LATERAL INHIBITION INDUCES ADAPTIVE SPARSITY RESULTING IN FORMATION OF INTERPRETABLE REPRESENTATIONS IN AUTOENCODERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Lateral inhibition in neurobiological systems achieves sparse activity through recurrent competitive dynamics. Research on sparse autoencoders (SAEs) shows that imposing sparsity facilitates the formation of interpretable features in neural representations. Conventional SAEs, however, typically enforce sparsity through externally specified regularization objectives or activation constraints. To investigate how recurrent lateral competition shapes the learning of interpretable features, we introduce the Lateral-Inhibition Sparse Autoencoder (LISAE), which uses lateral inhibition to generate sparse activity while learning interpretable features in its decoder. Across controlled MNIST datasets and Gemma-2 representations, we show that LISAE recovers interpretable features while enabling input-dependent sparsity, with the number of active neurons adapting to the input structure. Furthermore, we show that anti-Hebbian-trained lateral inhibitory connections capture the relational structure of the feature space, which reflects the underlying similarities among features. Together, these findings establish local lateral inhibition as an alternative mechanism for learning adaptive sparse and relational representations in autoencoders.

1 INTRODUCTION

Recent mechanistic interpretability research has shown a connection between sparse activation and the emergence of interpretable features in artificial neural networks (Templeton et al., 2024; Bricken et al., 2023; Huben et al., 2024; Makhzani & Frey, 2013; Gao et al., 2025). To achieve this, mainstream Sparse Autoencoders (SAEs) predominantly rely on static algorithmic constraints—most notably Top-K and its variants (Makhzani & Frey, 2013; Bussmann et al., 2024; Lieberum et al., 2024)—to enforce sparsity by strictly limiting activation to a predefined number of neurons. While highly effective in practice, these approaches impose a fixed sparsity level, limiting the number of simultaneously active features regardless of input structure. In contrast, neurobiological systems achieve sparsity not through hardcoded thresholds, but through a dynamic equilibrium governed by lateral inhibition—a ubiquitous computational motif where neurons compete to represent stimuli.

The role of lateral inhibition in sharpening neural tuning curves and enforcing sparse representation is well documented in neuroscience literature (Hartline et al., 1956; Blakemore & Tobin, 1972; Földiák, 1990). However, whether lateral inhibition can induce interpretable features in neural networks trained with backpropagation remains largely unexplored. In this work, we demonstrate that incorporating lateral-inhibition dynamics into an autoencoder provides capabilities beyond conventional sparsity-based feature extraction. We propose an architecture that integrates task-driven global backpropagation with a local inhibitory network trained via an anti-Hebbian rule. By forcing neurons to compete with each other through local recursive inhibition, the autoencoder learns a structured and interpretable latent space.

Here, beyond recovering interpretable features through the decoder weights as in conventional SAEs, we show that learned lateral inhibition provides two qualitatively different capabilities: adaptively adjusting the number of active hidden neurons according to input structure, and encoding geometric relationships among learned features in the inhibitory connections themselves. We demonstrate

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 on customized MNIST datasets and on activations from the Gemma-2-9b language model (LeCun et al., 1998; Team et al., 2024). Our main contributions are:

- We empirically show that lateral inhibition trained with an anti-Hebbian learning rule, together with homeostatic regulation of firing thresholds, effectively induces the formation of interpretable features in autoencoders trained with backpropagation.
- We show that the lateral inhibition layer enables the number of active neurons to adapt to the input structure on a spatially superposed MNIST dataset and a controlled language dataset.
- We report, for the first time, that the lateral inhibition connection itself can be interpretable when trained via anti-Hebbian learning within autoencoders, capturing both the structural similarity between digits and the semantic distance between English concepts.

2 Background and related work

Lateral inhibition. Lateral inhibition is a neural mechanism in which the firing of one neuron suppresses firing in other neurons through intermediate inhibitory connections $\boldsymbol{W}_{\mathrm{LI}} \in \mathbb{R}^{N\times N}_{\leq 0}$, typically among neurons within the same layer that exhibit overlapping tuning curves (Blakemore et al., 1970; Hartline et al., 1956); such competition can adaptively promote winner-take-all dynamics and sparse population responses (Földiák, 1990; Fukai & Tanaka, 1997; Xie et al., 2000; Mao & Massaquoi, 2007). In this paper we utilize a specific formulation described by the following recurrent dynamics (Földiák, 1990):

$$\frac{d\boldsymbol{h}}{dt} = -\boldsymbol{h} + \sigma(\boldsymbol{drive} + \boldsymbol{W}_{\mathrm{LI}}\boldsymbol{h} - \boldsymbol{t}), \quad \text{with } \boldsymbol{h}_0 = \sigma(\boldsymbol{drive}). \tag{1}$$

Intuitively, $\boldsymbol{h}$ represents the firing rates of the target neurons, shaped by the feedforward $\boldsymbol{drive}$ and recurrent lateral inhibition $\boldsymbol{W}_{\mathrm{LI}}$, with $\boldsymbol{t}$ specifying the firing threshold. We treat $\boldsymbol{drive}$ as the constant activation from the preceding layer, such that $\boldsymbol{h}_0$ is initialized by the feedforward drive and subsequently evolves through recurrent inhibition among target neurons. Neurons receiving stronger drive are therefore more likely to dominate the competition. We incorporate these dynamics into the hidden layer of an autoencoder, with details provided in Section 3.1.

Sparse autoencoder. Sparse coding has long been studied as a principle for efficient and structured representation learning in neural networks (Olshausen & Field, 1996; Ranzato et al., 2006; Bell & Sejnowski, 1997; Whittington et al., 2022). More recently, sparse autoencoders (SAEs) have become widely used in mechanistic interpretability, motivated in part by the *superposition hypothesis* (Mikolov et al., 2013; Elhage et al., 2022), which suggests that multiple semantic features are superposed within the latent neural activations of Large Language Models (LLMs). To disentangle these polysemantic representations, researchers increasingly rely on overcomplete SAEs to recover individual, more interpretable features. A common SAE uses an encoder-decoder architecture,

$$\boldsymbol{h} = \sigma(\boldsymbol{W}_e\boldsymbol{x}), \qquad \hat{\boldsymbol{x}} = \boldsymbol{W}_d\boldsymbol{h},$$

where $\boldsymbol{W}_e$ and $\boldsymbol{W}_d$ denote the encoder and decoder weights, respectively. The network is trained to reconstruct the input $\boldsymbol{x}$ in its output $\hat{\boldsymbol{x}}$ while enforcing sparsity on the latent representation $\boldsymbol{h}$ through either an ℓ_1 penalty or a hard-coded Top-K selection (Makhzani & Frey, 2013). Following the initial success of SAEs in extracting interpretable features from LLM activations (Bricken et al., 2023; Lieberum et al., 2024; Gao et al., 2025; Templeton et al., 2024), numerous Top-K-based variants and related sparsity mechanisms have been proposed, including BatchTopK (Leask et al., 2025), AbsTopK (Zhu et al., 2026), Matryoshka SAEs (Zaigrajew et al., 2025), AdaptiveK (Yao et al., 2026), and Gated SAEs (Rajamanoharan et al., 2024).

Motivated by these works, we study an autoencoder in which the hidden representation is shaped by recurrent lateral-inhibition dynamics, allowing competition among hidden neurons to emerge from learned inhibitory interactions rather than being imposed by a fixed sparsification operator. We then examine the resulting dynamics and learned representations.

3 Results

We now present our findings for the Lateral-Inhibition Sparse Autoencoder (LISAE). Section 3.1 outlines the architecture and optimization procedure, empirically verifying that LISAE achieves

the interpretability expected of sparse models, with decoder weights corresponding to interpretable input features. Crucially, we then demonstrate two computational mechanisms of lateral inhibition that differentiate LISAE from standard sparsity mechanisms. In Section 3.2, we show that lateral inhibition facilitates adaptive sparsity. Finally, in Section 3.3, we show that the lateral inhibitory connections themselves provide a structural representation of concept similarity.

3.1 Lateral-inhibition sparse autoencoder extracts interpretable features

Algorithm design. The architecture of the Lateral-Inhibition Sparse Autoencoder is illustrated in Figure 1a. Inside a standard autoencoder, the hidden representation $\boldsymbol{h}$ is shaped by recurrent lateral inhibition dynamics (Eq. 1), which are parameterized by the firing threshold vector $\boldsymbol{t}$ and the inhibitory weight matrix $\boldsymbol{W}_{\mathrm{LI}}$, whose diagonal entries are fixed at zero since self-inhibition is not allowed, and whose off-diagonal entries are nonpositive, i.e., $(\boldsymbol{W}_{\mathrm{LI}})_{ii} = 0, (\boldsymbol{W}_{\mathrm{LI}})_{ij} \leq 0 \quad \forall i \neq j$. Specifically, the encoder maps the input $\boldsymbol{x}$ to a feedforward $\boldsymbol{drive} = \sigma(\boldsymbol{W}_e\boldsymbol{x})$, which is then processed by the lateral inhibition dynamics (Eq. 1) to yield the equilibrium hidden representation $\boldsymbol{h}^*$. Then the decoder reconstructs the input as $\hat{\boldsymbol{x}} = \boldsymbol{W}_d\boldsymbol{h}^*$, and the reconstruction loss is computed as $\mathcal{L}_{\text{rec}} = \|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2^2$.

