LOCAL CREDIT ASSIGNMENT UNDER DALE CONSTRAINTS WITH PAIRED NON-NEGATIVE CHANNELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Biologically plausible credit-assignment models seek to explain how neural circuits can learn effectively under the constraints of real neurons. Despite significant progress, existing biologically motivated credit-assignment models generally do not jointly accommodate non-negative neuronal activity, fixed excitatory/inhibitory identity, and local synaptic learning. Under these constraints, neurons cannot directly represent negative activity or change the sign of their outgoing influence, making the propagation of signed learning signals particularly challenging. We introduce a biologically motivated architecture that represents signed quantities through paired non-negative channels, with sign encoded by channel identity rather than negative neuronal activity. A simple excitatory-inhibitory on-off motif implements this representation and is repeated throughout bottom-up and top-down pathways. Combined with a local activity-based learning rule, paired top-down channels provide opposing contributions to synaptic updates, yielding an effective signed credit-assignment signal using only local interactions and fixed-sign circuitry. We show theoretically that, when bottom-up and top-down connectivity are aligned, the resulting updates exactly recover backpropagation despite propagating only non-negative neural signals. Empirically, the model learns effectively across image-classification benchmarks, remains effective with non-aligned bottom-up and top-down pathways, and scales to Tiny ImageNet, showing that the proposed biological constraints need not come at the cost of effective representation or learning. These results provide a concrete mechanism for local credit assignment under non-negative activity and Dale-constrained connectivity.

1 INTRODUCTION

Deep learning has become an important framework for systems neuroscience, where architectures, objective functions, and learning rules serve as computational hypotheses about neural circuits in the brain (Richards et al., 2019). Backpropagation is the central mechanism behind modern deep learning, but a direct implementation of its computations in biological neural circuits remains difficult to reconcile with known properties of the brain (Rumelhart et al., 1986; Whittington & Bogacz, 2019; Lillicrap et al., 2020). Such an implementation would require mechanisms for global objective optimization, non-local credit assignment, closely matched forward and backward pathways, and the propagation of signed credit-assignment signals through the network. These challenges have motivated a large literature on biologically plausible learning, including predictive coding, equilibrium propagation, feedback alignment, target propagation, and dendritic learning models (Lee et al., 2015; Lillicrap et al., 2016; Scellier & Bengio, 2017; Whittington & Bogacz, 2017; Guerguiev et al., 2017; Sacramento et al., 2018; Payeur et al., 2021; Hinton, 2022; Abel & Ullman, 2024; Oliviers et al., 2026). A common goal across these approaches is to preserve the computational benefits of gradient-based credit assignment while replacing the mechanism with local, biologically plausible computations.

One central obstacle is the representation of signed quantities. In artificial networks, activations, errors, gradients, and synaptic weights are numerical variables that can be positive or negative. In biological circuits, however, neuronal activity is non-negative: neurons fire at non-negative rates and cannot directly transmit negative activity (Lillicrap et al., 2020). In addition, under the Dale constraint considered here, each neuron has a fixed excitatory or inhibitory identity, so its outgoing

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

synaptic effects share the same sign and learning may change synaptic magnitudes but not their signs (Strata & Harvey, 1999). That is, a single neuron cannot exert both positive and negative outgoing influence. Thus, performing credit assignment under the full combination of these constraints requires not only a local learning mechanism, but also a circuit-level representation of signed activations and learning signals using non-negative activity and fixed-sign synaptic connectivity.

Existing work addresses important subsets of this problem, but, to our knowledge, does not provide a supervised credit-assignment mechanism that jointly enforces non-negative neuronal activity, fixed excitatory/inhibitory synaptic identity, and local learning. Predictive coding and related models show that local error units can approximate, and in some settings recover, backpropagation-like updates (Whittington & Bogacz, 2017; Song et al., 2020; Salvatori et al., 2022). Recent refinements further impose non-negative firing rates on predictive-coding error neurons (Alonso & Neftci, 2021), but do not enforce the full Dale constraint considered here: individual neurons are not constrained to maintain a fixed excitatory or inhibitory identity across all of their outgoing synapses. In contrast, Dale-constrained neural networks explicitly enforce excitatory/inhibitory structure and have demonstrated strong computational capabilities (Cornford et al., 2020; Barranca et al., 2022; Haber & Schneidman, 2022; Li et al., 2023; Balwani et al., 2025), but are typically trained using backpropagation or other externally computed gradient-based procedures. Thus, biologically motivated learning methods address local credit assignment without fully enforcing Dale-constrained circuitry, while Dale-constrained networks generally rely on externally computed gradients for learning. A mechanism that jointly provides local credit assignment within Dale-constrained circuitry therefore remains missing.

One possible resolution is to represent signed learning signals through pairs of non-negative neural populations rather than through individual mixed-sign error units. Lillicrap et al. (2020) argue that cortical feedback may support credit assignment by inducing neural activities whose differences approximate error signals. Recent experimental evidence from Francioni et al. (2026) reports neuron-specific dendritic instructive signals with opposing positive and negative contributions to learning, providing evidence for vectorized signed learning-related signals in biological circuits. Together, these findings motivate the possibility that effective signed learning signals may emerge from structured interactions between non-negative neural activities, but a concrete neural circuit architecture and local learning mechanism for implementing this principle remain open.

In this work, we introduce an *on-off neural circuit motif* that is composed of excitatory and inhibitory neurons. The circuit produces two non-negative output channels corresponding to positive and negative components of a represented quantity. This motif is repeated throughout the network in a bottom-up and top-down structure (Gilbert & Li, 2013), so that both feedforward activations and feedback learning signals are represented by paired non-negative channels while preserving fixed excitatory/inhibitory separation.

We further propose a local Hebbian learning update that combines bottom-up activity with paired top-down feedback signals. One top-down population contributes positive synaptic updates, and another population contributes negative updates, so the effective signed learning signal is recovered through their combined influence. Thus, the model implements credit assignment through effective signed signals that emerge from paired positive neural populations while preserving fixed excitatory and inhibitory circuit structure. To our knowledge, this provides the first supervised credit-assignment mechanism that jointly combines non-negative neuronal activity, fixed-sign excitatory/inhibitory circuitry, and local synaptic learning. We show theoretically that, under aligned bottom-up and top-down connectivity, the proposed learning scheme recovers the backpropagation update despite propagating only non-negative error channels. Empirically, we show that the model learns effectively across image-classification benchmarks, remains effective when bottom-up and top-down pathways are not symmetric, and scales to Tiny ImageNet, where it outperforms comparable vanilla networks and approaches a backpropagation-trained paired-channel control. These results suggest that the paired representation can support effective computation rather than serving only as a mechanism for satisfying biological constraints. Together, our results show how effective credit assignment can be implemented under under the full set of circuit and learning constraints considered here.

2 BACKGROUND AND RELATED WORK

Core biological constraints. We focus on three constraints. First, neuronal activity is non-negative, so neurons cannot directly represent negative scalar values. Second, under the Dale constraint considered here, each neuron has a fixed excitatory or inhibitory identity: its outgoing synaptic effects share the same sign, and learning may change their magnitudes but not reverse their signs. Third, synaptic plasticity is local, so updates should depend on quantities available at or near the synapse rather than on globally computed gradients. Together, these constraints make supervised credit assignment challenging because signed learning signals cannot be represented directly by negative neuronal activity or changing synaptic signs.

Biologically plausible credit assignment. A large body of work seeks to replace backpropagation with more biologically plausible mechanisms while preserving gradient-based credit assignment. Predictive coding uses local prediction errors and recurrent inference dynamics (Rao & Ballard, 1999; Whittington & Bogacz, 2017), with later variants recovering exact backpropagation under additional assumptions (Song et al., 2020; Salvatori et al., 2022) or imposing non-negative firing rates (Alonso & Neftci, 2021). Equilibrium propagation learns from changes between network equilibria (Scellier & Bengio, 2017; Ernoult et al., 2019; Millidge et al., 2020); feedback alignment uses separate feedforward and feedback weights (Lillicrap et al., 2016; Nøkland, 2016; Akrout et al., 2019; Webster et al., 2020); target propagation transmits target activations rather than gradients (Lee et al., 2015; Meulemans et al., 2020; Ernoult et al., 2022); and dendritic and Hebbian models seek local learning through compartmentalized or pre/post-synaptic interactions (Hebb, 2005; Guerguiev et al., 2017; Sacramento et al., 2018; Payeur et al., 2021; Lagani et al., 2021; Abel & Ullman, 2024). Forward-only approaches instead avoid an explicit backward pass (Hinton, 2022; Chen et al., 2025). However, these approaches generally do not enforce the full Dale constraint considered here: non-negative neuronal activity, fixed excitatory/inhibitory synaptic identity, and local learning must hold simultaneously throughout credit assignment.

Dale-constrained networks. A complementary line of work studies neural networks that explicitly enforce fixed excitatory/inhibitory structure (Song et al., 2016; Barreiro et al., 2017; Cornford et al., 2020; Barranca et al., 2022; Haber & Schneidman, 2022; Li et al., 2023; Balwani et al., 2025). These studies show that Dale-constrained networks can support expressive computation, task learning, and robust dynamics. However, their training generally relies on backpropagation or other externally computed gradient-based updates, rather than a local neural mechanism that propagates signed credit-assignment signals under the same constraints. Thus, they address learning of Dale-constrained networks, whereas our focus is local credit assignment within Dale-constrained circuitry.

Paired on-off representations in biological vision Paired, complementary channels are a recurring principle in biological vision. In the retina, ON and OFF pathways respond preferentially to light increments and decrements (Kuffler, 1953; Schiller, 1992). In primary visual cortex, simple-cell receptive fields contain ON and OFF subregions, often organized through interactions between excitation and inhibition (Hubel & Wiesel, 1962; Ringach, 2004; Liu et al., 2010). These findings provide a biological precedent for representing opposing components of a signal through complementary non-negative populations, which motivates the coding principle used in our on-off motif.

