A GEOMETRIC VIEW FOR UNDERSTANDING CONCEPT LEARNING AND NEURON INTERPRETATION IN SPARSE AUTOENCODERS

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose a unified mathematical framework for a geometric and structural understanding of concept learning and neuron interpretation in sparse autoencoders (SAEs). While SAEs improve interpretability of neural networks by learning sparse feature representations, a principled definition of "concept" and "learning" remains unclear. We formalize concepts as sets of data points and cast concept learning as a set-alignment problem between human-defined and model-induced concepts. This formulation distinguishes three increasingly strong notions of learning (i.e., detection, separation, and approximation) and yields geometric conditions, error bounds, and capacity constraints for when concepts can be represented by individual neurons or multi-neuron units. It also provides a set-theoretic account for common SAE phenomena, including feature splitting, feature absorption, and hierarchical concepts. Finally, we connect concept learning and neuron interpretation through formal concept analysis, showing that the two directions need not agree and that their many-to-many structure can be organized by concept lattices. Experiments with ReLU and Top-K SAEs illustrate the theory and reveal the effects of SAE hyperparameters on concept learning.

1 INTRODUCTION

Large neural networks achieve remarkable performance, yet their internal mechanisms remain opaque, limiting scientific understanding (DeGrave et al., 2021; Simon & Zou, 2025) and reliable deployment (Anwar et al., 2024). Mechanistic interpretability seeks to understand how information is represented and transformed in model activations (Elhage et al., 2021; Olah, 2022). However, interpretability can be hindered by polysemanticity (Templeton et al., 2026), commonly attributed to superposition (Elhage et al., 2022; O'Neill et al., 2024): a neuron may activate for multiple concepts, and a concept may be distributed among multiple neurons. Sparse autoencoders (SAEs) aim to mitigate this problem with overcomplete sparse representations that are more monosemantic and interpretable (Ng et al., 2011; Cunningham et al., 2023).

SAEs are commonly motivated by the linear representation hypothesis (Park et al., 2024), under which concepts correspond to directions in activation space that combine approximately linearly. Yet a direction by itself does not specify an interpretable concept. In practice, SAE neurons are interpreted through the examples that activate them most strongly, by identifying patterns shared across those examples (Bills et al., 2023). Practically, then, the interpretation of an SAE neuron is determined by a *set of examples* rather than a vector.

This distinction points to a more fundamental ambiguity: the literature lacks a precise definition of *concept*, reflecting a long-standing divide in philosophy (Ayonrinde & Jaburi, 2025). A Platonic view treats a concept as an abstract entity that exists independently of its instances. A class-nominalist view instead identifies a concept with its *extension*, the set of instances to which it applies. Because interpretability methods typically infer and validate interpretations from activating examples, their operational notion of concept is closer to the class-nominalist view. Accordingly, here we define a concept by a set of examples. This leads to our central formulation:

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.

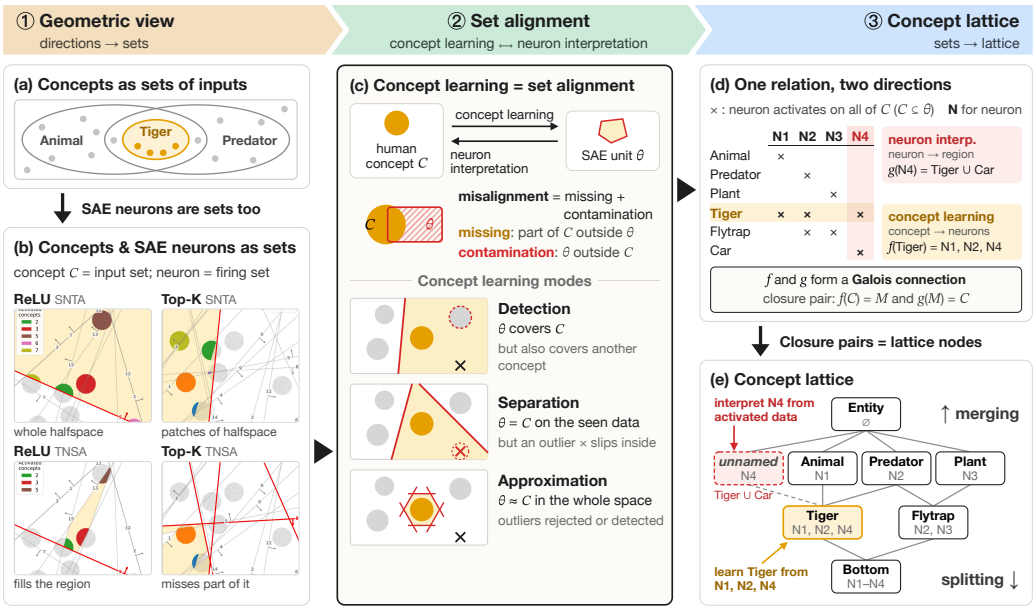


Figure 1: Overview of our framework. **(a,b)** Human concepts and SAE neurons are both sets of input data; (b) shows ReLU and Top-K SAE's activation pattern for single-neurons-total-activation (SNTA) and total-neuron-single-activation (TNSA). **(c)** Concept learning maps a concept C to a unit θ, three concept learning modes with increasing strength are shown; neuron interpretation maps θ back. **(d,e)** The two directions of concept learning and neuron interpretation are the rows (f) and columns (g) of one incidence relation, organized by a concept lattice, and need not agree: $f(\text{Tiger}) = \{N_1, N_2, N_4\}$, but $g(N_4) = \text{Tiger} \cup \text{Car}$.

> Concept learning and neuron interpretation are *set-alignment* problems between human-defined and model-induced concepts, where a concept is a set, and learning is alignment.

Under this formulation, learning a concept amounts to aligning a neuron-induced set with a human-defined set of examples. For instance, a neuron learns a concept when it selects examples that agree with the human-defined set. Interpreting a neuron amounts to aligning human-defined sets to the neuron-induced set. For instance, a neuron is polysemantic when it represents multiple sets.

Our contributions are the following. **(1)** We propose a unified geometric and set-theoretic framework that casts concept learning and neuron interpretation as set alignments. **(2)** We distinguish between three modes of concept learning (detection, separation, and approximation), deriving the necessary and sufficient conditions for each mode. **(3)** We show that concept learning and neuron interpretation are related but distinct, formally relating them through concept lattices (Ganter et al., 1999).

2 Preliminaries

Due to space limit, we delay the related work section to Appendix. A.2. We first review SAEs, focusing on ReLU SAE (Cunningham et al., 2023) and Top-K SAE (Gao et al., 2025). Let $x \in \mathbb{R}^n$, $n \geq 1$, be an activation vector from a neural network. An SAE maps x to a sparse activation vector $a \in \mathbb{R}^d$, where $d \geq 1$ is the number of SAE neurons, usually larger than n, and then reconstructs x:

$$z = \text{Enc}(x) = W_{\text{enc}}(x - b_{\text{pre}}) + b_{\text{enc}},$$
$$a = \text{Act}(z), \qquad \hat{x} = \text{Dec}(a) = W_{\text{dec}} a + b_{\text{dec}}.$$

Here z contains the pre-activation scores, a_i is neuron i's activation, and $\hat{x}$ is the reconstruction. The W terms are weight matrices, b_{pre} centers x, $b_{\text{enc}}, b_{\text{dec}}$ are biases, and Act is the activation function. ReLU uses $a_i = \max\{z_i, 0\}$ and loss $\|x - \hat{x}\|_2^2 + \lambda \|a\|_1$, where $\lambda \geq 0$ controls the sparsity penalty. Top-K keeps the k largest scores, applies ReLU to them, and sets the others to zero; its loss is $\|x - \hat{x}\|_2^2$. We take $1 \leq k \leq d$ and resolve ties by a fixed ordering of neuron indices, giving at most k positive activations. Writing $W = W_{\text{enc}}$ and $b = b_{\text{enc}} - W_{\text{enc}} b_{\text{pre}}$, we have $z_i(x) = \langle w_i, x \rangle + b_i$, where w_i is row i of W. Thus a linear direction, together with its bias and activation, determines a set of activating inputs.

3 MAIN FRAMEWORK

3.1 NOTATION AND DEFINITIONS

Concepts. Let $X \subseteq \mathbb{R}^n$ be a Borel data domain in the SAE input space. A human concept is a measurable set $C \in \mathcal{B}(X)$, and $\mathcal{C} \subseteq \mathcal{B}(X)$ is the family of concepts under study. We use Borel probability measures μ, ν on $\mathbb{R}^n$, with $\mu(X) = 1$. The measure μ evaluates alignment on the data domain, whereas ν evaluates alignment on the ambient space and may assign mass outside X.

SAE neurons. Write $[d] = \{1, \dots, d\}$. For neuron $i \in [d]$, define its positive- and nonpositive-score regions, using threshold zero, by

$$H_i^+ = \{x \in X : z_i(x) > 0\}, \qquad H_i^- = \{x \in X : z_i(x) \le 0\}.$$

Its actual *activation region* is $N_i = \{x \in X : a_i(x) > 0\}$, with $N_i^+ = N_i$ and $N_i^- = X \setminus N_i$. For ReLU, $N_i = H_i^+$; for Top-K, a positive score may not enter the top-k set, so $N_i \subseteq H_i^+$.

SAE activations. An *activation pattern* s assigns a sign $\sigma_{s;i} \in \{+, -\}$ to every neuron, recording whether it is active. Its *total neuron single activation* (TNSA) region is $R_s = \bigcap_{i \in [d]} N_i^{\sigma_{s;i}}$. Let $\mathcal{A}$ be the collection of all patterns, including those with empty regions. The *single-neuron total activation* (SNTA) is $N_i = \bigcup_{s \in \mathcal{A}:\, \sigma_{s;i} = +} R_s$: all inputs that activate neuron i, regardless of the other neurons.

For a neuron set $M \subseteq [d]$, define the *multi-neuron activation*, or *unit*, as $\theta_M = \bigcap_{i \in M} N_i$. It requires all selected neurons to activate, but leaves the others unrestricted. Write $\Theta = \{\theta_M : M \subseteq [d]\}$ for the model-learned concepts and set $\theta_\emptyset = X$. A single neuron N_i can cover several unrelated concepts, leading to *polysemanticity*, while R_s can be too small because it specifies every neuron's activity. By combining selected neurons, θ_M provides an intermediate level for concept learning and interpretation. Different concepts may use different numbers $|M|$ of neurons.

A tilde denotes the same activation evaluated on all of $\mathbb{R}^n$, rather than only X. Thus $N_i = X \cap \widetilde{N}_i$, $\theta_M = X \cap \widetilde{\theta}_M$, and $\widetilde{\theta}_\emptyset = \mathbb{R}^n$; the same convention applies to $H_i^\pm$.

3.2 SPARSE AUTOENCODER ARCHITECTURES

We categorize SAE architectures into *absolute activation*, where each neuron's activity is determined independently of other neurons, and *relative activation*, where it also depends on other neurons. ReLU is absolute activation: for $w_i \neq 0$, $\widetilde{N}_i$ is an open halfspace, one side of a flat boundary with the boundary excluded; for $w_i = 0$, it is empty or all of $\mathbb{R}^n$. A ReLU unit intersects these regions, so its ambient region is *convex*: it contains the line segment joining any two of its points. Its restriction to X need not be convex.

Top-K is relative activation: $N_i = \{x \in X : z_i(x) > 0,\ i \in \mathrm{TopK}_k(z(x))\}$, where TopK_k returns the selected indices. Its activation region can consist of several pieces inside the positive-score halfspace and need not be convex (Hindupur et al., 2026). Therefore, our geometric results focus on ReLU. Appendix A.9 analyzes *rank interference*: the removal of otherwise positive activations because competing neurons have higher-ranked scores.

3.3 CONCEPT LEARNING

For simplicity, we write θ for θ_M. **The goal of concept learning is to align human concepts $\mathcal{C}$ with model-learned concepts Θ.** We consider three modes: concept detection, concept separation, and concept approximation. For concept representations we require $M \neq \emptyset$; otherwise $\theta_\emptyset = X$ would make coverage automatic. Unless otherwise specified, we study ReLU SAE in the main text.

3.3.1 CONCEPT DETECTION

Concept detection asks whether a unit covers the target concept.

Definition 3.1 (Concept detection)**.** C is detected if some nonempty M satisfies $\mu(C \setminus \theta_M) = 0$.

This allows missed points of zero probability under μ. Detection alone permits many-to-many mappings: one concept may be covered by several units, and one unit may cover several concepts.

3.3.2 CONCEPT SEPARATION

Definition 3.2 (Concept separation)**.** A concept C is said to be separated by θ_M if (i) $x \in N_i$ for all $x \in C$ and $i \in M$, and (ii) for every $x' \in X \setminus C$, $x' \in N_j^-$ for some $j \in M$.

Equivalently, every concept point activates all selected neurons, and every non-concept point fails to activate at least one. Imperfect separation can be measured by $e_{\text{sep}}(C, \theta) = \mu(C \Delta \theta)$, where $C \Delta \theta = (C \setminus \theta) \cup (\theta \setminus C)$ is the *symmetric difference*: missed concept points together with incorrectly included points. Appendix A.6.5 gives details. Separation evaluates only X; the same unit may still include large blank regions outside it. This limits what separation alone tells us about previously unseen concepts.

3.3.3 CONCEPT APPROXIMATION

Definition 3.3 (Concept approximation)**.** For $\varepsilon \geq 0$, a concept C is said to be ε-approximated by θ_M if the total ν-measure of points violating the following conditions is at most ε: (i) $x \in \widetilde{N}_i$ for all $x \in C$ and $i \in M$, and (ii) for every $x' \in \mathbb{R}^n \setminus C$, $x' \in \widetilde{N}_j^-$ for some $j \in M$. Equivalently, $e_{app}(\Delta, \theta_M) \coloneqq \nu(C \Delta \widetilde{\theta}_M) \leq \varepsilon$.

The key difference between concept separation and concept approximation is the choice of space: the former is evaluated on X, whereas the latter is evaluated in $\mathbb{R}^n$. Intuitively, approximation asks the unit to fit the concept more tightly, penalizing included blank regions when ν gives them positive probability. This is relevant to concept discovery and anomaly detection (Ruff et al., 2018). The measures may differ, so small approximation error does not automatically imply small separation error (Appendix A.6.5).

