ORIENTATION SELECTIVE DENDRITIC FEATURE REPRESENTATION SUPPORTS ROBUST PATTERN IDENTIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Orientation-selective synaptic inputs in visual cortex motivate studying how feature selectivity is formed and represented within dendritic computations, and how dendritic representations influence network-level visual processing. We investigate this question using a Two-Stage Dendritic Network (TSDN), in which multiple dendritic branches process local inputs and their nonlinear responses are integrated at the soma. Hebbian and anti-Hebbian plasticity first learn feature detectors from unlabeled inputs; these detectors are then frozen while somatic integration and a classification readout are trained with supervision. Gabor analysis reveals orientation-selective feature structure, while ablations show that anti-Hebbian plasticity reduces detector redundancy and improves downstream classification relative to Hebbian-only learning. Across five image datasets, TSDN achieves higher mean classification accuracy than the tested backpropagation-trained baselines. It also exhibits improved robustness to Gaussian input noise on all tested datasets except CIFAR-10. Varying the anti-Hebbian learning rate reveals a discriminability–robustness trade-off, accompanied by non-monotonic changes in feature frequency content and branch-response sensitivity. These findings connect unsupervised dendritic feature formation with recognition and noise tolerance, highlighting plasticity strength as a factor shaping the computational properties of dendritic representations.

1 INTRODUCTION

Deep neural networks have achieved remarkable success on perceptual tasks such as image classification and speech recognition (Graves et al., 2013; Hinton et al., 2012; Krizhevsky et al., 2012). However, they differ from biological systems in two fundamental respects. First, mainstream architectures typically model neurons as point-like units, abstracting away the rich local computations performed by dendrites (Häusser et al., 2000; London & Häusser, 2005; Poirazi et al., 2003; Branco et al., 2010; Polsky et al., 2004; Stuart & Spruston, 2015). Second, their dominant training paradigm—end-to-end supervised learning with backpropagation (Rumelhart et al., 1986)—uses globally propagated error signals and large labeled datasets, whereas the brain develops much of its perceptual machinery from unlabeled sensory experience via local synaptic plasticity (Crick, 1989; Lillicrap et al., 2020). These discrepancies motivate a central question: how does unsupervised synaptic plasticity shape dendritic feature representations, and how do these representations support visual processing at the network level?

Experimental observations in visual cortex provide a biological starting point for specifying the dendritic representations of interest (Bonhoeffer & Grinvald, 1991; Ohki et al., 2005). Synaptic inputs to individual dendritic spines exhibit orientation tuning (Wilson et al., 2016; Iacaruso et al., 2017), and inputs with different orientation preferences can coexist within the dendritic tree of a single neuron (Jia et al., 2010). Dendritic branches can also act as nonlinear computational subunits, integrating local inputs before their contributions are combined at the soma (Polsky et al., 2004; Smith et al., 2013). Together, these observations motivate a computational view of dendritic branches as units that process selective visual features. From this perspective, both the selectivity of individual feature detectors and the differences among detectors are relevant properties of dendritic representations. This biological motivation guides our investigation of how unsupervised plasticity shapes

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.

these properties (Kirchner & Gjorgjieva, 2021) and how the resulting representations support visual processing at the network level.

A further biological constraint is that plasticity depends on a synapse's location within the dendritic tree (Harvey & Svoboda, 2007; Govindarajan et al., 2011; Losonczy et al., 2008; Cichon & Gan, 2015). Backpropagating action potentials, which can convey information about the neuron's output to its dendrites (Urbanczik & Senn, 2014), attenuate with distance from the soma (Sjöström & Häusser, 2006; Stuart & Sakmann, 1994). Motivated by this unequal access to output-related signals, we adopt a two-stage computational setting: dendritic features are first learned from unlabeled inputs without task supervision and then frozen, while a supervised readout is trained for downstream tasks (Illing et al., 2019). This separation allows us to examine the selectivity and diversity of features formed without task supervision, together with their utility for classification and robustness to input noise.

To study this question, we adopt a dendrite-inspired architecture motivated by spatially localized receptive fields in V1 (Hubel & Wiesel, 1962; Olshausen & Field, 1996). Each soma receives input from a localized region of the image and integrates responses from multiple dendritic branches, providing a computational setting for combining different feature detectors within a neuron. Building on established dendritic architectural templates (Poirazi et al., 2003; Chavlis & Poirazi, 2025), we use a Two-Stage Dendritic Network (TSDN) to investigate the formation and computational properties of dendritic feature representations. Within the two-stage setting described above, Hebbian and anti-Hebbian plasticity learn feature detectors from unlabeled input patches, allowing us to examine how their interaction shapes feature selectivity and diversity. We characterize the resulting representations through orientation and diversity analyses, and evaluate their utility for classification and robustness to input noise across multiple image datasets. Ablations of the plasticity components and comparisons with a backpropagation-trained MLP and an end-to-end backpropagation-trained dendritic network assess the contributions of unsupervised feature learning and the behavior of the resulting representations.

In summary, our contributions are:

- We show that unsupervised Hebbian and anti-Hebbian plasticity form orientation-selective dendritic features, with anti-Hebbian learning reducing redundancy and improving their utility for classification.
- We demonstrate that these features, combined through dendritic processing and somatic integration, support competitive classification across five datasets. Improved Gaussian-noise robustness over the tested backpropagation baselines is observed on all datasets except CIFAR-10.
- We identify a plasticity-dependent discriminability–robustness trade-off in dendritic representations, associated with changes in feature redundancy, spatial-frequency content, and branch-response sensitivity.