3 MODEL

We introduce a neural architecture that satisfies the biological constraints considered here in the following sense: (i) signed quantities are represented using only non-negative neuronal activity, (ii) each synapse has a fixed excitatory or inhibitory sign, and (iii) all outgoing synapses from a neuron have the same sign. The central component is a repeated *on-off neural circuit motif*: a small excitatory–inhibitory circuit that maps two non-negative inputs to a pair of non-negative output channels. The two channels encode positive and negative components of an underlying signed quantity, without requiring any individual neuron to carry negative activity. The same motif is used throughout the network in both the bottom-up (BU) stream, which computes activations, and the top-down (TD) stream, which propagates feedback signals for learning. Together with a local Heb-

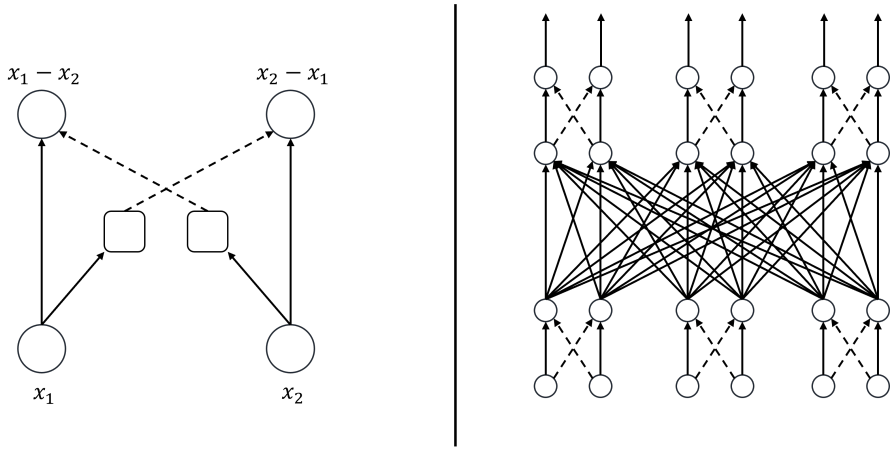

Figure 1: **On-off neural circuit motif. Left:** The basic motif receives two non-negative inputs, x_1 and x_2, and produces two non-negative output channels. Circles denote excitatory neurons and rectangles denote inhibitory neurons; solid and dashed arrows denote excitatory and inhibitory synapses, respectively. When all synaptic strengths are equal to one, the circuit encodes the signed difference between the inputs: the On channel fires when $x_1 > x_2$, while the Off channel fires when $x_2 > x_1$. Thus, a signed value is represented by two non-negative channels, with at most one active channel. **Right:** Networks are constructed by repeating this motif across layers. Inter-layer connectivity follows the standard pattern for the non-negative channels, while synaptic weights remain sign-constrained.

bian update rule, this construction enables signed credit assignment using only non-negative activity and fixed-sign Dale-constrained circuitry.

3.1 On-off neural circuit motif

The basic motif receives two non-negative inputs, $x_1, x_2 \geq 0$, and produces two non-negative output channels, denoted y^+ and y^-. The circuit contains excitatory input and output neurons, together with intermediate inhibitory neurons that mediate cross-channel inhibition (Figure 1). Each input excites one output channel and, through an inhibitory inter-neuron, suppresses the opposite output channel. To illustrate the computation performed by the motif, consider the special case in which all internal synaptic magnitudes are equal to one. The circuit then computes the signed contrast between its inputs, $d = x_1 - x_2$, and represents it through two complementary non-negative channels:

$$\begin{aligned} y^+ &= \mathrm{ReLU}(d) = \max(d, 0), \\ y^- &= \mathrm{ReLU}(-d) = \max(-d, 0). \end{aligned} \tag{1}$$

where ReLU denotes the Rectified Linear Unit function. Thus, at most one output channel is active, and the original signed quantity is recovered exactly as $d = y^+ - y^-$. A positive difference is represented by activity in the On channel, whereas a negative difference is represented with the same positive magnitude in the Off channel. The sign is therefore encoded by channel identity while all neuronal activities remain non-negative.

General sign-constrained motif. More generally, the internal synaptic magnitudes of the motif need not be fixed to one. They can be learned while preserving their excitatory or inhibitory signs. In this case, the motif computes a weighted difference of the input neurons. This relaxes the hard unit-magnitude wiring while maintaining the excitatory/inhibitory structure of the motif.

3.2 Network architecture

A full network is constructed by replacing each scalar hidden unit in a standard architecture with an on-off motif, preserving the original connectivity pattern while expanding each represented scalar into a pair of non-negative channels. Each motif outputs two channels, so a layer with n on-off

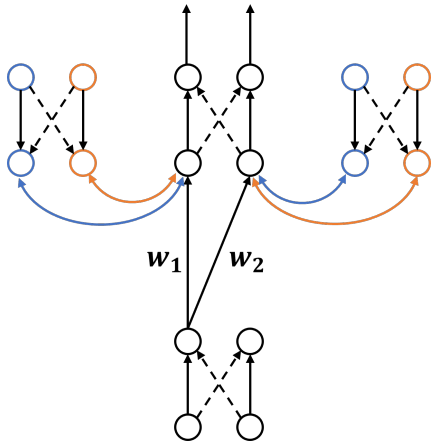

Figure 2: **Top-down streams and Hebbian learning.** The figure illustrates the update of inter-layer bottom-up weights connecting consecutive on-off layers. A bottom-up on-off motif receives feedback signals from two corresponding top-down motifs, associated with its On and Off channels. The top-down activities are non-negative, but are separated into channels corresponding to positive and negative synaptic updates (blue and orange). Learning is local: activity in the positive-update channel drives a Hebbian increase, whereas activity in the negative-update channel drives a Hebbian decrease, yielding an effective signed update proportional to the difference between paired top-down signals. For clarity, the figure shows only the connections relevant to the synaptic updates of a single bottom-up motif; a complete schematic is provided in Appendix A.1.

units contains $2n$ non-negative activity channels. Inter-layer connections follow the connectivity pattern of the original architecture and are implemented as excitatory connections, since the output neurons of each on-off motif are excitatory, as illustrated in Figure 1. Negative influences are not carried by negative inter-layer weights, but are mediated by the complementary on-off channels and the inhibitory circuitry within each motif. Learning therefore adjusts synaptic magnitudes while preserving excitatory or inhibitory identity, enforcing Dale-constrained connectivity while allowing flexible learned transformations.

3.3 Top-down propagation

The TD stream propagates feedback signals using the same on-off motif structure as the BU stream. For each BU motif, we associate two corresponding TD motifs, one for the BU On channel and one for the BU Off channel. Each TD motif contains two non-negative populations whose opposing contributions determine the sign of the local synaptic update.

TD propagation flows in the opposite direction to the BU, and combines three operations. First, lateral BU-TD connectivity gates the TD neural activity according to the active BU channel, suppressing the TD pathway associated with the inactive channel. This lateral BU-TD gating suppresses feedback through BU channels that were inactive during the forward computation, similar to the BU-TD connectivity proposed by Abel & Ullman (2024). For ReLU units, Section 4 shows that this operation implements the same gating pattern as the ReLU derivative appearing in backpropagation. Second, TD activity is processed within each layer by the on-off motif and cross-population connectivity between positive and negative feedback populations. Finally, feedback is propagated to lower layers through TD inter-layer connections, separately within each TD population. Thus, signed feedback is represented by differences between paired non-negative TD populations while all neuronal activities remain non-negative. A complete schematic of the TD flow is provided in Appendix A.1.3.

3.4 Local Hebbian learning scheme

Learning is implemented by a local Hebbian update rule that combines presynaptic BU activity with feedback from the corresponding TD channels. For each postsynaptic BU neuron i, the TD stream

provides a pair of non-negative feedback channels. Activity in one channel contributes a positive Hebbian update to the corresponding synapse, whereas activity in the other contributes a negative update, see Figure 2. Denoting the activities of these two TD channels by e_i^+ and e_i^-, respectively, with $e_i^+, e_i^- \geq 0$, the two contributions to a synapse with presynaptic activity x_j are

$$\Delta W_{ij}^+ = \eta x_j e_i^+, \qquad \Delta W_{ij}^- = -\eta x_j e_i^-. \tag{2}$$

Combining the two contributions gives

$$\Delta W_{ij} = \eta x_j (e_i^+ - e_i^-). \tag{3}$$

For mathematical analysis, we define the effective signed TD signal

$$\tilde{\delta}_i \coloneqq e_i^+ - e_i^-. \tag{4}$$

Importantly, $\tilde{\delta}_i$ is an effective mathematical quantity rather than the activity of a signed neuron: the network itself propagates only the two non-negative activities e_i^+ and e_i^-.

This rule is local: the update of a synapse depends only on the presynaptic BU activity and TD feedback signals arriving at the corresponding postsynaptic channel. The synaptic sign itself remains fixed; learning modifies synaptic magnitudes while preserving excitatory or inhibitory identity.

TD synapses follow the same Hebbian principle. Each TD weight is modified using its own presynaptic TD activity together with the corresponding postsynaptic BU-channel activity and the opposing TD-channel signal. As a result, corresponding BU and TD synapses receive the same update. Therefore, if the weights of the two pathways are aligned before the update, they remain aligned afterward.

The proposed architecture therefore provides a mechanism for representing activations, propagating feedback, and updating synapses using only non-negative neuronal activity and fixed-sign circuitry. In the next section, we analyze when this local learning rule produces gradient-based updates.

4 Theoretical analysis

We now show that, under aligned BU and TD connectivity, the local update exactly recovers gradient descent despite representing all propagated learning signals with non-negative activity.