3.4 NEURON INTERPRETATION

Concept learning and neuron interpretation are two directions of the same relation between concepts and units. Concept learning starts with a human concept and seeks a matching unit; neuron interpretation starts with a unit and seeks a human concept that describes its activating inputs. A neuron may cover several concepts, and a concept may be covered by several neurons. Section 4.4 uses formal concept analysis (FCA) (Ganter et al., 1999) to organize these relations by the neurons that inputs activate in common. Appendix A.4 describes common SAE phenomena in this framework.

4 THEORETICAL RESULTS

Here we present some theoretical results, which are existence results that allow the affine ReLU encoder parameters to be chosen freely and with enough neurons for each construction. They characterize representability, not whether training recovers the required neurons. Arbitrarily accurate approximation may require increasing width.

Assumption 4.1. The concept family $\mathcal{C}$ is finite and nonempty; each $C \in \mathcal{C}$ is nonempty and compact; and $X = \bigcup_{C \in \mathcal{C}} C$.

This includes finite samples and compact continuous concept regions, without requiring concepts to be convex or disjoint. It makes X compact, but $X \setminus C$ need not be closed.

4.1 CONCEPT SEPARATION

Write $D_C = X \setminus C$ for the non-concept region, Conv for convex hull, and $\overline{A}$ for ambient closure. Proofs are in Appendix A.6.

Theorem 4.2 (Concept separation with one neuron)**.** *A concept C can be separated from D_C by one neuron if and only if*

$$\mathrm{Conv}(C) \cap \overline{\mathrm{Conv}(D_C)} = \varnothing.$$

Intuitively, one affine boundary must place the entire target on one side and all non-target points on the other; taking convex hulls preserves these constraints.

Corollary 4.3. *All concepts can be separated by one neuron per concept if and only if the condition in Theorem 4.2 holds for every $C \in \mathcal{C}$. When it holds, the minimum number of representing neurons is $|\mathcal{C}|$.*

A neuron has only one activation region, so it cannot exactly represent two distinct concept sets.

Theorem 4.4 (Concept separation with multiple neurons)**.** *A concept C can be separated from D_C by finite unit if and only if*

$$\mathrm{Conv}(C) \cap \overline{D_C} = \varnothing.$$

To better understand, different neurons can exclude different non-target regions. However, any convex unit covering C must also cover its convex hull. The closure matters because a finite ReLU unit is open and cannot exclude non-target points that accumulate inside it.

Corollary 4.5. *All concepts can be separated by a finite units if and only if*

$$\mathrm{Conv}(C_i) \cap \overline{(C_j \setminus C_i)} = \varnothing \quad \textit{for all distinct } C_i, C_j \in \mathcal{C}.$$

Each target concept's convex hull must avoid the non-target portion of every other concept, including its limit points.

These results show that combining units can turn individually polysemantic neurons into a more selective unit. Increasing width provides more candidate boundaries, but the corollary also reveals a geometric limit: some concepts remain inseparable even with finitely many ReLU units. Thus, feature splitting can mitigate polysemanticity, but is not a universal solution, especially for hierarchical concepts that require separating a child from the remainder of its parent.

4.2 Concept Approximation

We now evaluate the ambient unit under e_{app}, rather than requiring exact equality on X.

Theorem 4.6 (Concept approximation condition)**.** *A concept C can be arbitrarily well approximated by a finite unit if and only if there exists a Borel convex set $K \subseteq \mathbb{R}^n$ such that $\nu(C \Delta K) = 0$.*

ReLU units can approximate convex regions, but cannot remove nonconvexity that remains visible to the evaluation measure. The condition allows differences on sets of zero ν-mass and holds even for atomic measures (Appendix A.7.1).

Proposition 4.7 (Concept approximation error rate)**.** *Let $L \subseteq \mathbb{R}^n$ be an r-dimensional affine subspace containing C, with $r \geq 2$. Suppose ν is supported on L and has a continuous, strictly positive density relative to r-dimensional Lebesgue measure. Let $K \subseteq L$ be a convex body (a compact convex set with nonempty relative interior) with C^2 boundary and positive Gaussian curvature, relative to L. For all sufficiently large m, some ReLU unit with $|M| \leq m$ satisfies*

$$e_{\mathrm{app}}(C, \widetilde{\theta}_M) \leq \nu(C \Delta K) + A_K m^{-2/(r-1)},$$

where A_K depends on K, the density, and r, but not on m.

Intuitively, each selected neuron supplies a flat boundary; more boundaries approximate a smooth convex region more closely, with a slower rate in higher dimension. This follows from weighted polytope approximation (Ludwig, 1999); the proof and the interval case $r = 1$ are in Appendix A.7.2.

The irreducible error is $e_{\mathrm{irr}} := \inf_{K \text{ Borel convex}} \nu(C \Delta K)$, equal to the infimum over finite ReLU units. It need not equal $\nu(\mathrm{Conv}(C) \setminus C)$: a better convex region may omit part of C. The displayed rate applies to a chosen smooth K, and has residual e_{irr} only if an optimal convex region meets the same regularity conditions. For $K = C$, the residual vanishes. Separately constructed units for multiple concepts use at most the sum of their sizes; neuron reuse can reduce this construction budget.

4.3 Concept Learning Capacity

Geometry constrains which concepts are representable; width also limits the number of candidate units. Let $k_c \geq 1$ be an integer budget on the number of neurons selected per concept, distinct from the mean activation sparsity L_0.

Theorem 4.8 (Concept learning capacity)**.** *Suppose distinct concepts are assigned distinct nonempty neuron sets $M_C \subseteq [d]$, with $|M_C| \leq k_c$. Then $|\mathcal{C}| \leq \sum_{j=1}^{\min(k_c,d)} \binom{d}{j}$. When $k_c/d \to 0$, this implies $d \geq (k_c! \, |\mathcal{C}|)^{1/k_c}$.*

Intuitively, combining neurons creates more candidate descriptions than using one neuron per concept. Exact separation forces this injective assignment, but counting subsets does not guarantee distinct activation regions or monosemantic individual neurons. Top-K additionally requires $|M_C| \leq k$ for a nonempty region: a larger dictionary cannot replace the need for the selected neurons to coactivate. The proof is in Appendix A.8.

4.4 Concept Learning and Neuron Interpretation

For a fixed SAE, the incidence relation $x \mathcal{I} i \iff x \in N_i$ defines a formal context with data points as objects and neurons as attributes (Ganter et al., 1999). For $A \subseteq X$ and $M \subseteq [d]$, define

$$f(A) = \{i \in [d] : A \subseteq N_i\}, \qquad g(M) = \bigcap_{i \in M} N_i = \theta_M, \quad g(\emptyset) = X.$$

Thus f collects all neurons covering a region, while g returns their common activation region.

Theorem 4.9 (Galois connection)**.** *The maps $f : \mathcal{P}(X) \to \mathcal{P}([d])$ and $g : \mathcal{P}([d]) \to \mathcal{P}(X)$ form an inclusion-reversing Galois connection:*

$$A \subseteq g(M) \iff M \subseteq f(A).$$

Intuitively, a region lies inside a unit exactly when every selected neuron covers the whole region. Enlarging the region can only remove common neurons; adding neurons can only shrink the unit.

The composition $g \circ f$ is a closure operator: $E_C := g(f(C))$ is the smallest available unit region containing C. An exact nonempty-unit representation exists precisely when $E_C = C$ and $f(C) \neq \emptyset$. This is the fixed-dictionary counterpart of the existence results: they ask whether suitable neurons can exist, whereas closure tests what the available neurons distinguish. A *formal concept* is a pair (E, M) with $E = g(M)$ and $M = f(E)$, called its *extent* and *intent*. Ordered by extent inclusion, these pairs form a concept lattice. Here "closed" means unchanged by $g \circ f$, not topologically closed. Full details of lattice operations are in Appendix A.5.

Different human concepts may share one closed region, and inclusion between regions reverses inclusion between their intents. The lattice therefore records sharing and hierarchy induced by the learned activation geometry, not necessarily by human semantics. Full-coverage f is distinct from selecting a best-F1 unit, and the label-closedness used in our experiments tests whole-label inclusion rather than equality $E_C = C$. Together, the geometric and structural views connect the encoder's activation regions, the arrangement of data concepts, their representation by intersections, and the shared or hierarchical interpretations of those representations.

5 Experiments

Our experiments ask how trained SAEs realize the concept regions characterized by our theory. We first study how data geometry, gating, and sparsity affect concept learning, then use FCA to connect these effects to concept learning and neuron interpretation.

5.1 Setup

Data and models. We use two-dimensional synthetic concepts represented by disjoint or partially overlapping clusters. For real-world data, we use the ten Imagenette classes with 768-dimensional DINOv3 embeddings (Howard, 2019; Siméoni et al., 2025), and Yelp review sentences with five star-rating labels and 1536-dimensional `text-embedding-3-small` embeddings (Yelp, 2024; OpenAI, 2024). We vary expansion factor (EF) and sparsity in ReLU and Top-K SAEs. Larger L_0, the mean number of active neurons per input, means denser activation. See Appendix A.3 for training and aggregation details.

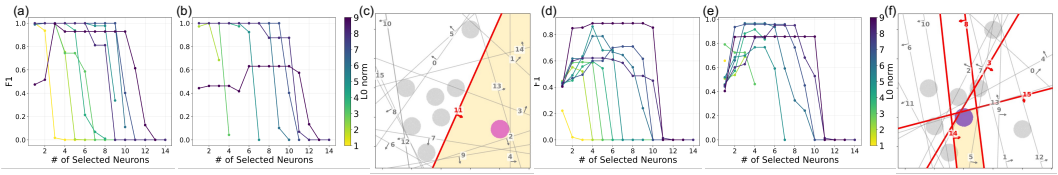


Figure 2: Concept learning on disjoint synthetic data at EF= 8. **(a,b)** F1 curves for an easy-to-separate concept using ReLU and Top-K SAEs; **(c)** its ReLU activation geometry. **(d–f)** Corresponding results for a harder, surrounded concept.

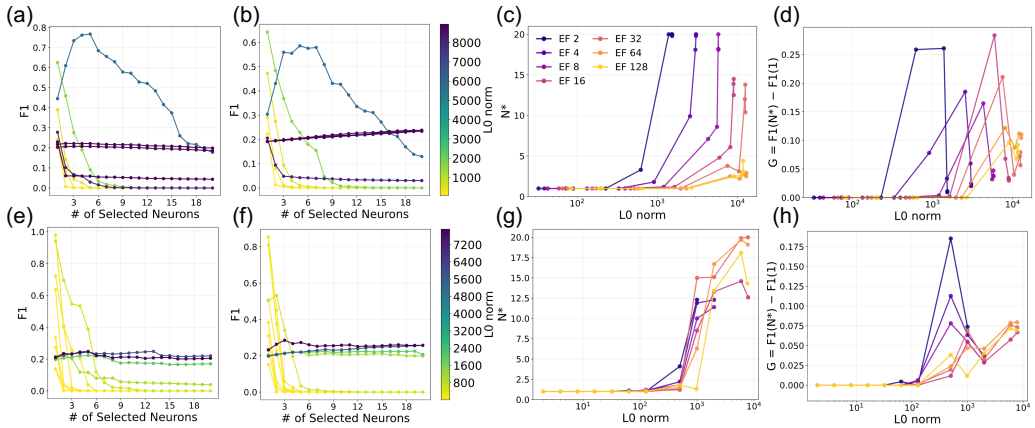


Figure 3: Concept learning on Imagenette. **(a,b)** ReLU F1 curves for tench and chain saw at EF= 16. **(c,d)** Mean N^* and G across concepts, varying L_0 and EF. **(e–h)** Corresponding Top-K results.

Concept learning. Following Hindupur et al. (2026), we measure alignment with F1. We rank neurons by individual F1 and intersect the top N activation regions: membership requires all selected neurons to activate. We sweep the permitted unit sizes, capped at 20 for real-world data. We report test score $F_1(N)$, the smallest best-scoring size N^* within the sweep, and gain $G = F_1(N^*) - F_1(1)$.

Concept lattice. We connect these findings to the concept lattice (Section 4.4) using the overlapping synthetic dataset. For each concept C, its intent $M_C = f(C)$ contains all neurons active on every training example of C. For nonempty intents, $E_C = g(M_C)$ is the smallest available unit region covering C on that sample. We evaluate its coverage on test points and summarize which whole concepts it contains. Concepts are *merged* when their full-coverage descriptions coincide; a *parent* description contains another whole concept, and a *child* has only the reverse relation. A concept is *label-closed* when its region covers C but no entire other concept, including both child and isolated states. An isolated concept is closed but not a child. Closedness need not mean $E_C = C$: partial overlap can remain. An *empty* intent instead indicates that no single activation constraint covers the whole concept, not that the SAE contains no information about it.

5.2 Synthetic Data: Concept Separation

Figure 2 contrasts a peripheral concept separable by one halfspace with a surrounded concept requiring several boundaries. For ReLU, this illustrates Theorems 4.2 and 4.4: individually nonselective halfspaces can jointly exclude different neighbours. For the harder concept, ReLU F1 at $L_0 = 9$ rises from 0.405 with one neuron to 0.887 with seven (119% relative gain). Top-K F1 at $L_0 = 7$ rises from 0.441 to 0.958 with three; its best score across sparsities is 0.966.

One neuron already reaches F1$= 1$ for the easier concept, but combinations can compensate for a poorly aligned dictionary: at $L_0 = 9$, three ReLU neurons improve F1 from 0.475 to 0.996. Thus, needing multiple neurons can reflect either concept geometry or the learned dictionary. A polysemantic neuron may still supply an essential boundary of a selective unit: useful concept representations need not coincide with individual dictionary entries. Additional details are in Appendix. A.10.

5.3 Real-World Data: Concept Separation

Figures 3 and 4 extend the analysis to image and text embeddings. Guided by the one-neuron separation criterion, we select easier concepts (tench and 1-star) and harder concepts (chain saw and 4-star) by their lowest and highest one-vs-rest linear SVM errors.

