A TESTABLE THEORY OF ATOMIC FEATURES

Anonymous authors
Paper under double-blind review

ABSTRACT

We develop and test a theory of language model representations in which there exist "atomic" features. Our main theoretical insight is that in such a model, sparse dictionaries (e.g., SAEs) of increasing size recover an increasing prefix of atomic features most prevalent in the training distribution. This "recovery principle" yields three testable predictions: many features in small SAEs are shared by all larger SAEs, SAEs trained on different distributions share features prevalent in both, and sufficiently large SAEs recover both parent and child features. In contrast to conventional wisdom that SAEs features are unstable and "split" as size increases, we find that these predictions hold on SAEs of sizes ranging from 512 to 131,072 trained on two large embedding models. From a theoretical perspective, our results suggest the promise of a scientific theory of representations based on atomic features. Practically, our results suggest the promise of scaling SAEs.

1 INTRODUCTION

Language models internally represent text inputs $x \in X$ as vectors $f(x) \in \mathbb{R}^d$. A natural goal is to better understand this representation. One prominent hypothesis, the linear representation hypothesis (LRH), is that $f(x)$ is a linear combination of individual features: that is, $f(x) = Az(x)$ for $A \in \mathbb{R}^{d \times M}$ and $z(x) \in \mathbb{R}^M_+$ is a "feature representation." We can imagine that $z_1(x)$ is how much x discusses cats, $z_2(x)$ is how much x discusses combinatorics, and so on.

Given the prominence of the LRH, we might wonder if it can serve as the foundation for a broader scientific theory. In this paper, we take a formal statement of the LRH as a postulate, and mathematically derive—and then empirically test—consequences. A first striking implication of the LRH is that the decomposition into A, z is unique, as we show in Theorem 1. That is, there exist *atomic* features. Therefore, we might call a scientific theory based on the LRH the *atomic theory*.

Of course, this does not prove the existence of atomic features. On the contrary, one might take the strength of Theorem 1's conclusion as evidence against the linear representation hypothesis. Indeed, there is much reason to be skeptical of the existence of atomic features. Sparse autoencoders (SAEs)—which are designed to recover the features posited by the LRH (Cunningham et al., 2023)—learn different features depending on the setting. Even SAEs in settings that vary only by random seed learn different features (Paulo & Belrose, 2025; Fel et al., 2025a). Moreover, coarse features in small SAEs can "split" into granular features in larger SAEs (Chanin et al., 2026; Bussmann et al., 2025; Templeton et al., 2024). How can we square variation in learned SAE features with the atomic theory?

Theorem 2, which we call the recovery principle, presents a solution. The recovery principle says that under the atomic theory, SAEs of increasing size trained on a data distribution $\mathcal{D}$ recover a growing prefix of atomic features: the atomic features most prevalent in $\mathcal{D}$. Theorem 3 further shows that the recovery principle holds even when features are hierarchically structured. The recovery principle is consistent with observed variation in learned SAE features, while also establishing a substantive connection between these observable SAE features and atomic features (see Figure 1).

Beyond reconciling existing empirical observations with the atomic theory, the recovery principle plays another major role in the theory: it implies empirically testable predictions.

- **(P1)** Many features in smaller SAEs have matches in all larger SAEs trained on the same data (corresponding to the most prevalent atomic features in the distribution).
- **(P2)** SAEs trained on different data have many matching features (corresponding to the atomic features that are prevalent in *both* distributions).

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.

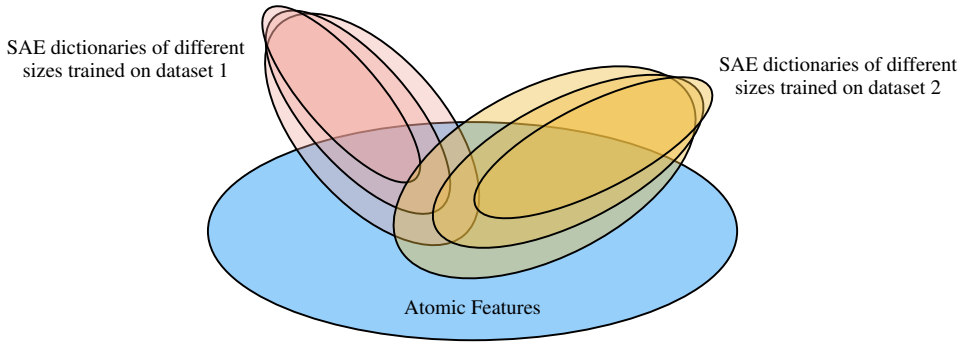


Figure 1: The recovery principle says that, under the atomic hypothesis, SAEs of increasing size recover a growing prefix of atomic features. This gives testable predictions.