Consider a standard feedforward network with

$$z_l = W_l h_{l-1}, \qquad h_l = \sigma(z_l), \tag{5}$$

where σ is ReLU in the hidden layers. Define the backpropagation descent signal

$$\delta_l = -\frac{\partial L}{\partial z_l}. \tag{6}$$

Then, backpropagation recursively computes

$$\delta_{l-1} = D_{l-1} W_l^T \delta_l, \qquad D_{l-1} = \operatorname{diag}(\sigma'(z_{l-1})), \tag{7}$$

and the gradient-descent update is

$$\Delta W_l = \eta \delta_l h_{l-1}^T = -\eta \frac{\partial L}{\partial W_l}. \tag{8}$$

A central challenge for a biological implementation is that δ_l is signed, whereas neuronal activity is non-negative. In the on-off model, the TD stream represents the learning signal using two non-negative activities, e_l^+ and e_l^-, with effective signed signal as defined in Eq. 4:

$$\tilde{\delta}_l = e_l^+ - e_l^-. \tag{9}$$

At the output layer, the descent signal is decomposed into its positive and negative components so that $\tilde{\delta}_L = \delta_L$.

Theorem 1. *Suppose that the positive and negative TD pathways are aligned with the corresponding BU connectivity,*

$$W_l^{\mathrm{TD},+} = W_l^{\mathrm{TD},-} = \left(W_l^{\mathrm{BU}}\right)^T, \tag{10}$$

and that $\tilde{\delta}_L = \delta_L$ at the output layer. Then $\tilde{\delta}_l = \delta_l$ at every layer, and the proposed local Hebbian update exactly recovers the gradient-descent update.

Proof sketch. The result follows by induction. During TD propagation, lateral BU-TD gating suppresses feedback through BU channels that were inactive during the forward computation. For ReLU units, this gating pattern is exactly the derivative gate D_{l-1} appearing in backpropagation. Therefore:

$$e_{l-1}^{+} = D_{l-1} W_l^{\mathrm{TD},+} e_l^{+}, \qquad e_{l-1}^{-} = D_{l-1} W_l^{\mathrm{TD},-} e_l^{-}. \tag{11}$$

Under the alignment condition, taking the difference between the two TD channels therefore gives

$$\tilde{\delta}_{l-1} = D_{l-1} \left(W_l^{\mathrm{BU}}\right)^T (e_l^{+} - e_l^{-}) = D_{l-1} \left(W_l^{\mathrm{BU}}\right)^T \tilde{\delta}_l, \tag{12}$$

which is exactly the backpropagation recursion. Therefore,

$$\Delta W_l^{\mathrm{BU}} = \eta \tilde{\delta}_l h_{l-1}^T = -\eta \frac{\partial L}{\partial W_l^{\mathrm{BU}}}. \tag{13}$$

A complete derivation is provided in Appendix A.2. The argument depends only on linear propagation, its corresponding transpose (adjoint) operation, and ReLU gating. It therefore applies directly to fully connected, locally connected, and convolutional layers, as well as linear skip connections, with the transpose understood as the corresponding adjoint operator.

Separate bottom-up and top-down weights. The exact equivalence above assumes aligned BU and TD connectivity, although the pathways are parameterized separately in the model. Corresponding BU and TD synapses receive matched local updates, so transpose alignment, when present at initialization, is preserved throughout learning without subsequent weight copying. When the pathways are not exactly aligned, exact equivalence to backpropagation is no longer guaranteed, and the resulting TD signal generally deviates from the exact backpropagation signal. We evaluate correlated and independently initialized BU-TD connectivity empirically in Section 5. The corresponding weight dynamics are derived in Appendix A.2 and measured directly in Appendix A.5.

5 Experiments

We evaluate the proposed on-off model in two regimes. First, controlled experiments on *MNIST* (LeCun et al., 1998), *Fashion-MNIST* (Xiao et al., 2017), and *CIFAR-10* (Krizhevsky et al., 2009) examine three questions: (i) whether the local Hebbian update matches backpropagation in the symmetric setting, (ii) whether the on-off representation is competitive with standard ReLU networks of comparable size, and (iii) how performance depends on bottom-up/top-down synaptic alignment and the parameterization of the on-off internal motif. Second, experiments on *Tiny ImageNet* (Le et al., 2015) evaluate whether the on-off architecture and local Hebbian learning scale to a more challenging visual benchmark.

5.1 Comparative experiments

Setup. We evaluate two-layer on-off MLPs on MNIST, Fashion-MNIST, and CIFAR-10 against two ReLU baselines. The first has the same number of effective hidden units as the on-off model, replacing every on-off motif with a standard ReLU neuron, while the second doubles the width to match the actual number of hidden neurons, since each on-off unit contains two non-negative channels. Unless otherwise stated, all methods use the same training configuration. Full architectural and optimization details are given in Appendix A.3.

Variants. We compare standard backpropagation on the on-off architecture (BP) with the proposed local rule under several BU-TD relationships: aligned, correlated, correlated with noisy TD updates, and independently initialized pathways. The independently initialized condition uses hyperparameters selected on CIFAR-10 and transferred unchanged to MNIST and Fashion-MNIST. We additionally evaluate learned variants of the internal on-off motif. Full experimental details, including the definitions of all model variants, are provided in Appendix A.3.

Table 1: Controlled experiments on MNIST, Fashion-MNIST, and CIFAR-10. We report test accuracy (%) averaged over 10 random seeds, with standard deviation.

Method	MNIST	Fashion-MNIST	CIFAR-10
Vanilla MLP	97.666 ± 0.066	87.891 ± 0.178	54.057 ± 0.274
Vanilla MLP $\times 2$	97.784 ± 0.066	88.077 ± 0.171	55.510 ± 0.423
On-Off BP	97.792 ± 0.067	88.100 ± 0.238	56.432 ± 0.372
On-Off Aligned	97.796 ± 0.074	88.147 ± 0.229	56.483 ± 0.318
On-Off Correlated	97.795 ± 0.079	87.888 ± 0.268	56.394 ± 0.377
On-Off Correlated + Noisy Updates	97.791 ± 0.094	87.734 ± 0.322	56.370 ± 0.578
On-Off Independent Init.[†]	97.932 ± 0.078	87.298 ± 0.416	55.604 ± 0.887
On-Off Learned Unit	97.765 ± 0.054	87.867 ± 0.276	56.434 ± 0.443
On-Off Shared Learned Unit	$\mathbf{98.027 \pm 0.051}$	$\mathbf{88.276 \pm 0.333}$	$\mathbf{56.484 \pm 0.346}$

[†] Hyperparameters for the independently initialized condition were selected on CIFAR-10 and transferred unchanged to MNIST and Fashion-MNIST. All other alignment conditions use the common hyperparameter configuration.

Results. Table 1 first confirms the theoretical result: when BU and TD pathways are aligned, the proposed local Hebbian rule closely matches backpropagation on the same on-off architecture. Importantly, effective learning does not require exact weight symmetry. Correlated pathways, including those with noisy TD updates, remain close to the aligned condition, while independently initialized pathways with approximately zero initial correlation reach $55.604 \pm 0.887\%$ on CIFAR-10, compared with $56.483 \pm 0.318\%$ under aligned initialization; the same configuration transfers effectively to MNIST and Fashion-MNIST without further dataset-specific tuning. Alignment diagnostics in Appendix A.5 show that independently initialized BU and TD pathways become increasingly correlated as matched local updates accumulate, while the resulting local updates become increasingly aligned with the true backpropagation gradient. Thus, exact equivalence requires aligned connectivity, but effective local credit assignment extends well beyond this idealized setting.

The learned-motif variants show that the on-off circuit does not need to enforce a fixed hard difference with weights ± 1. The model learns effectively when the internal motif weights are themselves learned under sign constraints, starting from the unit-magnitude initialization. The Shared Learned Unit variant achieves the highest mean accuracy on all three datasets, suggesting that sharing the learned on-off internal motif computation within a layer may provide a useful regularization effect.

5.2 Scaling to Tiny ImageNet

Setup. We next evaluate scalability on Tiny ImageNet (Le et al., 2015), a 200-class benchmark that is substantially more challenging than CIFAR-10 and provides a useful stress test for biologically motivated learning methods, whose performance can degrade markedly in this regime (Chen et al., 2025). Following the recent SCFF evaluation, we use a similar five-convolution-layer architecture. The on-off model is trained with the proposed local Hebbian rule in the aligned BU-TD setting and compared with the same two vanilla ReLU controls used above: one matched by feature units and one by scalar channel count.

We also include CReLU (Shang et al., 2016) as a representational control. CReLU preserves paired positive and negative responses but is trained with standard backpropagation and does not impose the fixed excitatory/inhibitory constraints considered here. This isolates the representational benefit of paired channels from the proposed circuit and learning mechanism. Full details are provided in Appendix A.3.

Results. Table 2 reports final-epoch test accuracy, showing that the on-off model scales beyond the small-scale fully connected setting. On the same five-convolution-layer scale as SCFF, the on-off model reaches 42.31% top-1 and 67.70% top-5 accuracy, outperforming SCFF by a large margin. It also substantially outperforms both vanilla convolutional baselines: the matched-feature-width model reaches 35.58% top-1, while the matched-channel-count $\times 2$ model reaches 37.10%. Thus, the performance of the on-off model cannot be attributed simply to the increased number of scalar

Table 2: Tiny ImageNet results. We report top-1 and top-5 accuracy (%). Results are averaged over 5 random seeds. Reference results are reported from (Chen et al., 2025), '–' means no reported results.

Method	Top-1	Top-5
Hebb-based (Lagani et al., 2021)	–	37.0
DFA (Webster et al., 2020)	32.1	–
SCFF (Chen et al., 2025)	35.7	59.8
Vanilla Conv	35.58 ± 0.44	60.69 ± 0.32
Vanilla Conv $\times 2$	37.10 ± 0.57	62.64 ± 0.33
CReLU (BP)	$\mathbf{43.90 \pm 0.08}$	$\mathbf{68.95 \pm 0.28}$
On-Off (Ours)	42.31 ± 0.41	67.70 ± 0.31

channels introduced by the paired representation. More importantly, these results show that enforcing the proposed paired non-negative representation together with fixed excitatory/inhibitory separation does not impose a representational disadvantage relative to standard ReLU networks.

The CReLU control further shows that preserving complementary positive and negative responses is itself an effective representational strategy. Its slightly higher performance also helps separate this representational benefit from the main contribution of our model: representing signed quantities using non-negative neural activity while maintaining fixed excitatory/inhibitory circuitry and learning through local credit assignment. To further analyze this representation, Appendix A.4 visualizes the learned activations of opposing on-off channels.