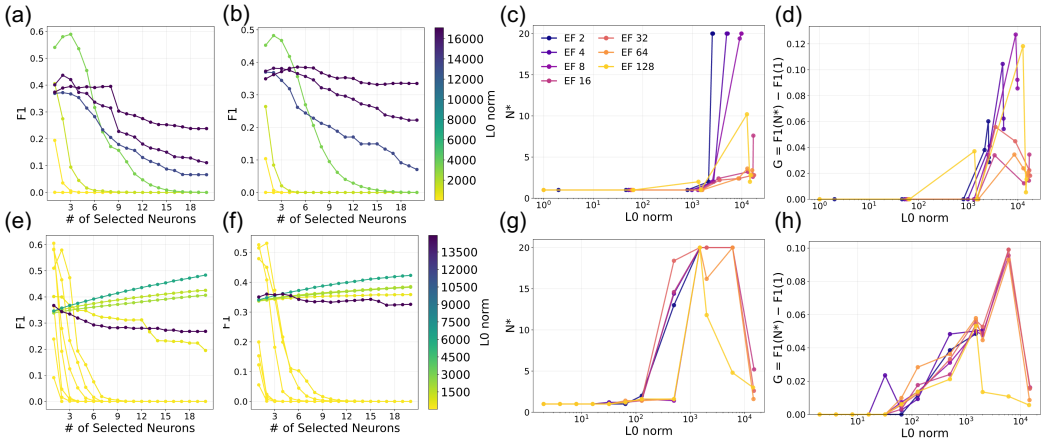


Figure 4: Concept learning on Yelp, with the same organization as Figure 3. The case studies use 1-star and 4-star sentences.

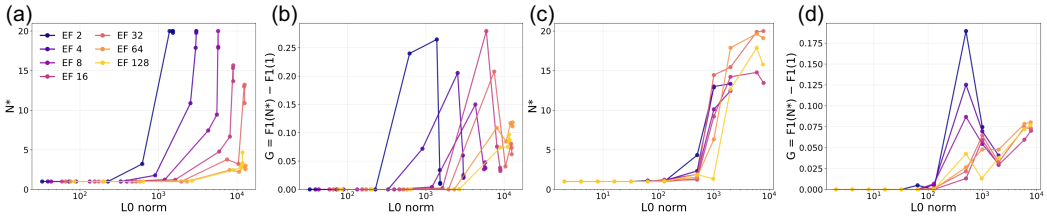


Figure 5: Concept learning on Imagenette without tench. (ab) N^* and G for ReLU SAEs aggregated over other concepts. (cd) similar to (ab) but use Top-K SAEs.

The curves typically fall in sparse SAEs, become hump-shaped at intermediate sparsity, and rise more slowly in denser SAEs. Within a fixed dictionary, each additional neuron can remove both contamination and target examples. Sparse features may cover different portions of a concept, making their intersection lose recall; at intermediate sparsity, several boundaries can remove contamination before recall deteriorates; broad, strongly coactive features may require many intersections to become selective. For example, five ReLU neurons improve tench F1 from 0.446 to 0.768, while three improve Yelp 1-star F1 from 0.541 to 0.590 at EF$= 16$.

Across concepts, the mean within-EF Spearman correlation ρ_s between L_0 and N^* is 0.76–0.82. At comparable L_0, wider ReLU SAEs tend to need fewer neurons ($\rho_s = -0.68$ on Imagenette and -0.40 on Yelp). For Top-K, the mean bias-corrected η^2 values for N^*, G, and curve shape are 0.47–0.81 when grouped by L_0, but near or below zero when grouped by EF. ReLU transitions also shift with width, suggesting that active fraction L_0/d is useful alongside L_0 when comparing dictionaries. These results distinguish combinatorial capacity (Theorem 4.8) from usable concept representations: width supplies candidates, but their geometry and coactivation determine whether they combine usefully. Additional details including the full statistics table are in Appendix. A.10.

Importantly, the largest gain need not produce the highest final F1. Across our sweep, $13/15$ concepts attain their best ReLU score with multiple neurons, whereas $12/15$ attain their best score across both architectures with a single Top-K neuron. This is consistent with our relative-gating analysis (Appendix. A.9): one Top-K activation region can exploit boundaries imposed by competing neurons. A single-neuron interpretation therefore need not imply geometrically simple or independent computation. Architecture and sparsity should be chosen jointly with interpretation granularity, evaluating final F1 for both single neurons and units, not just G.

5.4 Concept learning on unseen concepts: concept approximation

To probe beyond the training support (Section 4.2), we withhold tench during SAE training and reintroduce it for neuron selection and evaluation. The aggregate N^* and G trends for other concepts, as shown in Fig. 5 remain similar to full-data training Fig. 3. This suggests that learned activation regions can correctly exclude concepts absent from SAE training, rather than requiring every concept to acquire a dedicated feature during fitting. Additional details can be found in Appendix. A.10.

5.5 FCA: FROM ACTIVATION GEOMETRY TO CONCEPT STRUCTURE

Figure 6 shows a progression from empty or merged descriptions at low L_0, through parent–child structure, toward more closed concepts at higher L_0. Near $L_0 = 2$, 59% of ReLU and 76% of Top-K concept descriptions are merged; near $L_0 = 64$, 75% and 73% are closed. Greater width often resolves sharing, but its benefit is not uniform at matched sparsity. This is consistent with feature specialization/splitting: denser coactivation can yield finer *joint* descriptions even when individual neurons remain broad.

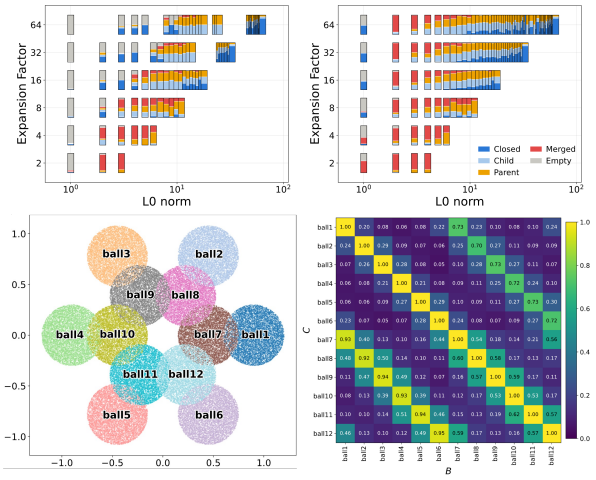


Figure 6: FCA on overlapping concepts. Top: states for ReLU (left) and Top-K (right). Bottom: data geometry and mean directional inclusion $\ell(C, B)$ over SAEs with nonempty intents.

Crucially, the labels contain no parent–child hierarchy. The induced hierarchy instead depends on spatial arrangement. Let $\ell(C, B) = |B \cap E_C|/|B|$. Across SAEs, inner concepts' regions cover 92–95% of their outward radial neighbours, versus 70–73% in reverse; this asymmetry persists in dense SAEs. A plausible explanation is a directional bias in the learned gates: boundaries preserving an inner concept can also admit points farther outward along the same direction. Thus, apparent hierarchical structure or feature families can partly reflect the geometry of the learned gates, rather than a semantic hierarchy recovered from the data.

These relations also explain variation in alignment: maximum inclusion correlates with best greedy F1 at $\rho_s \approx -0.74$ for both architectures. In the complementary greedy-selection analysis, we select a unit for C and ask which concept best matches that unit. Parent states disagree in 83% of ReLU and 80% of Top-K cases, compared with 2.5% and 4.8% for child states. Among disagreements with nonempty intents, the reverse label is the neighbour with largest $\ell(C, B)$ in 70–74% of cases. The same geometric entanglement therefore helps explain both difficulty recovering a target and which alternative interpretation is preferred.

Taken together, the experiments suggest viewing an SAE as a system of reusable activation constraints, rather than a one-to-one dictionary of human concepts. Data geometry and SAE hyperparameters shape those constraints; their intersections determine concept learning; and the FCA lattice organizes the resulting shared, hierarchical, and distinct interpretations. Evaluating this structure alongside F1 can reveal whether an apparently interpretable feature captures the intended concept or an architecture-dependent grouping of concepts.

6 DISCUSSION

We present a set-theoretic framework for concept learning and neuron interpretation in SAEs. Geometrically, it characterizes when individual neurons or multi-neuron units can represent concepts; structurally, FCA organizes the many-to-many relation between concepts and activation regions. Together with our experiments, the framework shows how data geometry, and SAE architectural parameters shape concept learning, and why combining features can improve concept alignment without guaranteeing a unique neuron interpretation.

Our analysis focuses primarily on ReLU SAEs and does not fully address other architectures, generalization, training dynamics or label-free concept discovery. Extending the framework to richer SAE architectures and developing methods for bidirectional concept–neuron alignment are natural directions for future work.

AI USE STATEMENT

In this work, we used generative AI tools for language editing and presentation polishing, for assisting with the implementation of a subset of the code used to analyze experimental results, and for independently checking the correctness of mathematical proofs. Generative AI tools were not used to generate research ideas, design experiments, develop the main experimental codebase, conduct data or statistical analyses, interpret experimental results, or draft the substantive scientific content of the paper. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The code are provided as in the supplementary materials.

REFERENCES

Charalambos D. Aliprantis and Kim C. Border. *Infinite Dimensional Analysis: a Hitchhiker's Guide*. Springer, Berlin; London, 2006. ISBN 9783540326960 3540326960. doi: 10.1007/3-540-29587-9.

Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, et al. Foundational challenges in assuring alignment and safety of large language models. *arXiv preprint arXiv:2404.09932*, 2024.

Kola Ayonrinde and Louis Jaburi. A mathematical philosophy of explanations in mechanistic interpretability–the strange science part I.i. *arXiv preprint arXiv:2505.00808*, 2025.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6541–6549, 2017.

Usha Bhalla, Thomas Fel, Can Rager, Sheridan Feucht, Tal Haklay, Daniel Wurgaft, Siddharth Boppana, Matthew Kowal, Vasudev Shyam, Jack Merullo, et al. Do sparse autoencoders capture concept manifolds? *arXiv preprint arXiv:2604.28119*, 2026.

Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. Language models can explain neurons in language models. `https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html`, 2023.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2(5):6, 2023.

Bart Bussmann, Noa Nabeshima, Adam Karvonen, and Neel Nanda. Learning multi-level features with matryoshka sparse autoencoders. In *Forty-second International Conference on Machine Learning*, 2025. URL `https://openreview.net/forum?id=m25T5rAy43`.

Tue M Cao, Hoang X Nhat, Raed Alharbi, Phi Le Nguyen, and My T Thai. Tree sae: Learning hierarchical feature structures in sparse autoencoders. *arXiv preprint arXiv:2605.07922*, 2026.

David Chanin. Are sparse autoencoder benchmarks reliable? *arXiv preprint arXiv:2605.18229*, 2026. URL `https://arxiv.org/abs/2605.18229`.

David Chanin and Adrià Garriga-Alonso. SynthSAEBench: Evaluating sparse autoencoders on scalable realistic synthetic data. *arXiv preprint arXiv:2602.14687*, 2026. URL `https://arxiv.org/abs/2602.14687`.

David Chanin, James Wilken-Smith, Tomáš Dulka, Hardik Bhatnagar, Satvik Golechha, and Joseph Bloom. A is for absorption: Studying feature splitting and absorption in sparse autoencoders. *Advances in Neural Information Processing Systems*, 38:82318–82355, 2026.

Valérie Costa, Thomas Fel, Ekdeep S Lubana, Bahareh Tolooshams, and Demba Ba. From flat to hierarchical: Extracting sparse representations with matching pursuit. *Advances in Neural Information Processing Systems*, 38:34381–34424, 2026.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

Alex J DeGrave, Joseph D Janizek, and Su-In Lee. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, 2021.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE, 2009.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.

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.

Dominik Endres and Peter Földiák. Interpreting the neural code with formal concept analysis. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou (eds.), *Advances in Neural Information Processing Systems*, volume 21, pp. 441–448. Curran Associates, Inc., 2008. URL https://proceedings.neurips.cc/paper/2008/hash/eb160de1de89d9058fcb0b968dbbbd68-Abstract.html.

Josh Engels, Eric Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are one-dimensionally linear. In *International Conference on Learning Representations*, volume 2025, pp. 84591–84622, 2025.

Soham Gadgil, Chris Lin, and Su-In Lee. Ensembling sparse autoencoders. *arXiv preprint arXiv:2505.16077*, 2025.

Bernhard Ganter, Rudolf Wille, and Rudolf Wille. *Formal concept analysis*, volume 150. Springer, 1999.

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.

Yehoram Gordon, Mathieu Meyer, and Shlomo Reisner. Volume approximation of convex bodies by polytopes—a constructive method. *Studia Mathematica*, 111(1):81–95, 1994. URL http://eudml.org/doc/216121.

Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. Finding neurons in a haystack: Case studies with sparse probing. *arXiv preprint arXiv:2305.01610*, 2023.

Thomas Heap, Tim Lawson, Lucy Farnik, and Laurence Aitchison. Sparse autoencoders can interpret randomly initialized transformers. *arXiv e-prints*, pp. arXiv–2501, 2025.

Sai Sumedh R Hindupur, Ekdeep S Lubana, Thomas Fel, and Demba Ba. Projecting assumptions: The duality between sparse autoencoders and concept geometry. *Advances in Neural Information Processing Systems*, 38:11649–11699, 2026.

Johannes Hirth and Tom Hanika. Conceptual views of neural networks: A framework for neuro-symbolic analysis. *arXiv preprint arXiv:2209.13517*, 2022. URL https://arxiv.org/abs/2209.13517v2. Revised 2026.

Jeremy Howard. Imagenette: A smaller subset of 10 easily classified classes from imagenet, March 2019. URL https://github.com/fastai/imagenette.

Vladimir Itskov, Alexander Kunin, and Zvi Rosen. Hyperplane neural codes and the polar complex. In *Topological Data Analysis: The Abel Symposium 2018*, pp. 343–369. Springer, 2020.

R Amzi Jeffs. Morphisms of neural codes. *SIAM Journal on Applied Algebra and Geometry*, 4(1): 99–122, 2020.

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 *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=qrU3yNfX0d.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pp. 5338–5348. PMLR, 2020.

Monika Ludwig. Asymptotic approximation of smooth convex bodies by general polytopes. *Mathematika*, 46(1):103–125, 1999. doi: 10.1112/S0025579300007609.

Yifan Luo, Yang Zhan, Jiedong Jiang, Tianyang Liu, Mingrui Wu, Zhennan Zhou, and Bin Dong. From atoms to trees: Building a structured feature forest with hierarchical sparse autoencoders. *arXiv preprint arXiv:2602.11881*, 2026.

Eric J Michaud, Liv Gorton, and Tom McGrath. Understanding sparse autoencoder scaling in the presence of feature manifolds. *arXiv preprint arXiv:2509.02565*, 2025.

Guido Montúfar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. On the number of linear regions of deep neural networks. *Advances in neural information processing systems*, 27, 2014.

Jesse Mu and Jacob Andreas. Compositional explanations of neurons. *Advances in Neural Information Processing Systems*, 33:17153–17163, 2020.