2 Model

2.1 Architecture

As illustrated in Figure 1, the TSDN follows the general template of dendrite-inspired models: each computational unit (a soma) receives input from a fixed, spatially localized receptive field rather than from the full image, mimicking the local connectivity of dendritic branches. Learning is organized into two stages. First, a dendritic feature layer learns feature detectors from unlabeled data using local Hebbian and anti-Hebbian plasticity and is then frozen. Second, a somatic classification layer is trained supervised on top of the frozen dendritic responses.

Dendritic feature layer (unsupervised). The network contains n_{soma} independent somas. Each soma i is assigned a fixed receptive-field location (h_i, w_i) corresponding to a 10×10 patch; (h_i, w_i) is sampled i.i.d. uniformly at random before training and held fixed thereafter (receptive fields may overlap). Each soma has $n_{dendrite}$ dendritic branches. Each branch applies a learned feature detector to the local patch via an inner product followed by a ReLU nonlinearity, producing a branch response. There are no lateral connections between somas. After unsupervised training, the

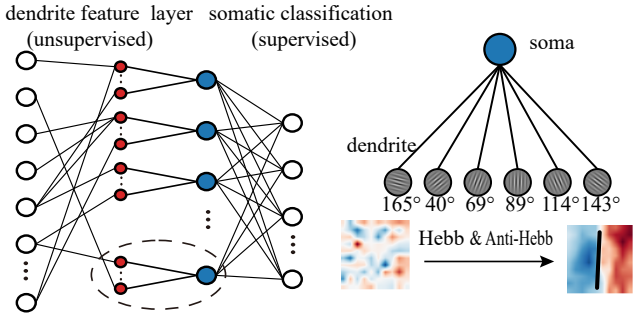

Figure 1: Schematic of the TSDN. Left: Dendritic branches learn local features through unsupervised plasticity, while a supervised somatic readout produces class logits. Right: Branches within a soma develop diverse orientation preferences (top), and their features are updated by Hebbian and anti-Hebbian learning (bottom).

feature-layer weights are frozen and the layer outputs a representation $X \in \mathbb{R}^{n_{\text{soma}} \times n_{\text{dendrite}}}$. Unless otherwise stated, we use n_{soma} = 500 and $n_{dendrite}$ = 10 (selected via the parameter scan in Fig. 4A to balance accuracy and parameter efficiency).

Somatic classification layer (supervised). Classification proceeds in two steps: (i) dendritic integration within each soma and (ii) a fully connected readout across somas. For integration, soma i computes a weighted sum of its branch responses using learnable weights $W_{\text{d}} \in \mathbb{R}^{n_{\text{soma}} \times n_{\text{dendrite}}}$, followed by ReLU, yielding soma activities $s \in \mathbb{R}^{n_{\text{soma}}}$. A final linear layer with weights $W_{\text{fc}} \in \mathbb{R}^{C \times n_{\text{soma}}}$ maps s to C class logits. Only (W_{d}, W_{fc}) are trained in the supervised stage (cross-entropy with Adam).

2.2 Local learning rules and training

Unsupervised learning updates the feature-detector matrix $\Phi = [\phi_1, \ldots, \phi_M]^\top$ using Hebbian and anti-Hebbian plasticity (Földiák, 1990), where ϕ_m is the weight vector of the m-th detector. At each iteration, we sample a 10×10 patch x from a training image and normalize it to $\hat{x}$. Feature activations are sparsified with a k-WTA mechanism that keeps the top 20% most active units, producing a_{sparse}. We then apply the update

$$\Phi(t+1) = \Phi(t) + \eta_{hebb}\, a_{sparse}\, \hat{x}^\top - \eta_{anti}\, C\, \Phi(t) \tag{1}$$

The update is followed by per-feature L2 normalization. The co-activation matrix C in the anti-Hebbian term suppresses correlated features:

$$\begin{aligned} \phi_m &\leftarrow \frac{\phi_m}{\|\phi_m\|_2 + \delta}, \quad \delta = 10^{-8}, \\ C &= a_{\text{sparse}} a_{\text{sparse}}^\top, \quad \operatorname{diag}(C) = 0. \end{aligned} \tag{2}$$

Unless otherwise stated, we set $\eta_{\text{hebb}} = 0.01$ and $\eta_{\text{anti}} = 0.0025$. Unsupervised training runs for 60,000 iterations, after which Φ is frozen.

Feature extraction (frozen dendrites). For soma i and branch j, we extract the local patch x_i at (h_i, w_i) and compute $d_{i,j} = \text{ReLU}(\phi_{i,j}^\top \text{x}_i)$, where $\phi_{i,j}$ denotes the feature detector associated with branch j of soma i. $d_{i,j}$ represents the response of dendritic branch j in soma i.

Supervised classification (trainable soma/readout). Branch responses are integrated and mapped to logits via

$$s_i = \text{ReLU}(\sum_j W_{\text{d}i,j} d_{i,j}), \hat{y} = W_{\text{fc}} s + b \tag{3}$$

where $\hat{y} \in \mathbb{R}^C$ are class logits and $b \in \mathbb{R}^C$ is a bias, s_i represents the response of soma i. We train $(W_{\mathrm{d}}, W_{\mathrm{fc}}, b)$ for 10 epochs using Adam with learning rate 10^{-3}.