6 Discussion

We introduced an on-off architecture for local credit assignment under non-negative neuronal activity and Dale-constrained connectivity, where neural activations and learning signals are represented through paired non-negative channels. This provides a concrete mechanism for a central problem in biologically plausible credit assignment: how to represent and propagate signed learning signals when neuronal activity is non-negative and synaptic influences have fixed excitatory or inhibitory signs. In our model, two TD populations play opposing instructive roles, with one driving Hebbian increases and the other driving decreases of synaptic weights. The use of paired opposing channels is consistent with a broader biological coding principle observed in sensory systems, including ON/OFF organization in early vision (Hubel & Wiesel, 1962; Schiller, 1992) and opponent representations at higher levels of visual processing (Valentine, 1991; Skinner & Benton, 2010).

The model is also in line with theoretical proposals that feedback may induce error-like differences between neural activities (Lillicrap et al., 2020), and recent evidence for vectorized instructive signals in cortical dendrites (Francioni et al., 2026). The structure of the model is relatively simple: the same on-off motif is repeated across the network and used in both BU and TD streams, composing signed signals using a two-channel circuit-level representation based on excitatory and inhibitory interactions. The experiments suggest that this paired representation is not merely a costly doubling of network channels. The on-off model achieves higher mean accuracy than the matched vanilla ReLU architectures across the evaluated settings, while the CReLU control shows that explicitly preserving complementary positive and negative responses can itself provide a strong representational advantage. Notably, the on-off model approaches the performance of this backpropagation-trained representational control while implementing the same coding principle within non-negative, fixed-sign circuitry and using the proposed local learning mechanism. While the structured BU-TD correspondence and gating remain idealized abstractions of biological circuitry, the model provides a concrete mechanism for implementing local credit assignment under these constraints. Together, these results suggest that paired channels encoding positive and negative components provide a useful representational primitive that can support local credit assignment under Dale-constrained connectivity.

AI USE STATEMENT

We used generative AI tools to assist with editing and refining the writing of the manuscript, including improving clarity, structure, and presentation. We also used generative AI for limited assistance with experimental implementation of some experiments. In addition, generative AI was used to help identify and summarize potentially relevant literature and related resources.

All AI-assisted code and experimental suggestions were reviewed, tested, and verified by the authors before use. Mathematical arguments, experimental results, and their interpretation were independently checked by the authors, and literature suggested by generative AI was verified against the original sources. Generative AI was not used as an autonomous source of experimental results or evidence. We take responsibility for the final content of this work, including all text, claims, code, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide the details needed to reproduce the proposed method and experiments throughout the paper and appendices. Experimental architectures, hyperparameters, training procedures, and variant definitions are provided in Appendix A.3. Source code with a README containing instructions for reproducing the experiments is included in the supplementary material. The code repository will be made publicly available with the final version of the paper.

REFERENCES

Roy Abel and Shimon Ullman. Biologically inspired learning model for instructed vision. *Advances in Neural Information Processing Systems*, 37:45315–45358, 2024.

Mohamed Akrout, Collin Wilson, Peter C Humphreys, Timothy Lillicrap, and Douglas Tweed. Deep learning without weight transport. *arXiv preprint arXiv:1904.05391*, 2019.

Nicholas Alonso and Emre Neftci. Tightening the biological constraints on gradient-based predictive coding. In *International conference on neuromorphic systems 2021*, pp. 1–9, 2021.

Aishwarya Balwani, Alex Q Wang, Farzaneh Najafi, and Hannah Choi. Constructing biologically constrained rnns via dale's backpropagation and topologically informed pruning. *Science Advances*, 11(50):eadw4970, 2025.

Victor J Barranca, Asha Bhuiyan, Max Sundgren, and Fangzhou Xing. Functional implications of dale's law in balanced neuronal network dynamics and decision making. *Frontiers in neuroscience*, 16:801847, 2022.

Andrea K Barreiro, J Nathan Kutz, and Eli Shlizerman. Symmetries constrain dynamics in a family of balanced neural networks. *The Journal of Mathematical Neuroscience*, 7(1):10, 2017.

Xing Chen, Dongshu Liu, Jérémie Laydevant, and Julie Grollier. Self-contrastive forward-forward algorithm. *Nature Communications*, 16(1):5978, 2025.

Jonathan Cornford, Damjan Kalajdzievski, Marco Leite, Amélie Lamarquette, Dimitri M Kullmann, and Blake Richards. Learning to live with dale's principle: Anns with separate excitatory and inhibitory units. *BioRxiv*, pp. 2020–11, 2020.

Maxence Ernoult, Julie Grollier, Damien Querlioz, Yoshua Bengio, and Benjamin Scellier. Updates of equilibrium prop match gradients of backprop through time in an rnn with static input. *Advances in neural information processing systems*, 32, 2019.

Maxence M Ernoult, Fabrice Normandin, Abhinav Moudgil, Sean Spinney, Eugene Belilovsky, Irina Rish, Blake Richards, and Yoshua Bengio. Towards scaling difference target propagation by learning backprop targets. In *International Conference on Machine Learning*, pp. 5968–5987. PMLR, 2022.

Valerio Francioni, Vincent D Tang, Enrique HS Toloza, Zilan Ding, Norma J Brown, and Mark T Harnett. Vectorized instructive signals in cortical dendrites. *Nature*, 652(8112):1254–1263, 2026.

Charles D Gilbert and Wu Li. Top-down influences on visual processing. *Nature Reviews Neuroscience*, 14(5):350–363, 2013.

Jordan Guerguiev, Timothy P Lillicrap, and Blake A Richards. Towards deep learning with segregated dendrites. *elife*, 6:e22901, 2017.

Adam Haber and Elad Schneidman. The computational and learning benefits of daleian neural networks. *Advances in Neural Information Processing Systems*, 35:5194–5206, 2022.

Donald Olding Hebb. *The organization of behavior: A neuropsychological theory*. Psychology Press, 2005.

Geoffrey Hinton. The forward-forward algorithm: Some preliminary investigations. *arXiv preprint arXiv:2212.13345*, 2022.

David H Hubel and Torsten N Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1):106, 1962.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Stephen W Kuffler. Discharge patterns and functional organization of mammalian retina. *Journal of neurophysiology*, 16(1):37–68, 1953.

Gabriele Lagani, Fabrizio Falchi, Claudio Gennaro, and Giuseppe Amato. Hebbian semi-supervised learning in a sample efficiency setting. *Neural Networks*, 143:719–731, 2021.

Yann Le, Xuan Yang, et al. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Dong-Hyun Lee, Saizheng Zhang, Asja Fischer, and Yoshua Bengio. Difference target propagation. In *Joint european conference on machine learning and knowledge discovery in databases*, pp. 498–515. Springer, 2015.

Pingsheng Li, Jonathan Cornford, Arna Ghosh, and Blake Richards. Learning better with dale's law: A spectral perspective. *Advances in Neural Information Processing Systems*, 36:944–956, 2023.

Timothy P Lillicrap, Daniel Cownden, Douglas B Tweed, and Colin J Akerman. Random synaptic feedback weights support error backpropagation for deep learning. *Nature communications*, 7(1): 1–10, 2016.

Timothy P Lillicrap, Adam Santoro, Luke Marris, Colin J Akerman, and Geoffrey Hinton. Backpropagation and the brain. *Nature Reviews Neuroscience*, 21(6):335–346, 2020.

Bao-hua Liu, Pingyang Li, Yujiao J Sun, Ya-tang Li, Li I Zhang, and Huizhong Whit Tao. Intervening inhibition underlies simple-cell receptive field structure in visual cortex. *Nature neuroscience*, 13(1):89–96, 2010.

Alexander Meulemans, Francesco Carzaniga, Johan Suykens, João Sacramento, and Benjamin F Grewe. A theoretical framework for target propagation. *Advances in Neural Information Processing Systems*, 33:20024–20036, 2020.

Beren Millidge, Alexander Tschantz, Christopher L Buckley, and Anil Seth. Activation relaxation: A local dynamical approximation to backpropagation in the brain. *arXiv preprint arXiv:2009.05359*, 2020.

Arild Nøkland. Direct feedback alignment provides learning in deep neural networks. *Advances in neural information processing systems*, 29, 2016.

Gaspard Oliviers, Mufeng Tang, and Rafal Bogacz. Bidirectional predictive coding. In *The Fourteenth International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=HbRihpurRr`.

Alexandre Payeur, Jordan Guerguiev, Friedemann Zenke, Blake A Richards, and Richard Naud. Burst-dependent synaptic plasticity can coordinate learning in hierarchical circuits. *Nature neuroscience*, 24(7):1010–1019, 2021.

Rajesh PN Rao and Dana 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.

Blake A Richards, Timothy P Lillicrap, Philippe Beaudoin, Yoshua Bengio, Rafal Bogacz, Amelia Christensen, Claudia Clopath, Rui Ponte Costa, Archy de Berker, Surya Ganguli, et al. A deep learning framework for neuroscience. *Nature neuroscience*, 22(11):1761–1770, 2019.

Dario L Ringach. Mapping receptive fields in primary visual cortex. *The Journal of Physiology*, 558 (3):717–728, 2004.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

João Sacramento, Rui Ponte Costa, Yoshua Bengio, and Walter Senn. Dendritic cortical microcircuits approximate the backpropagation algorithm. *Advances in neural information processing systems*, 31, 2018.

Tommaso Salvatori, Yuhang Song, Zhenghua Xu, Thomas Lukasiewicz, and Rafal Bogacz. Reverse differentiation via predictive coding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 8150–8158, 2022.

Benjamin Scellier and Yoshua Bengio. Equilibrium propagation: Bridging the gap between energy-based models and backpropagation. *Frontiers in computational neuroscience*, 11:24, 2017.

Peter H Schiller. The on and off channels of the visual system. *Trends in neurosciences*, 15(3): 86–92, 1992.