Mark Muchane, Sean Richardson, Kiho Park, and Victor Veitch. Incorporating hierarchical semantics in sparse autoencoder architectures. *arXiv preprint arXiv:2506.01197*, 2025.

Andrew Ng et al. Sparse autoencoder. *CS294A Lecture notes*, 72(2011):1–19, 2011.

Tuomas Oikarinen, Ge Yan, and Tsui-Wei Weng. Evaluating neuron explanations: A unified framework with sanity checks. *arXiv preprint arXiv:2506.05774*, 2025.

Chris Olah. Mechanistic interpretability, variables, and the importance of interpretable bases. *Transformer Circuits Thread*, 2(4), 2022.

Bruno A Olshausen and David J Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision research*, 37(23):3311–3325, 1997.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Charles O'Neill, Christine Ye, Kartheik Iyer, and John F Wu. Disentangling dense embeddings with sparse autoencoders. *arXiv preprint arXiv:2408.00657*, 2024.

OpenAI. New embedding models and api updates, January 2024. URL https://openai.com/index/new-embedding-models-and-api-updates/.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 39643–39666. PMLR, 2024. URL https://proceedings.mlr.press/v235/park24c.html.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. In *International Conference on Learning Representations*, volume 2025, pp. 76441–76463, 2025.

Razvan Pascanu, Guido Montufar, and Yoshua Bengio. On the number of response regions of deep feed forward networks with piece-wise linear activations. *arXiv preprint arXiv:1312.6098*, 2013.

Gonçalo Santos Paulo, Alex Troy Mallen, Caden Juang, and Nora Belrose. Automatically interpreting millions of features in large language models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=EemtbhJOXc.

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*, 2024a.

Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. Jumping ahead: Improving reconstruction fidelity with jumprelu sparse autoencoders. *arXiv preprint arXiv:2407.14435*, 2024b.

Alexandre Rocchi, Thomas Fel, and Gianni Franchi. A geometric unification of concept learning with concept cones. *arXiv preprint arXiv:2512.07355*, 2025.

R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *International conference on machine learning*, pp. 4393–4402. PMLR, 2018.

Raphaël Sarfati, Eric Bigelow, Daniel Wurgaft, Jack Merullo, Atticus Geiger, Owen Lewis, Tom McGrath, and Ekdeep Singh Lubana. The shape of beliefs: Geometry, dynamics, and interventions along representation manifolds of language models' posteriors. *arXiv preprint arXiv:2602.02315*, 2026.

Or David Shafran, Shaked Ronen, Omri Fahn, Shauli Ravfogel, Atticus Geiger, and Mor Geva. From directions to regions: Decomposing activations in language models via local geometry. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=Tz7n7pX9cO.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. URL https://arxiv.org/abs/2508.10104.

Elana Simon and James Zou. InterPLM: discovering interpretable features in protein language models via sparse autoencoders. *Nature methods*, 22(10):2107–2117, 2025.

Chandan Singh, Aliyah R Hsu, Richard Antonello, Shailee Jain, Alexander G Huth, Bin Yu, and Jianfeng Gao. Explaining black box text modules in natural language with language models. *arXiv preprint arXiv:2305.09863*, 2023.

Richard P Stanley et al. An introduction to hyperplane arrangements. *Geometric combinatorics*, 13: 389–496, 2007.

Ye Su, Huayi Tang, Zixuan Gong, and Yong Liu. Sparsity is combinatorial depth: Quantifying moe expressivity via tropical geometry. *arXiv preprint arXiv:2602.03204*, 2026.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, et al. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *arXiv preprint arXiv:2605.29358*, 2026.

Bo Xiong. The lattice representation hypothesis of large language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=5K1FG92m5s.

Bo Xiong and Steffen Staab. From tokens to lattices: Emergent lattice structures in language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=md9qolJwLl.

Yelp. Yelp open dataset, 2024. URL https://business.yelp.com/data/resources/open-dataset/.

A Appendix

A.1 List of Contents

We organize the appendix as follows.

- **Related Work** A.2
- **Sparse Autoencoder Phenomena** A.4 describes how to use our framework to explain and formulate several major SAE phenomena.
- **Experiment Details** A.3 describes datasets and SAE training hyperparameters.
- **A Complete View of Concept Learning and Neuron Interpretation** A.5 includes proofs for Section 4.4.
- **Proofs for Concept Separation** A.6 contains proofs for Section 4.1.
- **Proofs for Concept Approximation** A.7 contains proofs for Section 4.2.
- **Proofs for Concept Learning Capacity** A.8 contains the proof for Section 4.3.
- **Additional Discussion on Top-K SAE** A.9 discusses the additional rank-interference effect induced by Top-K gating.
- **Additional Experiments** A.10 contains more experiment results about concept learning on both synthetic and real-world datasets and unseen concept learning.

A.2 Related Work

Mechanistic Interpretability and Sparse Autoencoders. Mechanistic interpretability studies models' internal computations (Elhage et al., 2021; Olah, 2022; Olsson et al., 2022). SAEs address superposition by learning sparse dictionaries that reconstruct activations and expose interpretable features (Elhage et al., 2022; Olshausen & Field, 1997; Ng et al., 2011; Cunningham et al., 2023; O'Neill et al., 2024; Templeton et al., 2026). Findings on random models and scaling motivate explicit criteria for concept alignment (Heap et al., 2025; Michaud et al., 2025).

SAE Architectures and Phenomena. Feature splitting, families, and absorption reveal relationships among SAE features (Bricken et al., 2023; Chanin et al., 2026). Gated, JumpReLU, Top-K, matching-pursuit, ensemble, and hierarchical variants modify sparsity and feature structure (Rajamanoharan et al., 2024a;b; Gao et al., 2025; Costa et al., 2026; Gadgil et al., 2025; Muchane et al., 2025; Bussmann et al., 2025; Luo et al., 2026; Cao et al., 2026). Hindupur et al. (2026) relate architecture to receptive-field geometry and design geometry-adapted encoders. Building on this duality, we characterize when intersections of ReLU activation regions can separate concepts that no single ReLU neuron can, and distinguish separation on the data support from ambient approximation.

Concept Learning and Neuron Interpretation. Interpretation infers descriptions from activations (Bills et al., 2023; Singh et al., 2023; Paulo et al., 2025), whereas concept-based analysis identifies features for supplied concepts (Gurnee et al., 2023; Koh et al., 2020; Bau et al., 2017; Mu & Andreas, 2020). Evaluating these associations (Oikarinen et al., 2025; Karvonen et al., 2025; Chanin & Garriga-Alonso, 2026; Chanin, 2026) motivates a common set-alignment formulation. Geometric accounts study directions and manifolds (Park et al., 2024; 2025; Engels et al., 2025; Sarfati et al., 2026). Shafran et al. (2026) learn Gaussian-region decompositions; we analyze regions induced by SAE gates. In concurrent work, Bhalla et al. (2026) study manifold reconstruction by compact decoder groups, as well as tiling by features whose active subsets vary across samples. We instead analyze the target coverage and non-target exclusion of a fixed activation conjunction; reconstruction alone does not establish these membership properties.

Generator cones and membership intersections differ. Rocchi et al. (2025) use nonnegative spans of dictionary directions and evaluate directional and activation agreement. Our units intersect encoder membership regions. For affine-threshold gates, their non-strict homogeneous ambient counterparts are positive duals of cones generated by encoder normals, which need not equal decoder atoms (Rockafellar, 1970). With biases, these gates yield affine polyhedra; relative gates need not yield halfspaces. This distinction motivates direct coverage, separation, and approximation criteria.

Formal Concept Analysis and Lattice Representations. These many-to-many associations motivate formal concept analysis (FCA), whose extent–intent construction (Ganter et al., 1999) has been applied to neural responses and language models (Endres & Földiák, 2008; Hirth & Hanika, 2022; Xiong & Staab, 2025). Concurrent work by Xiong (2026) derives lattices from thresholded attribute directions, overlapping with our affine-gating case. We adopt the same established algebra for SAE activation predicates, including relative gates. That work studies semantic hierarchy and concept composition; we study SAE membership alignment, its geometric feasibility, approximation error, and neuron budgets.

Network Capacity and Hyperplane Arrangements. This geometric view also connects to capacity analysis: hyperplane arrangements relate activation patterns to expressivity (Montúfar et al., 2014; Pascanu et al., 2013; Stanley et al., 2007; Itskov et al., 2020), including Top-K expert selection (Su et al., 2026). Our positive-support units correspond to preimages of the neural-code trunks studied by Jeffs (2020), providing a combinatorial basis for neuron budgets. Classical polytope approximation results (Gordon et al., 1994; Ludwig, 1999) supply the smooth-convex approximation rate; we translate facet counts into neuron budgets for ambient concept approximation.

A.3 Experiment Details

Synthetic data. For each cluster, we sample 10,000 points with uniform density. The disjoint-concept dataset contains 8 clusters, and the overlapping-concept dataset contains 12 clusters. In the overlapping dataset, the density in an overlap region is kept the same as the density in a non-overlap region. This mimics realistic concept densities: for example, the density of "red car" should not be double-counted simply because a point belongs to both "red" and "car." We approximate the data-supported measure μ by the empirical measure on observed data. For concept approximation, we additionally sample 10,000 probe points from the blank region inside the convex hull of all observed data, i.e., $\mathrm{Conv}(X) \setminus X$, and include these points when empirically estimating the ambient/probe measure ν. For ReLU SAEs, we use expansion factors $\{1, 2, 4, 8, 16, 32, 64\}$ and L_1 regularization coefficients 10^{-5}, $5{\times}10^{-5}$, 10^{-4}, $5{\times}10^{-4}$, 10^{-3}, $5{\times}10^{-3}$, 10^{-2}, 5×10^{-2}, 10^{-1}, 5×10^{-1}, 1, 3, 5, 7, and 10. For Top-K SAEs, K must not exceed the SAE width. Since the input dimension is two, the SAE width is $2 \cdot \mathrm{EF}$. We therefore use $K \in \{1, \ldots, 2 \cdot \mathrm{EF}\}$ for $\mathrm{EF} \in \{1, 2, 4, 8, 16\}$; $K \in \{1, \ldots, 35, 40, 45, 50, 55, 60, 64\}$ for $\mathrm{EF} = 32$; and $K \in \{1, \ldots, 35, 40, 45, 50, 55, 60, 64, 80, 100, 128\}$ for $\mathrm{EF} = 64$.

For both architectures, we use learning rates in $\{0.1, 0.01, 0.001, 0.0003\}$ and train for 200 epochs, which is sufficient for convergence in this synthetic setting. Since our theory is constructive, we run 10 random seeds, $\{0, 1, 42, 3407, 8347, 19285, 657306, 482910, 915673, 2746089\}$, and report the best concept-learning metric across seeds. Each run is trained on one NVIDIA A40 GPU and takes approximately 5–20 minutes, depending on the expansion factor.

Realistic data. Imagenette dataset Howard (2019) is a subset of the ImageNet dataset Deng et al. (2009). Imagenette contains 10 classes with around 13k images in each class. We use the 320px version to do embedding. DINOv3 Siméoni et al. (2025) is an image embedding model, which tokenizes the input image to 1 CLS token, 4 register tokens, 196 patch tokens. DINOv3 first downsamples 320px input images to 224px and produces 201 tokens for our SAE training. The Yelp sentence dataset is constructed from the Yelp Open Dataset Yelp (2024), which contains user reviews and their associated 1–5 star ratings. We split the reviews into individual sentences and assign each sentence the star rating of its source review, resulting in five sentiment classes. We train SAEs with expansion factors $\mathrm{EF} \in 2, 4, 8, 16, 32, 64, 128$. For ReLU SAEs, we use L_1 regularization coefficients $10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 0.3, 0.6, 1.0$. For Top-$K$ SAEs, we use $K \in 2, 4, 8, 16, 32, 64, 128$. For Imagenette, we additionally consider $K \in 500, 1000, 2000, 6000, 8000$, and train all SAEs for 50 epochs to ensure convergence. For the Yelp sentence dataset, we additionally use $K \in 500, 1500$ for $\mathrm{EF} = 2$; $K \in 500, 1500, 2000$ for $\mathrm{EF} = 4$; $K \in 500, 1500, 2000, 6000$

for EF $= 8$; and $K \in 500, 1500, 2000, 6000, 15000$ for EF $\in 16, 32, 64, 128$. All SAEs on Yelp are trained for 200 epochs to ensure convergence. The reason that we train denser Top-K SAEs is to match the sparsity with ReLU SAEs in each expansion factor for a more complete comparison between SAE architectures. Each run is trained on one NVIDIA H200 GPU and takes approximately 2-5 hours, depending on the expansion factor.

A.4 Sparse Autoencoder Phenomena

We now formulate and explain several common SAE phenomena in our framework.

Polysemanticity and monosemanticity. Polysemanticity means that a neuron is related to multiple concepts. In terms of $\mathcal{R}$, this means that a neuron is associated with multiple unrelated concepts. Monosemanticity means that this reverse relation is concentrated on one concept, and thus requires f to be injective: different concepts should not be assigned to the same neuron. Therefore, the number of available neurons must be at least the number of concepts when using single neuron. This is made precise in Section 4.3.

Feature splitting. Feature splitting means that a broad neuron in a smaller SAE is split into several more specific neurons in a larger SAE. Formally, if θ is a broad neuron and $\theta_1, \ldots, \theta_r$ are more specific neurons, then feature splitting can be expressed as $\theta \approx \bigcup_{j=1}^{r} \theta_j$ with $\theta_j \cap \theta_l \approx \varnothing$ for $j \neq l$. The approximate disjointness is consistent with sparsity: if the split neurons overlapped heavily, then data in the overlap would activate many neurons at once and violate the sparsity constraint.

Feature absorption. Feature absorption is a failure mode of hierarchical learning. Suppose $C_i \subset C_j$, where C_i is the child concept and C_j is the parent concept. Ideally, data in C_i should also activate the model concept for C_j. Absorption occurs when $\mu(C_i \cap \theta_{C_j}^c) > 0$. That is, the parent feature fails to activate on part of its child concept. This can happen because activating both parent and child features increases the sparsity cost.

Feature family. A feature family consists of several neurons that tend to activate together. Formally, a family $\theta_1, \ldots, \theta_r$ has nontrivial co-activation if $\bigcap_{l=1}^{r} \theta_l \neq \varnothing$. These neurons may represent different aspects of the same concept or nearby concepts in a semantic family.

