

While our model has been mapped onto the canonical six-layered structure of the neocortex, it operates with only three layers. This raises an intriguing consideration: the evolutionarily conserved three-layered structures found in other brain regions, such as the hippocampus and piriform cortex in mammals, as well as in the cortices of other species like turtles[57], may represent the foundational blueprint for self-supervised learning. This suggests that a three-layer structure might serve as the fundamental building block for self-supervised learning, which was subsequently elaborated upon throughout evolution, firstly by increasing the number of layers and then secondly by expanding the primary locus of self-supervised learning in L2/3. Consistent with this view, we also find that sparsity increases as the size of L2/3 increases (Fig. 5b). L2/3 is greatly expanded in human evolution, even when compared to other layers[43,44]. Our results indicate that such expansion may improve network function, including expanding the predictive capabilities of the human neocortex.

In our model, we derive prediction errors directly from optimization principles. Consequentially, our model predicts the need for feedback between neocortical layers that carries information about error signals. This provides a form of credit assignment within the neocortical microcircuit. Although models of the neocortex often disregarded feedback pathways, numerous experimental studies demonstrate their existence[46,47,58]. Despite this, these pathways are understudied and their organization is not well understood, in contrast to feedforward which often shows highly organized subnetwork architectures[59–61]. Our findings indicate that both structured, i.e. reciprocal, as well as sparse, random feedback enable learning, with the former potentially advantageous for more complex tasks. A further question is how feedback error signals may be computed in a biologically plausible manner. Recent work shows how this can be achieved using dendritic compartments and interneuron cell types[62–64]. Different interneuron sub-types, with distinct connectivity, control feedforward and feedback processing in the L2/3-L5 circuitry[65–68], which may underlie distinct aspects of the self-supervised learning proposed here.

While these studies suggest biological mechanisms through which self-supervised learning may emerge, is there evidence that it occurs within the brain? Recent experimental studies support the ability of the neocortex to perform self-supervised learning[69], while deep networks trained using self-supervised learning better capture experimentally observed representations compared to networks trained via supervised learning[13]. For example, training deep networks using self-supervised predictive error functions yields representations that resemble visual cortical features[8,16,70–73]. Taking a step towards understanding the underlying learning mechanisms, recent research has introduced a combination of Hebbian and predictive synaptic plasticity[12]. This body of work supports the notion that sensory cortices engage in self-supervised learning, yet the specific circuit-level computations facilitating this process have remained unclear. Our work ties self-supervised learning to specific neocortical layers, suggesting that L2/3 and L5 provide complementary roles for implementing self-supervised learning. Consistent with these findings, the L2/3-to-L5 pathway is highly conserved across cortical regions[74,75] and behavioral studies have highlighted its importance for learning[76]. In future work, it would be of interest to test our theory by performing layer-specific experiments[77,78].

Predictive coding has provided a framework by which to learn sensory representations in the brain[79–81]. It postulates that the brain learns an internal model of the world from sensory streams by directly updating neuronal dynamics through prediction errors[1,7]. In temporal predictive coding, the neural networks constantly attempt to predict the incoming stimulus. Lotter et al.[82] demonstrated that a deep convolutional network that is trained using predictive coding learned sensory representations useful for downstream tasks. This contrasts with our model, where prediction errors lead to plasticity, rather than changes in neuronal dynamics directly. Temporal predictive coding models are also often relatively abstract and do not consider how predictive coding is implemented. A notable exception is the work of Bastos et al.[1] in which it was proposed that L5 encodes input expectation while L2/3 encodes positive and negative prediction errors in separate populations. This is in contrast with our model in which L2/3 predicts the incoming input while L5 encodes the current sensory input and computes prediction errors locally, which in turn explains a range of experimental observations. A feature of our model which goes beyond existing predictive coding models is the fact that our model jointly predicts the incoming sensory input in L2/3 together with representing the current input in L5, in line with recent developments in deep learning[83]. In future work, it would be of interest to explicitly contrast our model with existing predictive coding frameworks.

Overall, our work suggests that circuit motifs found throughout the neocortex implement efficient self-supervised predictive learning in the brain.

Materials and Methods

We model the dynamics of the neocortical circuitry using a set of connected neuronal layers. The architecture consists of distinct layers corresponding to the L4, L2/3, and L5 layers of the neocortex. The connectivity between layers is all-to-all unless otherwise stated.

In our model, L4 neurons receive direct thalamic input, x_t, and their activity is given by

$$\mathbf{z}_4^t = \sigma(W_{\text{Thal.}\to L4} \cdot x_t) \tag{1}$$

where σ is the sigmoid function and $W_{\text{Thal.}\to L4}$ is the weight matrix that models the connectivity from the thalamus to all L4 neurons.

We then model the neuronal and synaptic delay introduced by L4 as information flows onto L2/3 by one timestep. This means that L2/3 integrates past inputs from L4 (i.e. from time step $t-1$) with top-down inputs from higher-order cortical areas at time step t, I_{td}^t. L2/3 is modelled as,

$$\mathbf{z}_{2/3}^t = \sigma(W_{L4\to L2/3} \cdot \mathbf{z}_4^{t-1} + W_{td\to L2/3} \cdot I_{td}^t) \tag{2}$$

where $\mathbf{z}_{23}$ is a vector with all neurons in L2/3 and $W_{L4\to L2/3}$ is the weight matrix from L4 to L2/3. As above, all neurons are subject to the sigmoid non-linearity σ. L5 receives direct thalamic input and L2/3 input. It is modeled as,

$$\mathbf{z}_5^t = \sigma(\alpha W_{L2/3\to L5} \cdot \mathbf{z}_{23}^t + W_{\text{Thal.}\to L5} \cdot x_t) \tag{3}$$

where $\mathbf{z}_5$ is a vector with all L5 neurons, $W_{L2/3\to L5}$ and $W_{\text{Thal.}\to L5}$ are the weight matrices from L2/3-to-L5 and thalamus-to-L5, respectively. α is a constant that models the dendritic-to-somatic attenuation of L2/3-to-L5 input. We set $\alpha = 0.3$, but the exact value does not qualitatively change our results.

In our network, the weight matrices $W_{L2/3\to L5}$, $W_{L4\to L2/3}$, $W_{\text{Thal.}\to L4}$, $W_{\text{Thal.}\to L5}$ and $W_{td\to L2/3}$ are subject to optimization through gradient descent. The learning rules for these connections follow cost/error functions inspired by those commonly used in machine self-supervised learning. In particular, we use a combination of two cost functions,

$$\mathcal{C}_{\text{total}} = \underbrace{\lambda_p \mathcal{C}_{L_{23}\to L_5}}_{\text{predictive}} + \underbrace{\lambda_r \mathcal{C}_{L_5}}_{\text{reconstruction}} \tag{4}$$

λ_p and λ_r are hyperparameters that scale the predictive and reconstruction costs, respectively. The first component of $\mathcal{C}_{\text{total}}$ is the temporal self-supervised cost in which L2/3 *predictions* given L4 input at time $t-1$ are compared with L5 inputs at t

$$\mathcal{C}_{L_{23}\to L_5} = \frac{1}{2}(\mathbf{z}_5^t - \underbrace{W_{L_{23}\to L_5} \cdot \mathbf{z}_{23}^t}_{\text{prediction, } \hat{\mathbf{z}}_5^t})^2 \tag{5}$$

The second component of $\mathcal{C}_{\text{total}}$ encourages the model to learn non-trivial representations through *reconstruction* of L5 thalamic input given its own activity, as follows

$$\mathcal{C}_{L_5} = \frac{1}{2}(W_{\text{decoder}} \cdot \mathbf{z}_5^t - x_t)^2 \tag{6}$$

This reconstruction cost is only used to optimise $W_{\text{Thal.}\to L5}$ and W_{decoder} weights. To ensure this, we block the resulting error signals (i.e. gradients) from adjusting $W_{L2/3\to L5}$ weights and the remaining weights.