To jointly optimize the reconstruction objective while leveraging winner-takes-all dynamics, LISAE employs a synergistic three-pathway optimization scheme. Rather than backpropagating the reconstruction loss through the lateral parameters, we update the task parameters $\boldsymbol{W}_e, \boldsymbol{W}_d$ using standard gradient descent, while simultaneously updating the inhibitory weights $\boldsymbol{W}_{\mathrm{LI}}$ with an anti-Hebbian learning rule that encourages decorrelation among neurons (Wu et al., 2022). The thresholds $\boldsymbol{t}$ are updated using a homeostatic rule to maintain an average firing rate p for each neuron, thereby facilitating learning and mitigating inactive neurons. Both the lateral weights and thresholds are updated using local plasticity rules evaluated at the equilibrium state $\boldsymbol{h}^*$. The resulting mechanism promotes sparse representations without requiring backpropagation through the lateral dynamics's parameters. Additionally, we found adding an amplitude-preserving gate will make the training easier to converge without affecting other components of the model. Our local plasticity rules are adapted from the implementation of Foldiak (Földiák, 1990), with modifications to accommodate the backpropagation training. We refer to the original draft for a detailed discussion of the motivation behind these plasticity rules. Throughout the paper, we use ReLU as the activation function σ. The whole training algorithm is summarized as:

Algorithm 1 Lateral-inhibition sparse autoencoder training

Require: Dataset $\mathcal{X}$, target firing rate p, learning rates $\eta_{task}, \eta_W, \eta_t$, gate threshold ϵ

Initialize $\boldsymbol{W}_e, \boldsymbol{W}_d, \boldsymbol{W}_{\mathrm{LI}}, \boldsymbol{t}$
for each batch $\boldsymbol{x} \in \mathcal{X}$ **do**
  $\boldsymbol{drive} \leftarrow \mathrm{ReLU}(\boldsymbol{W}_e\boldsymbol{x})$ ▷ Feedforward drive
  Solve lateral dynamics (Eq. 1) until convergence to obtain equilibrium states $\boldsymbol{h}^*$
  $\boldsymbol{z} \leftarrow \mathbb{I}(\boldsymbol{h}^* > \epsilon) \odot \boldsymbol{drive}$
  $\hat{\boldsymbol{x}} \leftarrow \boldsymbol{W}_d\boldsymbol{z}$ ▷ Reconstruction
  Update task parameters via backpropagation:
  $\boldsymbol{W}_e, \boldsymbol{W}_d \leftarrow \mathrm{Step}(\nabla\mathcal{L}_{\text{rec}}(\boldsymbol{x}, \hat{\boldsymbol{x}}), \eta_{task})$
  Update lateral-inhibition layer's parameters via local rules:
  $\boldsymbol{W}_{\mathrm{LI}} \leftarrow \boldsymbol{W}_{\mathrm{LI}} - \eta_W \frac{1}{B} \sum \left(\boldsymbol{h}^*(\boldsymbol{h}^*)^T - p^2\right)$ ▷ Anti-Hebbian
  $(\boldsymbol{W}_{\mathrm{LI}})_{ii} \leftarrow 0 \quad \text{and} \quad (\boldsymbol{W}_{\mathrm{LI}})_{ij} \leftarrow \min\left\{0, (\boldsymbol{W}_{\mathrm{LI}})_{ij}\right\}, \quad \forall i \neq j$ ▷ Forced non-positivity
  $\boldsymbol{t} \leftarrow \boldsymbol{t} + \eta_t \frac{1}{B} \sum (\boldsymbol{h}^* - p)$ ▷ Firing-threshold homeostasis
end for

Trained LISAE learns interpretable features. We first evaluate the learned features on the MNIST dataset. As shown in Figure 1a, the decoder weights $\boldsymbol{W}_d$ learned by LISAE correspond to interpretable digits, similar to those learned by a standard Top-K SAE (Makhzani & Frey, 2013). Although a standard autoencoder achieves more accurate reconstruction, it utilizes a dense coding strategy and the learned features are less interpretable. Besides, these results also show more inactive neurons in the standard autoencoder and Top-K SAE, motivating our subsequent analysis of homeostatic regulation of firing thresholds. In Appendix A, we ablate the anti-Hebbian learning of the

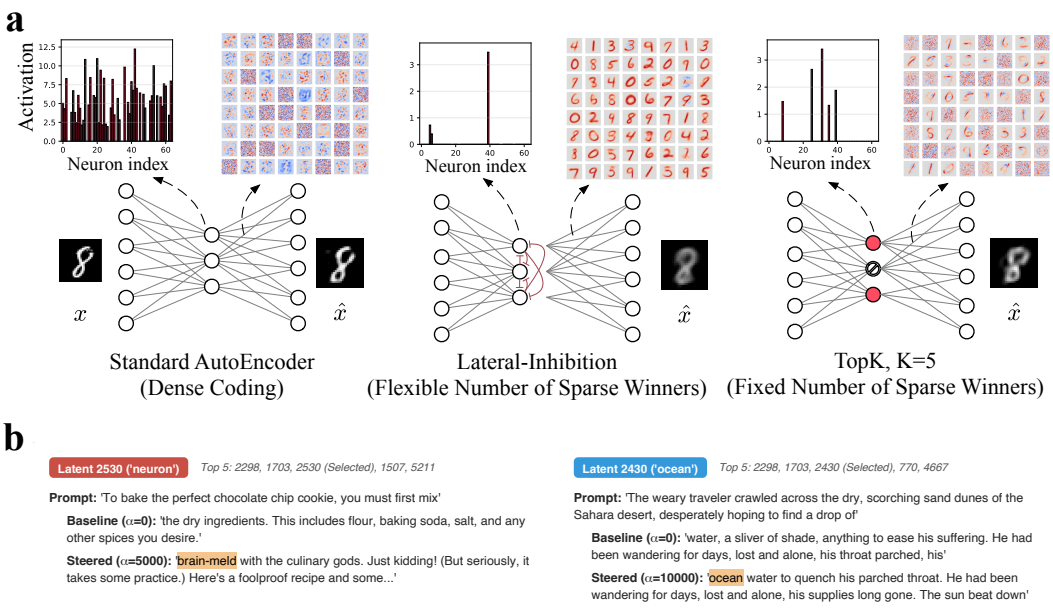


Figure 1: **Lateral-inhibition sparse autoencoder learns interpretable features.** **(a)** The decoder weights $\boldsymbol{W}_d$ learned by LISAE correspond to interpretable digit features as a result of sparse activation (middle, $\mathcal{L}_{\text{rec}}$=0.042), similar to those learned by a standard Top-K sparse autoencoder (right, $\mathcal{L}_{\text{rec}}$=0.028), whereas a standard autoencoder using dense coding achieves better reconstruction but learns less interpretable features (left, $\mathcal{L}_{\text{rec}}$=0.006). A digit 8 is used here as an example to illustrate the reconstruction and activation profile; different inputs will produce different activation patterns. **(b)** LISAE trained with hidden activations from layer 20 of the Gemma-2-9b-it model learns monosemantic features that can be manipulated to steer the model's generation toward specific concepts ($\mathcal{L}_{\text{rec}}$=4.30, average activation 34.3 / 7168).

lateral inhibition and the homeostatic regulation of thresholds by keeping $\boldsymbol{W}_{\text{LI}}$ and $\boldsymbol{t}$, respectively, fixed at their initial values. We find that anti-Hebbian updates to $\boldsymbol{W}_{\text{LI}}$ are necessary for obtaining sparse activation, while homeostatic regulation of $\boldsymbol{t}$ reduces the number of inactive neurons in LISAE. We further find that homeostatic regulation of thresholds has a similar effect in the standard Top-K SAE, offering a stabilization mechanism that could also be adapted to improve traditional Top-K training regimes.

To evaluate whether LISAE can capture manipulable monosemantic features from LLM representations, we then trained a 7168 dimension LISAE on the hidden activations from layer 20 of the Gemma-2-9b-it model (Team et al., 2024) using 500,000 tokens sampled from the Fineweb-edu dataset (Penedo et al., 2024). As illustrated in Figure 1b, intervening on specific learned latents corresponding to concepts such as neuron or ocean predictably steers the model's generation toward the respective concept. This demonstrates that LISAE can extract semantically meaningful features that are directly manipulable through latent interventions. Experimental details are provided in Appendix B.1.