2.3 Orientation analysis via Gabor fitting

To quantify orientation selectivity, we fit a 2D Gabor function to each learned feature detector ϕ_m (Jones & Palmer, 1987) and extract its preferred orientation θ_m:

$$G(x, y) = A \exp\left[-\frac{1}{2}\left(\frac{x'^2}{\sigma_x^2} + \frac{y'^2}{\sigma_y^2}\right)\right] \cos(2\pi f x' + \psi) \tag{4}$$

with the rotated coordinates

$x^{'} = (x - c_x)\cos\theta + (y - c_y)\sin\theta$, $y^{'} = -(x - c_x)\sin\theta + (y - c_y)\cos\theta$, $\theta \in [0, \pi]$, where A, f, and ψ denote the amplitude, spatial frequency, and phase of the sinusoidal carrier, respectively, and σ_x, σ_y control the Gaussian envelope. We fit Gabor functions using multi-start nonlinear least squares (12 initializations) and quantify fit quality by the Pearson correlation r between the learned filter ϕ_m and its fitted counterpart. Orientation θ_m is retained only when $r \geq 0.6$; otherwise, it is marked as missing. For visualization, we report edge orientation as $\theta_{\mathrm{display}} = (\theta_m + 90^\circ) \mathrm{mod} 180^\circ$

2.4 Diversity metrics

We quantify feature diversity using two complementary measures.

$$\cos = \frac{2}{M(M-1)} \sum_{1 \leq m < l \leq M} \frac{|\phi_m^\top \phi_l|}{\|\phi_m\| \|\phi_l\|}, \qquad \mathrm{PR} = \frac{(\sum_i \lambda_i)^2}{\sum_i \lambda_i^2}. \tag{5}$$

Mean cosine similarity (cos). We quantify feature redundancy by averaging the absolute pairwise cosine similarity among M learned detectors. Lower values indicate more diverse feature detectors.

Participation ratio (PR). PR measures the effective dimensionality of the dendritic response representation. Here, λ_i denotes the i-th eigenvalue of the covariance matrix computed from dendritic branch responses across input.

2.5 Robustness evaluation

We evaluate on MNIST, FashionMNIST, KMNIST, EMNIST, and CIFAR-10. Each experiment is repeated with five random seeds; we report mean and standard deviation. For noise robustness, we corrupt only the test set with additive Gaussian noise:

$$\widetilde{X} = \mathrm{clip}(X + \epsilon, 0, 1), \qquad \epsilon \sim \mathcal{N}(0, \sigma^2). \tag{6}$$

with noise intensity $\sigma \in \{0.0, 0.1, \ldots, 1.0\}$. Training always uses clean images.

We additionally quantify the intrinsic noise sensitivity of feature responses. Given clean inputs X and their noisy counterparts $\widetilde{X}$, the noise sensitivity is defined as the normalized change in branch-level feature responses:

$$S_{noise} = \frac{\|d(\widetilde{X}) - d(X)\|_2}{\|d(X)\|_2 + \delta}, \delta = 10^{-8} \tag{7}$$

where $d(X)$ represents the concatenated responses of all feature branches. Lower S_{noise} indicates that the learned features are less sensitive to input perturbations.

2.6 Spatial-frequency content of learned features

To characterize the spatial-frequency content of each learned feature detector, we compute its 2D power spectrum via FFT and define the high-frequency (HF) energy ratio as the fraction of spectral

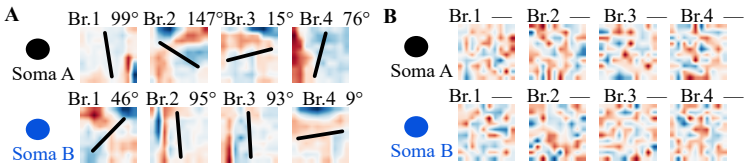

Figure 2: Orientation selectivity of dendritic branches. (A) Unsupervised training produces diverse orientation preferences within and across somas. (B) Random features show no clear orientation structure. Lines and numbers indicate fitted orientations; "—" denotes an invalid Gabor fit.

energy lying at high spatial frequencies. Specifically, frequencies whose radial distance from the DC component exceeds half the Nyquist frequency (equivalently, greater than one-quarter of the patch size in frequency-bin units) are classified as HF. We average the HF ratio across all feature detectors to obtain a single HF value per model configuration; a larger HF value indicates finer, more rapidly varying spatial structure in the learned features.

3 Results

Building on the two-stage design—an unsupervised dendritic feature layer followed by a supervised somatic classification layer (Figure 1)—we first examine the orientation selectivity of the learned dendritic features and assess how anti-Hebbian plasticity shapes feature diversity. We then evaluate the classification performance and noise robustness of the resulting representations relative to end-to-end supervised baselines.

3.1 Formation of Dendritic Feature Representations

To test whether the TSDN develops orientation-selective dendritic feature, we analyzed the learned dendritic feature detectors after the unsupervised stage. For each detector, we fit a two-dimensional Gabor function and quantified fit quality using the Pearson correlation coefficient r. Detectors with $r \geq 0.6$ were treated as valid Gabor-like features, and the fitted preferred orientation θ was extracted.

As shown in Figure 2(A), unsupervised learning produces localized feature detectors with clear orientation selectivity. Moreover, branches within the illustrated somas exhibit different preferred orientations, showing distinct feature preferences within individual somas. This pattern is consistent with experimental findings that synaptic inputs on different dendritic branches can exhibit distinct orientation preferences in visual cortex (Jia et al., 2010; Wilson et al., 2016). As a control, we applied the same analysis to randomly initialized (untrained) feature detectors. In this case, most detectors did not exhibit clear orientation selectivity (Figure 2(B)). Quantitatively, trained detectors achieved substantially higher fit quality than random detectors (trained: mean r = 0.861; random: mean r = 0.425; Mann–Whitney U test, U = 16384, $p = 8.57 \times 10^{-44}$), indicating that orientation selectivity is not an incidental consequence of architecture or initialization but emerges from local unsupervised learning.