The key predictive learning rule is the one that governs $W_{L_{23}\to L_5}$ weights. It can be derived from the total cost as,

$$\begin{aligned}\Delta W_{L_{23}\to L_5} &= -\eta\left(\frac{\partial \mathcal{C}_{\text{total}}}{\partial W_{L_{23}\to L_5}}\right) \\ &= \eta(\mathbf{z}_5^t - W_{L_{23}\to L_5} \cdot \mathbf{z}_{23}^t)\mathbf{z}_{23}^t\end{aligned} \tag{7}$$

where η denotes the learning rate ($\eta = 0.001$). Note that in order to train our model efficiently we use batch update and one step of backpropagation to train $W_{L_4\to L_{23}}$ and we do not allow gradients to flow backwards in time.

Similarly, the learning rule for $W_{L_4\to L_{23}}$ is given by the derivative of the cost function with respect to this weight matrix,

$$\begin{aligned}\Delta W_{L_4\to L_{23}} &= -\eta\left(\frac{\partial\mathcal{C}_{\text{total}}}{\partial W_{L_4\to L_{23}}}\right)\\ &= -\eta\left(\frac{\partial\mathcal{C}_{\text{total}}}{\partial \mathbf{z}_{L23}}\frac{\partial \mathbf{z}_{L23}}{\partial W_{L4\to L2/3}}\right)\\ &= -\eta(\mathbf{z}_5^t - W_{L_{23}\to L_5}\cdot\mathbf{z}_{23}^t)\cdot W_{L_{23}\to L_5}^{T/\text{random}}\sigma'(W_{L_4\to L_{23}}\cdot\mathbf{z}_4^{t-1})\mathbf{z}_4^{t-1}\end{aligned} \tag{8}$$

where $W_{L_{23}\to L_5}^{T/\text{random}}$ is the transpose of $W_{L_{23}\to L_5}$ or a random matrix, depending on the experiments we performed (see Fig. 6).

Finally, the learning rules for $W_{\text{Thal.}\to L4}$ and $W_{\text{Thal.}\to L5}$ are given by,

$$\begin{aligned}\Delta W_{\text{Thal.}\to L_4} &= -\eta\left(\frac{\partial\mathcal{C}_{\text{total}}}{\partial W_{\text{Thal.}\to L_4}}\right)\\ &= -\eta\left(\frac{\partial\mathcal{C}_{\text{total}}}{\partial \mathbf{z}_{L23}}\frac{\partial \mathbf{z}_{L23}}{\partial \mathbf{z}_{L4}}\frac{\partial \mathbf{z}_{L4}}{\partial W_{\text{Thal.}\to L4}}\right)\\ \Delta W_{\text{Thal.}\to L_5} &= -\eta\left(\frac{\partial\mathcal{C}_{\text{total}}}{\partial \mathbf{z}_5}\frac{\partial \mathbf{z}_5}{\partial W_{\text{Thal.}\to L_5}}\right)\\ &= \eta(\mathbf{z}_5^t - W_{L_{23}\to L_5}\cdot\mathbf{z}_{23}^t)x_t\end{aligned} \tag{9}$$

Tasks

Gabor contextual-temporal task

In this task we generate synthetic sequential data. Each data point is a 28x28 Gabor patch with frequencies sampled from $\mathcal{N}(0.2, 0.1)$ with variability along x and y axis from $\mathcal{U}(3, 8)$ and fix orientations for each class with $\theta = [0, 18^\circ, 36^\circ, ..., 162^\circ]$. The top-down inputs can take values of $[-18^\circ, 0^\circ, +18^\circ]$. At each timestep, we randomly sample a datapoint x_t with orientations θ_i where i denotes the index number from the θ list, a top-down contextual input I_{td}^t and generate the next input x_{t+1} by sampling a datapoint with the orientation $\theta_i + I_{td}^t$. This setup means that for each input at timestep t, three subsequent orientations are possible, except angles 0° and $+162^\circ$, which only have two possible successors.

This task is designed to investigate how Gabor patches at t can be predicted when their orientation is determined by a top-down variable. This top-down variable takes values of -18°, 0°, or $+18^\circ$, dictating how a Gabor patch in the recent past (x_{t-1}) leads to a second (transformed) Gabor patch, x_t). This results in sequences where the orientation shifts to the left, shifts to the right, or remains the same (see examples in Fig. 2b).

According to our model, at each timestep, the L5 network receives x_t as its sensory input. Simultaneously, the L2/3 network processes the output of L4 (from previous timestep, $t-1$) combined with the top-down contextual input.

Noise and occlusion tests

These experiments assess the model's robustness to input degradation (Fig. 4). We focus on two forms of degradation: noise and occlusion. For noise, Gaussian noise is added to the input as $x_t^* = x_t + \lambda\mathcal{N}(0, 1)$, where λ scales the noise level (values from 0 to 1, in increments of 0.2). To examine the importance of the $L2/3 \to L5$ connection, we selectively disable the self-supervised cost during training. This prevents updates to L2/3 and L4 parameters through this loss, isolating the effects of this connection. Reconstruction performance across layers is measured using mean squared error between the reconstructed and the original (denoised) input.

For occlusion, random image sections are obscured with a dark patch (pixel values set to zero). After training, a Support Vector Machine is used to classify outputs of L2/3 and L5 based on the x_t label. Classification accuracy on a held-out test set indicates how well the model copes with occlusions.

Visuomotor task

Simulating the Experimental Setup: To closely replicate the visuomotor task from Jordan and Keller[19], we have developed a method to generate synthetic sensorimotor data with modelled visual flow and motor speed. In our model, each vector dimension encapsulates a distinct aspect of visual flow, which is essential for simulating the sensory inputs typical in motion perception tasks. In particular, visual flow was calculated as $\mathbf{x}_t$ at any given time t following

$$\mathbf{x}_t = f\left(s_t\right) + \epsilon_t$$

where, f denotes a function that converts speed into visual flow. We set $f(s_t) = s_t$ to model a linear relationship (Fig. 7,8) or $f(s_t) = \sin(s_t)$ to model a non-linear interaction (Fig. S7c). The term s_t represents the speed at time t. To mimic more realistic conditions we also add Gaussian noise, $\epsilon_t = \mathcal{N}(0, 1)$. In our simulations, we model speed following a random walk. At each timestep, the speed s_t is determined with equal probability between the following options

- Decreasing by 1: $s_t = s_{t-1} - 1$
- Remaining the same: $s_t = s_{t-1}$
- Increasing by 1: $s_t = s_{t-1} + 1$

This approach simulates the natural fluctuations of running speed, where an individual might slightly accelerate, decelerate, or maintain pace from moment to moment. After sampling speeds, we generate visual flow input $x_{t-1} = f(s_{t-1})$ and $x_t = f(s_t)$, with noise as defined above.