We further test LISAE on two more controlled superposition settings with known underlying features to demonstrate the generality of the algorithm. As shown in Appendix C, LISAE successfully separates the superposed digits and Foldiak bar stimuli into decorrelated features, providing controlled evidence that the proposed local plasticity and competitive dynamics can recover structured features under superposition.

Together with the MNIST results, our experiments empirically demonstrate LISAE's ability to learn interpretable features across both simple visual inputs and high-dimensional language-model representations, enabled by the complementary roles of local plasticity rules to adjust inhibitory synaptic weights and firing thresholds, while using gradient-based optimization for the main task.

3.2 ANTI-HEBBIAN TRAINED LATERAL INHIBITION ENABLES ADAPTIVE SPARSITY

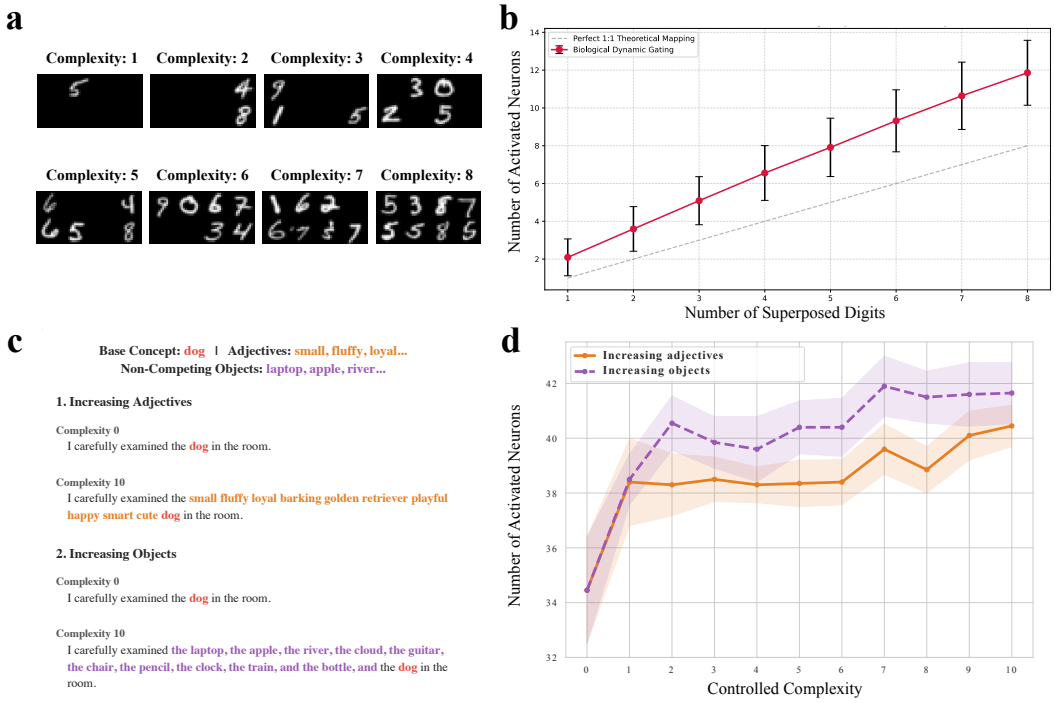


Figure 2: **Lateral inhibition enables input-dependent adaptive sparsity. (a)** Customized spatially superposed MNIST dataset. **(b)** The number of hidden neurons in LISAE activated on a single trial is proportional to the number of superposed digits in the input. Through dynamic competition, the network automatically expands its active pool, scaling from a mean of 2.09 (±0.98 SD) activated neurons at complexity 1 up to 11.86 (±1.72 SD) neurons at complexity 8 out of total 200 hidden neurons. **(c)** Customized text dataset with superposed concepts. **(d)** The number of hidden neurons in the trained LISAE activated on a single trial is proportional to the number of superposed concepts in the input text, demonstrating that lateral inhibition enables adaptive sparsity across language concepts. Lines show mean activation; shading shows standard deviation.

Lateral inhibition is known to implement winner-takes-all dynamics (Xie et al., 2000), where only a subset of winner neurons maintains firing after the network converges when responding to a given stimulus. We first state that every subset of target neurons can occur as the active set of an equilibrium for an appropriate input.

Theorem 1. *Consider the fixed-point equation* $\boldsymbol{h}^* = \mathrm{ReLU}(\boldsymbol{drive} + \boldsymbol{W}_{LI}\boldsymbol{h}^* - \boldsymbol{t})$, *where* $\boldsymbol{h}^* \in \mathbb{R}^N_{\geq 0}$, $\boldsymbol{drive} \in \mathbb{R}^N$, $\boldsymbol{t} \in \mathbb{R}^N$, *and* $\boldsymbol{W}_{LI} \in \mathbb{R}^{N\times N}$ *satisfies* $(\boldsymbol{W}_{LI})_{ii} = 0, \quad (\boldsymbol{W}_{LI})_{ij} \leq 0 \quad$ *for* $i \neq j$. *Then, for every subset* $S \subseteq \{1, \ldots, N\}$, *there exists a nonempty open region of input* $\boldsymbol{drive}$ *for which equation 1 admits an equilibrium* $\boldsymbol{h}^*$ *satisfying* $\boldsymbol{h}^*_i > 0 \quad$ *for* $i \in S, \quad \boldsymbol{h}^*_i = 0 \quad$ *for* $i \notin S$. *Consequently, every possible active-neuron configuration is realizable, including every cardinality* $k = |S| \in \{0, \ldots, N\}$.

Full proof is shown in Appendix D. Thus, for a fixed lateral-inhibition connectivity, different input drives can realize arbitrary subsets of neurons as equilibrium active sets—a qualitatively different capability from a k-sparse autoencoder.

To illustrate the network's ability to adapt to input structure, we first create a customized spatial multiplexing dataset by linearly superposing between one and eight randomly sampled MNIST digits into a single composite image (Fig. 2a). After training a LISAE with 200 hidden neurons, indeed, the number of activated hidden neurons is proportional to the number of superposed digits (Fig. 2b). For instance, an image containing eight multiplexed digits naturally recruits a correspondingly larger

set of active hidden neurons to fully explain the stimulus. This adaptive scaling also indicates that the network has learned a decorrelated feature space, which is shown in Appendix E.

When applied to the intermediate activations of the Gemma-2-9b-it language model, our lateral inhibition architecture exhibits the same dynamic capacity scaling observed in the MNIST task, where the number of activated hidden neurons increases systematically with controlled semantic structure of the input text (Fig. 2c, d). We modulated the conceptual structure of the inputs by adjusting the number of discrete objects and modifying adjectives within the prompts. As the sentences become more conceptually dense by incorporating additional objects and descriptive adjectives, the latent layer dynamically expands the number of its active neurons to capture the increasing semantic component. This again confirms that lateral inhibition could adapt the active latent number to match the input structure. Experimental details are provided in Appendix B.2.

3.3 The lateral inhibitory connections encode feature similarity

Beyond the conventional observation that SAE decoder weights can be interpretable, we report another interesting observation: the lateral inhibition matrix $\boldsymbol{W}_{\mathrm{LI}}$ itself emerges as an interpretable structure that encodes geometric relationships among learned features (Figure 3). We analyze $\boldsymbol{W}_{\mathrm{LI}}$ from the same experiments as Section 3.1, where the lateral inhibition connections are learned with an anti-Hebbian rule and the model has learnt corresponding features (Figure 1).

In our experiments on the plain MNIST dataset, first, we define how each latent neuron is assigned to a specific digit class. Let $\boldsymbol{h}^*(\boldsymbol{x})$ be the latent activation vector for a given input image $\boldsymbol{x}$, and $y \in \{0, \ldots, 9\}$ be the corresponding ground-truth label. The preferred class c_u for a given neuron u is determined by finding the class that yields the maximum expected activation over the dataset:

$$c_u = \arg \max_{k \in \{0,\ldots,9\}} \mathbb{E}_{\boldsymbol{x}|y=k}[\boldsymbol{h}^*_u(\boldsymbol{x})].$$

Once the neurons are partitioned into class-specific subpopulations $\mathcal{N}_i = \{u \mid c_u = i\}$, the average inhibitory connection strength between class i and class j is computed as the absolute mean of the lateral weights $\boldsymbol{W}_{\mathrm{LI}}$ connecting the two subpopulations:

$$I_{ij} = \left| \frac{1}{|\mathcal{N}_i||\mathcal{N}_j|} \sum_{u \in \mathcal{N}_i} \sum_{v \in \mathcal{N}_j} (\boldsymbol{W}_{\mathrm{LI}})_{uv} \right|.$$

After normalized to $[0, 1]$, plotting this learned inter-digit inhibition against independently measured pixel-level similarity reveals a strong positive correlation (Figure 3a). This result suggests that the network learns stronger inhibitory interactions between neurons representing digits that look similar, potentially preventing highly overlapping features from collapsing into a shared representation.