Wenling Shang, Kihyuk Sohn, Diogo Almeida, and Honglak Lee. Understanding and improving convolutional neural networks via concatenated rectified linear units. In *international conference on machine learning*, pp. 2217–2225. PMLR, 2016.

Andrew L Skinner and Christopher P Benton. Anti-expression aftereffects reveal prototype-referenced coding of facial expressions. *Psychological Science*, 21(9):1248–1253, 2010.

H Francis Song, Guangyu R Yang, and Xiao-Jing Wang. Training excitatory-inhibitory recurrent neural networks for cognitive tasks: a simple and flexible framework. *PLoS computational biology*, 12(2):e1004792, 2016.

Yuhang Song, Thomas Lukasiewicz, Zhenghua Xu, and Rafal Bogacz. Can the brain do backpropagation?—exact implementation of backpropagation in predictive coding networks. *Advances in neural information processing systems*, 33:22566, 2020.

Piergiorgio Strata and Robin Harvey. Dale's principle. *Brain research bulletin*, 50(5-6):349–350, 1999.

Tim Valentine. A unified account of the effects of distinctiveness, inversion, and race in face recognition. *The Quarterly Journal of Experimental Psychology Section A*, 43(2):161–204, 1991.

Matthew Bailey Webster, Jonghyun Choi, et al. Learning the connections in direct feedback alignment. 2020.

James CR Whittington and Rafal Bogacz. An approximation of the error backpropagation algorithm in a predictive coding network with local hebbian synaptic plasticity. *Neural computation*, 29(5): 1229–1262, 2017.

James CR Whittington and Rafal Bogacz. Theories of error back-propagation in the brain. *Trends in cognitive sciences*, 23(3):235–250, 2019.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pp. 818–833. Springer, 2014.

A Appendix

A.1 Complete on-off model connectivity

This appendix provides a complete schematic description of the connectivity and computational flow of the on-off model. The main text presents the basic on-off motif and the local update rule in simplified form. Here we describe how bottom-up motifs, top-down motifs, gating connections, and inter-layer weights are arranged in the full model.

A.1.1 Bottom-up and top-down motifs

Figure 3 illustrates the organization of representative on-off motifs in two consecutive layers. Each layer contains multiple bottom-up (BU) on-off motifs through which feedforward activity propagates upward. For every BU motif, the model includes two corresponding top-down (TD) motifs: one associated with the BU On channel and one associated with the BU Off channel.

Each TD motif is itself composed of two non-negative channels, corresponding to two TD populations of positive and negative synaptic updates. In the figure, the positive-update population is shown in blue and the negative-update population in orange. Thus, while every BU motif contains two output channels, its corresponding TD structure contains four channels, derived from two TD motifs. This organization allows the model to represent signed feedback signals using only non-negative neural activity.

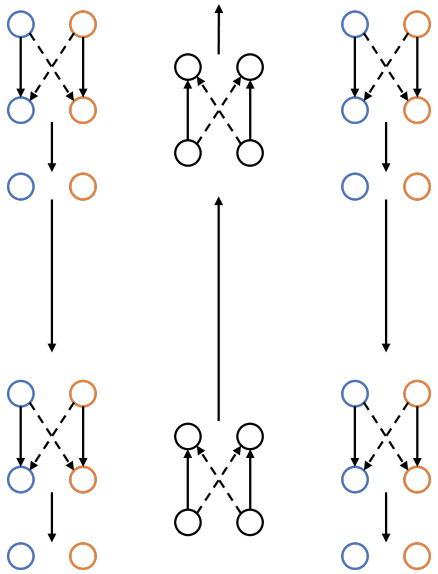

Figure 3: **Bottom-up and top-down motif organization.** The figure shows two consecutive on-off layers, each containing one bottom-up (BU) on-off motif and two corresponding top-down (TD) motifs. BU activity flows upward through the network. For each BU motif, one TD motif corresponds to the BU On channel and the other to the BU Off channel. Each TD motif contains two non-negative feedback channels: a positive-update population (blue) and a negative-update population (orange). TD activity flows downward and provides the feedback signals used for local Hebbian updates.

A.1.2 Bottom-up connectivity

The BU computation is illustrated in Figure 4. Within each on-off motif, the two output neurons are passed through the on-off rectifying nonlinearities, denoted by ReLU^+ and ReLU^-. As a result, at most one output neuron of each BU motif is active at a time. The active output channel represents the sign and magnitude of the encoded scalar quantity.

Inter-layer BU connectivity is applied over the individual output channels of the previous layer and the individual input neurons of the next layer. Thus, if a layer contains n on-off motifs, it produces

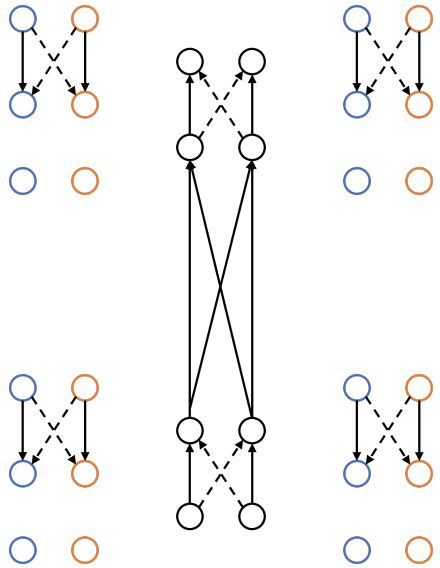

Figure 4: **Bottom-up connectivity and flow.** The BU stream propagates upward through consecutive on-off motifs. Each motif applies ReLU^+ and ReLU^- nonlinearities, so that at most one output channel is active. The two output neurons of a lower-layer motif are connected to the input neurons of motifs in the next layer through standard inter-layer connectivity. Thus, inter-layer BU weights operate over individual non-negative channels, while pairs of channels jointly form on-off motifs.

$2n$ non-negative output neurons, and a fully connected BU layer maps these $2n$ neurons to the input neurons of the motifs in the next layer. In this sense, the network preserves the standard fully connected structure over neurons, while grouping pairs of neurons into on-off motifs.

A.1.3 Top-down connectivity and gating

Figure 5 shows the TD connectivity. Although the full diagram is dense, the computation consists of four simple components.

First, TD propagation is gated by the corresponding BU activity. Since each BU motif has at most one active output channel, the TD motif corresponding to the inactive BU channel is suppressed. This gating operation is implemented through lateral BU-TD connections, shown by dashed colored arrows in the figure. Functionally, this operation plays the same role as the derivative of a rectifying nonlinearity: feedback is propagated only through the channel that participated in the BU computation.

Second, each TD motif applies the same on-off motif computation as the BU stream. The motif computes a weighted difference between its non-negative input channels and applies the on-off rectifying nonlinearities. Therefore, TD signals are also represented by paired non-negative channels rather than signed neurons.

Third, the two TD motifs associated with the same BU motif interact through cross-channel connectivity. The figure includes four intermediate TD neurons between the upper-layer TD motifs and the lower-layer TD inputs. These neurons implement the cross-channel structure: signals can cross both between the On and Off BU-associated motifs and between the positive-update and negative-update TD populations. This cross-channel connectivity allows signed credit-assignment information to be represented as the difference between paired non-negative TD channels.

Fourth, the intermediate TD neurons are connected to the inputs of the next lower TD layer through inter-layer TD connectivity. The positive-update TD population is connected to the corresponding positive-update population in the next layer, and the negative-update population is connected analogously. Consequently, the model contains three inter-layer weight sets: one for the BU stream, one for the positive-update TD stream, and one for the negative-update TD stream. Although the TD stream contains twice as many neurons as the BU stream, the positive and negative TD populations

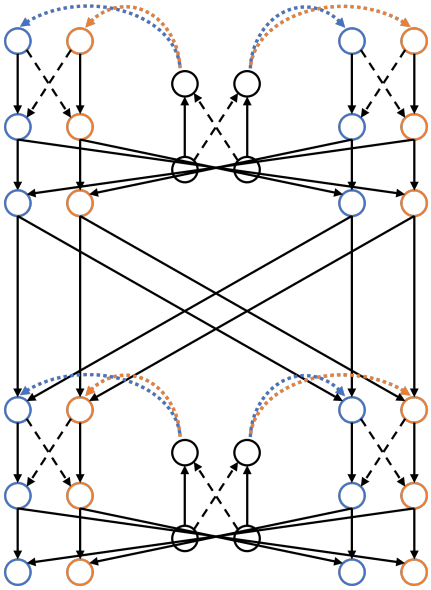

Figure 5: **Top-down connectivity and flow.** The TD stream propagates feedback downward through two TD motifs associated with each BU motif. The computation consists of four operations. First, lateral BU-TD connections gate TD activity according to the active BU channel. Second, each TD motif applies an on-off transformation, representing signed feedback through paired non-negative channels. Third, cross-channel TD connectivity links the two TD motifs associated with the same BU motif, allowing feedback to cross between On/Off-associated motifs and between positive-update and negative-update populations. Fourth, inter-layer TD connectivity propagates feedback to the next lower layer, separately for the positive-update and negative-update populations. This yields separate inter-layer weight sets for the BU stream, the positive-update TD stream, and the negative-update TD stream.

are connected separately; therefore, each TD inter-layer weight matrix has the same shape as the BU inter-layer weight matrix, up to transposition.

A.1.4 Computational flow of learning

The full learning algorithm consists of three stages. First, the network performs a BU pass. The input is propagated upward through the on-off motifs, and each motif produces a pair of non-negative output channels. Due to the on-off nonlinearities, at most one channel in each motif is active.

Second, the network performs a TD pass. The output-layer descent signal is decomposed into positive and negative components and represented by the corresponding non-negative TD channels. This signal is propagated downward through the TD stream. At each layer, lateral BU-TD gating restricts feedback to the TD motif corresponding to the active BU channel. The TD stream therefore carries feedback signals that are consistent with the BU computation while representing signed error information through paired non-negative channels.