Having established orientation-selective structure in the learned dendritic features, we next examine how Hebbian and anti-Hebbian plasticity shape feature diversity. In particular, does the anti-Hebbian component reduce feature redundancy relative to Hebbian learning alone? To isolate the contribution of each component, we conducted an ablation study on FashionMNIST with three unsupervised conditions: random (untrained) features, Hebbian-only learning, and Hebbian & anti-Hebbian learning. We additionally varied the anti-Hebbian learning rate η_{anti} to quantify how strongly anti-Hebbian modulation controls the outcome.

Under Hebbian-only learning, different branches gradually converged toward similar orientation preferences during training. For example, two representative detectors (d24 and d87) converged to nearly identical preferred orientations after 10,000 iterations (86° and 88°, respectively; Figure 3(A)). In contrast, with the addition of anti-Hebbian plasticity, the corresponding detectors developed distinct orientation preferences, converging to 3° and 86°, respectively (Figure 3(A)). To track the evolution of branch representations during training, we measured the mean pairwise cosine similarity among detectors over iterations. Under Hebbian-only learning, cosine similarity grad-

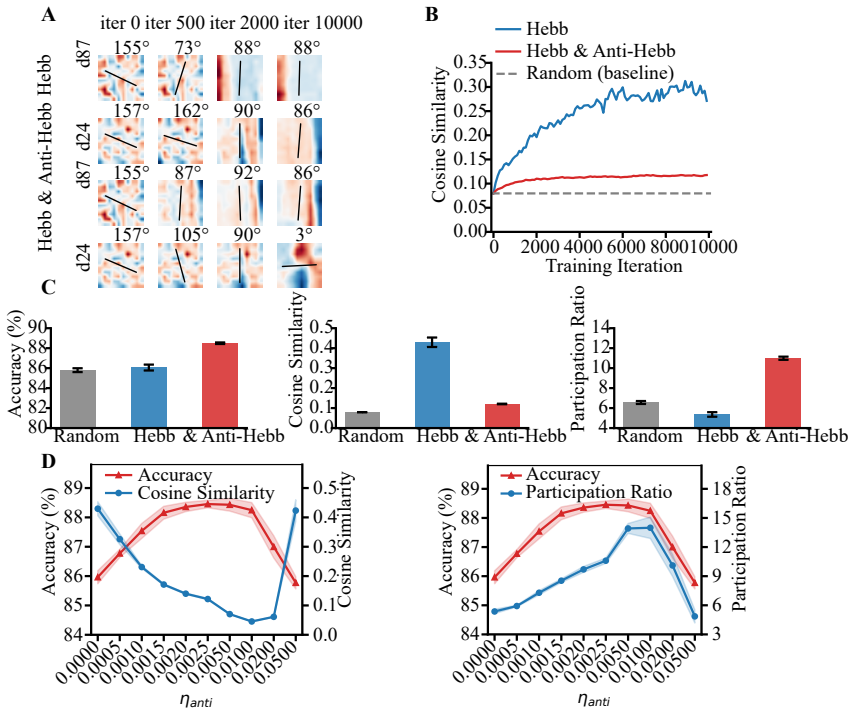

Figure 3: Anti-Hebbian plasticity shapes dendritic feature representations. (A) Training trajectories of two representative detectors (d24 and d87) under Hebbian-only (top) and Hebbian & anti-Hebbian learning (bottom). Anti-Hebbian plasticity preserves orientation diversity. (B) Mean pairwise cosine similarity during training. (C) Accuracy, cosine similarity, and participation ratio for random, Hebbian, and Hebbian & anti-Hebbian features. (D) Effects of the anti-Hebbian learning rate η_{anti} on accuracy and feature diversity. Error bars and shaded regions denote standard deviations across random seeds.

ually increased throughout training (Figure 3(B)), indicating that branch representations became increasingly similar over time. In contrast, adding anti-Hebbian plasticity maintained lower cosine similarity, preventing the progressive accumulation of redundant features.

To quantify the effects of different learning rules on learned representations, we compared downstream classification accuracy together with the mean cosine similarity between learned feature detectors and the participation ratio of the resulting feature space across random features, Hebbian-only learning, and Hebbian & anti-Hebbian learning (Figure 3(C)). Hebbian-only learning slightly improved classification accuracy compared with random features (86.06% vs.85.80%), indicating that Hebbian plasticity can extract task-relevant structure from unlabeled inputs. However, this improvement was accompanied by increased mean cosine similarity and reduced participation ratio, revealing a more redundant and lower-dimensional representation. Adding the anti-Hebbian component improved classification accuracy to 88.50% while promoting a more diverse representation. Specifically, anti-Hebbian learning increased the participation ratio and reduced mean cosine similarity compared with Hebbian-only learning, indicating reduced feature overlap among branches. These results suggest that anti-Hebbian plasticity suppresses excessive co-activation and promotes branch specialization without requiring additional parameters.

The effect of anti-Hebbian plasticity on learned representations depended non-monotonically on the anti-Hebbian learning rate η_{anti} (Figure 3(D)). When η_{anti} was too small, the reduction in representation overlap was limited, resulting in less diverse branch representations. In contrast, excessively large η_{anti} imposed excessive anti-Hebbian modulation: cosine similarity increased toward the Hebbian-only regime, participation ratio decreased, indicating increased redundancy among branch representations. Overall, Hebbian learning extracts meaningful local features, while anti-Hebbian plasticity—at an appropriate strength—maintains feature diversity; together they produce

high-quality unsupervised dendritic representations. Unless otherwise noted, all subsequent experiments use the full TSDN with $\eta_{\text{anti}} = 0.0025$.