The speed variable provides top-down context, an important factor in our model that provides contextual information which help the model (L2/3) to predict the incoming visual flow given past visual flow and the current speed.

Mismatch simulation and measurement: After training, we obtain a baseline error signal by averaging the *prediction error* for each neuron across the dataset. Next, we simulate a mismatch (i.e. breaking the coupling between locomotion and visual feedback) by randomly setting the visual flow input to zero. Each mismatch period lasted for K timesteps (K=600 timesteps). We subsequently record the average prediction error signals for each neuron under simulated sensorimotor mismatches. To isolate the mismatch response for each neuron z_i, we use the formula: $PE_i = PE_{i,baseline} - PE_{i,mismatch}$. Here, PE_i denotes the prediction errors, modeled as described below.

L2/3 and L5 prediction errors: The prediction errors for L2/3 and L5 that we analyse in Figs. 7 and 8 were calculate using the gradients of the cost function with respect to L2/3 and L5 neurons, respectively. Therefore, L2/3 prediction errors are calculated as:

$$PE_{L_{23}} = -\frac{\partial \mathcal{C}_{\text{total}}}{\partial \mathbf{z}_{L23}} \tag{10}$$

Whereas the L5 prediction errors are calculated as:

$$PE_{L_5} = -\frac{\partial \mathcal{C}_{\text{total}}}{\partial \mathbf{z}_{L_5}} \tag{11}$$

Note that because these prediction errors were calculated after learning converged, this means that the reconstruction cost was effectively zero.

Sparsity metric

To measure activity sparsity of each layer in our model, we used the Treves-Rolls metric[42,84]. The population sparseness, S, of each layer for a single stimulus was measured as:

$$S = \frac{[\sum_{i=1}^{N} r_i / N]^2}{\sum_{i=1}^{N} [r_i^2 / N]}$$

where N is the number of neurons, and r_i the firing rate of neuron i. To get the average population sparseness for the entire sequence, we simply average S over a trial.

Feedback and feedforward connection probabilities

In our work, we have tested the importance of feedback from L5 to L2/3 for learning. As part of this, we tested a range of connection probabilities P for this feedback pathway. The feedback connection from L5 to L2/3 are dropped with the probability of $(1 - P_{\text{connectivity}}$, such that when the connection probability are set to 0, all the feedback connections are set to 0. . The connection probability of all forward connections was set to 1.

Data availability

Data to reproduce the simulated data can be generated using our code (see below).

Code availability

The source code for the model proposed here and the respective analysis is available at https://github.com/neuralml/neoSSL.

Acknowledgements

We would like to thank the Neural & Machine Learning group, Randy Bruno, Nicol Harper, Andrew King and Katherine Willard for useful feedback. K.K.N. was funded by the UKRI Centre for Doctoral Training in Interactive Artificial, L.H. by the DFG (460088091), P.G.A by the Academy of Medical Sciences (SBF006\1047), Brain and Behaviour Research Foundation (YI Award 29300) and BBSRC (BB/X016331/1) and R.P.C. by the Medical Research Council (MR/X006107/1), BBSRC (BB/X013340/1) and a ERC-UKRI Frontier Research Guarantee Grant (EP/Y027841/1). We would like to thank Shuzo Sakata and Kenneth Harris for sharing their layer-specific sparsity data. This work made use of the HPC system Blue Pebble at the University of Bristol, UK. We would like to thank Dr Stewart for a donation that supported the purchase of GPU nodes embedded in the Blue Pebble HPC system.

Author contributions

K.K.N. developed the computational framework with guidance from L.H. and R.P.C. K.K.N. performed all simulations and data analysis. K.K.N., P.A, L.H. and R.P.C. jointly wrote the manuscript. R.P.C supervised the project with help from L.H.

References

[1] A. M. Bastos, W. M. Usrey, R. A. Adams, G. R. Mangun, P. Fries, and K. J. Friston. Canonical microcircuits for predictive coding. *Neuron*, 76(4):695–711, 2012.

[2] H. Barlow. What is the computational goal of the neocortex. *Large-scale neuronal theories of the brain*, pages 1–22, 1994.

[3] A. Banerjee, R. V. Rikhye, and A. Marblestone. Reinforcement-guided learning in frontal neocortex: emerging computational concepts. *Current Opinion in Behavioral Sciences*, 38:133–140, 2021.

[4] H. Adesnik and A. Naka. Cracking the function of layers in the sensory cortex. *Neuron*, 100(5):1028–1043, 2018.

[5] H. B. Barlow. Unsupervised learning. *Neural computation*, 1(3):295–311, 1989.

[6] P. Dayan, M. Sahani, and G. Deback. Unsupervised learning. *The MIT encyclopedia of the cognitive sciences*, pages 857–859, 1999.

[7] Y. Huang and R. P. Rao. Predictive coding. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(5):580–593, 2011.

[8] A. v. d. Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[9] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33: 21271–21284, 2020.

[10] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[11] X. Chen and K. He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021.

[12] M. S. Halvagal and F. Zenke. The combination of hebbian and predictive plasticity learns invariant object representations in deep sensory networks. *Nature Neuroscience*, 26(11):1906–1915, 2023.

[13] C. Zhuang, S. Yan, A. Nayebi, M. Schrimpf, M. C. Frank, J. J. DiCarlo, and D. L. Yamins. Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3), 2021.

[14] T. Konkle and G. A. Alvarez. A self-supervised domain-general learning framework for human ventral stream representation. *Nature communications*, 13(1):491, 2022.

[15] R. Geirhos, K. Narayanappa, B. Mitzkus, T. Thieringer, M. Bethge, F. A. Wichmann, and W. Brendel. Partial success in closing the gap between human and machine vision. *Advances in Neural Information Processing Systems*, 34:23885–23899, 2021.

[16] A. Nayebi, R. Rajalingham, M. Jazayeri, and G. R. Yang. Neural foundations of mental simulation: Future prediction of latent representations on dynamic scenes. *arXiv preprint arXiv:2305.11772*, 2023.

[17] R. J. Douglas, K. A. Martin, and D. Whitteridge. A canonical microcircuit for neocortex. *Neural computation*, 1(4):480–488, 1989.

[18] K. D. Harris and G. M. Shepherd. The neocortical circuit: themes and variations. *Nature neuroscience*, 18(2):170–181, 2015.

[19] R. Jordan and G. B. Keller. Opposing influence of top-down and bottom-up input on excitatory layer 2/3 neurons in mouse primary visual cortex. *Neuron*, 108(6):1194–1206, 2020.

[20] S. Zhang, M. Xu, T. Kamigaki, J. P. Hoang Do, W.-C. Chang, S. Jenvay, K. Miyamichi, L. Luo, and Y. Dan. Long-range and local circuits for top-down modulation of visual cortex processing. *science*, 345(6197):660–665, 2014.

[21] S. Zhang, M. Xu, W.-C. Chang, C. Ma, J. P. Hoang Do, D. Jeong, T. Lei, J. L. Fan, and Y. Dan. Organization of long-range inputs and outputs of frontal cortex for top-down control. *Nature neuroscience*, 19(12):1733–1742, 2016.

[22] K. F. Jensen and H. P. Killackey. Terminal arbors of axons projecting to the somatosensory cortex of the adult rat. i. the normal morphology of specific thalamocortical afferents. *Journal of Neuroscience*, 7(11):3529–3543, 1987.