Third, all weights are updated using a local Hebbian rule. For each inter-layer BU weight, the update depends on the product of the presynaptic BU activity and the relevant postsynaptic TD feedback channels. Activity in the positive-update TD channel contributes a Hebbian increase, whereas activity in the negative-update TD channel contributes a Hebbian decrease. The effective update is therefore proportional to the difference between paired TD channels. The corresponding TD weights are updated analogously, producing matched local updates across BU and TD pathways, with transpose orientation at the matrix level. The same local learning principle is also applied to the intra-motif and intermediate TD connections, so all learnable synapses in the model are updated through local Hebbian interactions.

The overall learning flow is summarized as:

1. **Bottom-up pass:** propagate the input through BU on-off motifs.
2. **Top-down pass:** propagate feedback through gated TD on-off motifs using positive-update and negative-update channels.
3. **Local update:** update BU and TD weights using Hebbian products of local neural activities.

A.2 Full proof of gradient recovery

This section provides the full derivation of the theoretical result in Section 4. We show that, under aligned BU-TD connectivity, the effective signed signal carried by the paired non-negative TD channels follows the same recursion as the backpropagation descent signal, and that the resulting local Hebbian update exactly recovers the gradient-descent update.

Backpropagation. Consider a feedforward network with pre-activations z_l, activations h_l, and weights W_l:

$$z_l = W_l h_{l-1}, \qquad h_l = \sigma(z_l), \tag{14}$$

where σ is ReLU in the hidden layers. We define the backpropagation descent signal as

$$\delta_l \coloneqq -\frac{\partial L}{\partial z_l}. \tag{15}$$

Defining

$$D_{l-1} \coloneqq \mathrm{diag}(\sigma'(z_{l-1}))\,, \tag{16}$$

the backpropagation recursion is

$$\delta_{l-1} = D_{l-1} W_l^T \delta_l. \tag{17}$$

With the descent-signal convention in Eq. equation 15, the corresponding gradient-descent update is

$$\Delta W_l = \eta \delta_l h_{l-1}^T = -\eta \frac{\partial L}{\partial W_l}. \tag{18}$$

Non-negative representation of the learning signal. In the on-off model, let W_l^{BU} denote the bottom-up weights corresponding to the forward weights W_l in the standard formulation above. At each layer, the TD stream carries two non-negative activity vectors, e_l^+ and e_l^-. Their effective signed TD signal is

$$\tilde{\delta}_l \coloneqq e_l^+ - e_l^-, \qquad e_l^+, e_l^- \geq 0. \tag{19}$$

The quantity $\tilde{\delta}_l$ is used only for mathematical analysis; the neural circuit itself propagates only the two non-negative TD-channel activities.

At the output layer, the TD activities are set by decomposing the descent signal into its positive and negative components:

$$e_L^+ = \mathrm{ReLU}(\delta_L), \qquad e_L^- = \mathrm{ReLU}(-\delta_L). \tag{20}$$

Consequently,

$$\tilde{\delta}_L = e_L^+ - e_L^- = \delta_L. \tag{21}$$

Top-down propagation and BU gating. Let $W_l^{\mathrm{TD},+}$ and $W_l^{\mathrm{TD},-}$ denote the inter-layer weights of the positive and negative TD pathways, respectively. These matrices map activity from layer l to layer $l-1$ and therefore have the transpose orientation of the corresponding BU weights. As described in Section 3, TD propagation is gated by the corresponding BU activity, such that feedback is propagated only through the BU channel that was active during the forward computation. Let $\mathbf{Gate}_{l-1}$ denote the corresponding diagonal gating matrix.

After the within-layer on-off and cross-channel TD computation described in Appendix A.1, let e_l^+ and e_l^- denote the resulting positive- and negative-update activities propagated between layers. The two TD channels propagate according to

$$e_{l-1}^+ = \mathbf{Gate}_{l-1} W_l^{\mathrm{TD},+} e_l^+, \tag{22}$$

$$e_{l-1}^- = \mathbf{Gate}_{l-1} W_l^{\mathrm{TD},-} e_l^-. \tag{23}$$

For each ReLU channel, the corresponding BU gate is active exactly when its pre-activation is positive. Therefore,

$$\mathbf{Gate}_{l-1} = \mathrm{diag}(\mathbf{1}[z_{l-1} > 0]) = \mathrm{diag}(\sigma'(z_{l-1})) = D_{l-1}, \tag{24}$$

where we take $\sigma'(0) = 0$, consistent with suppressing an inactive BU channel. Thus, the BU-dependent gating operation implemented by the model is equivalent to the ReLU derivative gate appearing in the backpropagation recursion.

Theorem 2. *Consider the paired non-negative TD activities defined in Eq. equation 19. Suppose that, for every layer l, the positive and negative TD inter-layer weights are aligned with the corresponding BU connectivity,*

$$W_l^{\mathrm{TD},+} = W_l^{\mathrm{TD},-} = \left(W_l^{\mathrm{BU}}\right)^T, \tag{25}$$

and that the output TD channels satisfy Eq. equation 21. Then

$$\tilde{\delta}_l = \delta_l \tag{26}$$

at every layer. Consequently, the proposed local Hebbian update exactly recovers the gradient-descent update.

Proof. The result holds at the output layer by Eq. equation 21. Suppose that $\tilde{\delta}_l = \delta_l$ at some layer l. Using Eqs. equation 22–equation 23, the gate equivalence in Eq. equation 24, and the alignment condition in Eq. equation 25, we obtain

$$\begin{aligned}
\tilde{\delta}_{l-1} &= e_{l-1}^+ - e_{l-1}^- && (27)\\
&= D_{l-1} W_l^{\mathrm{TD},+} e_l^+ - D_{l-1} W_l^{\mathrm{TD},-} e_l^- && (28)\\
&= D_{l-1} \left(W_l^{\mathrm{BU}}\right)^T e_l^+ - D_{l-1} \left(W_l^{\mathrm{BU}}\right)^T e_l^- && (29)\\
&= D_{l-1} \left(W_l^{\mathrm{BU}}\right)^T (e_l^+ - e_l^-) && (30)\\
&= D_{l-1} \left(W_l^{\mathrm{BU}}\right)^T \tilde{\delta}_l && (31)\\
&= D_{l-1} \left(W_l^{\mathrm{BU}}\right)^T \delta_l && (32)\\
&= \delta_{l-1}, && (33)
\end{aligned}$$

where the final equality follows from the backpropagation recursion in Eq. equation 17, with W_l^{BU} corresponding to the forward weight matrix W_l in the standard formulation. Therefore, by induction, $\tilde{\delta}_l = \delta_l$ at every layer.

The local Hebbian update to the BU weights is

$$\Delta W_l^{\mathrm{BU}} = \eta(e_l^+ - e_l^-)h_{l-1}^T = \eta\tilde{\delta}_l h_{l-1}^T. \tag{34}$$

Using $\tilde{\delta}_l = \delta_l$ and Eq. equation 18,

$$\Delta W_l^{\mathrm{BU}} = \eta\delta_l h_{l-1}^T = -\eta\frac{\partial L}{\partial W_l^{\mathrm{BU}}}. \tag{35}$$

Thus, under aligned BU-TD connectivity, the proposed local Hebbian rule exactly recovers the gradient-descent update. □

The argument depends only on linear propagation, its corresponding transpose (adjoint) operation, and ReLU gating. It therefore applies directly to fully connected, locally connected, and convolutional layers, as well as linear skip connections, with the transpose understood as the corresponding adjoint operator.

Separate bottom-up and top-down weights. The theorem assumes aligned BU and TD connectivity, but the pathways are parameterized separately in the model rather than explicitly tied. Corresponding BU and TD synapses receive matched local updates.

Throughout the analysis, W denotes the effective synaptic weight used by the neural computation; fixed excitatory/inhibitory identity is represented separately from the underlying parameterization used to enforce the sign constraint.

Because the corresponding matrices describe propagation in opposite directions, their matrix updates satisfy

$$\Delta W_l^{\mathrm{TD},\pm}(t) = \left(\Delta W_l^{\mathrm{BU}}(t)\right)^T. \tag{36}$$

It follows that

$$W_l^{\mathrm{TD},\pm}(t) - \left(W_l^{\mathrm{BU}}(t)\right)^T = W_l^{\mathrm{TD},\pm}(0) - \left(W_l^{\mathrm{BU}}(0)\right)^T. \tag{37}$$

Thus, if the pathways are aligned at initialization, $W_l^{\mathrm{TD},\pm}(0) = (W_l^{\mathrm{BU}}(0))^T$, they remain exactly aligned throughout learning without requiring subsequent weight copying.

More generally, let $U_l(t)$ denote the accumulated BU update up to time t. The separately parameterized weights can then be written as

$$W_l^{\mathrm{BU}}(t) = W_l^{\mathrm{BU}}(0) + U_l(t), \tag{38}$$

and

$$W_l^{\mathrm{TD},\pm}(t) = W_l^{\mathrm{TD},\pm}(0) + U_l(t)^T. \tag{39}$$

Therefore, independently initialized pathways do not become identical under perfectly matched updates: their absolute deviation from transpose alignment remains fixed according to Eq. equation 37. Nevertheless, if the shared learned component becomes large relative to the initial pathway-specific components, relative alignment measures such as correlation or cosine similarity can increase even though the absolute difference remains unchanged. Exact equivalence to backpropagation is guaranteed only in the aligned setting; correlated and independently initialized pathways are examined empirically in Section 5. This increase in relative alignment is observed directly in the independently initialized experiments in Appendix A.5.

A.3 Experimental details

All experiments were run on a single NVIDIA A10 GPU.

A.3.1 Controlled experiments

For the controlled experiments on MNIST, Fashion-MNIST, and CIFAR-10, we used two-layer fully connected networks. For MNIST and Fashion-MNIST, the on-off network contained two hidden layers with 250 on-off units per layer. For CIFAR-10, the network contained two hidden layers with 500 on-off units per layer. Since each on-off unit consists of two non-negative channels, these correspond to 500 and 1000 scalar hidden channels per layer, respectively.