Hierarchical concepts. Hierarchical concepts correspond to set inclusion. If $C_i \subset C_j$, then C_i is more specific than C_j. For example, "mammal" is contained in "animal". Ideally, the learned model concepts should satisfy $\theta_{C_i} \subset \theta_{C_j}$. However, maintaining this hierarchy can increase the sparsity cost because data in the child concept may need to activate both child and parent neurons.

A.5 A Complete View of Concept Learning and Neuron Interpretation

Although neuron interpretation has not been specified in the previous sections, the theoretical results above still provide insight into it. In particular, using individual neurons or multi-neuron units for concept learning leads to different levels of interpretation quality, such as different degrees of monosemanticity. We now state the link between concept learning and neuron interpretation in an algebraic way.

Define U as the power set of all data, $U = \mathcal{P}(X)$, so that the human concept family $\mathcal{C}$ is a finite subset of U. Define $\mathcal{N} = [d]$ as the set of neurons. Recall that $M \in \mathcal{P}(\mathcal{N})$ is a set of neurons and $\theta_M \subseteq X$ is the activation region of M. For a single neuron $N \in \mathcal{N}$, we write $\theta_N := \theta_{\{N\}}$. Since each θ_M is a subset of X, we also regard it as an element of U.

We define a binary relation $R \subseteq U \times \mathcal{N}$ by

$$CRN \iff C \subseteq \theta_N.$$

Intuitively, CRN means that neuron N is active on the whole data region C. When $C \in \mathcal{C}$, this says that N covers the whole human concept region C. This relation is not assumed to be a function: a concept may be represented by multiple neurons, and a neuron may be related to multiple concepts.

The relation R induces two maps

$$f : U \to \mathcal{P}(\mathcal{N}), \qquad f(C) = \{N \in \mathcal{N} : CRN\},$$
$$g : \mathcal{P}(\mathcal{N}) \to U, \qquad g(M) = \bigcap_{N \in M} \theta_N,$$

with the convention that $g(\emptyset) = X$. The map f sends a data region to the set of neurons that are active on the entire region. Thus, f corresponds to the concept-to-neuron direction. The map g sends a set of neurons to their common activation region. Thus, g corresponds to the neuron-to-region direction used in neuron interpretation.

Both U and $\mathcal{P}(\mathcal{N})$ are partially ordered by set inclusion. We write these posets as $(U, \subseteq)$ and $(\mathcal{P}(\mathcal{N}), \subseteq)$. This order-theoretic formulation is useful because it allows us to study hierarchical structure among concepts and neuron-defined regions. The larger the data region C is, the fewer neurons can be active on all of it; the larger the neuron set M is, the smaller its common activation region becomes. This is captured by the following Galois connection.

Theorem A.1 (Galois Connection Between Human Concepts and Model Neurons)**.** *The maps f and g form a contravariant Galois connection between U and $\mathcal{P}(\mathcal{N})$. That is, for every $C \in U$ and every $M \in \mathcal{P}(\mathcal{N})$,*

$$C \subseteq g(M) \iff M \subseteq f(C).$$

Proof. Fix $C \in U$ and $M \in \mathcal{P}(\mathcal{N})$. By the definition of g,

$$C \subseteq g(M) \iff C \subseteq \bigcap_{N \in M} \theta_N.$$

This holds if and only if $C \subseteq \theta_N$ for every $N \in M$. By the definition of R, this is equivalent to saying that CRN for every $N \in M$. By the definition of $f(C)$, this holds if and only if every $N \in M$ belongs to $f(C)$, or equivalently $M \subseteq f(C)$. Therefore, $C \subseteq g(M) \iff M \subseteq f(C)$. □

Intuitively, the theorem says that a data region C lies inside the common activation region of a neuron set M if and only if every neuron in M is active on the whole region C. When $C \in \mathcal{C}$, we interpret C as a human concept. The word "contravariant" reflects the reversal of inclusion: adding more neurons to M shrinks $g(M)$, while enlarging C can only remove neurons from $f(C)$. Therefore, the concept-learning direction f and the neuron-interpretation direction g are two sides of the same order-theoretic structure.

This Galois connection naturally induces closure operators $g \circ f$ on U and $f \circ g$ on $\mathcal{P}(\mathcal{N})$. For a data region $C \in U$, the region $g(f(C))$ is the smallest neuron-closed region, relative to the current neuron family, that contains C. If $C = g(f(C))$, then C is closed under the concept-to-neuron-to-concept operation and is called a *fixed point*. Similarly, a neuron set M is a fixed point if $M = f(g(M))$. Studying fixed points is useful for two reasons. First, different regions or neuron sets with the same closure collapse to the same canonical object, which makes redundancy visible. Second, fixed points carry lattice operations, giving a clean way to compare broader and narrower model-learned regions.

From fixed points, we define *formal concepts* as follows: a formal concept is a pair (C, M) satisfying $f(C) = M$ and $g(M) = C$. Equip formal concepts with the order

$$(C_1, M_1) \leq (C_2, M_2) \iff C_1 \subseteq C_2.$$

Equivalently, the neuron-side order is reversed: $(C_1, M_1) \leq (C_2, M_2)$ if and only if $M_2 \subseteq M_1$. The meet and join operations are

$$\begin{aligned}(C_1, M_1) \wedge (C_2, M_2) \\ \quad = (C_1 \cap C_2, f(C_1 \cap C_2)), \\ (C_1, M_1) \vee (C_2, M_2) \\ \quad = (g(M_1 \cap M_2), M_1 \cap M_2).\end{aligned}$$

Intuitively, the meet gives a more specific region, $C_1 \cap C_2$, and then collects the neurons active on all of that smaller region. The join keeps the neurons shared by the two descriptions, $M_1 \cap M_2$, and returns the broader closed region described by those shared neurons. The collection of formal concepts with this order is the *concept lattice*. The U-side components C are extents, and the $\mathcal{P}(\mathcal{N})$-side components M are intents. Extents that are not in $\mathcal{C}$ are not necessarily meaningless; rather, they are closed regions that are not named by the chosen human concept family, and may therefore appear uninterpretable under that vocabulary. Here "closed" refers to Galois closure. There are at most 2^d formal concepts, with top $(X, f(X))$ and bottom $(g(\mathcal{N}), \mathcal{N})$, so the finite lattice is complete.

This algebraic framework is an abstraction of the geometric framework. It gives a more general view of SAE phenomena and helps explain why concept learning can be cast as a set alignment problem between human concepts and model-learned regions.

Polysemanticity/monosemanticity. One geometric signature of polysemanticity is that a neuron covers disjoint nonempty $C_1, C_2 \in \mathcal{C}$ regarded as unrelated at the chosen semantic granularity, with $C_1 \subseteq g(\{N\})$ and $C_2 \subseteq g(\{N\})$. If $\{N\}$ is already closed, the corresponding lattice node has extent $g(\{N\})$ and intent $\{N\}$; in general, its intent is $f(g(\{N\}))$. Its extent contains $C_1 \cup C_2$, so this neuron alone does not distinguish the two concepts. Lattice operations inside this fixed context can only recombine regions induced by the available neurons; they cannot create a new boundary inside θ_N unless some neuron already encodes that boundary. At the chosen semantic granularity, ideal monosemantic alignment with a target $C \in \mathcal{C}$ means $g(\{N\}) = C$, up to the chosen notion of approximation. If the existing dictionary contains no regions that distinguish the target concepts, resolving them requires changing the dictionary or training objective. This is what feature splitting attempts to do.

Feature splitting. Enlarging the SAE changes the neuron family and may provide new activation regions; separately trained dictionaries need not be nested. The training objective, through reconstruction under sparsity, may learn new neurons $N_1, N_2, \ldots$ with $\theta_{N_i} \approx C_i$. Once these neurons are present, a coarse region θ_N that previously covered several concepts can be resolved into purer regions. The actual split from coarse to fine is performed by optimization in the larger dictionary. After training, the refined lattice records the result through nodes whose extents approximate the C_i; a human concept C_i itself is an extent only if it is closed. If the old coarse feature remains, or if the refined dictionary contains shared neurons describing it, the broad region can appear above the finer ones when their extents are contained in it. Otherwise, the relation $\theta_N \approx \bigcup_i \theta_{N_i}$ should be read as an approximate geometric relation, not necessarily as an exact join inside a single intersection-based lattice.

Concept Hierarchy. This is natural in the poset. If C_{large} is more general than C_{small}, then $C_{\text{small}} \subseteq C_{\text{large}}$, and by antitonicity the neuron sets are reverse-ordered:

$$f(C_{\text{large}}) \subseteq f(C_{\text{small}}).$$

Thus, moving to a more specific concept can only add neuron constraints, while moving to a more general concept removes them. However, under a sparsity constraint, a standard SAE tends to learn a relatively flat set of features and does not explicitly represent these subconcept/superconcept containments. Capturing such hierarchy may require lattice- or graph-structured SAEs.

Feature family. A feature family can be viewed as a collection of lattice nodes whose extents have nonempty intersection. For a neuron family M, this common extent is $g(M)$, and $(g(M), f(g(M)))$ is the meet of its members' nodes. This captures related features without forcing each member of the family to be represented by a single neuron.

Feature absorption. In the exact lattice, the hierarchy $C_{\text{small}} \subseteq C_{\text{large}}$ implies

$$f(C_{\text{large}}) \subseteq f(C_{\text{small}}).$$

That is, any neuron N that is active on the whole parent concept C_{large} must also be active on the child concept C_{small}. This gives the ideal implication

$$x \in C_{\text{small}} \Longrightarrow N \text{ activates on } x,$$
$$\text{for } N \in f(C_{\text{large}}).$$

Feature absorption is a mismatch between a detector's intended parent label and its actual coverage: a detector intended for C_{large} fails to activate on a non-negligible part of C_{small}. In that case $N \notin f(C_{\text{small}})$ and therefore also $N \notin f(C_{\text{large}})$. The exact Galois implication remains valid; the intended parent label does not establish its premise. This suggests a diagnostic: for neurons assigned a parent label, check the expected child-to-neuron implications against empirical activation data and flag failures of coverage.

A.6 Proofs for Concept Separation

Throughout this subsection, $C \in \mathcal{C}$ is a nonempty compact concept and

$$D_C := X \setminus C = \bigcup_{C' \in \mathcal{C},\, C' \neq C} (C' \setminus C)$$

is the non-target region. We assume $|\mathcal{C}|$ is finite and $X \subset \mathbb{R}^n$ is bounded. All closures are taken in the ambient input space, and $\mathrm{Conv}(\cdot)$ denotes convex hull, with $\mathrm{Conv}(\varnothing) = \varnothing$. The neurons may be freely chosen, and a unit may use any finite, nonempty set of neurons.

A.6.1 Proof of Theorem 4.2

We first recall two standard convexity facts.

Lemma A.2 (Carathéodory's theorem Aliprantis & Border (2006))**.** *In an n-dimensional vector space, every vector in the convex hull of a nonempty set can be written as a convex combination of at most $n+1$ points from that set.*

Corollary A.3. *Aliprantis & Border (2006) The convex hull of a compact subset of a finite-dimensional vector space is compact.*

Proof. For nonempty compact $C \subseteq \mathbb{R}^n$, Carathéodory's theorem writes its convex hull as the image of the compact set $C^{n+1} \times \Delta_n$ under the continuous map $((x_j),(\lambda_j)) \mapsto \sum_{j=1}^{n+1} \lambda_j x_j$, where $\Delta_n = \{\lambda \geq 0 : \sum_j \lambda_j = 1\}$. The image is therefore compact. The empty case is immediate. □

Theorem 4.2. A concept C can be separated from D_C with one neuron if and only if

$$\mathrm{Conv}(C) \cap \overline{\mathrm{Conv}(D_C)} = \varnothing.$$

Proof. If $D_C = \varnothing$, the condition is immediate and a sufficiently large bias makes any nonzero neuron positive on the bounded set C. We therefore assume $D_C \neq \varnothing$.

By Corollary A.3, $K := \mathrm{Conv}(C)$ is compact and convex. Since $D_C \subset X$ and X is bounded, $\mathrm{Conv}(D_C)$ is bounded; hence $L := \overline{\mathrm{Conv}(D_C)}$ is compact by the Heine–Borel theorem. The set L is also convex because the closure of a convex set is convex.

Recall that separation by one neuron means that there are $w \in \mathbb{R}^n \setminus \{0\}$ and $b \in \mathbb{R}$ such that

$$w^\top x + b > 0 \quad (x \in C), \qquad w^\top y + b \leq 0 \quad (y \in D_C). \tag{1}$$

Equivalently, with $\beta := -b$, we require $w^\top x > \beta$ on C and $w^\top y \leq \beta$ on D_C. The usual hyperplane-separation theorem almost gives this result, but here the positive side is strict while the negative side is weak. We therefore give the direct proof.

(⇐) Assume $K \cap L = \varnothing$. Since K and L are nonempty compact sets, the continuous map $(u,v) \mapsto \|u - v\|$ attains its minimum over $K \times L$ at some (u^*, v^*). Let

$$\delta := \|u^* - v^*\|, \qquad w := u^* - v^*.$$

Because $K \cap L = \varnothing$, we have $\delta > 0$ and $w \neq 0$.

For any $u \in K$ and $t \in [0,1]$, convexity gives $u^* + t(u - u^*) \in K$. By minimality,

$$\|u^* + t(u - u^*) - v^*\|^2 \geq \|u^* - v^*\|^2.$$

Expanding, dividing by $t > 0$, and sending $t \to 0^+$ gives

$$w^\top u \geq w^\top u^* \qquad (u \in K). \tag{2}$$

Similarly, varying $v \in L$ while keeping u^* fixed yields

$$w^\top v \leq w^\top v^* \qquad (v \in L). \tag{3}$$

Moreover,

$$w^\top u^* - w^\top v^* = w^\top (u^* - v^*) = \|w\|^2 = \delta^2 > 0. \tag{4}$$

Set $\beta := \frac{1}{2}(w^\top u^* + w^\top v^*)$ and $b := -\beta$. For $x \in C \subseteq K$, ref 2 and 4 imply

$$w^\top x + b \geq w^\top u^* - \beta = \tfrac{1}{2}\delta^2 > 0.$$

For $y \in D_C \subseteq \mathrm{Conv}(D_C) \subseteq L$, equations 3 and 4 imply

$$w^\top y + b \leq w^\top v^* - \beta = -\tfrac{1}{2}\delta^2 < 0.$$

Thus (w, b) satisfies 1.

($\Rightarrow$) Conversely, suppose (w, b) satisfies 1, and set $\beta := -b$. Then $w^\top x > \beta$ for all $x \in C$ and $w^\top y \leq \beta$ for all $y \in D_C$.