A similar pattern emerges in the Gemma-2-9b-it model. We extract the lateral inhibition connection $\boldsymbol{W}_{\mathrm{LI}}$ and identify the top 7 most actively firing latent neurons for each concept from four distinct semantic categories, including animals, neuroscience, finance, and geography. We then perform unsupervised clustering using only the pairwise inhibitory weights among these neurons, without access to their feedforward activations or decoder weights. Interestingly, the resulting clusters align with the corresponding semantic categories, indicating that the lateral connections capture meaningful structure in the relationships among learned language concepts (Figure 3b). Thus, even in a high-dimensional language representation, the inhibitory connectivity alone provides an interpretable relational structure. Experimental details are provided in Appendix B.3.

This emergent relational interpretability is a direct consequence of the synergy between the global and local learning rules. As the task-driven backpropagation shapes the decoder weights into discrete, interpretable concepts, the local anti-Hebbian learning rule simultaneously acts on the latent space to decorrelate these features. Because the anti-Hebbian update penalizes co-activation, it selectively strengthens the inhibitory weights between neurons that frequently receive overlapping feedforward drive. Consequently, $\boldsymbol{W}_{\mathrm{LI}}$ instantiates a map of feature co-occurrence and semantic similarity, where the magnitude of inhibition reflects the conceptual proximity of the encoded features. Lastly, beyond representation learning, our findings also offer a testable hypothesis for systems neuroscience. While excitatory neurons are traditionally viewed as the primary drivers of semantic tuning, our model predicts that the synaptic weights of local inhibitory microcircuits physically encode a conceptual distance metric.

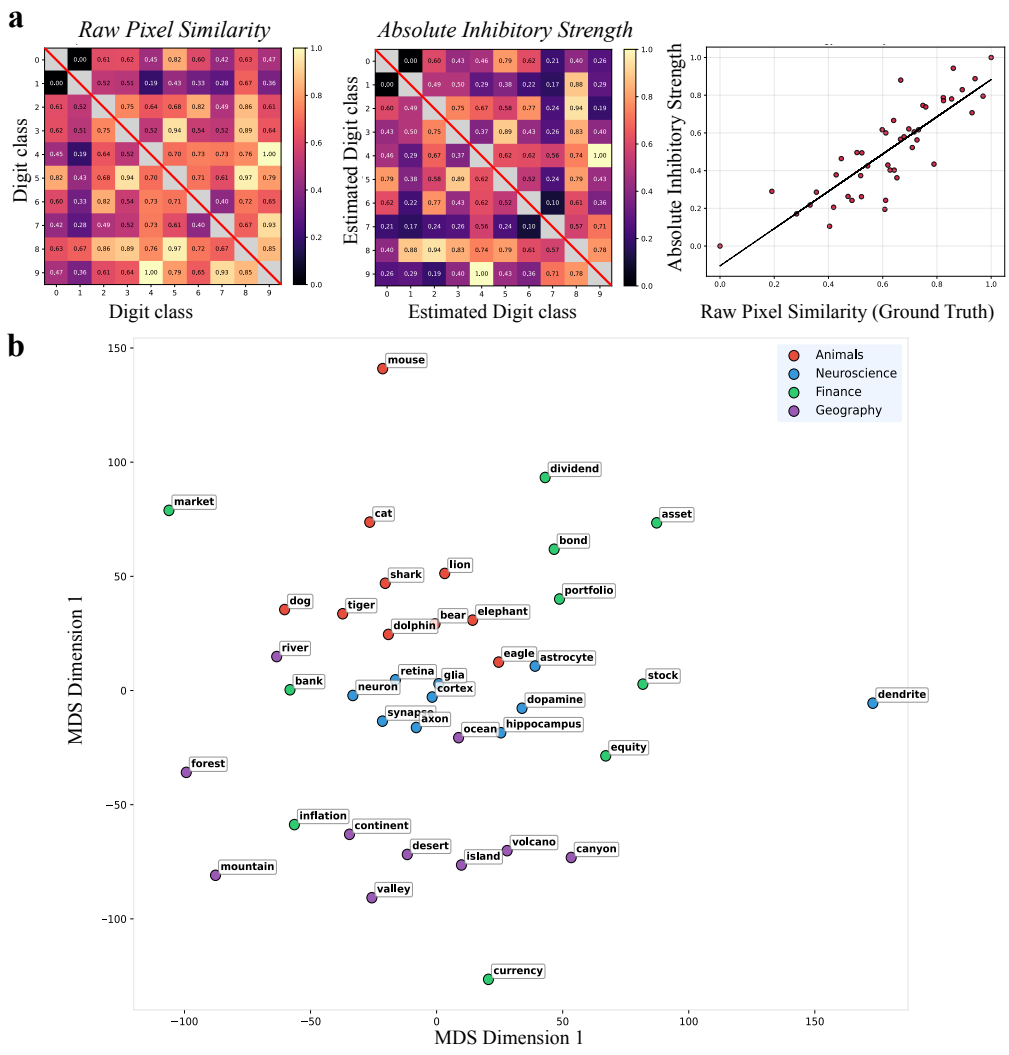


Figure 3: **The lateral inhibition connections intrinsically encode feature similarity.** **(a)** On MNIST, the normalized learned inhibitory weights between digit-selective neurons (middle) strongly correlate with the baseline pixel-level similarity (left) of the respective digit classes (right, $R = 0.78$). **(b)** On Gemma-2-9b-it, clustering the top-7 most highly activated latent neurons with Multi-Dimensional Scaling algorithm purely by their pairwise inhibitory weights recovers coherent semantic groups, separating concepts into distinct domains (e.g., animals, neuroscience, finance, and geography).

4 Discussion

Here, we introduced the Lateral-Inhibition Sparse Autoencoder (LISAE) that integrates lateral inhibition dynamics into a standard autoencoder architecture. We demonstrate that local plasticity rules can be synergistically integrated with backpropagation to unlock complementary representational capabilities. A primary characteristic of this approach is its departure from rigid algorithmic sparsity constraints. While recent advancements such as AdaptiveK (Yao et al., 2026) successfully achieve dynamic sparsity by training an explicit auxiliary probe to estimate input complexity, we demonstrate that this adaptive capacity does not strictly require external estimators. Instead, input-driven latent allocation inherently emerges from the continuous-time, competitive dynamics of lateral inhibition (Section 3.2). However, we also acknowledge that determining the upper bound of input complexity that the network can accommodate before the interpretability of the learned

sparse features begins to degrade remains an important open question. Moreover, the emergence of concept clusters from inhibitory connections provide additional opportunities for inspecting the neural network representations, as it shows that the connections of a neural network can have distinct functions beyond feedforward information passing.

We acknowledge that our current experiments represent an initial proof of concept rather than a fully scaled deployment. Our language model experiments were conducted on a single consumer-grade GPU, limiting the scale of our evaluation. A fundamental limitation of LISAE is the $\mathcal{O}(N^2)$ parameter scaling of the lateral inhibition matrix relative to the SAE width, which currently poses challenges for scaling to the model widths required for large-scale LLM feature extraction. Future work will explore scaling strategies, such as sparse or low-rank approximations of the lateral matrix, with more GPUs to enable comprehensive quantitative evaluations on standardized frameworks like SAEBench (Karvonen et al., 2025). Lastly, we hope our results will inspire the development of neural architectures that incorporate computational motifs from neuroscience and harness their synergy with backpropagation.

5 Conclusion

In this paper, we introduced the Lateral-Inhibition Sparse Autoencoder, integrating biologically motivated lateral inhibition with gradient-based representation learning. We showed the model can recover interpretable features while providing input-dependent sparsity, allowing the number of active neurons to adapt to input structure rather than being constrained by external regularization objectives. Beyond the representations themselves, we found that the learned lateral inhibitory connections also capture interpretable relationships among features, from pixel-level similarity between digits to semantic distance in language representations. Together, these results suggest that lateral inhibition can provide complementary computational and interpretability benefits beyond conventional sparse autoencoders, while offering a testable prediction about the functional role of lateral inhibitory connectivity in neurobiological systems.

AI use statement

In this work, we used generative AI tools to formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, implement methods, and clean and reformat datasets. We have not used generative AI tools to generate synthetic datasets, help develop theoretical models or conceptual frameworks, propose or refine hypotheses, design or provide feedback on research methodology or experiments, or interpret results. Support for qualitative and thematic data analysis and assistance with translation are not applicable to this work. Additionally, we used generative AI tools to create or modify scientific figures or images and to create or edit software code. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

Ethics statement

The research does not involve human subjects or personally identifiable information, and the datasets used are publicly available and widely used in the research community. No specific ethical concerns are raised by this work.

Reproducibility statement

All experimental details are provided in Appendix B. Our code for experiments is provided in the supplementary material for reproducibility and will be made publicly available on GitHub upon de-anonymization.