3.2 Validating Model Performance across Multiple Datasets

To evaluate classification performance and generalization, we compared the TSDN against two end-to-end backpropagation baselines on five datasets: MNIST, FashionMNIST, KMNIST, EMNIST, and CIFAR-10. BP-MLP is a conventional fully connected network with two hidden layers trained end-to-end by backpropagation, without dendritic structure or localized receptive fields. BP-DN, following the dendrite-inspired baseline of (Chavlis & Poirazi, 2025), uses the same overall architecture as TSDN but trains the dendritic feature detectors end-to-end with backpropagation rather than with unsupervised local plasticity. For each dataset, we tuned model size independently for each method and report results for the best-performing configuration.

Table 1: Classification accuracy (%) on five datasets. Values are mean ± standard deviation over five random seeds. The best result in each column is shown in bold.

Method	MNIST	FashionMNIST	KMNIST	EMNIST	CIFAR-10
BP-MLP	97.78±0.31	88.72±0.22	90.74±0.51	84.92±0.17	50.37±0.78
BP-DN	98.64±0.09	89.03±0.28	91.74±0.55	85.83±0.12	54.23±0.87
TSDN	**98.71±0.07**	**89.98±0.22**	**92.20±0.19**	**85.96±0.10**	**54.96±0.43**

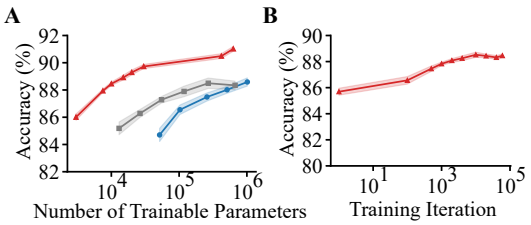

Figure 4: Classification performance and parameter efficiency of the TSDN. (A) Test accuracy versus the number of gradient-trained parameters, comparing TSDN, BP-DN, and BP-MLP. (B) Test accuracy of TSDN versus the number of unsupervised feature-learning iterations. Shaded regions indicate standard deviation across random seeds.

As shown in Table 1, the TSDN achieves higher accuracy than both backpropagation baselines across all five datasets, indicating that dendritic features learned via local unsupervised plasticity can match—and in our experiments exceed—the performance of end-to-end training across diverse visual recognition tasks. We next examined parameter efficiency on FashionMNIST (Figure 4(A) and Figure S1(A)). We swept the number of gradient-trained parameters by scaling the hidden-layer width for BP-MLP and scaling the number of somas and branches for TSDN/BP-DN, and recorded the resulting accuracies. Across a wide range of model sizes, TSDN achieves higher accuracy with substantially fewer gradient-trained parameters, requiring orders of magnitude fewer parameters than BP-based models. For example, TSDN achieves 89.98% accuracy with ~ 30K gradient-trained parameters, whereas BP-DN and BP-MLP require ~ 2.0 M and ~ 20.0 M gradient-trained parameters, respectively, to achieve comparable accuracy. These results show that the learned dendritic features support accurate classification with a smaller supervised parameter budget.

With identical supervised training on frozen features at each checkpoint, additional unsupervised learning increased classification accuracy and reduced test loss before performance saturated (Figure 4(B) and Figure S1(B)). These gains reflect improvements in the learned dendritic representations used for downstream classification.

3.3 Robustness Testing

To evaluate how the learned dendritic representations support classification under input noise, we compared the test accuracy of TSDN, BP-MLP, and BP-DN across Gaussian-noise levels. As shown in Figure 5 and Figure S2, TSDN consistently achieved higher accuracy under noise on MNIST,

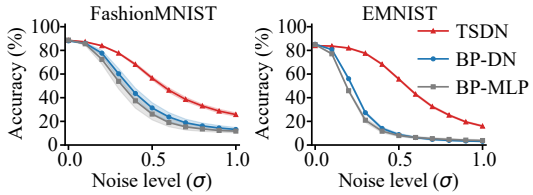

Figure 5: Noise robustness of the TSDN. Classification accuracy as a function of the intensity of input Gaussian noise on FashionMNIST and EMNIST, comparing the TSDN, BP-DN, and BP-MLP. Shaded areas indicate the standard deviation across multiple random seeds.

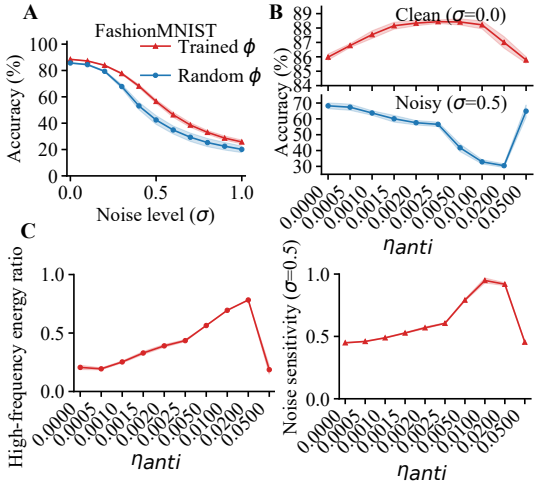

Figure 6: Factors shaping TSDN noise robustness and the robustness–discriminability trade-off. (A) With frozen feature layers, learned features (Trained ϕ) outperform random features (Random ϕ) across noise levels. (B) Accuracy versus η_{anti} shows an inverted-U trend for clean inputs (top, $\sigma = 0$) and a U-shaped trend for noisy inputs (bottom, $\sigma = 0.5$). (C) High-frequency energy (left) and feature noise sensitivity (right, $\sigma = 0.5$) increase together from low to intermediate η_{anti} but decrease at $\eta_{\text{anti}} = 0.05$. Shading denotes standard deviation across seeds.