By Lemma A.2, each $u \in \mathrm{Conv}(C)$ is a finite convex combination $u = \sum_j \lambda_j x_j$ with $x_j \in C$, $\lambda_j \geq 0$, and $\sum_j \lambda_j = 1$. Therefore

$$w^\top u = \sum_j \lambda_j w^\top x_j \geq \min_j w^\top x_j > \beta,$$

where the final strict inequality is valid because the minimum is over finitely many terms. Hence

$$w^\top u > \beta \qquad (u \in \mathrm{Conv}(C)). \tag{5}$$

The same finite-convex-combination argument gives $w^\top v \leq \beta$ for all $v \in \mathrm{Conv}(D_C)$. By continuity, this extends to

$$w^\top v \leq \beta \qquad (v \in \overline{\mathrm{Conv}(D_C)}). \tag{6}$$

If a point p belonged to both $\mathrm{Conv}(C)$ and $\overline{\mathrm{Conv}(D_C)}$, equation. 5 and 6 would give the contradiction $w^\top p > \beta$ and $w^\top p \leq \beta$. Thus the two sets are disjoint. □

A.6.2 Proof of Corollary 4.3

Corollary 4.3.

1. All concepts in $\mathcal{C}$ can be separated from each other by one neuron per concept if and only if
$$\mathrm{Conv}(C_i) \cap \overline{\mathrm{Conv}(D_i)} = \varnothing \qquad (C_i \in \mathcal{C}),$$
where $D_i = X \setminus C_i$ denotes the non-target region for C_i.
2. If perfect one-neuron separation is possible for all distinct, nonempty concepts in a finite $\mathcal{C}$, then at least $|\mathcal{C}|$ neurons are necessary and $|\mathcal{C}|$ neurons are sufficient.

Proof. The first claim follows by applying Theorem 4.2 to each concept C_i with $D_i = X \setminus C_i$.

For the second claim, sufficiency follows from the first claim: each concept can be assigned one separating neuron. For necessity, suppose a neuron with activation region H^+ perfectly represents two concepts C_i and C_j. Perfect separation gives $H^+ \cap X = C_i$ and also $H^+ \cap X = C_j$, so $C_i = C_j$. Thus distinct concepts cannot share the same one-neuron representation, and at least $|\mathcal{C}|$ neurons are required. □

A.6.3 Proof of Theorem 4.4

We first record a strict point-versus-convex-set separation lemma.

Lemma A.4 (Strict separation of a point)**.** *Let $K \subseteq \mathbb{R}^n$ be nonempty, compact, and convex, and let $y \notin K$. Then there exist $w \neq 0$ and $\beta \in \mathbb{R}$ such that*

$$w^\top y < \beta, \qquad w^\top u > \beta \quad (u \in K).$$

Proof. Since K is compact, the map $u \mapsto \|u - y\|$ attains its minimum at some $u^* \in K$. Because $y \notin K$, $\delta := \|u^* - y\| > 0$. Let $w := u^* - y$. For $u \in K$ and $t \in [0, 1]$, convexity gives $u^* + t(u - u^*) \in K$, and minimality gives

$$\|u^* + t(u - u^*) - y\|^2 \geq \|u^* - y\|^2.$$

Expanding and letting $t \to 0^+$ yields $w^\top (u - u^*) \geq 0$, so

$$w^\top u \geq w^\top u^* = w^\top y + \|w\|^2 \qquad (u \in K).$$

Choosing $\beta := w^\top y + \frac{1}{2}\|w\|^2$ proves the claim. □

Theorem 4.4. A concept C can be separated from D_C by a unit if and only if

$$\mathrm{Conv}(C) \cap \overline{D_C} = \varnothing.$$

Proof. A unit is a finite intersection of neuron activation regions. We use $\widetilde{\theta}$ for the ambient region and $\theta = X \cap \widetilde{\theta}$ for its restriction to the data space. Given neurons $(w_i, b_i)_{i=1}^m$, $m \geq 1$, write

$$\widetilde{\theta} := \bigcap_{i=1}^{m} \{x \in \mathbb{R}^n : w_i^\top x + b_i > 0\}.$$

Separating C from D_C by this unit means

$$C \subseteq \widetilde{\theta}, \qquad D_C \cap \widetilde{\theta} = \varnothing, \tag{7}$$

or equivalently, every point in C activates all selected neurons, while every point in D_C fails to activate at least one selected neuron. If $D_C = \varnothing$, one neuron with sufficiently large bias covers the bounded set $C = X$. Henceforth assume $D_C \neq \varnothing$.

($\Leftarrow$) Assume $\mathrm{Conv}(C) \cap \overline{D_C} = \varnothing$. Let $K := \mathrm{Conv}(C)$, which is nonempty, compact, and convex by Corollary A.3. The set $\overline{D_C}$ is compact because $D_C \subseteq X$ and X is bounded.

For each $y \in \overline{D_C}$, Lemma A.4 gives $w_y \neq 0$ and β_y such that

$$w_y^\top y < \beta_y, \qquad w_y^\top u > \beta_y \quad (u \in K). \tag{8}$$

The open halfspaces $U_y := \{x : w_y^\top x < \beta_y\}$ form an open cover of $\overline{D_C}$. By compactness, choose a finite subcover $U_{y_1}, \ldots, U_{y_m}$. Define neurons by $(w_i, b_i) := (w_{y_i}, -\beta_{y_i})$ and set

$$\widetilde{\theta} = \bigcap_{i=1}^{m} \{x : w_i^\top x + b_i > 0\}.$$

For every $x \in C \subseteq K$, the second inequality in 8 gives $w_i^\top x + b_i > 0$ for all i, so $C \subseteq \widetilde{\theta}$. For every $y \in D_C \subseteq \overline{D_C}$, the finite subcover gives some i with $y \in U_{y_i}$, hence $w_i^\top y + b_i < 0$ and $y \notin \widetilde{\theta}$. Therefore $D_C \cap \widetilde{\theta} = \varnothing$.

($\Rightarrow$) Conversely, suppose a finite unit $\widetilde{\theta}$ satisfies 7. We first show that $\mathrm{Conv}(C) \subseteq \widetilde{\theta}$. For any $u \in \mathrm{Conv}(C)$, write $u = \sum_j \lambda_j x_j$ with $x_j \in C$, $\lambda_j \geq 0$, and $\sum_j \lambda_j = 1$. For each selected neuron i,

$$w_i^\top u + b_i = \sum_j \lambda_j (w_i^\top x_j + b_i) > 0,$$

because every score is positive and the nonnegative coefficients sum to one. Hence $u \in \widetilde{\theta}$.

If $p \in \mathrm{Conv}(C) \cap \overline{D_C}$, then $p \in \widetilde{\theta}$. Since $\widetilde{\theta}$ is open and $p \in \overline{D_C}$, there exists a sequence $y_k \in D_C$ with $y_k \to p$ and eventually $y_k \in \widetilde{\theta}$, contradicting $D_C \cap \widetilde{\theta} = \varnothing$. Thus $\mathrm{Conv}(C) \cap \overline{D_C} = \varnothing$. □

A.6.4 Proof of Corollary 4.5

Corollary 4.5. Assume $\mathcal{C}$ is finite. All concepts in $\mathcal{C}$ can be separated from each other by units if and only if

$$\mathrm{Conv}(C_i) \cap \overline{(C_j \setminus C_i)} = \varnothing \qquad (C_i, C_j \in \mathcal{C},\ C_j \neq C_i).$$

Proof. Fix $C_i \in \mathcal{C}$ and write

$$D_i := X \setminus C_i = \bigcup_{j \neq i} (C_j \setminus C_i).$$

By Theorem 4.4, C_i can be separated from D_i by a unit if and only if

$$\mathrm{Conv}(C_i) \cap \overline{D_i} = \varnothing.$$

Since $\mathcal{C}$ is finite,

$$\overline{D_i} = \overline{\bigcup_{j \neq i} (C_j \setminus C_i)} = \bigcup_{j \neq i} \overline{(C_j \setminus C_i)}.$$

Therefore

$$\begin{aligned}&\mathrm{Conv}(C_i) \cap \overline{D_i} = \varnothing \\ &\iff \\ &\mathrm{Conv}(C_i) \cap \overline{(C_j \setminus C_i)} = \varnothing \quad \text{for all } j \neq i.\end{aligned}$$

Taking the conjunction over all i proves the equivalence. □

Neuron-budget remark. The equivalence above guarantees that each concept can be separated by some finite unit, but the number of halfspaces obtained from the compact-cover proof depends on the geometry of the sets. When $|\mathcal{C}| \geq 2$, a geometry-independent bound of $|\mathcal{C}|(|\mathcal{C}| - 1)$ neurons follows under the stronger pairwise one-neuron condition

$$\mathrm{Conv}(C_i) \cap \overline{\mathrm{Conv}(C_j \setminus C_i)} = \varnothing \qquad (i \neq j).$$

Indeed, Theorem 4.2 then gives one neuron that is positive on C_i and non-positive on $C_j \setminus C_i$ for each ordered pair (i, j). Intersecting the $|\mathcal{C}| - 1$ neurons associated with a fixed i separates C_i from $X \setminus C_i$. Across all concepts this uses at most $|\mathcal{C}|(|\mathcal{C}|-1)$ neurons before any reuse. Under the weaker condition of Corollary 4.5, the correct general statement is finiteness, not a universal $|\mathcal{C}|(|\mathcal{C}| - 1)$ bound.

When $|\mathcal{C}| = 1$, one neuron suffices to cover the bounded set X. These constructions explain why units can be more monosemantic than individual neurons: neurons can be reused across several units, and each unit can combine several halfspaces to exclude different non-target regions.

A.6.5 Separation error and evaluation measures

Definition A.5 (Separation error)**.** For $C, \theta \subseteq X$ measurable, define

$$e_{\mathrm{sep}}(C, \theta) = \mu(C \Delta \theta) = \underbrace{\mu(\theta \setminus C)}_{\text{contamination mass}} + \underbrace{\mu(C \setminus \theta)}_{\text{missing mass}}.$$

The equality follows because the two differences are disjoint. Under Assumption 4.1, $\theta \setminus C = \theta \cap \bigcup_{C' \neq C}(C' \setminus C)$, recovering the union-of-non-targets expression. With $\mu(C) + \mu(\theta) > 0$, the measure-based F1 score satisfies

$$\mathrm{F1}(C, \theta) = \frac{2\mu(C \cap \theta)}{\mu(C) + \mu(\theta)} = 1 - \frac{e_{\mathrm{sep}}(C, \theta)}{\mu(C) + \mu(\theta)}.$$

Thus F1 equals one precisely at zero separation error, but F1 and raw symmetric-difference error need not rank imperfect units identically because the denominator varies with θ.

Zero separation error means agreement up to a μ-null set, not necessarily pointwise equality. For example, let $X = [0, 2]$, $C = [0, 1]$, and let $\mu = \nu$ be uniform on X. Exact ReLU-unit separation fails because $\overline{X \setminus C} = [1, 2]$ intersects C. Nevertheless, the open interval $\widetilde{\theta} = (-1, 1)$ has zero approximation and separation error, missing only the null endpoint 1 on X. This also explains why nonempty compact concepts cannot equal open ambient ReLU units exactly, while zero measure-based error can still occur.

For arbitrary evaluation measures, approximation does not imply detection or separation error control. If $\mu \ll \nu$, then $\nu(C \Delta \widetilde{\theta}) = 0$ implies $\mu(C \Delta \theta) = 0$, because $C \Delta \theta = X \cap (C \Delta \widetilde{\theta})$ and $\mu(X) = 1$. More quantitatively, if $\mu(A) \leq B\nu(A)$ for all Borel A, then $e_{\mathrm{sep}} \leq B e_{\mathrm{app}}$. No such relation between the two measures is otherwise assumed.

The choice of evaluation measure also determines whether the approximation problem is nontrivial. If $\nu(C) = 0$, the empty activation region has zero approximation error. In particular, a purely continuous probe measure gives zero mass to a finite set of target samples. To evaluate both observed examples and blank-region probes, one possible design is $\nu = \alpha\mu + (1 - \alpha)\pi$ with $0 < \alpha \leq 1$ and a probe distribution π. This is an illustrative design, not a claim about the implementation of our experiments. It gives $e_{\mathrm{sep}} \leq e_{\mathrm{app}}/\alpha$. Such empirical mixtures are allowed by Theorem 4.6, but need not satisfy the density assumptions of Proposition 4.7.

A.7 Proofs for Concept Approximation

Approximation is evaluated using the ambient activation region $\widetilde{\theta}_M$, an intersection of finitely many open halfspaces. The encoder parameters and the finite number of selected neurons may vary with the desired accuracy. The approximation condition below holds for every Borel probability measure ν on $\mathbb{R}^n$, including atomic measures; density and smoothness assumptions are needed only for the quantitative rate.

A.7.1 Proof of Theorem 4.6

Theorem 4.6. A concept $C \in \mathcal{C}$ can be arbitrarily well approximated under the approximation error by units if and only if C is convex up to a ν-null set.

Proof. Here "convex up to a ν-null set" means that there exists a convex Borel set K such that $\nu(C\Delta K) = 0$.

($\Leftarrow$) Suppose such a set K exists. If $\nu(K) = 0$, an empty activation region, realized by $x_1 > 1$ and $x_1 < 0$, has zero error. Otherwise, fix $\varepsilon > 0$. By inner regularity of finite Borel measures on $\mathbb{R}^n$, there is a nonempty compact set $F \subseteq K$ with $\nu(K \setminus F) < \varepsilon/2$. Its convex hull $Q := \mathrm{Conv}(F)$ is compact and satisfies $F \subseteq Q \subseteq K$.

We next approximate Q from outside by open polyhedra. Take a closed box B whose interior contains Q. For each integer $j \geq 1$, the set

$$E_j := \{y \in B : \mathrm{dist}(y, Q) \geq 1/j\}$$

is compact and disjoint from Q. Lemma A.4 and a finite subcover give finitely many open halfspaces whose intersection contains Q and excludes E_j. Intersecting with the interior of B gives an open polyhedron O_j such that

$$Q \subseteq O_j \subseteq \{x : \mathrm{dist}(x, Q) < 1/j\}.$$

If E_j is empty, take O_j to be the interior of B. Set $P_j = \bigcap_{\ell=1}^{j} O_\ell$. These open polyhedra decrease to Q, so continuity from above gives $\nu(P_j \setminus Q) \to 0$. Since each P_j is an intersection of finitely many strict affine inequalities, it is an ambient unit, and for all sufficiently large j with $\nu(P_j \setminus Q) < \varepsilon/2$,

$$\nu(C\Delta P_j) \leq \nu(C\Delta K) + \nu(K \setminus Q) + \nu(P_j \setminus Q) < \varepsilon.$$

Thus C can be arbitrarily well approximated.