References

Anthony J. Bell and Terrence J. Sejnowski. The "independent components" of natural scenes are edge filters. *Vision Research*, 37(23):3327–3338, 1997.

Colin Blakemore and Elisabeth A Tobin. Lateral inhibition between orientation detectors in the cat's visual cortex. *Experimental brain research*, 15(4):439–440, 1972.

Colin Blakemore, Roger HS Carpenter, and Mark A Georgeson. Lateral inhibition between orientation detectors in the human visual system. *Nature*, 228(5266):37–39, 1970.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL `https://transformer-circuits.pub/2023/monosemantic-features/index.html`.

Bart Bussmann, Patrick Leask, and Neel Nanda. BatchTopK Sparse Autoencoders. *arXiv preprint arXiv:2412.06410*, 2024.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

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

Tomoki Fukai and Shigeru Tanaka. A simple neural network exhibiting selective activation of neuronal ensembles: from winner-take-all to winners-share-all. *Neural computation*, 9(1):77–97, 1997.

Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *International Conference on Learning Representations*, volume 2025, pp. 26721–26754, 2025.

H K Hartline, Henry G Wagner, and Floyd Ratliff. Inhibition in the eye of Limulus. *The Journal of general physiology*, 39(5):651–673, 1956.

Robert Huben, Hoagy Cunningham, Logan Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, volume 2024, pp. 7827–7845, 2024.

Adam Karvonen, Can Rager, Johnny Lin, Curt Tigges, Joseph Isaac Bloom, David Chanin, Yeu-Tong Lau, Eoin Farrell, Callum Stuart Mcdougall, Kola Ayonrinde, Demian Till, Matthew Wearden, Arthur Conmy, Samuel Marks, and Neel Nanda. SAEBench: A comprehensive benchmark for sparse autoencoders in language model interpretability. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 29223–29264, 2025.

Patrick Leask, Bart Bussmann, Michael Pearce, Joseph Bloom, Curt Tigges, Noura Al Moubayed, Lee Sharkey, and Neel Nanda. Sparse autoencoders do not find canonical units of analysis. In *International Conference on Learning Representations*, volume 2025, pp. 53617–53642, 2025.

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.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on Gemma 2. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 278–300, 2024.

Alireza Makhzani and Brendan Frey. K-sparse autoencoders. *arXiv preprint arXiv:1312.5663*, 2013.

Zhi-Hong Mao and Steve G. Massaquoi. Dynamics of winner-take-all competition in recurrent neural networks with lateral inhibition. *IEEE transactions on neural networks*, 18(1):55–69, 2007.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 746–751, 2013.

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.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The FineWeb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37:30811–30849, 2024.

Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Tom Lieberum, Vikrant Varma, János Kramár, Rohin Shah, and Neel Nanda. Improving dictionary learning with gated sparse autoencoders. *arXiv preprint arXiv:2404.16014*, 2024.

Marc'Aurelio Ranzato, Christopher Poultney, Sumit Chopra, and Yann LeCun. Efficient learning of sparse representations with an energy-based model. *Advances in Neural Information Processing Systems*, 19, 2006.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, Alex Tamkin, Esin Durmus, Tristan Hume, Francesco Mosconi, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*, 2024. URL `https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html`.

James CR Whittington, Will Dorrell, Surya Ganguli, and Timothy EJ Behrens. Disentanglement with biological constraints: A theory of functional cell types. *arXiv preprint arXiv:2210.01768*, 2022.

Yue Kris Wu, Christoph Miehl, and Julijana Gjorgjieva. Regulation of circuit organization and function through inhibitory synaptic plasticity. *Trends in Neurosciences*, 45(12):884–898, 2022.

Xiaohui Xie, Richard Hahnloser, and H. Sebastian Seung. Learning winner-take-all competition between groups of neurons in lateral inhibitory networks. *Advances in Neural Information Processing Systems*, 13, 2000.

Yifei Yao, Hanrong Zhang, and Mengnan Du. AdaptiveK: Complexity-driven sparse autoencoders for interpretable language model representations. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 23702–23728, 2026.

Vladimir Zaigrajew, Hubert Baniecki, and Przemyslaw Biecek. Interpreting CLIP with hierarchical sparse autoencoders. *arXiv preprint arXiv:2502.20578*, 2025.

Xudong Zhu, Mohammad Mahdi Khalili, and Zhihui Zhu. AbsTopK: Rethinking sparse autoencoders for bidirectional features. In *International Conference on Learning Representations*, volume 2026, pp. 50325–50352, 2026.

A EFFECTS OF ABLATING ANTI-HEBBIAN LEARNING OF LATERAL INHIBITION AND HOMEOSTATIC REGULATION OF FIRING THRESHOLD

To evaluate the individual contributions of our local plasticity mechanisms, we conducted an ablation study on the MNIST dataset, examining the impact of removing either the anti-Hebbian learning rule that controls adjustment of inhibitory weights or the homeostatic regulation of firing threhold separately.

A.1 REMOVING ANTI-HEBBIAN LEARNING OF LATERAL INHIBITION ALONE LEADS TO DENSE CODING

When the lateral inhibition matrix $\boldsymbol{W}_{\text{LI}}$ is frozen at its initial state and exempt from anti-Hebbian updates, the network loses its ability to dynamically decorrelate features. As illustrated in Figure 4, the absence of local anti-Hebbian plasticity causes the latent space to degrade into a dense, unspecialized representation just like in standard SAEs (Figure 1a left panel), confirming that task-driven backpropagation alone is insufficient to organize structured, orthogonal feature maps; rather, learnt lateral inhibition is necessary for interpretable representations to form in LISAE.

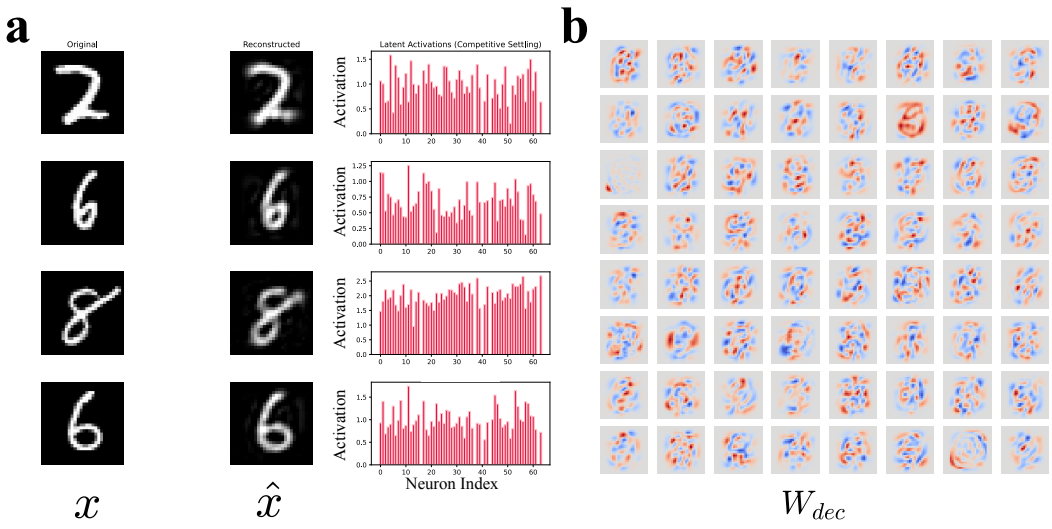


Figure 4: **Removing anti-Hebbian updates of lateral inhibition leads to dense coding. (a)** Reconstructions and activation for the model. **(b)** Decoder weights of LISAE trained without anti-Hebbian updates of lateral inhibition weights, resulting in less interpretable features.

A.2 REMOVING HOMEOSTATIC REGULATION OF THRESHOLD ALONE INCREASES INACTIVE NEURONS

When the network is trained without the homeostatic regulation of firing threshold alone ($\eta_t = 0$), the latents suffer from severe capacity waste. As shown in Figure 5 through the neuron utilization metrics, a significant portion of the neurons become completely inactive in response to all stimuli because their thresholds fail to dynamically lower to match the activation distribution. Comparing to Figure 1, both decoders learn interpretable features, while the model lacking homeostatic regulation of threshold exhibits more inactive neurons. Activation frequency denotes the percentage of dataset samples that elicit a non-zero response from a given latent neuron.