[23] R. J. Douglas and K. Martin. A functional microcircuit for cat visual cortex. *The Journal of physiology*, 440(1):735–769, 1991.

[24] J. H. Maunsell and J. R. Gibson. Visual response latencies in striate cortex of the macaque monkey. *journal of Neurophysiology*, 68(4):1332–1344, 1992.

[25] L. Petreanu, T. Mao, S. M. Sternson, and K. Svoboda. The subcellular organization of neocortical excitatory connections. *Nature*, 457(7233):1142–1145, 2009.

[26] C. M. Constantinople and R. M. Bruno. Deep cortical layers are activated directly by thalamus. *Science*, 340(6140):1591–1594, 2013.

[27] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[28] P. J. Sjöström and M. Häusser. A cooperative switch determines the sign of synaptic plasticity in distal dendrites of neocortical pyramidal neurons. *Neuron*, 51(2):227–238, 2006.

[29] D. L. Ringach. Spatial structure and symmetry of simple-cell receptive fields in macaque primary visual cortex. *Journal of neurophysiology*, 88(1):455–463, 2002.

[30] J. P. Gavornik and M. F. Bear. Learned spatiotemporal sequence recognition and prediction in primary visual cortex. *Nature neuroscience*, 17(5):732–737, 2014.

[31] M. Leinweber, D. R. Ward, J. M. Sobczak, A. Attinger, and G. B. Keller. A sensorimotor circuit in mouse cortex for visual flow predictions. *Neuron*, 95(6):1420–1432, 2017.

[32] R. Zhang, P. Isola, and A. A. Efros. Colorful image colorization. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 649–666. Springer, 2016.

[33] P. Bachman, R. D. Hjelm, and W. Buchwalter. Learning representations by maximizing mutual information across views. *Advances in neural information processing systems*, 32, 2019.

[34] S. Pluta, A. Naka, J. Veit, G. Telian, L. Yao, R. Hakim, D. Taylor, and H. Adesnik. A direct translaminar inhibitory circuit tunes cortical output. *Nature neuroscience*, 18(11):1631–1640, 2015.

[35] R. J. Rabinovich, D. D. Kato, and R. M. Bruno. Learning enhances encoding of time and temporal surprise in mouse primary sensory cortex. *Nature Communications*, 13(1):5504, 2022.

[36] W. E. Vinje and J. L. Gallant. Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, 287(5456):1273–1276, 2000.

[37] B. A. Olshausen and D. J. Field. Sparse coding of sensory inputs. *Current opinion in neurobiology*, 14(4):481–487, 2004.

[38] S. Crochet, J. F. Poulet, Y. Kremer, and C. C. Petersen. Synaptic mechanisms underlying sparse coding of active touch. *Neuron*, 69(6):1160–1175, 2011.

[39] A. L. Barth and J. F. Poulet. Experimental evidence for sparse firing in the neocortex. *Trends in neurosciences*, 35(6):345–355, 2012.

[40] E. Froudarakis, P. Berens, A. S. Ecker, R. J. Cotton, F. H. Sinz, D. Yatsenko, P. Saggau, M. Bethge, and A. S. Tolias. Population code in mouse v1 facilitates readout of natural scenes through increased sparseness. *Nature neuroscience*, 17(6):851–857, 2014.

[41] S. Sakata and K. D. Harris. Laminar structure of spontaneous and sensory-evoked population activity in auditory cortex. *Neuron*, 64(3):404–418, 2009.

[42] B. Willmore and D. J. Tolhurst. Characterizing the sparseness of neural codes. *Network: Computation in Neural Systems*, 12(3):255, 2001.

[43] J. Berg, S. A. Sorensen, J. T. Ting, J. A. Miller, T. Chartrand, A. Buchin, T. E. Bakken, A. Budzillo, N. Dee, S.-L. Ding, et al. Human neocortical expansion involves glutamatergic neuron diversification. *Nature*, 598(7879):151–158, 2021.

[44] A. Galakhova, S. Hunt, R. Wilbers, D. B. Heyer, C. de Kock, H. D. Mansvelder, and N. A. Goriounova. Evolution of cortical neurons supporting human cognition. *Trends in cognitive sciences*, 26(11):909–922, 2022.

[45] N. A. Cayco-Gajic, C. Clopath, and R. A. Silver. Sparse synaptic connectivity is required for decorrelation and pattern separation in feedforward networks. *Nature communications*, 8(1):1116, 2017.

[46] X. Jiang, S. Shen, C. R. Cadwell, P. Berens, F. Sinz, A. S. Ecker, S. Patel, and A. S. Tolias. Principles of connectivity among morphologically defined cell types in adult neocortex. *Science*, 350(6264):aac9462, 2015.

[47] T. A. Hage, A. Bosma-Moody, C. A. Baker, M. B. Kratz, L. Campagnola, T. Jarsky, H. Zeng, and G. J. Murphy. Synaptic connectivity to l2/3 of primary visual cortex measured by two-photon optogenetic stimulation. *Elife*, 11:e71103, 2022.

[48] T. P. Lillicrap, D. Cownden, D. B. Tweed, and C. J. Akerman. Random synaptic feedback weights support error backpropagation for deep learning. *Nature communications*, 7(1):13276, 2016.

[49] G. B. Keller, T. Bonhoeffer, and M. Hübener. Sensorimotor mismatch signals in primary visual cortex of the behaving mouse. *Neuron*, 74(5):809–815, 2012.

[50] Z. Padamsey and N. L. Rochefort. Defying expectations: how neurons compute prediction errors in visual cortex. *Neuron*, 108 (6):1016–1019, 2020.

[51] A. Agmon and B. Connors. Thalamocortical responses of mouse somatosensory (barrel) cortexin vitro. *Neuroscience*, 41(2-3): 365–379, 1991.

[52] A. Pitas, A. L. Albarracín, M. Molano-Mazón, and M. Maravall. Variable temporal integration of stimulus patterns in the mouse barrel cortex. *Cerebral Cortex*, 27(3):1758–1764, 2017.

[53] A. Ayaz, A. Stäuble, M. Hamada, M.-A. Wulf, A. B. Saleem, and F. Helmchen. Layer-specific integration of locomotion and sensory information in mouse barrel cortex. *Nature communications*, 10(1):2585, 2019.

[54] S. Vanni, H. Hokkanen, F. Werner, and A. Angelucci. Anatomy and physiology of macaque visual cortical areas v1, v2, and v5/mt: bases for biologically realistic models. *Cerebral Cortex*, 30(6):3483–3517, 2020.

[55] U. Hasson, E. Yang, I. Vallines, D. J. Heeger, and N. Rubin. A hierarchy of temporal receptive windows in human cortex. *Journal of Neuroscience*, 28(10):2539–2550, 2008.

[56] D. L. Yamins and J. J. DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19 (3):356–365, 2016.

[57] G. M. Shepherd. The microcircuit concept applied to cortical evolution: from three-layer to six-layer cortex. *Frontiers in neuroanatomy*, 5:30, 2011.

[58] S. Lefort, C. Tomm, J.-C. F. Sarria, and C. C. Petersen. The excitatory neuronal network of the c2 barrel column in mouse primary somatosensory cortex. *Neuron*, 61(2):301–316, 2009.

[59] Y. Yoshimura, J. L. Dantzker, and E. M. Callaway. Excitatory cortical neurons form fine-scale functional networks. *Nature*, 433 (7028):868–873, 2005.