($\Rightarrow$) Conversely, choose ambient units $\widetilde{\theta}_j$ such that $\nu(C\Delta\widetilde{\theta}_j) < 2^{-j}$. For each integer $q \geq 1$, the union bound gives

$$\nu\left(\bigcup_{j\geq q}(C\Delta\widetilde{\theta}_j)\right) \leq \sum_{j\geq q} 2^{-j} \longrightarrow 0.$$

Thus, outside a ν-null set, membership in $\widetilde{\theta}_j$ eventually agrees with membership in C. Define

$$K := \liminf_{j\to\infty} \widetilde{\theta}_j = \bigcup_{q=1}^{\infty} \bigcap_{j\geq q} \widetilde{\theta}_j.$$

Each intersection on the right is convex, and these intersections increase with q, so their union is convex. The set K is Borel and agrees with C outside the null set above. Hence $\nu(C\Delta K) = 0$. □

A.7.2 Proof of Proposition 4.7

Proposition 4.7. Let $L \subseteq \mathbb{R}^n$ be an r-dimensional affine subspace containing C, where $r \geq 2$, and suppose ν is supported on L with continuous strictly positive density ρ_ν relative to r-dimensional Lebesgue measure. Let $K \subset L$ be a convex body with C^2 boundary and positive Gaussian curvature, all relative to L. There are constants A_K and m_0 such that, for every $m \geq m_0$, some unit with $|M| \leq m$ satisfies

$$e_{\mathrm{app}}(C, \widetilde{\theta}_M) \leq \nu(C\Delta K) + A_K m^{-\frac{2}{r-1}}.$$

In particular, the residual vanishes when $\nu(C\Delta K) = 0$. The constant A_K depends on K, ν, and r.

Proof. Identify L isometrically with $\mathbb{R}^r$. Theorem 3 of Ludwig (1999), equation (1.8), bounds the infimum of weighted symmetric-difference error over polytopes with at most m facets. Choosing a polytope within $m^{-2/(r-1)}$ of that infimum and enlarging the constant gives a constant A_K such that, for all sufficiently large m, there is a bounded, full-dimensional polytope $P_m \subset L$ with at most m facets such that

$$\nu(K \Delta P_m) \le A_K m^{-\frac{2}{r-1}}.$$

For large m this error is smaller than $\nu(K) > 0$, so P_m cannot be lower-dimensional, since any lower-dimensional polytope is ν-null. The relative interior of P_m is defined by at most m strict affine inequalities on L. To extend them, write $L = x_0 + V\mathbb{R}^r$ with $V^\top V = I_r$. An inequality $u^\top y + c > 0$ in the coordinates $x = x_0 + Vy$ extends to $u^\top V^\top (x - x_0) + c > 0$ on $\mathbb{R}^n$. These extensions realize the relative interior as $L \cap \widetilde{\theta}_M$. Since ν is supported on L and the polytope boundary has zero r-dimensional volume, it is ν-null. Therefore

$$\nu(K \Delta \widetilde{\theta}_M) = \nu(K \Delta P_m) \le A_K m^{-\frac{2}{r-1}}.$$

The triangle inequality now gives

$$\nu(C \Delta \widetilde{\theta}_M) \le \nu(C \Delta K) + \nu(K \Delta \widetilde{\theta}_M) \le \nu(C \Delta K) + A_K m^{-\frac{2}{r-1}}.$$

□

For $r = 1$, under the same density assumption relative to the line L, a nondegenerate compact interval $K \subset L$ is represented, up to its ν-null endpoints, by two strict affine inequalities. Thus the analogous upper bound is $e_{\mathrm{app}}(C, \widetilde{\theta}_M) \le \nu(C \Delta K)$ with $|M| \le 2$.

The residual $\nu(C \Delta K)$ is the error of the chosen convex surrogate, not generally the least attainable error. The true irreducible error is

$$e_{\mathrm{irr}} := \inf_{K \text{ convex Borel}} \nu(C \Delta K).$$

Every finite ambient ReLU unit is Borel and convex, so e_{irr} is no larger than the infimum over such units. Conversely, the construction in the proof of Theorem 4.6 approximates any Borel convex K by a finite ambient unit $\widetilde{\theta}$ with $\nu(K \Delta \widetilde{\theta}) < \varepsilon$. Hence

$$\inf_{\text{finite ReLU units } \theta} \nu(C \Delta \widetilde{\theta}) \le \nu(C \Delta K) + \varepsilon.$$

Taking the infimum over K and then letting $\varepsilon \downarrow 0$ proves equality of the two infima. Theorem 4.6 then shows that this common value vanishes exactly when C agrees with a convex Borel set up to a ν-null set. If a minimizing convex set satisfies the smoothness hypotheses above, the rate holds with residual e_{irr}. Without that additional hypothesis, the displayed rate is only asserted for each fixed admissible K. When $K = \mathrm{Conv}(C)$ is admissible, its residual is $\nu(\mathrm{Conv}(C) \setminus C)$, an upper bound on e_{irr} that can be strict.

A.8 Proofs for Concept Learning Capacity

Theorem 4.8. Let $d \ge 1$ be the number of available neurons and let $k_c \ge 1$ be an integer. Suppose distinct concepts have distinct nonempty representing sets $M_C \subseteq [d]$, with $|M_C| \le k_c$. Then

$$|\mathcal{C}| \le \sum_{r=1}^{\min(k_c, d)} \binom{d}{r}.$$

In particular, along any sequence with $k_c/d \to 0$,

$$d \ge (1 - o(1))(k_c! \, |\mathcal{C}|)^{1/k_c}.$$

Proof. The selected assignment $h(C) := M_C$ is injective by assumption.

There are $\sum_{r=1}^{\min(k_c,d)} \binom{d}{r}$ nonempty subsets of $[d]$ with at most k_c elements, proving the exact bound. For $1 \le k_c < d$, use $\binom{d}{r} \le d^r/r!$ and compare successive terms to obtain

$$\begin{aligned} |\mathcal{C}| &\le \sum_{r=1}^{k_c} \binom{d}{r} \le \sum_{r=0}^{k_c} \frac{d^r}{r!} \\ &\le \frac{d^{k_c}}{k_c!} \sum_{j=0}^{k_c} \left(\frac{k_c}{d}\right)^j \le \frac{d^{k_c}}{k_c!(1-k_c/d)}. \end{aligned}$$

Indeed, the term indexed by $r = k_c - j$ divided by $d^{k_c}/k_c!$ is a product of j factors, each at most k_c/d. Rearranging gives

$$d \ge (1 - k_c/d)^{1/k_c} (k_c!\,|\mathcal{C}|)^{1/k_c}.$$

Since $(1 - k_c/d)^{1/k_c} \to 1$ whenever $k_c/d \to 0$, the stated asymptotic bound follows. □

A.9 Additional Discussion on Top-K SAE

For Top-K SAE, write $z_i(x) = \langle w_i, x\rangle + b_i$, and let TopK_k select exactly k indices with the fixed tie-breaking convention from Section 2. The activation region of neuron i is

$$N_i^{\mathrm{topk}} = \{x \in X : z_i(x) > 0 \text{ and } i \in \mathrm{TopK}_k(z(x))\}.$$

Thus $N_i^{\mathrm{topk}} \subseteq H_i^+$. For a set of neurons M,

$$\theta_M^{\mathrm{topk}} = \bigcap_{j \in M} N_j^{\mathrm{topk}}.$$

If $|M| > k$, then $\theta_M^{\mathrm{topk}} = \varnothing$. If $|M| \le k$, then θ_M^{topk} equals the absolute-gating region $\bigcap_{j\in M} H_j^+$ after removing points where at least one neuron in M fails to enter the top-k set. This is the relative-gating effect.

Let

$$P_M := \bigcap_{j \in M} \widetilde{H}_j^+, \qquad Q_M := \widetilde{\theta}_M^{\mathrm{topk}}.$$

These are ambient regions, so the following identity applies both to μ (supported on X) and to ν. Then $Q_M \subseteq P_M$. Define the rank-interference region

$$I_M := P_M \setminus Q_M.$$

For any concept C and any measure η equal to either μ or ν,

$$\eta(C \Delta Q_M) = \eta(C \Delta P_M) + \eta(C \cap I_M) - \eta((P_M \setminus C) \cap I_M).$$

Indeed, $Q_M = P_M \setminus I_M$, so replacing P_M by Q_M adds false negatives on $C \cap I_M$ and removes false positives on $(P_M \setminus C) \cap I_M$.

Therefore, Top-K gating can hurt by removing true target points, but it can also help by removing false positives outside the target. The identity gives $\eta(C\Delta Q_M) \le \eta(C\Delta P_M) + \eta(C \cap I_M)$, so error upper bounds transfer when target rank loss is controlled. Convexity-based necessary conditions do not transfer: Top-K units need not be convex.

Geometry of the Top-K pieces. Fix the index ordering used to resolve ties, and say that j outranks i when $z_j > z_i$, or when $z_j = z_i$ and j has higher priority. The condition that i is selected is that at most $k-1$ indices outrank it. Fixing the set of outranking indices gives an intersection of affine strict or weak comparisons. Taking the union over such sets of size at most $k-1$ therefore gives a finite union of these regions; intersecting with $z_i > 0$ gives $\widetilde{N}_i$. These pieces are measurable and convex individually, but their union need not be convex. For example, in one dimension take $d = 3$, $k = 2$, and scores $z_1(x) = 1$, $z_2(x) = x + 2$, $z_3(x) = -x + 2$. Neuron 1 is active at $x = -2$ and $x = 2$, but inactive at $x = 0$, regardless of the tie convention. Thus its ambient activation region is not convex. Comparisons with the absolute-gating regions above keep the same scores fixed; they do not identify the encoders learned by separately trained ReLU and Top-K SAEs.

A.10 Additional Experiments

A.10.1 Synthetic Data

We present additional results on the two synthetic datasets: the disjoint-concept dataset in Fig. 7 and the partially overlapping dataset in Fig. 8. On the partially overlapping dataset, we evaluate concept approximation.

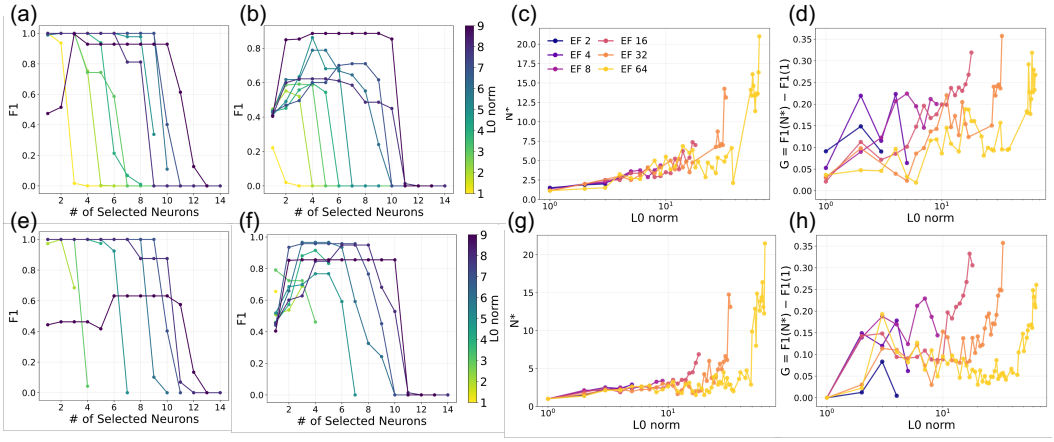


Figure 7: Concept learning on the synthetic disjoint-concept dataset. **(a,b)** ReLU F1 curves for an easy and a hard concept at EF$= 8$. **(c,d)** Mean N^* and G across concepts as L_0 and EF vary. **(e–h)** Corresponding Top-K results.

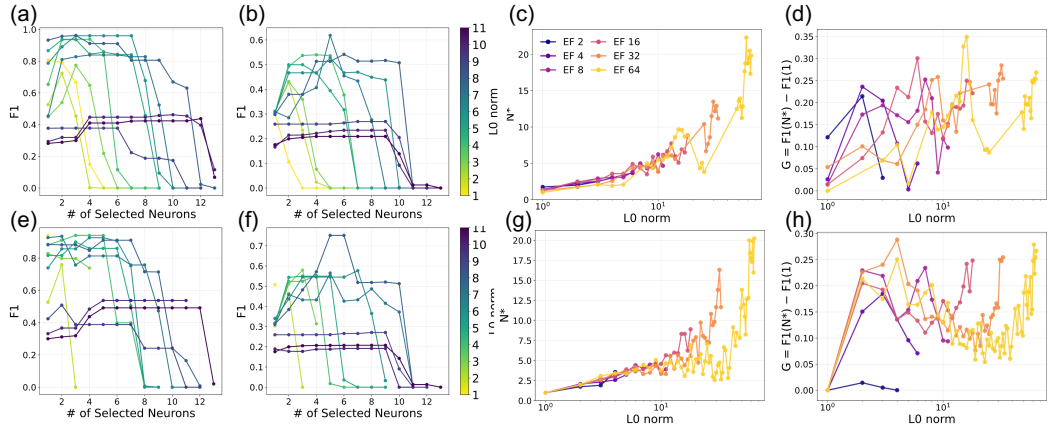


Figure 8: Concept learning on the synthetic partially overlapping dataset. **(a,b)** ReLU F1 curves for an easy and a hard concept at EF$= 8$. **(c,d)** Mean N^* and G across concepts as L_0 and EF vary. **(e–h)** Corresponding Top-K results.

The blue region is the part of the convex hull of all concepts that lies outside the seen concept regions. We use a soft boundary: points within the 95th-percentile distance to a concept cluster are treated as seen. Panels (a) and (e) show an easy concept that overlaps with only one other concept. For the ReLU SAE at $L_0 = 6$, F1 increases from 0.710 with a single neuron to 0.9281 with a 4-neuron unit; for the Top-K SAE at $L_0 = 4$, it increases from 0.7815 to 0.9089 with a 5-neuron unit. Panels (b) and (f) show a harder-to-approximate concept that overlaps with three other concepts. For the ReLU SAE at $L_0 = 7$, F1 increases from 0.2829 with a single neuron to 0.6261 with a 5-neuron unit. For the Top-K SAE, the best multi-neuron unit achieves a substantially higher F1 than any single neuron, but it still falls short of the F1 achieved on the easy concept. These results show that a single neuron is generally insufficient for concept approximation, and that multi-neuron units improve approximation but do not eliminate the difficulty caused by heavy overlap.