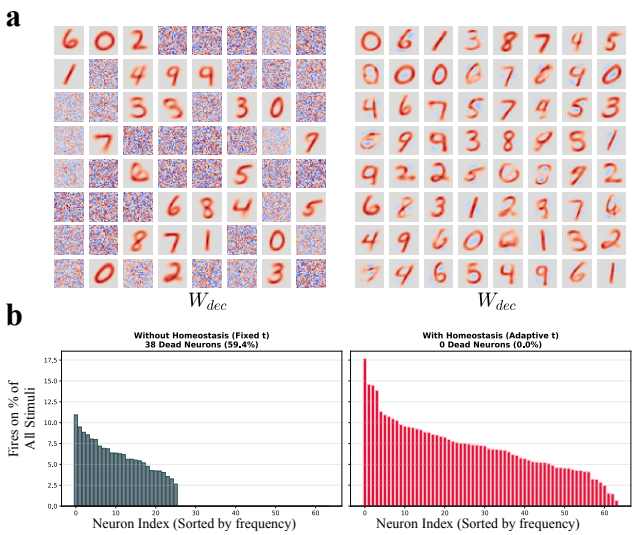


Figure 5: **Removing homeostatic regulation of firing threshold increases inactive neurons in LISAE. (a)** Effect of homeostatic regulation of firing threshold on decoder weights. Disabling this mechanism (left) results in significantly more inactive neurons compared to the regulated baseline (right). **(b)** Neuron utilization profiles demonstrating the necessity of homeostatic threshold regulation. Ablating this homeostatic regulation (left) causes more inactive neurons, leaving 59.4% of neurons inactive across all stimuli, whereas homeostatic regulation (right) prevents inactive neurons and ensures full capacity utilization.

A.3 Homeostatic regulation of threshold decreases inactive neurons also in traditional top-k sparse autoencoders

We integrated our homeostatic regulation of firing threshold into a traditional Top-K Sparse Autoencoder. In Figure 6, adding homeostatic regulation to Top-K training successfully decreases inactive neurons compared to the model in Figure 1a. Figure 6b shows that adding homeostatic regulation of threshold is sufficient to prevent inactive neurons, whereas Figure 5 shows that homeostatic regulation of threshold is necessary to prevent inactive neurons in LISAE; collectively, these results show that homeostatic regulation of threshold encourages utilization of all neurons.

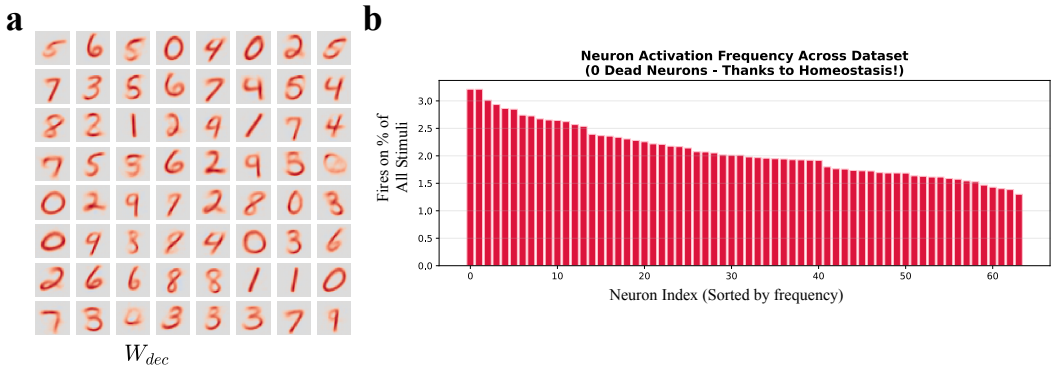


Figure 6: **Homeostatic regulation of firing threshold decreases inactive neuron formation in traditional top-k sparse autoencoders. (a)** Visualization of decoder weights of a hybrid Top-K SAE augmented with homeostatic regulation of firing threshold. **(b)** Top-K SAE trained with homeostatic regulation of firing threshold avoids inactive neurons (compare to right panel of Fig 1a). Neuron utilization improves significantly across the latent capacity.

B EXPERIMENT DETAILS

All experiments were carried out on a single NVIDIA GeForce RTX 5090 GPU using PyTorch.

B.1 EXPERIMENTAL DETAILS FOR SECTION 3.1

MNIST experiment. We utilized the standard MNIST dataset, normalizing pixel values to the range $[0, 1]$. The Lateral-Inhibition Sparse Autoencoder (LISAE) was trained using Algorithm 1 with the hyperparameters detailed in Table 1. To visualize the learned representations, we extracted the decoder weights ($\boldsymbol{W}_d$) connected to each individual hidden neuron and reshaped them into 28×28 spatial feature maps.

Gemma-2 experiment. We used the Gemma-2-9b-it model from Hugging Face (Team et al., 2024) and processed 500,000 tokens sampled from the Fineweb-edu dataset (Penedo et al., 2024). We extracted the hidden multilayer perceptron activations specifically from Layer 20 to serve as the training dataset for our LISAE. The model was trained using Algorithm 1 with hyperparameter specified in Table 2. Furthermore, we conducted concept steering experiments to validate feature extraction (Bricken et al., 2023). To perform concept steering, we first identified the latent dimension k that maximally responded to a target concept (e.g., "ocean", "neuron") using a targeted discovery prompt. We then extracted its corresponding decoder weight vector and L_2-normalized it to obtain the structural steering direction $\mathbf{v}_{steer}$:

$$\mathbf{v}_{steer} = \frac{\boldsymbol{W}_d[:, k]}{\|\boldsymbol{W}_d[:, k]\|_2}.$$

During the forward pass of the language model on a structurally unrelated test context (e.g., injecting the "neuron" feature into a driving prompt), we intervened at the target layer by aggressively modifying the model's intermediate representations $\mathbf{a}$:

$$\mathbf{a}' = \mathbf{a} + \alpha \mathbf{v}_{steer},$$

where $\alpha \in [0, 20000]$ modulates the intervention strength, forcing the network to dynamically synthesize the injected concept into the generated output.

Table 1: Hyperparameters for digit experiment

Hyperparameter	Value
Input Dimension	784
Target Average Intensity (p)	0.05
Task Learning Rate (η_{task})	0.001 (Adam)
Lateral Weight Learning Rate (η_W)	1
Threshold Learning Rate (η_t)	0.1
ϵ	1e-4
Batch Size	512
Latent Dimension (D_{SAE})	64
Epochs	100

Table 2: Hyperparameters for large language model experiments

Hyperparameter	Value
Model Layer Extracted	20
Input Dimension (D_{model})	3584
Expansion Factor	4
Latent Dimension (D_{SAE})	7168
Target Average Intensity (p)	0.01
Task Learning Rate (η_{task})	0.001 (Adam)
Lateral Weight Learning Rate (η_W)	100
Threshold Learning Rate (η_t)	1
Batch Size	512
ϵ	1e-6
Epochs	20

B.2 Experimental details for section 3.2

All training hyperparameters remain identical to those described in Appendix B.1, except that the latent dimension is set to 200 for the spatial superposed MNIST experiment.

Customized spatial superposition MNIST. To test adaptive capacity, we generated a custom dataset by linearly superposing between 1 and 8 randomly selected MNIST digits into a single composite image. The network's capacity to dynamically scale its active latent neurons was measured against the known number of spatially superposed inputs.

Concept superposition. To evaluate adaptive sparsity in language representations, we tested the network against multi-concept prompts with increasing semantic numbers, as shown in Figure 2c. Table 3 details the 20 domains and axes used.

Table 3: Concept superposition dataset: 20 base domains tested for adaptive sparsity.

Target	Increasing adjectives	Increasing objects
dog	small, fluffy, loyal, barking...	laptop, apple, river, cloud...
market	busy, global, financial, volatile...	bird, bicycle, mountain, coffee...
car	new, fast, electric, red...	banana, book, ocean, telescope...
apple	fresh, crisp, red, sweet...	clock, train, neuron, sword...
neuron	small, biological, spiking, cortical...	car, pizza, rainbow, guitar...
ocean	cold, deep, blue, vast...	keyboard, tiger, bread, statue...
mountain	steep, tall, snowy, rocky...	laptop, cat, spaghetti, watch...
king	old, wise, powerful, wealthy...	satellite, mango, river, train...
coffee	hot, dark, bitter, strong...	lion, planet, shoe, bridge...
sword	heavy, sharp, steel, glowing...	apple, ocean, truck, camera...
house	big, cozy, wooden, suburban...	bird, pizza, lake, helmet...
computer	fast, modern, digital, quantum...	tree, banana, mountain, dog...
book	thick, old, dusty, hardcover...	car, cloud, tiger, coffee...
tree	tall, green, leafy, ancient...	laptop, hamburger, sword, phone...
planet	large, rocky, distant, alien...	guitar, apple, shoe, house...
guitar	loud, acoustic, wooden, vintage...	tree, ocean, burger, train...
river	long, winding, muddy, rapid...	laptop, cat, pizza, sword...
city	dense, loud, crowded, modern...	apple, bird, guitar, lake...
clock	old, ticking, golden, antique...	tree, ocean, dog, car...
bird	small, blue, flying, singing...	laptop, mountain, pizza, sword...

B.3 Experimental details for Section 3.3

All training hyperparameters are identical to those in Appendix B.1.