[60] N. A. Morgenstern, J. Bourg, and L. Petreanu. Multilaminar networks of cortical neurons integrate common inputs from sensory thalamus. *Nature neuroscience*, 19(8):1034–1040, 2016.

[61] T. Otsuka and Y. Kawaguchi. Cortical inhibitory cell types differentially form intralaminar and interlaminar subnetworks withexcitatory neurons. *Journal of Neuroscience*, 29(34):10533–10540, 2009.

[62] J. Sacramento, R. Ponte Costa, Y. Bengio, and W. Senn. Dendritic cortical microcircuits approximate the backpropagation algorithm. *Advances in neural information processing systems*, 31, 2018.

[63] B. A. Richards, T. P. Lillicrap, P. Beaudoin, Y. Bengio, R. Bogacz, A. Christensen, C. Clopath, R. P. Costa, A. de Berker, S. Ganguli, et al. A deep learning framework for neuroscience. *Nature neuroscience*, 22(11):1761–1770, 2019.

[64] W. Greedy, H. W. Zhu, J. Pemberton, J. Mellor, and R. Ponte Costa. Single-phase deep learning in cortico-cortical networks. *Advances in Neural Information Processing Systems*, 35:24213–24225, 2022.

[65] P. G. Anastasiades, A. Marques-Smith, D. Lyngholm, T. Lickiss, S. Raffiq, D. Kätzel, G. Miesenböck, and S. J. Butt. Gabaergic interneurons form transient layer-specific circuits in early postnatal neocortex. *Nature communications*, 7(1):10584, 2016.

[66] T. K. Berger, R. Perin, G. Silberberg, and H. Markram. Frequency-dependent disynaptic inhibition in the pyramidal network: a ubiquitous pathway in the developing rat neocortex. *The Journal of physiology*, 587(22):5411–5425, 2009.

[67] A. Naka, J. Veit, B. Shababo, R. K. Chance, D. Risso, D. Stafford, B. Snyder, A. Egladyous, D. Chu, S. Sridharan, et al. Complementary networks of cortical somatostatin interneurons enforce layer specific control. *Elife*, 8:e43696, 2019.

[68] A. J. Apicella, I. R. Wickersham, H. S. Seung, and G. M. Shepherd. Laminarly orthogonal excitation of fast-spiking and low-threshold-spiking interneurons in mouse motor cortex. *Journal of Neuroscience*, 32(20):7021–7033, 2012.

[69] L. Zhong, S. Baptista, R. Gattoni, J. Arnold, D. Flickinger, C. Stringer, and M. Pachitariu. Distinct streams for supervised and unsupervised learning in the visual cortex. *bioRxiv*, pages 2024–02, 2024.

[70] Y. Singer, Y. Teramoto, B. D. Willmore, J. W. Schnupp, A. J. King, and N. S. Harper. Sensory cortex is optimized for prediction of future input. *elife*, 7:e31557, 2018.

[71] S. Bakhtiari, P. Mineault, T. Lillicrap, C. Pack, and B. Richards. The functional specialization of visual cortex emerges from training parallel pathways with self-supervised predictive learning. *Advances in Neural Information Processing Systems*, 34, 2021.

[72] A. Nayebi, N. C. Kong, C. Zhuang, J. L. Gardner, A. M. Norcia, and D. L. Yamins. Unsupervised models of mouse visual cortex. *bioRxiv*, 2021.

[73] Y. Singer, L. Taylor, B. D. Willmore, A. J. King, and N. S. Harper. Hierarchical temporal prediction captures motion processing along the visual pathway. *Elife*, 12:e52599, 2023.

[74] B. Hooks, S. A. Hires, Y.-X. Zhang, D. Huber, L. Petreanu, K. Svoboda, and G. M. Shepherd. Laminar analysis of excitatory local circuits in vibrissal motor and sensory cortical areas. *PLoS biology*, 9(1):e1000572, 2011.

[75] D. P. Collins, P. G. Anastasiades, J. J. Marlin, and A. G. Carter. Reciprocal circuits linking the prefrontal cortex with dorsal and ventral thalamic nuclei. *Neuron*, 98(2):366–379, 2018.

[76] T. Otsuka and Y. Kawaguchi. Pyramidal cell subtype-dependent cortical oscillatory activity regulates motor learning. *Communications Biology*, 4(1):495, 2021.

[77] C. A. Olman, N. Harel, D. A. Feinberg, S. He, P. Zhang, K. Ugurbil, and E. Yacoub. Layer-specific fmri reflects different neuronal computations at different depths in human v1. *PloS one*, 7(3):e32536, 2012.

[78] P. Kok, L. J. Bains, T. van Mourik, D. G. Norris, and F. P. de Lange. Selective activation of the deep layers of the human primary visual cortex by top-down feedback. *Current Biology*, 26(3):371–376, 2016.

[79] R. P. Rao and D. H. Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1):79–87, 1999.

[80] G. B. Keller and T. D. Mrsic-Flogel. Predictive processing: a canonical cortical computation. *Neuron*, 100(2):424–435, 2018.

[81] R. P. Rao, D. C. Gklezakos, and V. Sathish. Active predictive coding: A unified neural framework for learning hierarchical world models for perception and planning. *arXiv preprint arXiv:2210.13461*, 2022.

[82] W. Lotter, G. Kreiman, and D. Cox. Deep predictive coding networks for video prediction and unsupervised learning. *arXiv preprint arXiv:1605.08104*, 2016.

[83] R. Balestriero and Y. LeCun. Learning by reconstruction produces uninformative features for perception. *arXiv preprint arXiv:2402.11337*, 2024.

[84] A. Treves and E. T. Rolls. What determines the capacity of autoassociative memories in the brain? *Network: Computation in Neural Systems*, 2(4):371, 1991.

Supplementary material

Experimental details

We model each layer of the neocortical microcircuit as a linear transformation of their input followed by a non-linearity. The network layers are specified with the following default neuron counts: Layer 4 (L4) and Layer 2/3 (L2/3) each have 128 neurons, and Layer 5 (L5) has 16 neurons. In some experiments, we varied the number of neurons in each layer to explore the effect of neuronal density on the outcomes being evaluated. For experiments aligning with those conducted by Jordan and Keller[19], we adjusted the neuron counts in L2/3 and L5 to 32 and 16, respectively, to more closely match the proportion of neurons recorded from each layer during their experiments.

Network parameters are optimized using the standard ADAM optimizer with a learning rate of 0.001, and beta parameters (beta1 = 0.9, beta2 = 0.999). To achieve efficient learning, we employ standard stochastic gradient descent with a batch size of 32 and continue training until convergence, typically around 1000 epochs.