FashionMNIST, KMNIST, and EMNIST. These results indicate that the unsupervised dendritic representations support more robust classification under Gaussian input noise than those learned by the tested end-to-end supervised baselines.

On the more complex CIFAR-10 dataset, however, TSDN did not show a consistent robustness advantage over the tested baselines (Figure S2), motivating further investigation of dendritic feature learning on complex natural images. To examine the contribution of unsupervised feature learning, we compared TSDN with a random-projection baseline on FashionMNIST (Figure 6(A)). This baseline retained the same dendritic architecture and supervised training protocol, but used frozen, randomly initialized feature detectors instead of learned ones. TSDN achieved higher accuracy at all tested noise levels, indicating that unsupervised feature learning improves classification under Gaussian noise beyond what is provided by the architecture with random features.

Building on the above results that unsupervised learning provides robustness to noisy inputs, we next investigated how the balance between Hebbian and anti-Hebbian plasticity shapes model robustness. Across different anti-Hebbian settings, we observed a trade-off between discriminative performance and noise tolerance: settings that achieved higher accuracy on clean inputs (top) tended to show lower accuracy under noisy conditions (bottom, Figure 6(B)). To further investigate the origin of this trade-off, we analyzed the frequency properties of the learned features and their sensitivity to input noise (Figure 6(C)). Specifically, we quantified the proportion of high-frequency energy in the learned features and measured their sensitivity to input noise under different anti-Hebbian learning rates. As η_{anti} increased from low to intermediate values, the high-frequency energy of the

learned features increased, accompanied by increased noise sensitivity (Figure 6(C)). This indicates that stronger anti-Hebbian modulation promotes the extraction of more fine-grained, high-frequency structures, which improves discriminative ability but also increases susceptibility to input perturbations. Together, these results indicate that the anti-Hebbian learning rate controls the balance between discriminative feature extraction and noise robustness.

4 DISCUSSION

We use the TSDN to investigate how unsupervised plasticity shapes dendritic feature representations and their computational properties. Anti-Hebbian plasticity plays a central role by reducing redundancy among learned feature detectors. Notably, the orientation preferences of these detectors are not prescribed in advance but emerge through unsupervised learning. These representations support competitive classification and improved Gaussian-noise robustness over the tested end-to-end baselines, with CIFAR-10 as the robustness exception. The parameter comparison shows a reduced supervised optimization budget, while the plasticity analysis reveals a discriminability–robustness trade-off associated with the anti-Hebbian learning rate.

TSDN differs from prior dendrite-inspired networks (Chavlis & Poirazi, 2025; Guerguiev et al., 2017; Sacramento et al., 2018) primarily in how features are learned. Many such models are trained using task-dependent signals. In contrast, TSDN learns dendritic features through unsupervised plasticity, freezes them, and restricts supervision to somatic integration and the readout. Within this framework, we investigate how locally learned dendritic features are organized and how their properties relate to network-level computation. The orientation analysis reveals that individual dendritic branches develop selective feature preferences, providing evidence for structured branch-level representations. The plasticity ablation experiments further show that anti-Hebbian modulation reduces feature redundancy and alters the diversity of learned features, which is associated with changes in classification performance and noise robustness. From this perspective, TSDN provides a computational framework for studying how local dendritic plasticity can shape selective feature representations and how these branch-level organizations influence the resulting network behavior.

Our results also suggest a mechanistic origin for the observed discriminability–robustness trade-off. Hebbian learning favors dominant variance directions in the input (e.g., principal-component-like solutions under related rules; (Oja, 1982)), which tends to produce smoother features that average over neighboring inputs. Such features can suppress independent noise and are therefore relatively robust, but may be less discriminative. Anti-Hebbian modulation, in contrast, promotes less redundant and often higher-frequency structure: this can improve discriminability by capturing finer details, while increasing sensitivity to input perturbations.

A broader implication of these findings is that efficient robustness may be attainable without end-to-end error propagation, provided that unsupervised feature learning includes an explicit mechanism for controlling redundancy. In TSDN, anti-Hebbian modulation acts as a local inductive bias that shapes the geometry of the representation—encouraging a distributed set of complementary features—while the two-stage protocol preserves the learned feature detectors during supervised training. This suggests a general design principle for local-learning systems: combine (i) local competitive feature formation (Hebbian-like plasticity) with (ii) local redundancy suppression (decorrelation/anti-Hebbian terms) and (iii) a lightweight readout, yielding representations that are both parameter-efficient and robust. More broadly, the ability to tune the discriminability–robustness balance through a single plasticity parameter points to a tractable route for adapting local-learning models to different deployment regimes (e.g., noise-tolerant sensing vs. fine-grained recognition) without increasing model size or relying on large labeled datasets.

This work has two main limitations. First, limited robustness on CIFAR-10 suggests that hierarchical structures, such as convolutional or multi-layer receptive fields, may be needed for complex natural images. Second, although dendritic learning is local and unsupervised, the somatic readout still relies on gradient descent. Future work could replace it with local, compartmentalized teaching signals, such as apical-dendrite-like feedback (Guerguiev et al., 2017; Sacramento et al., 2018), extending local learning to somatic and output-level plasticity.

REFERENCES

Tobias Bonhoeffer and Amiram Grinvald. Iso-orientation domains in cat visual cortex are arranged in pinwheel-like patterns. *Nature*, 353(6343):429–431, 1991. doi: 10.1038/353429a0.