Extracting feature similarity from lateral inhibition matrix (MNIST). To establish the baseline similarity between two digit classes i and j, we computed the cosine similarity of their average 28×28 image templates ($\bar{\boldsymbol{x}}$):

$$S_{ij} = \frac{\bar{\boldsymbol{x}}^{(i)} \cdot \bar{\boldsymbol{x}}^{(j)}}{\|\bar{\boldsymbol{x}}^{(i)}\|_2 \|\bar{\boldsymbol{x}}^{(j)}\|_2}.$$

We then calculated the learned inhibitory connection strength between the discrete subpopulations of active neurons ($\mathcal{N}_i, \mathcal{N}_j$) representing those specific digits:

$$I_{ij} = \left| \frac{1}{|\mathcal{N}_i||\mathcal{N}_j|} \sum_{u \in \mathcal{N}_i} \sum_{v \in \mathcal{N}_j} (\boldsymbol{W}_{\mathrm{LI}})_{uv} \right|.$$

To quantify this relationship, we performed a statistical regression, computing the Pearson correlation coefficient (R) between the flattened upper-triangular elements of the ground-truth pixel similarity matrix S and the normalized learned biological confusion matrix I.

Relational geometry in language (Gemma-2). To map the conceptual distance of the LLM representations, we defined 40 distinct concepts distributed across four domains: Animals, Neuroscience, Finance, and Geography. For each concept, we used the prompt, "The text discusses the properties of a {selected concept} in detail," to extract the corresponding activations and identify the seven most highly activated latent dimensions. By extracting the structural lateral inhibition strictly between these resulting composite vectors, we constructed a pairwise distance matrix. We then applied Multi-Dimensional Scaling to project this inhibitory geometry into a 2D coordinate space, visualizing the intrinsic semantic clusters learned by the lateral connections. Table 4 lists the specific domains and concepts used to map this geometry.

Table 4: Relational geometry dataset: 40 distinct concepts categorized across four semantic domains used to construct the pairwise inhibitory distance matrix.

Domain	Concepts (10 per domain)
Animals	dog, cat, lion, tiger, elephant, mouse, dolphin, shark, eagle, bear
Neuroscience	neuron, synapse, axon, dendrite, cortex, hippocampus, astrocyte, glia, retina, dopamine
Finance	market, stock, bond, equity, inflation, currency, bank, portfolio, asset, dividend
Geography	ocean, mountain, river, valley, desert, forest, continent, island, canyon, volcano

C TWO ADDITIONAL SUPERPOSED CASES

In this section, we provide complementary results evaluating the LISAE on two multiplexed datasets to provide controlled evidence that the proposed local plasticity and competitive dynamics can recover structured features under superposition.

C.1 OVERLAPPING DIGITS

To evaluate the network's capacity to untangle spatially overlapping features, we constructed a superposed MNIST dataset. Each image was generated by linearly superposing two randomly selected digits on a single canvas, with pixel intensities limited to a maximum of 1.0. We trained the LISAE using a latent dimension of 60 and a target average firing rate of $p = 0.01$. The model was trained for 30 epochs with a batch size of 256. For the optimization, the task parameters were updated using Adam with a learning rate of $\eta_{task} = 0.002$, while the biological lateral weights and homeostatic thresholds were updated via SGD with learning rates of $\eta_W = 20$ and $\eta_t = 1$, respectively.

As illustrated by the activation profiles in Figure 7, the network learned distinct digit representations in its decoder despite being trained exclusively on superposed digit pairs. This demonstrates that the

LISAE can recover underlying features from spatially overlapping inputs. Moreover, the continuous-time lateral inhibition dynamics allow the network to dynamically recruit distinct subsets of neurons, whose decoder weights correspond to the underlying digit representations.

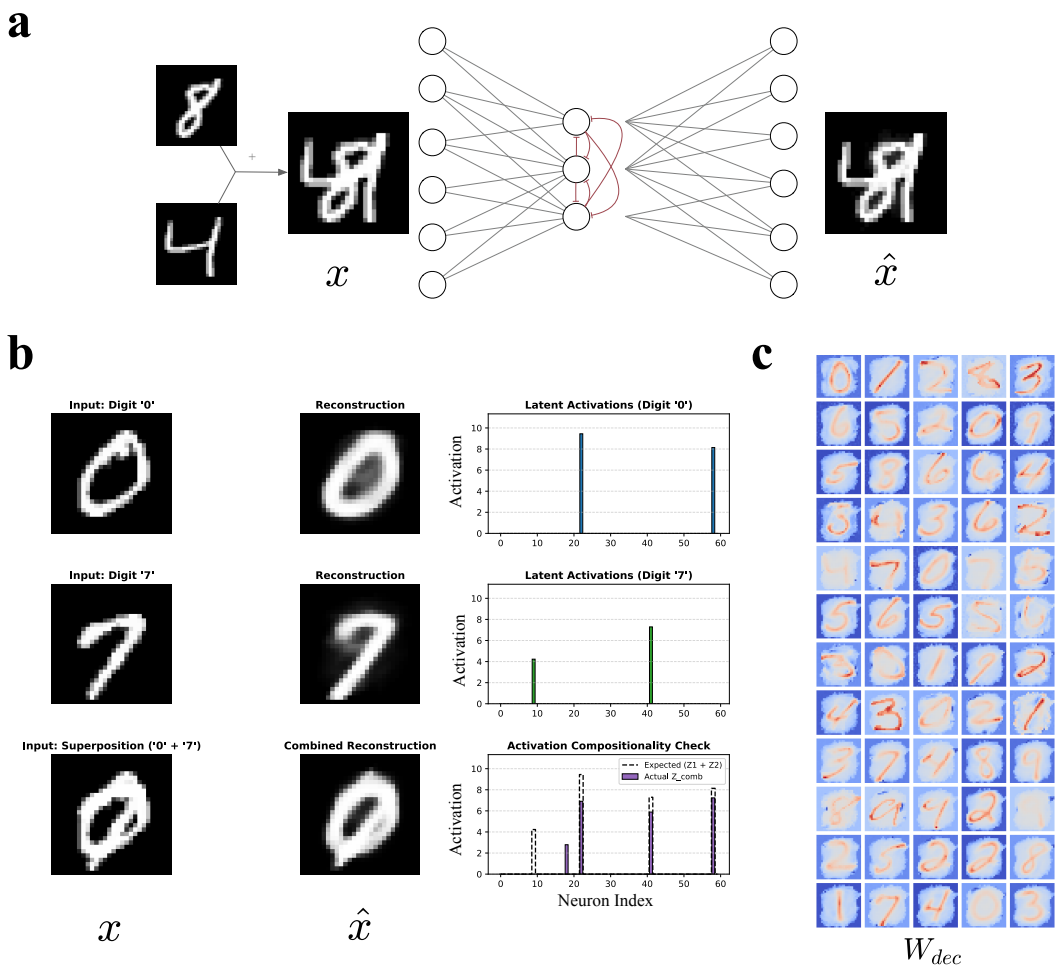


Figure 7: **Reconstructions and learnt representations for the superposed MNIST dataset. (a)** Architecture schematic illustrating the superposition of multiple digits (e.g., 8 and 4) into a single input x and the corresponding reconstruction $\hat{x}$. **(b)** Separate and superposed digit inputs (left), corresponding reconstructions (middle), and latent activation profiles (right). The top and middle rows show the responses to individual digits ('0' and '7'), while the bottom row shows their superposition ('0' + '7'). The final panel compares the observed activation for the superposed input with the expected linear sum of the individual activations, showing that the network activates the corresponding individual digit representations for the superposed input. **(c)** The learned decoder weights (W_{dec}), showing that the network disentangles the mixtures into clean, individual digit features.

C.2 Split-screen Foldiak bars

We further tested the architecture on a spatial variant of the classic Foldiak bars task Földiák (1990). The inputs consist of a 16×16 canvas logically divided into two independent 8×8 regions (top-left and bottom-right). For each sample, between 2 and 8 horizontal or vertical bars were placed non-linearly (where intersecting pixel values are overwritten to 1.0 rather than summed) into one of the randomly selected regions.

The model was configured with an expanded latent dimension of 100 and a target activation of $p = 0.5$ to account for the higher combinatorial density of the intersecting bars. Training was conducted over 100 epochs with a batch size of 256. The task parameters were optimized via Adam ($\eta_{task} = 0.005$), while the lateral weights and thresholds were updated using SGD ($\eta_W = 0.5$, $\eta_t = 0.1$). To ensure convergence of the competing discrete groups, the lateral settling process was extended to 50 iterations per pass ($\alpha = 0.4$).

As shown in Figure 8, despite the spatial superposition and the nonlinearity, the local anti-Hebbian-trained lateral inhibition successfully forces the network to parse the independent bar features and structurally organize the decoder weights into distinct functional groups.