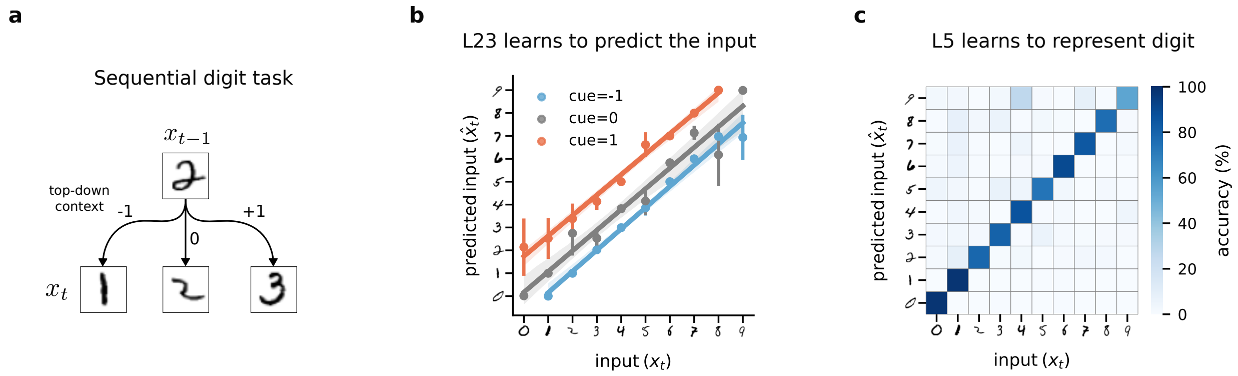
ANN parameters are initialized using an uniform distribution $\mathcal{U}(-\sqrt{k}, \sqrt{k})$, where k represents the number of neurons per layer. We use backpropagation for network training, except in experiments shown in Figure 6, where we explore alternative ways of setting the feedback weights. All results reflect an average across 5 random seeds.

Experiments were performed on the BluePebble supercomputer (University of Bristol), primarily using GeForce RTX 2080 Ti GPUs, with occasional CPU usage.

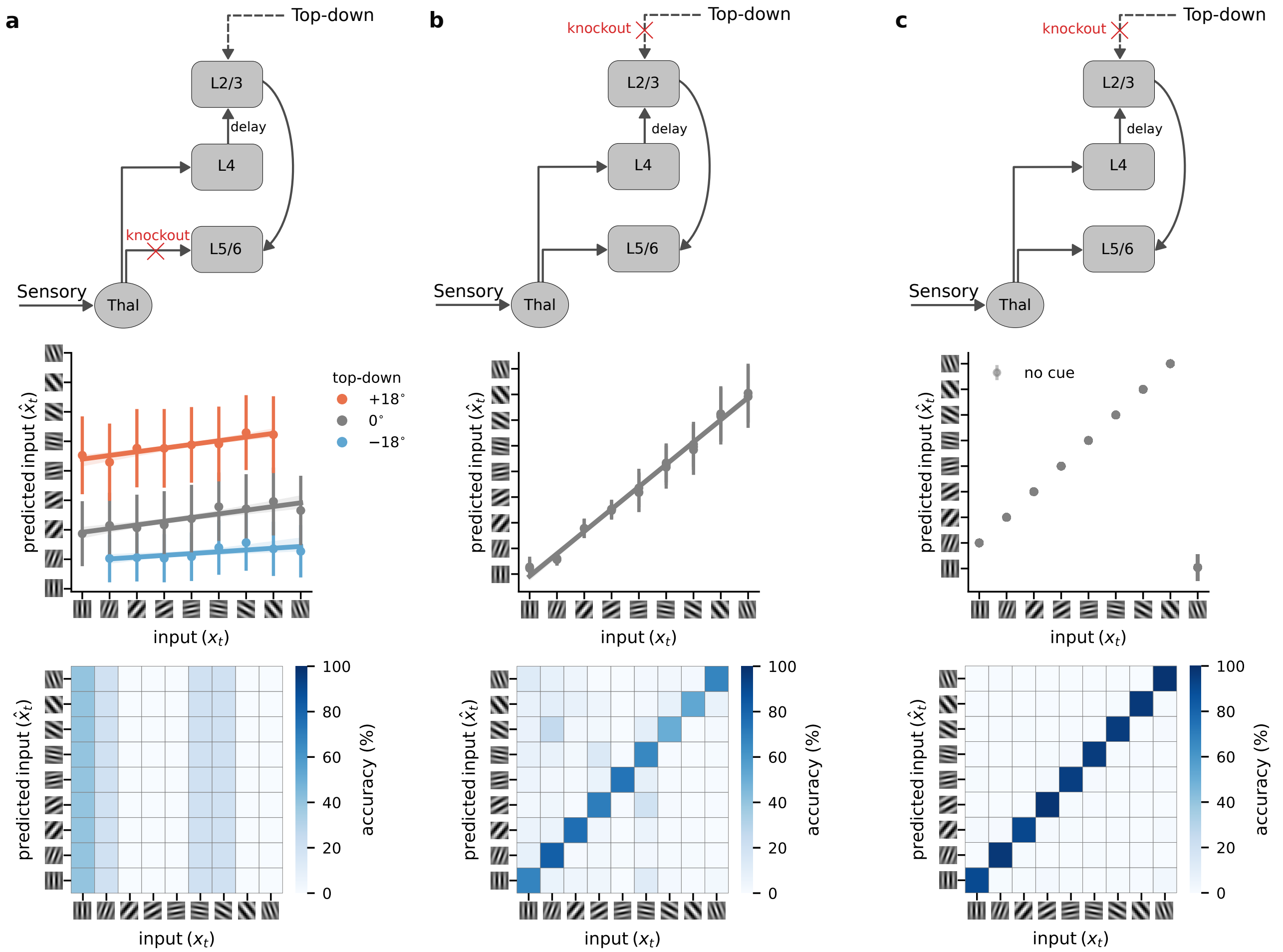
L23-to-L5 feedback experiments

This experiment investigates the importance of the feedback connection from L5 to L2/3 in model performance. Instead of using the optimal feedback weight matrix ($W_{L5\rightarrow L2/3} = W^T_{L2/3\rightarrow L5}$), as derived using the backpropagation algorithm, we replace it with random weights in line with a variant of backpropagation known as feedback alignment[48]. In addition, we introduce a probability variable (P) to control the density of the feedback connections ($W_{L5\rightarrow L2/3}$) compared to the forward connections ($W_{L2/3\rightarrow L5}$). Each feedback connection has a probability $1-P$ of being set to zero. For the results shown in figs$_N N/.\, 6 and S6 wecreatedmultiplemodelvariants, eachwithvaryingfeedbackconnectiondensitiesdeterminedbythisprobability.$