Tiago Branco, Beverley A. Clark, and Michael Häusser. Dendritic discrimination of temporal input sequences in cortical neurons. *Science*, 329(5999):1671–1675, 2010. doi: 10.1126/science.1189664.

Spyridon Chavlis and Panayiota Poirazi. Dendrites endow artificial neural networks with accurate, robust and parameter-efficient learning. *Nature Communications*, 16:943, 2025. doi: 10.1038/s41467-025-56297-9.

Joseph Cichon and Wen-Biao Gan. Branch-specific dendritic Ca^{2+} spikes cause persistent synaptic plasticity. *Nature*, 520:180–185, 2015. doi: 10.1038/nature14251.

Francis Crick. The recent excitement about neural networks. *Nature*, 337(6203):129–132, 1989. doi: 10.1038/337129a0.

Peter Földiák. Forming sparse representations by local anti-hebbian learning. *Biological Cybernetics*, 64(2):165–170, 1990. doi: 10.1007/BF02331346.

Arvind Govindarajan, Inbal Israely, Shu-Ying Huang, and Susumu Tonegawa. The dendritic branch is the preferred integrative unit for protein synthesis-dependent LTP. *Neuron*, 69(1):132–146, 2011. doi: 10.1016/j.neuron.2010.12.008.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.

Jordan Guerguiev, Timothy P. Lillicrap, and Blake A. Richards. Towards deep learning with segregated dendrites. *eLife*, 6:e22901, 2017. doi: 10.7554/eLife.22901.

Christopher D. Harvey and Karel Svoboda. Locally dynamic synaptic learning rules in pyramidal neuron dendrites. *Nature*, 450(7173):1195–1200, 2007. doi: 10.1038/nature06416.

Michael Häusser, Nelson Spruston, and Greg J. Stuart. Diversity and dynamics of dendritic signaling. *Science*, 290(5492):739–744, 2000. doi: 10.1126/science.290.5492.739.

Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012. doi: 10.1109/MSP.2012.2205597.

David H. Hubel and Torsten N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, 160(1):106–154, 1962. doi: 10.1113/jphysiol.1962.sp006837.

M. Florencia Iacaruso, Ioana T. Gasler, and Sonja B. Hofer. Synaptic organization of visual space in primary visual cortex. *Nature*, 547:449–452, 2017. doi: 10.1038/nature23019.

Bernd Illing, Wulfram Gerstner, and Johanni Brea. Biologically plausible deep learning—but how far can we go with shallow networks? *Neural Networks*, 118:90–101, 2019. doi: 10.1016/j.neunet.2019.06.001.

Hongbo Jia, Nathalie L. Rochefort, Xiaowei Chen, and Arthur Konnerth. Dendritic organization of sensory input to cortical neurons in vivo. *Nature*, 464(7293):1307–1312, 2010. doi: 10.1038/nature08947.

Judson P. Jones and Larry A. Palmer. An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, 58(6):1233–1258, 1987. doi: 10.1152/jn.1987.58.6.1233.

Jan H. Kirchner and Julijana Gjorgjieva. Emergence of local and global synaptic organization on cortical dendrites. *Nature Communications*, 12:4005, 2021. doi: 10.1038/s41467-021-23557-3.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012.

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. doi: 10.1038/s41583-020-0277-3.

Michael London and Michael Häusser. Dendritic computation. *Annual Review of Neuroscience*, 28 (1):503–532, 2005.

Attila Losonczy, Judit K. Makara, and Jeffrey C. Magee. Compartmentalized dendritic plasticity and input feature storage in neurons. *Nature*, 452(7186):436–441, 2008. doi: 10.1038/nature06725.

Kenichi Ohki, Sooyoung Chung, Yeang H. Ch'ng, Prakash Kara, and R. Clay Reid. Functional imaging with cellular resolution reveals precise micro-architecture in visual cortex. *Nature*, 433 (7026):597–603, 2005. doi: 10.1038/nature03274.

Erkki Oja. Simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, 15(3):267–273, 1982. doi: 10.1007/BF00275687.

Bruno A. Olshausen and David J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996. doi: 10.1038/381607a0.

Panayiota Poirazi, Terrence Brannon, and Bartlett W. Mel. Pyramidal neuron as two-layer neural network. *Neuron*, 37(6):989–999, 2003.

Alon Polsky, Bartlett W. Mel, and Jackie Schiller. Computational subunits in thin dendrites of pyramidal cells. *Nature Neuroscience*, 7(6):621–627, 2004. doi: 10.1038/nn1253.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986. doi: 10.1038/323533a0.

João Sacramento, Rui Ponte Costa, Yoshua Bengio, and Walter Senn. Dendritic cortical microcircuits approximate the backpropagation algorithm. In *Advances in Neural Information Processing Systems*, 2018.

P. Jesper Sjöström and Michael Häusser. A cooperative switch determines the sign of synaptic plasticity in distal dendrites of neocortical pyramidal neurons. *Neuron*, 51(2):227–238, 2006. doi: 10.1016/j.neuron.2006.06.017.

Spencer L. Smith, Ikuko T. Smith, Tiago Branco, and Michael Häusser. Dendritic spikes enhance stimulus selectivity in cortical neurons in vivo. *Nature*, 503(7474):115–120, 2013. doi: 10.1038/nature12600.

Greg J. Stuart and Bert Sakmann. Active propagation of somatic action potentials into neocortical pyramidal cell dendrites. *Nature*, 367(6458):69–72, 1994. doi: 10.1038/367069a0.