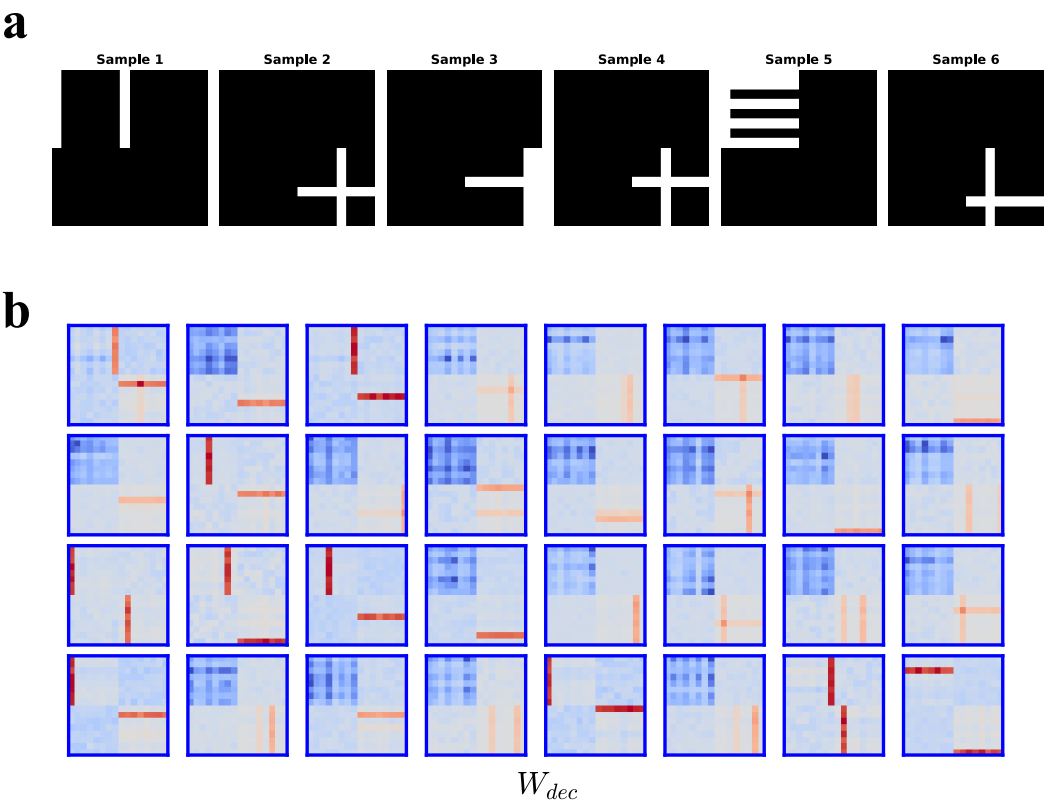


Figure 8: **Evaluation on the split-screen non-linear foldiak bars task. (a)** Six dataset samples demonstrating severe feature entanglement and non-linear occlusion across two spatial regions. **(b)** Part of the learned decoder weights (W_{dec}), showing that the network demixes the complex inputs into horizontal and vertical bar features localized to their specific quadrants.

D PROOF OF THEOREM 1

Proof. With the global stability result established in (Xie et al., 2000), we characterize the active sets of such equilibria.

Let $S \subseteq \{1, \ldots, N\}$ be an arbitrary subset, and let S^c denote its complement. We construct an input for which the active set is exactly S.

First consider the case $1 \leq |S| \leq N - 1$. Choose any vector

$$\boldsymbol{h}_S^* \gg 0.$$

Define the feedforward input to the active neurons by

$$\boldsymbol{drive}_S = \boldsymbol{t}_S + (I - \boldsymbol{W}_{LI}^{SS})\boldsymbol{h}_S^*.$$

Then

$$\boldsymbol{drive}_S + \boldsymbol{W}_{LI}^{SS}\boldsymbol{h}_S^* - \boldsymbol{t}_S = \boldsymbol{h}_S^* \gg 0,$$

so the ReLU is in its active regime and the neurons in S satisfy the fixed-point equation exactly.

For the inactive neurons, choose the feedforward input such that

$$\boldsymbol{drive}_{S^c} < \boldsymbol{t}_{S^c} - \boldsymbol{W}_{LI}^{S^cS}\boldsymbol{h}_S^*.$$

Since $\boldsymbol{h}_{S^c}^* = 0$, the corresponding ReLU arguments satisfy

$$\boldsymbol{drive}_{S^c} + \boldsymbol{W}_{LI}^{S^cS}\boldsymbol{h}_S^* - \boldsymbol{t}_{S^c} < 0.$$

Therefore,

$$\boldsymbol{h}_{S^c}^* = 0.$$

Combining the two conditions, the constructed state satisfies

$$\boldsymbol{h}_i^* > 0 \quad \text{for } i \in S, \qquad \boldsymbol{h}_i^* = 0 \quad \text{for } i \in S^c.$$

Thus S is exactly the active set of the equilibrium.

Moreover, both the active and inactive conditions are strict. Since the corresponding ReLU arguments depend continuously on $\boldsymbol{drive}$, these strict inequalities remain satisfied under sufficiently small perturbations of $\boldsymbol{drive}$. Hence the set of inputs realizing the active set S contains a nonempty open region.

For $S = \varnothing$, any input satisfying

$$\boldsymbol{drive} < \boldsymbol{t}$$

componentwise yields the equilibrium $\boldsymbol{h}^* = 0$.

For $S = \{1, \ldots, N\}$, thus $|S| = N$, choose any $\boldsymbol{h}^* \gg 0$ and define

$$\boldsymbol{drive} = \boldsymbol{t} + (\boldsymbol{I} - \boldsymbol{W}_{LI})\boldsymbol{h}^*.$$

Then

$$\boldsymbol{drive} + \boldsymbol{W}_{LI}\boldsymbol{h}^* - \boldsymbol{t} = \boldsymbol{h}^* \gg 0,$$

so all neurons are active.

Therefore, every subset S of neurons can be realized as the active set of an equilibrium for an open region of input space. □

E LEARNED DECODER FOR SPATIALLY SUPERPOSED DIGITS

In Figure 9 we visualize the features learned by the decoder for the spatially superposed MNIST dataset. Because the network successfully decomposes a collage comprising multiple digits into these discrete, single-digit representations, it is able to dynamically scale its active capacity based on the number of digits present in the stimulus. This structural disentanglement directly enables the adaptive sparsity performance reported in Figure 2.

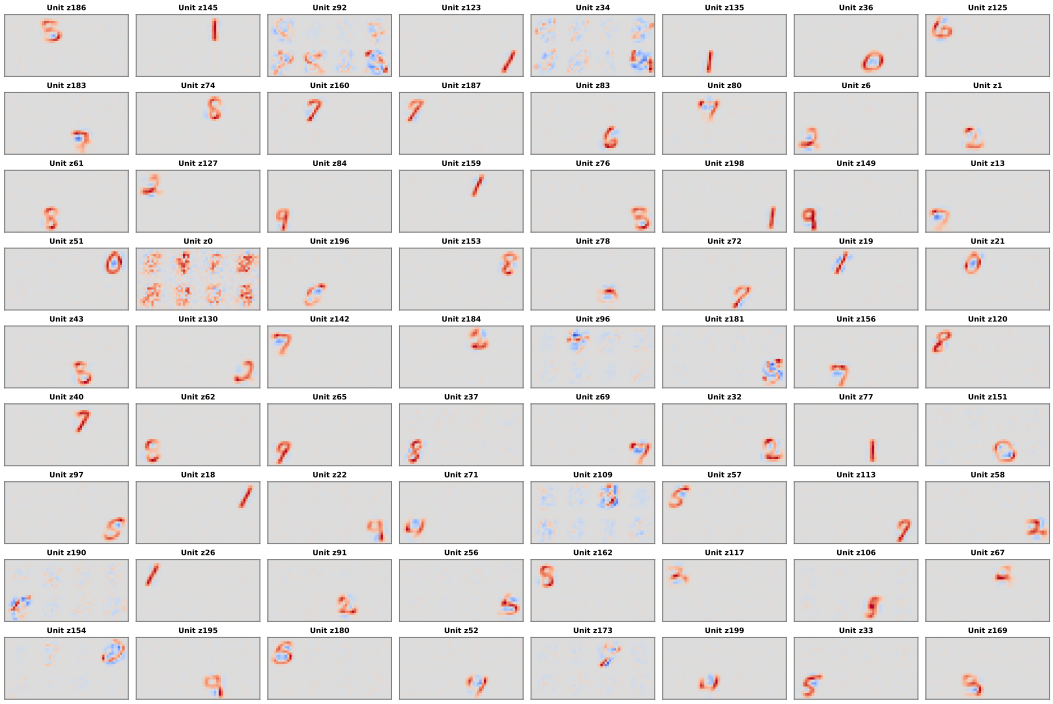


Figure 9: **Part of the learned decoder weights for the spatially superposed MNIST task.** By isolating individual digit features, the network allows the continuous-time lateral inhibition dynamics to dynamically activate the number of latent neurons required to reconstruct the input in proportion to its structural complexity.