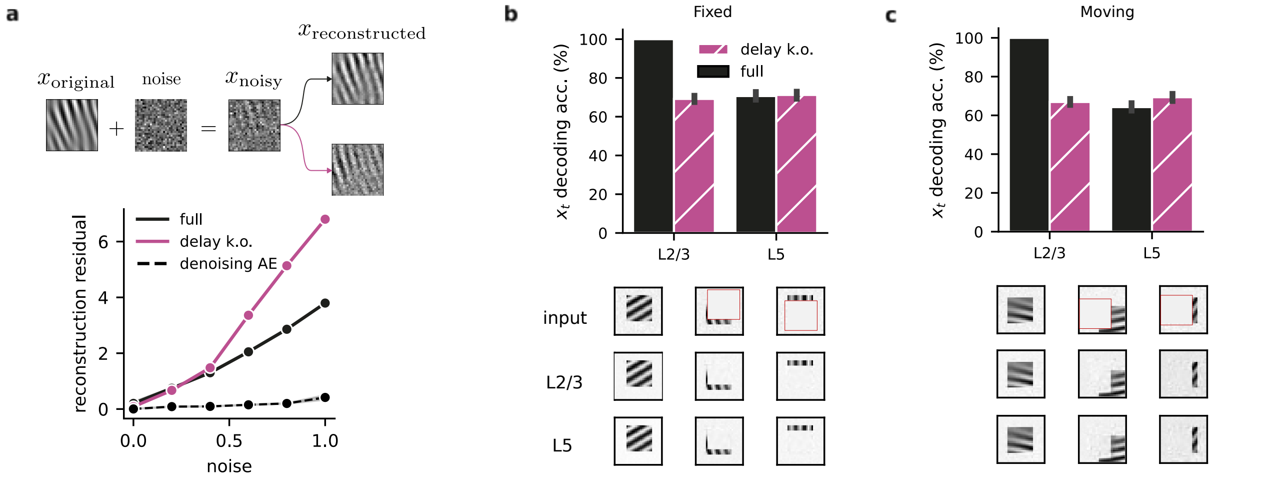
Supplementary figures



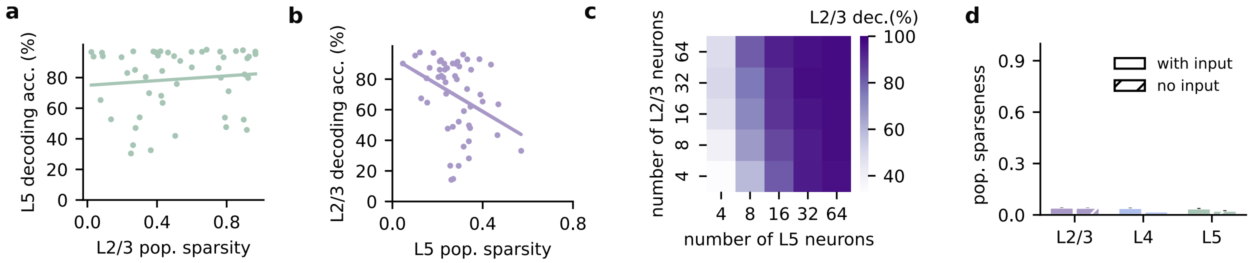
Supplementary Figure S1. Learning a sequential digit task in a self-supervised canonical microcircuit. a, Sequential digit task used for training. The generative factor which is passed as top-down context determines the next digit. **b,** Prediction accuracy of a linear model trained on the output of L2/3. For a given input, L2/3 predicts the next possible input with high accuracy for all three cues (denoted by different colors). **c,** Confusion matrix showing the classification accuracy of a linear model trained on the output of L5.



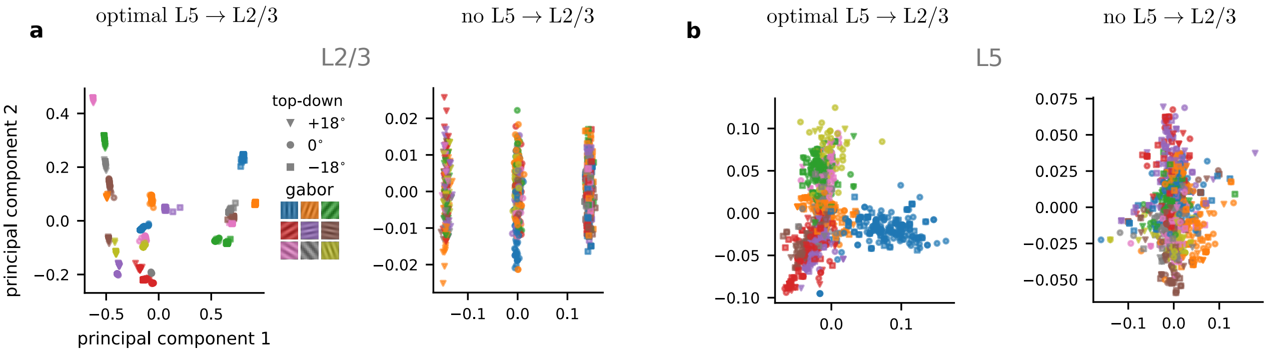
Supplementary Figure S2. Prediction accuracy in L2/3 and L5 when connections are knocked-out. a, Connections from Thalamus to L5 are necessary for learning in both L2/3 and L5. **b,** Top-down input to L2/3 is crucial for L2/3 to predict the incoming input, while L5 performance is not significantly impacted when the top-down signal is deleted. **c,** Removal of top-down input to L2/3 in a deterministic task in which inputs always rotate clockwise (i.e. $+18°$), top-down input is not required for task. In this case removing the top-down does not have any effect on the task performance.



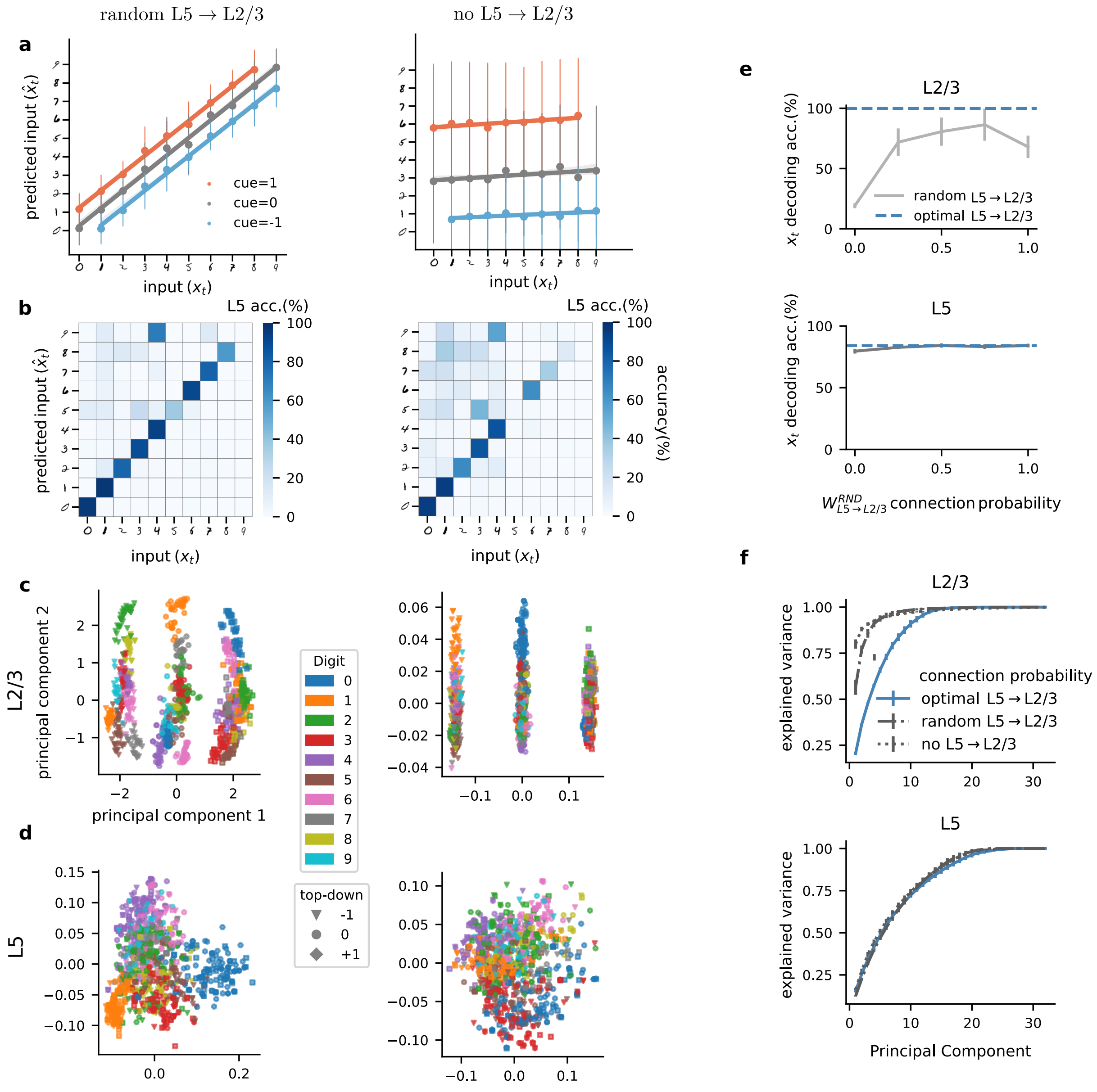
Supplementary Figure S3. L4 to L2/3 delay increases denoising and robustness to occluded stimuli. a, L4 to L2/3 delay promotes noise suppression in L5 representations. Top: Schematic of noise added to the original inputs. Bottom: Noise-corrupted input samples lead to higher L5 reconstruction residuals ($\hat{x}_t - x_t$) when the L4-to-L2/3 delay is ablated (purple) compared to the full model (solid black). The dashed line represents the reconstruction residual for an autoencoder explicitly trained to denoise the input. **b,** Top: Decoding accuracy with and without L4 to L2/3 delay for a Gabor task with occlusion. Bottom: Three examples depicting L2/3's losing it's ability to recover occluded information, similar to L5's incomplete reconstructions (top row: original occluded input; middle row: L2/3 reconstruction; bottom row: L5 reconstruction). **c,** Top: Accuracy with and without L4-to-L2/3 delay for a task in which Gabor patches move (top). Bottom: Examples further illustrate the robustness with moving Gabor patches (top row: original input with motion; middle row: L2/3 reconstruction; bottom row: L5 reconstruction).