All controlled experiments included bias terms, and used Cross-Entropy loss. The internal on-off motif used fixed unit-magnitude difference weights unless evaluating the learnable motif variants. The common training configuration used stochastic gradient descent with learning rate 10^{-2}, batch size 64, no weight decay, and no learning-rate decay. MNIST and Fashion-MNIST models were trained for 20 epochs, while CIFAR-10 models were trained for 50 epochs. Except for the independently initialized BU-TD condition described below, we use this common configuration without extensive hyperparameter tuning. Results are reported as test accuracy at the best epoch during training, averaged over 10 random seeds.

For vanilla baselines, we replaced each on-off pair by standard ReLU channels and trained using standard backpropagation. The first baseline uses one ReLU channel for each on-off pair, matching the number of represented feature units but using half as many scalar hidden channels. The second doubles the ReLU width, matching the number of scalar hidden channels in the on-off model.

Bottom-up/top-down alignment variants. The *Aligned* condition initializes corresponding BU and TD weights identically and applies matched local Hebbian updates. The *Correlated* condition relaxes exact alignment by initializing TD weights as noisy copies of the corresponding BU weights, using independent elementwise Gaussian noise with standard deviation 0.01. The *Correlated + Noisy Updates* condition uses the same correlated initialization and additionally adds independent elementwise Gaussian noise with standard deviation 0.01 to the TD updates, so that the two pathways do not receive identical updates. The *Independent Init.* condition initializes corresponding BU and TD weights independently, yielding approximately zero initial correlation.

The first three conditions use the common training configuration described above. For *Independent Init.*, we selected the weight initialization scale and the corresponding optimization configuration using CIFAR-10 only. The selected initialization scale multiplies the default weight initialization by 0.1, reducing the magnitude of the independently initialized pathway-specific components. We therefore pair it with a higher learning rate of 0.04 and longer training of 100 epochs, allowing the accumulated learned component to reach a comparable scale. This configuration was then transferred unchanged to MNIST and Fashion-MNIST, without further dataset-specific tuning.

For completeness, Table 3 also reports the independently initialized model using the common hyperparameter configuration. The CIFAR-selected configuration yields substantially higher accuracy, particularly on CIFAR-10, while also transferring successfully to MNIST and Fashion-MNIST.

Table 3: Independent BU-TD initialization under the common and CIFAR-selected hyperparameter configurations. Results are test accuracy (%) averaged over 10 random seeds, with standard deviation.

Configuration	MNIST	Fashion-MNIST	CIFAR-10
Common hyperparameters	95.244 ± 0.492	79.468 ± 1.814	41.707 ± 2.071
CIFAR-selected	97.932 ± 0.078	87.298 ± 0.416	55.604 ± 0.887

The independently initialized condition provides an empirical test outside the aligned regime for which exact equivalence to backpropagation is guaranteed. With approximately zero initial BU-TD correlation, the propagated TD signals and resulting local updates are not expected to initially match the corresponding backpropagation signals and gradients. Nevertheless, the model learns effectively without requiring the BU and TD weights to be copied or aligned at initialization. With the CIFAR-selected configuration, performance approaches that of the aligned model, showing that effective learning can extend substantially beyond the conditions required for exact theoretical equivalence.

on-off motif variants. In the default setting, the internal motif weights are fixed to unit magnitude, implementing the idealized difference operation. In the *Learned Unit* setting, the internal motif weights are initialized to one and learned independently for each unit while preserving their signs. In the *Shared Learned Unit* setting, the internal motif weights are initialized to one and learned, but a single set of motif weights is shared across all units within a layer.

A.3.2 Tiny ImageNet experiments

For Tiny ImageNet, we used a convolutional architecture following the scale of the Self-Contrastive Forward–Forward Tiny ImageNet setup. The model was trained on Tiny ImageNet-200 images of size 64×64. We report both top-1 and top-5 accuracy, averaged over 5 random seeds.

The on-off convolutional network consists of five convolutional stages followed by a linear output layer. Each convolutional stage applies a convolution over the current on-off channel representation, followed by the on-off motif and $\mathrm{ReLU}^{\pm}$ activation. Max pooling is applied after stages 1, 2, and 5. The architecture is summarized in Table 4. The number of on-off pairs matches the corresponding vanilla SCFF-style architecture, while the actual number of channels is doubled due to the on-off representation.

For the vanilla convolutional baselines, each on-off unit is replaced by a standard ReLU channel. The matched-effective-width vanilla baseline uses the vanilla channel counts shown in Table 4. The matched-neuron-count baseline doubles these channel counts: $3 \to 128$, $128 \to 384$, $384 \to 768$, $768 \to 512$, and $512 \to 512$.

The on-off Tiny ImageNet model was trained in the aligned BU-TD setting using the proposed local Hebbian rule. We used Cross-Entropy loss, AdamW with learning rate 2.5×10^{-4}, weight decay 10^{-5}, batch size 256, and 25 epochs. The learning rate followed a cosine schedule with warmup: warmup fraction 0.1, peak learning-rate multiplier 1.1, and final learning-rate multiplier 0.1.

For the CReLU control, we use the same convolutional architecture and optimization configuration as the on-off model. The only architectural change is the paired activation mechanism. Whereas an on-off unit receives two non-negative inputs, forms their difference $d = x_1 - x_2$, and

Table 4: Tiny ImageNet convolutional architectures. The on-off model uses the same number of effective channel pairs as the vanilla SCFF-style architecture, but represents each channel by an on-off pair, doubling the actual number of scalar channels. Max pooling is applied after stages 1, 2, and 5.

Stage	Kernel	Pool	Vanilla channels	on-off actual channels
Conv1	5×5, pad 2	yes	$3 \to 64$	$6 \to 128$
Conv2	3×3, pad 1	yes	$64 \to 192$	$128 \to 384$
Conv3	3×3, pad 1	no	$192 \to 384$	$384 \to 768$
Conv4	3×3, pad 1	no	$384 \to 256$	$768 \to 512$
Conv5	3×3, pad 1	yes	$256 \to 256$	$512 \to 512$

outputs $[\mathrm{ReLU}(d), \mathrm{ReLU}(-d)]$, CReLU operates on a single scalar pre-activation x and outputs $[\mathrm{ReLU}(x), \mathrm{ReLU}(-x)]$. Thus, both models propagate paired positive and negative channels with the same resulting channel widths, but CReLU generates the pair directly from a conventional scalar pre-activation rather than through the Dale-constrained on-off circuit. CReLU is trained using standard backpropagation. All other optimization settings, including the optimizer, learning rate, weight decay, batch size, number of epochs, and learning-rate schedule, are identical to those used for the on-off model.

A.4 Learned on-off representations

To further inspect the representations learned by the on-off architecture, we visualize activation maps from the convolutional model trained on Tiny ImageNet. We follow the activation-maximization visualization approach of Zeiler & Fergus (2014). We focus on two representative on-off units from the second convolutional layer. For each unit, we score every image in the Tiny ImageNet test set by the maximum spatial activation of each channel:

$$s_k^+(x) = \max_{h,w} h_k^+(x,h,w), \qquad s_k^-(x) = \max_{h,w} h_k^-(x,h,w), \tag{40}$$

where h_k^+ and h_k^- denote the post-$\mathrm{ReLU}^{\pm}$ On and Off activation maps. We then select the top 8 images for the On channel and the top 8 images for the Off channel, forming a set of 16 images per unit. For each selected image, we display both the On and Off activation maps together with the original image.

The normalized map is overlaid on the original image using a heatmap, so bright regions indicate where the corresponding channel responded most strongly. Because maps are normalized separately, the visualization indicates spatial selectivity within each image, rather than absolute activation magnitude across images or channels.

Figure 6 shows two representative examples of learned on-off units. The first unit exhibits a boundary-sensitive representation. Across the selected images, the On and Off channels respond to complementary parts of object boundaries. This behavior arises because the two channels are associated with opposite effective kernels: a visual pattern that produces a positive response in one channel produces the complementary response in the other. As a result, one channel tends to activate on one side or polarity of an edge, while the opposite channel activates on the reversed edge polarity. For example, when an object boundary contains a transition in one direction, the On channel may highlight that transition. Because opposite contrast transitions often occur at different sides of an object boundary, the On and Off channels can highlight complementary portions of the same object's contour. This arises from basic properties of natural images where often a transition from background visual semantics to object are symmetric across spatial locations of the object. Together, the two channels provide a richer description of object shape than either channel alone, because the pair can represent both directions of local contrast within the same unit.

The second unit illustrates a different form of complementarity, related to color and foreground–background contrast. Here too, the paired $\mathrm{ReLU}^+/\mathrm{ReLU}^-$ channels learn opposite effective kernels. One channel often responds strongly to regions associated with the object or foreground, especially where color or texture differs from the surrounding image context. The opposite channel tends to emphasize complementary regions, such as background structure or regions with the oppo-