Greg J. Stuart and Nelson Spruston. Dendritic integration: 60 years of progress. *Nature Neuroscience*, 18(12):1713–1721, 2015.

Robert Urbanczik and Walter Senn. Learning by the dendritic prediction of somatic spiking. *Neuron*, 81(3):521–528, 2014. doi: 10.1016/j.neuron.2013.11.030.

David E. Wilson, David E. Whitney, Benjamin Scholl, and David Fitzpatrick. Orientation selectivity and the functional clustering of synaptic inputs in primary visual cortex. *Nature Neuroscience*, 19 (8):1003–1009, 2016. doi: 10.1038/nn.4323.

A AI USE STATEMENT

We used ChatGPT and Codex to assist with translation, language editing, and formatting checks, and to obtain feedback on manuscript presentation and analytical approaches. The authors reviewed the suggestions and take full responsibility for the content of the paper.

B SUPPLEMENTARY RESULTS

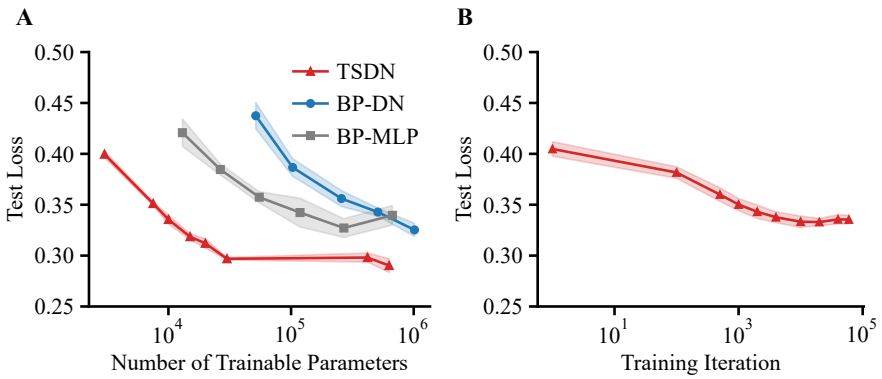

Figure S1: Additional classification-performance analyses on FashionMNIST. (A) Test loss versus the number of gradient-trained parameters for TSDN, BP-DN, and BP-MLP. (B) Test loss of TSDN versus the number of unsupervised feature-learning iterations. Shaded regions denote standard deviations across five random seeds.

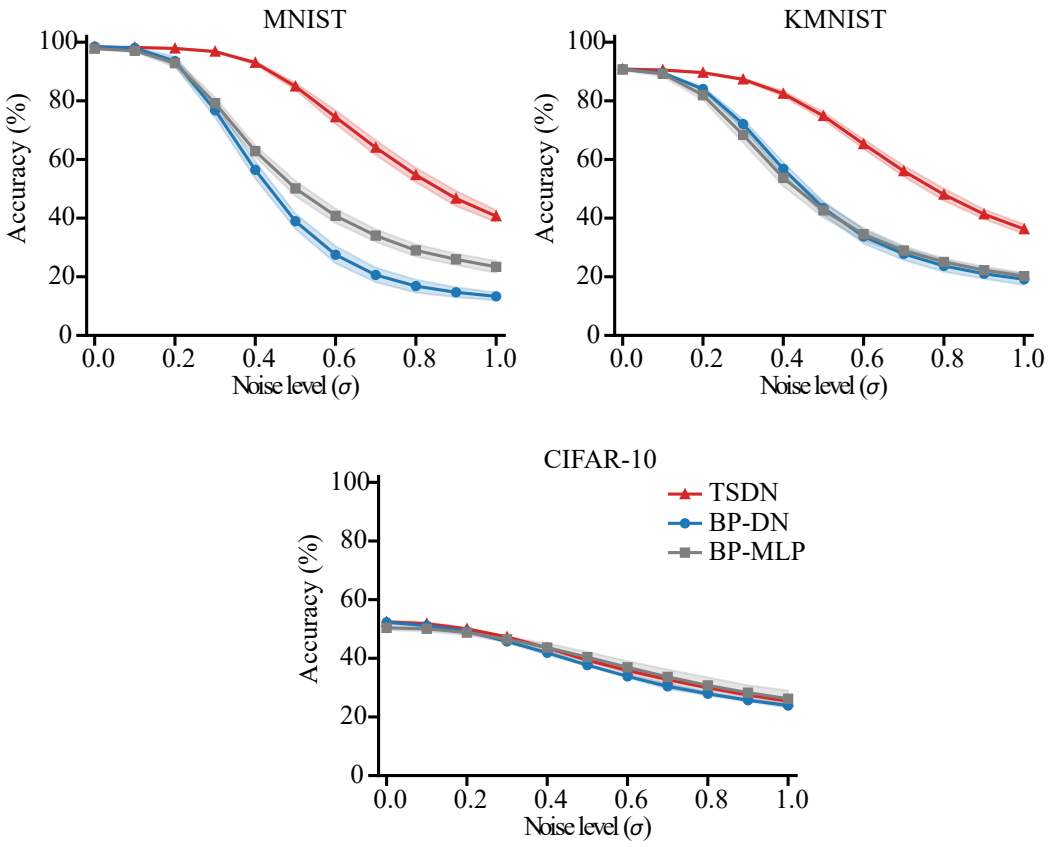

Figure S2: Noise robustness on additional datasets. Classification accuracy versus Gaussian-noise intensity σ on MNIST, KMNIST, and CIFAR-10, comparing TSDN, BP-DN, and BP-MLP. Shaded regions denote standard deviations across five random seeds.