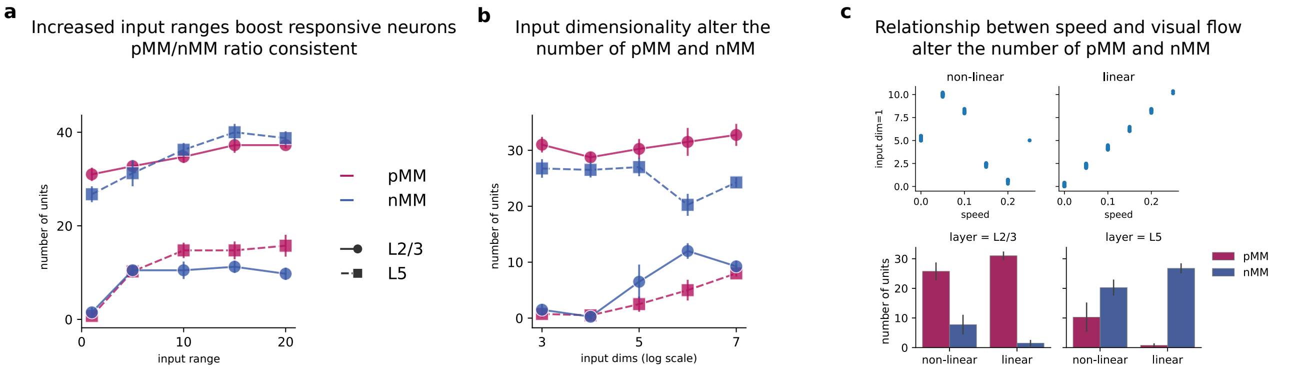
Supplementary Figure S4. Sparsity and decodability of current input in layer 2/3 and layer 5. a, L5 decoding accuracy as a function of L2/3 population sparsity. **b,** L2/3 decoding accuracy as a function of L5 population sparsity. **c,** L5 decoding accuracy as a function of different numbers of neurons in L2/3 and L5. **d,** Effect of input removal on the sparsity of neocortical layers before training.



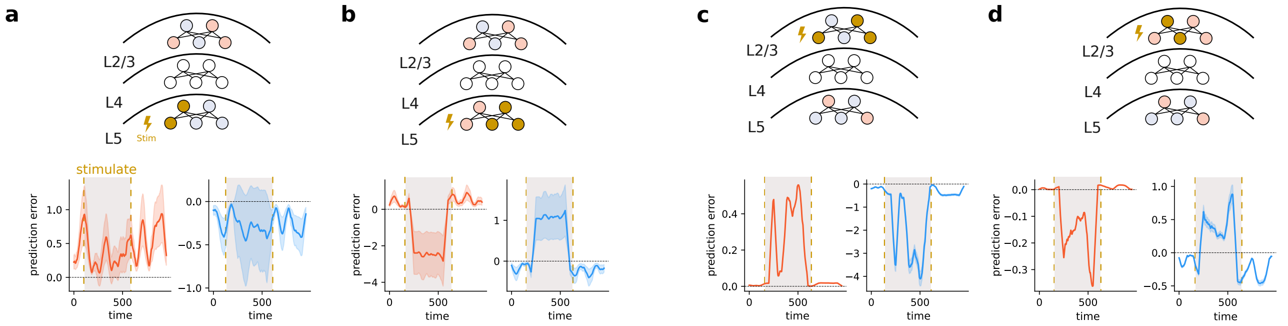
Supplementary Figure S5. Low dimensional representations with optimal L5-to-L2/3 feedback in the Gabor temporal task. a, L2/3 representations are separated by context and Gabor orientation for optimal feedback (left) but are only grouped by context when the feedback is absent entirely (right). **b,** L5 representations between optimal feedback (left) and no feedback (right).



Supplementary Figure S6. Role of feedback connections from L5 to L2/3 in learning a sequential digit task. a, L2/3 learns to predict the input with random feedback (left) but failed without feedback (right). **b,** L5 learns a good representation of the task with random feedback (left) and without feedback (right). **c,** L2/3 representations are separated by context and digit class for random feedback (left) but are only grouped by context when the feedback is absent entirely (right). **d,** L5 representation does not show a significant difference between random feedback (left) and no feedback (right). **e,** L2/3 linear prediction accuracy drops to chance level when the feedback is removed (top) while L5 classification accuracy is not impacted (bottom). **f,** The representation in L2/3 is mostly explained by the first few PCs for the 'no feedback' condition (top) while more PCs are required to explain the variance for the case with random feedback. L5 PCs explained variance is very similar in both conditions (bottom).



Supplementary Figure S7. Mismatch errors across different model parameters. a, L2/3 and L5 prediction errors remain consistent across varying input values (pMM: positive mismatch errors; nMM: negative mismatch errors). **b,** L2/3 and L5 prediction errors are robust to the dimension of the input. **c,** L2/3 and L5 prediction errors persist with both linear and non-linear visual flow relationships to speed.



Supplementary Figure S8. Mismatch errors in L2/3 and L5 after stimulating groups of neurons in L5 and L2/3, respectively. a, Increasing the activity of L5 neurons with positive errors during sensorimotor mismatch has variable effect on mismatch errors in L2/3. **b,** Modulating the L5 neurons with negative errors inverts the errors in L2/3. **c,** Scaling the output of L2/3 neurons with positive errors signals enhances the mismatch errors in both L5 neurons with negative and positive errors. **d,** Modulation of L2/3 neurons with negative mismatch errors flips the sign of errors in L5.