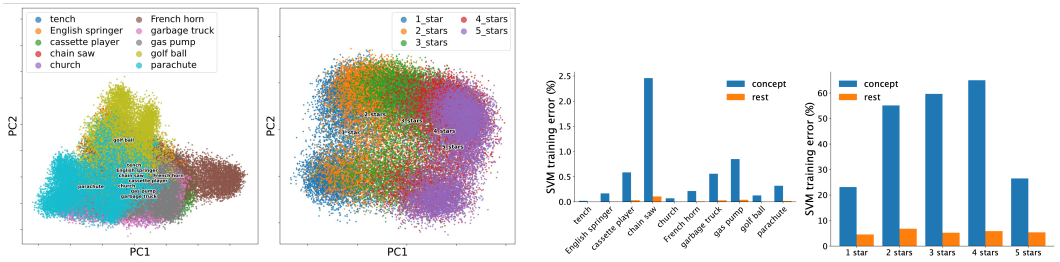


Figure 10: Visualization and linear separability of the real-world datasets. Left two panels: 2D PCA projections of Imagenette and Yelp. Right two panels: one-vs-rest linear SVM training error on Imagenette and Yelp.

The concept learning capacity results in panels (c,d,g,h) of Fig. 7 and Fig. 8 are similar to those on the real-world datasets, which we analyze next.

To illustrate how SAEs exclude unseen data in the partially overlapping dataset, we provide a visualization in Fig. 9.

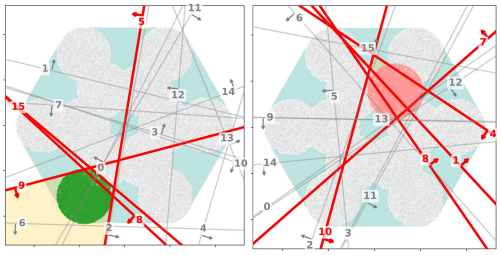


Figure 9: Visualization of case studies on how SAEs approximate concepts in the partially overlapping dataset. The easy concept lies in the outer circle and the hard concept in the inner circle, matching the concepts in Fig. 8. Light blue points denote unseen data.

A.10.2 Real-World Data

The left two panels of Fig. 10 visualize the real-world datasets via 2D PCA projections. We test whether the convex hulls of two concepts' activations intersect. Since two finite point sets have disjoint convex hulls if and only if some hyperplane strictly separates them, this is equivalent to testing whether the two concepts are linearly separable. For every pair of concepts, we fit a linear SVM with a large regularization constant C, which approximates a hard-margin SVM. Before fitting, we center the activations of each pair at the midpoint of the two class means and divide them by the pair's median inter-class distance. This similarity transform preserves separability and relative margins, and it makes C scale-free. We start from $C = 10^4$ and multiply C by 100 as long as training errors remain and their number keeps changing, for at most four fits. A pair with zero training error is declared *disjoint*, and we report its separation $2/\|w\|$. To ensure that this separation does not depend on the choice of C, we solve the hard-margin dual on an active set of near-boundary points and certify the solution with a primal–dual bound. A pair that still has training errors after the final fit is declared *overlapping*, and we summarize the overlap by the fraction of each concept's points that the SVM misclassifies. We repeat the same test for each concept against the union of all other concepts. One-vs-rest separability is strictly stronger than pairwise separability, because it requires a single separating hyperplane rather than an intersection of halfspaces. We report the separability results in the right two panels of Fig. 10.

Tables 1 and 2 quantify the concept learning curves in Fig. 3 and Fig. 4, respectively.

Table 1: Statistical quantification (mean $\pm$ standard deviation) of the phenomena observed in Fig. 3: the behavior of the unit concept learning curves (Fig. 3(a,b,e,f)) and the relationship of N^* and G to the expansion factor (EF) and L_0. Shape trend encodes each curve's shape as 0 (fall), 1 (hump), or 2 (rise).

Quantity	SAE	η^2 L_0	adj η^2 L_0	ρ_s (log L_0)	η^2 EF	adj η^2 EF	ρ_s (log EF)
N*	ReLU	0.46 ± 0.07	0.37 ± 0.09	0.76 ± 0.24	0.22 ± 0.05	0.13 ± 0.06	-0.68 ± 0.35
	Top-K	0.78 ± 0.17	0.75 ± 0.19	0.76 ± 0.12	0.06 ± 0.04	-0.02 ± 0.04	-0.24 ± 0.52
Shape trend	ReLU	0.53 ± 0.08	0.44 ± 0.10	0.74 ± 0.28	0.13 ± 0.04	0.03 ± 0.04	-0.60 ± 0.44
	Top-K	0.78 ± 0.16	0.75 ± 0.19	0.74 ± 0.13	0.04 ± 0.03	-0.04 ± 0.03	-0.18 ± 0.52
G	ReLU	0.27 ± 0.09	0.15 ± 0.10	0.56 ± 0.19	0.06 ± 0.03	-0.05 ± 0.04	-0.47 ± 0.50
	Top-K	0.54 ± 0.12	0.47 ± 0.14	0.75 ± 0.12	0.03 ± 0.02	-0.05 ± 0.02	-0.07 ± 0.53

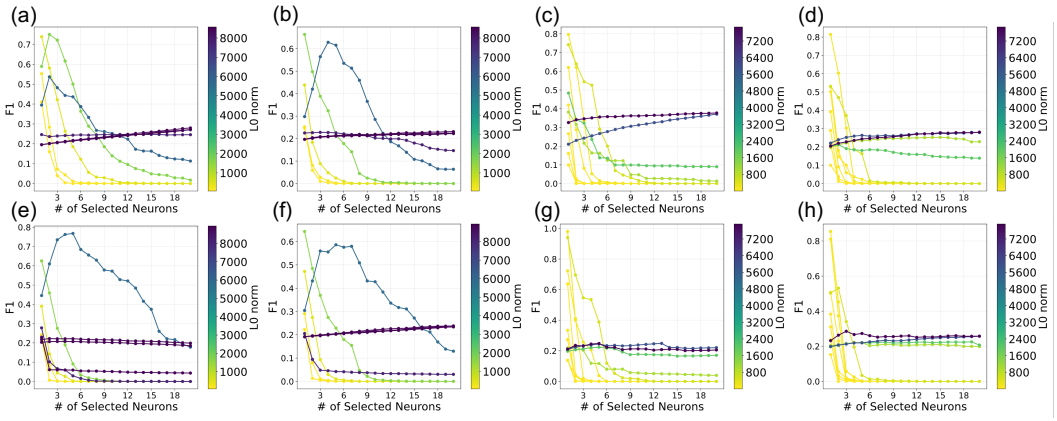


Figure 11: Case study of concept learning quality on chain saw and tench for SAEs trained with or without tench held out. **(a,b)** Concept learning quality of tench and chain saw for ReLU SAEs trained on all classes. **(c,d)** The same for Top-K SAEs. **(e–h)** Same as (a–d), but for SAEs trained with tench held out.

Table 2: Statistical quantification (mean $\pm$ standard deviation) of the phenomena observed in Fig. 4: the behavior of the unit concept learning curves (Fig. 4(a,b,e,f)) and the relationship of N^* and G to the expansion factor (EF) and L_0. Shape trend encodes each curve's shape as 0 (fall), 1 (hump), or 2 (rise).

Quantity	SAE	η^2 L_0	adj η^2 L_0	ρ_s (log L_0)	η^2 EF	adj η^2 EF	ρ_s (log EF)
N*	ReLU	0.38 ± 0.02	0.27 ± 0.03	0.82 ± 0.13	0.21 ± 0.02	0.10 ± 0.02	-0.40 ± 0.70
	Top-K	0.83 ± 0.07	0.81 ± 0.09	0.76 ± 0.13	0.03 ± 0.01	-0.05 ± 0.02	-0.28 ± 0.46
Shape trend	ReLU	0.44 ± 0.08	0.34 ± 0.10	0.69 ± 0.33	0.12 ± 0.04	-0.01 ± 0.04	-0.45 ± 0.70
	Top-K	0.79 ± 0.05	0.76 ± 0.06	0.71 ± 0.16	0.04 ± 0.02	-0.05 ± 0.02	-0.23 ± 0.46
G	ReLU	0.48 ± 0.08	0.39 ± 0.09	0.70 ± 0.14	0.14 ± 0.04	0.02 ± 0.05	-0.25 ± 0.65
	Top-K	0.75 ± 0.06	0.71 ± 0.07	0.74 ± 0.14	0.04 ± 0.02	-0.04 ± 0.02	-0.17 ± 0.52

Fig. 13 provides evidence for the shape trend reported in the main text, where we label the curve shapes fall, hump, and rise as 0, 1, and 2, respectively. The curve shape shifts from fall at high sparsity to a mixture of hump and rise at lower sparsity, where rise is more frequent than hump.

A.10.3 Unseen Concept Learning

In the main text, we hold out the tench class of Imagenette during SAE training to test concept approximation, and we report concept learning quality aggregated over the remaining concepts only. Here, we examine how well SAEs learn tench even though they never see it during training. Fig. 12 shows the results aggregated over all concepts, including tench (cf. Fig. 5, which excludes tench), and Fig. 11 presents a case study of chain saw and tench in this setting. For completeness, Table 3 reports the corresponding statistical quantification.

Comparing the two rows of Fig. 12, we observe that holding out tench has little effect on the overall concept learning capacity and quality. The behavior of Top-K SAEs changes more than that of ReLU SAEs, indicating that ReLU SAEs are less sensitive to changes in the training data. This may be due to the geometry of Top-K gating, which can exclude more data through its rank-interference region (Section A.9). In the case study (Fig. 11), the concept learning quality of chain saw is largely unaffected by the training data, whereas the overall F1 score for tench is lower but remains within a reasonable range. This suggests that SAEs have the potential to learn, or at least detect, unseen concepts, pointing to possible uses in anomaly detection and the discovery of novel scientific concepts: if one concept shows lower concept learning quality while the others remain unaffected, it may be a novel concept.

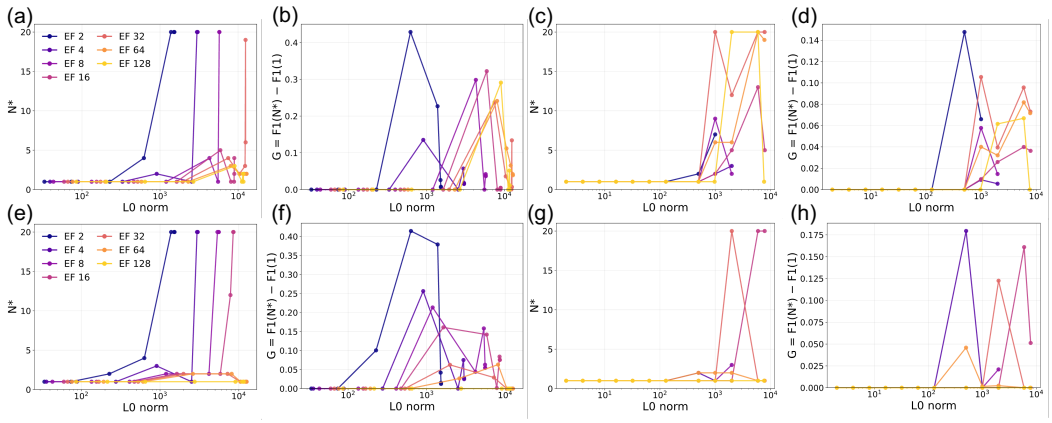


Figure 12: Concept learning quality and capacity over all concepts (including tench) for SAEs trained on Imagenette with or without tench held out; when held out, tench is added back for neuron selection. **(a,b)** N^* and G for ReLU SAEs trained on all classes. **(c,d)** N^* and G for Top-K SAEs trained on all classes. **(e–h)** Same as (a–d), but for SAEs trained with tench held out.

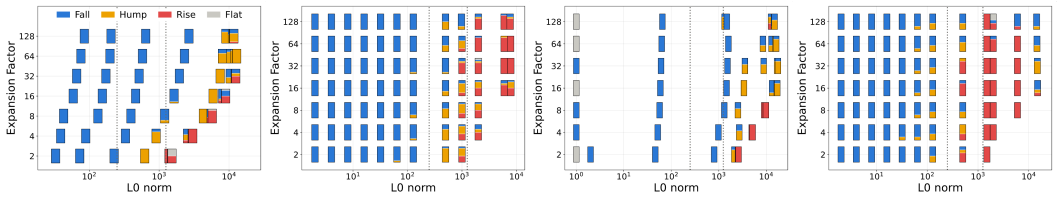


Figure 13: Shape trend analysis. From left to right: ReLU and Top-K SAEs on Imagenette, and ReLU and Top-K SAEs on Yelp.

Table 3: Statistical quantification (mean $\pm$ standard deviation) of the phenomena observed in Fig. 12: the behavior of the unit concept learning curves (Fig. 12(a,b,e,f)) and the relationship of N^* and G to the expansion factor (EF) and L_0. Shape trend encodes each curve's shape as 0 (fall), 1 (hump), or 2 (rise).

Quantity	SAE	η^2 L_0	adj η^2 L_0	ρ_s (log L_0)	η^2 EF	adj η^2 EF	ρ_s (log EF)
N*	ReLU	0.41 ± 0.08	0.30 ± 0.09	0.75 ± 0.26	0.19 ± 0.05	0.10 ± 0.06	-0.70 ± 0.34
	Top-K	0.71 ± 0.28	0.66 ± 0.33	0.75 ± 0.16	0.06 ± 0.03	-0.02 ± 0.04	-0.27 ± 0.48
Shape trend	ReLU	0.51 ± 0.10	0.42 ± 0.12	0.73 ± 0.30	0.13 ± 0.08	0.03 ± 0.09	-0.65 ± 0.42
	Top-K	0.72 ± 0.27	0.68 ± 0.32	0.72 ± 0.17	0.04 ± 0.02	-0.04 ± 0.02	-0.31 ± 0.46
G	ReLU	0.33 ± 0.11	0.21 ± 0.13	0.55 ± 0.21	0.08 ± 0.05	-0.02 ± 0.06	-0.42 ± 0.51
	Top-K	0.48 ± 0.16	0.40 ± 0.18	0.73 ± 0.14	0.04 ± 0.02	-0.04 ± 0.02	-0.03 ± 0.59