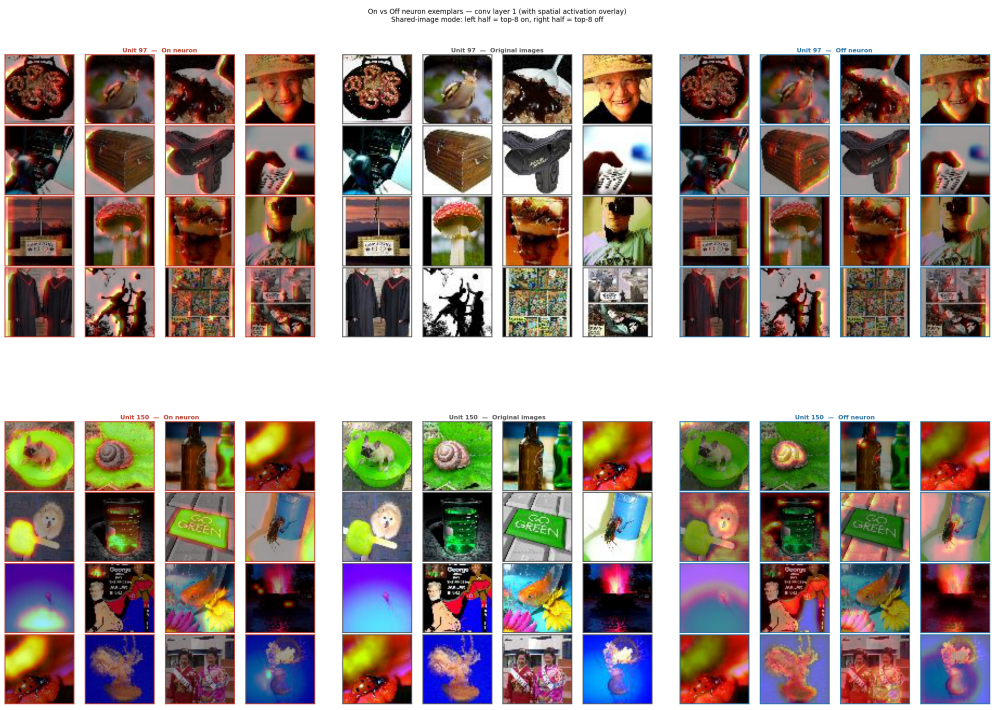

Figure 6: **Learned complementary on-off activations.** Activation maps from two representative on-off units sampled from the second convolutional layer of the Tiny ImageNet model. Activation maps are visualized using the activation-based method of Zeiler & Fergus (2014). For each unit, images are selected as the union of the top 8 test images ranked by maximum On-channel activation and the top 8 test images ranked by maximum Off-channel activation. For every selected image, the original image is shown together with both corresponding activation maps, allowing comparison between the On and Off responses for the same input. Activation maps are post-$\mathrm{ReLU}^{\pm}$ responses, bilinearly upsampled to the input resolution, and normalized independently per image for visualization. **Top:** A boundary-sensitive unit, where the paired channels respond to complementary edge or boundary patterns. **Bottom:** A color- or contrast-sensitive unit, where the paired channels respond to complementary object and background structure. These examples illustrate that paired on-off channels can learn complementary visual representations rather than redundant duplicated features.

site color/contrast relationship. Thus, the pair separates complementary visual evidence along the same learned feature dimension, rather than simply duplicating the same detector twice.

These qualitative examples support the view that paired on-off channels are not redundant copies. Instead, the $\mathrm{ReLU}^{+}/\mathrm{ReLU}^{-}$ construction encourages each unit to learn a feature together with its opposite-sign counterpart, yielding complementary responses that jointly encode a more complete visual feature: opposite edge polarity in the first example, and complementary color or foreground–background structure in the second. Importantly, this structure emerges from the on-off motif and the paired rectifying channels without explicit supervision encouraging complementary channel behavior. This is in line with biological observations of on-off and opponent channels in sensory systems in low-level parts of the visual system (Hubel & Wiesel, 1962; Schiller, 1992), as well as in higher-level, e.g., in face and expression perception (Valentine, 1991; Skinner & Benton, 2010).

A.5 BU-TD ALIGNMENT DYNAMICS

The independently initialized condition in Section 5 shows that effective learning does not require BU and TD pathways to be aligned at initialization. We next examine how the relation between the two pathways evolves during learning and whether this evolution is reflected in the quality of the resulting local updates.

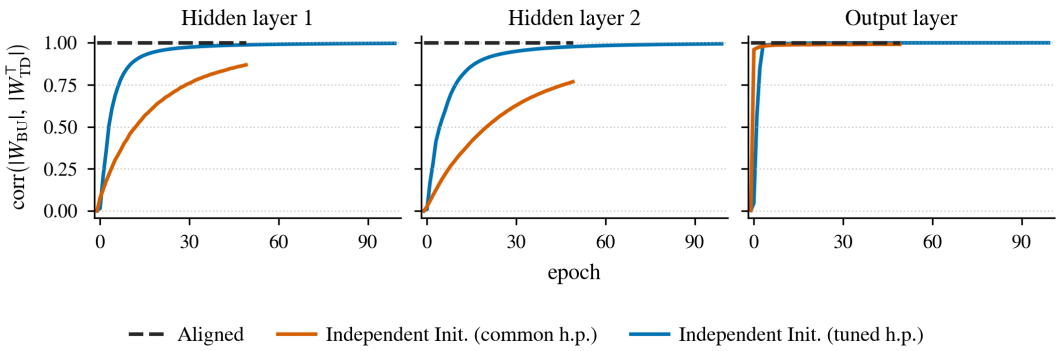

Figure 7: **BU-TD weight alignment during learning.** Pearson correlation between corresponding BU and TD effective weights, $|W_l^{\mathrm{BU}}|$ and $|W_l^{\mathrm{TD}}|^T$, on CIFAR-10, shown separately for the first hidden layer, second hidden layer, and output layer (mean ± 1 s.d. over 5 seeds). Independently initialized pathways begin with approximately zero correlation and become increasingly aligned as matched local updates accumulate. Under the configuration selected for independent initialization, correlation approaches one in all layers. The aligned condition remains at one throughout training. The independently initialized configurations use different training budgets, so their curves terminate at different epochs.

Setup. We use the two-hidden-layer on-off MLP on CIFAR-10 with 500 on-off units per layer and independently parameterized BU and TD weights. We compare three conditions: *Aligned*, using the common training configuration with identical BU/TD initialization; *Independent Init. (common h.p.)*, using independent BU/TD initialization with the same common configuration; and *Independent Init. (selected h.p.)*, using the CIFAR-10 configuration selected for the independently initialized regime. The latter scales the initialization of both pathways by 0.1, uses learning rate 0.04, batch size 512, and trains for 100 epochs, whereas the common configuration uses the default initialization scale, learning rate 0.01, batch size 64, and 50 epochs. Results are averaged over 5 random seeds, with shaded regions indicating ± 1 standard deviation. These diagnostic runs are separate from the 10-seed accuracy results reported in Table 1.

Weight alignment. For each weighted layer, we compare the BU and corresponding TD weights in the common BU orientation. Because the proposed update acts on non-negative synaptic magnitudes, alignment is measured in effective-weight space, comparing $|W_l^{\mathrm{BU}}|$ with $|W_l^{\mathrm{TD}}|^T$. We compute the Pearson correlation between the corresponding matrix entries at initialization and after every epoch. Independently initialized pathways therefore begin with approximately zero correlation, whereas transpose-aligned pathways have correlation one.

Figure 7 shows that independently initialized pathways become increasingly aligned during learning. Their initial Pearson correlations are approximately zero in all layers. Under the common hyperparameter configuration, final correlations reach 0.869, 0.768, and 0.991 in the first hidden, second hidden, and output layers, respectively. Under the configuration selected for independent initialization, the corresponding correlations reach 0.997, 0.994, and 1.000. The aligned condition remains at correlation 1 throughout training.

These dynamics are consistent with the matched-update structure of the model. Corresponding BU and TD pathways accumulate a shared learned component while retaining their independently initialized components. Their absolute initial mismatch therefore need not vanish; rather, their *relative* alignment increases as the shared learned component becomes large relative to the initialization-specific components. Reducing the initialization scale makes this shared component dominant earlier, explaining the substantially stronger alignment in the selected configuration. Importantly, no step of this process reads or copies the corresponding BU weights into the TD pathway.

Alignment with the backpropagation update. We next test whether improved BU-TD alignment translates into more accurate credit assignment. At selected training checkpoints, we measure the cosine similarity between the local Hebbian update produced by the model and the gradient-descent

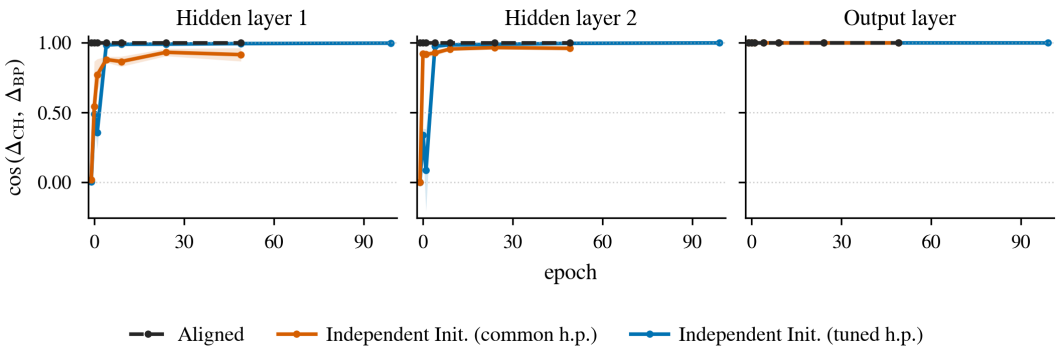

Figure 8: **Alignment of local updates with backpropagation.** Cosine similarity between the local Hebbian update and the gradient-descent update produced by backpropagation for the same batch, shown separately for each layer (mean ± 1 s.d. over 5 seeds). Backpropagation gradients are computed only for this diagnostic and are never used for training. Independently initialized hidden layers begin near zero alignment and become increasingly aligned during learning; under the selected independent-initialization configuration, cosine similarity reaches 0.997 and 0.999 in the two hidden layers. The output layer remains at 1 by construction.

update that standard backpropagation would produce for the same batch. The backpropagation gradient is computed only for this diagnostic and is never used to update the model. Measurements are performed on a fixed 256-example batch at checkpoints throughout training.

As expected from the theoretical result, the aligned condition has update cosine similarity 1 in every layer throughout training. In the independently initialized conditions, the two hidden layers begin with essentially uncorrelated local and backpropagation updates. Under the common hyperparameters, their cosine similarities increase to 0.915 and 0.960 by the end of training. Under the configuration selected for independent initialization, they reach 0.997 and 0.999, respectively. The output-layer update has cosine similarity 1 in all conditions by construction, since its TD signal is initialized directly from the loss gradient.

Thus, the increase in relative BU-TD weight alignment is accompanied by a corresponding increase in gradient alignment of the local learning rule. Together with the performance results in Table 1, these diagnostics show that independently initialized feedback pathways can develop highly accurate credit-assignment signals through the proposed local learning dynamics, without explicit weight copying.