(P3) There exist hierarchies where sufficiently large SAEs recover both parent *and* child features.

These predictions are consistent with existing results but run against the conventional wisdom that SAE features are unstable and become strictly more granular as SAE size increases. P3 suggests a particularly surprising phenomenon: parent features that are lost as SAE size increases (as in feature splitting) should eventually return as we scale even further. We call this *mesoscale* feature splitting.

We test P1, P2, and P3 by training SAEs of sizes ranging in powers of two from 512 to 131,072 on two large embedding models: gemini-embedding-2 and llama-embed-nemotron-8b. Testing P1, we find that about 70% of features recovered by size-4K SAEs simultaneously have matches in SAEs of size 8K, 16K, 32K, 65K, and 131K. Testing P2, we find that 26-59% of features are shared across pairs of SAEs trained on five different distributions, compared to mostly 0% and at most 15% for the PCA baseline. Testing P3, we find that on three hierarchies (prefixes, geography, and species), recovery at every level is maximized at size 65K or 131K. In fact, we find instances of the predicted mesoscale splitting, where parent recovery initially decreases, but increases again at larger widths.

These results suggest promise of the atomic theory as a scientific theory of language model representations. A practical implication is that by scaling SAEs, we should expect to recover more and more stable atomic features, including both coarse features and granular features.

2 Overview of the Framework: Theory to Experiments

We begin by presenting the conceptual ideas of the atomic theory, taking care to show how the mathematical formalism yields testable predictions. We start by motivating our formalization of the linear representation hypothesis that serves as the foundation of the theory.

We study language model representations, which we formalize as functions $f : X \to \mathbb{R}^d$. X is an input space (e.g., texts) and $f(x)$ is a representation of the input (e.g., neuron activations or an embedding). The main idea of the linear representation hypothesis is to write

$$f(x) = \sum_{i=1}^{M} z_i(x) a_i, \tag{1}$$

where we call $z_1, z_2, \cdots, z_M : X \to \mathbb{R}^M$ feature activations (which we think of as measuring the presence of individual concepts in the input) and $a_1, a_2, \cdots, a_M \in \mathbb{R}^d$ feature directions (directions in the representation space that correspond to each concept). Concretely, we can think of $z_1(x)$ as measuring the extent to which x discusses cats, and a_1 to be the direction in which this information is represented within $f(x)$. Taking $a_1, a_2, \cdots, a_M$ to be the columns of $A \in \mathbb{R}^{d \times M}$ and defining the "feature representation" $z : x \mapsto (z_1(x), z_2(x), \cdots, z_M(x))$, we can simply write $f(x) = Az(x)$.

We now motivate two more natural assumptions. The first is the sparsity of $z(x)$, which is intuitively motivated: while the total number of concepts must be enormous, the number of concepts needed to describe a particular input is small. Sparsity motivates an analogy to natural language where A, z is a

dictionary and the directions a_i are the words.[1] Given a "situation" represented by $f(x)$, the idea is to write it as the combination of a small number of words (say, at most K). Then z_i specifies in what situations to use the direction a_i (it can be viewed, therefore, as providing a "definition"). In the setting here, z_i also specifies "how much" the word should be used. We call M the dictionary size (the total number of features) and K the description length (the maximum number of features needed to describe an input).

To motivate the second assumption, observe that we should expect features to be *accessible* from a representation. If $f(x)$ is a sufficiently expressive representation of the input x, then $z_1(x), z_2(x), \cdots, z_M(x)$ should be accessible from $f(x) = Az(x)$ alone. For example, given $x =$ "Djokovic Overwhelms Nadal for the Wimbledon Title", we should be able to learn a number of properties from $f(x) = Az(x)$: that x discusses tennis, that x mentions Novak Djokovic, that x mentions Rafael Nadal, that x discusses a championship match, that x describes an event in England, that x describes a decisive victory, and so on. Formally, we expect there to exist a function $g : \mathbb{R}^d \to \mathbb{R}^M$ such that

$$g(Az(x)) = z(x), \tag{2}$$

i.e., mapping representations to features. However, M greatly exceeds d (embeddings typically have d in the thousands, and there are many more concepts), so $A \in \mathbb{R}^{d \times M}$ is not invertible in general. However, classical results in compressed sensing say that it is possible when $z(x)$ is sparse when A is well-behaved (which we make precise in Section 3). Therefore, the second assumption is that A is well-behaved. [2]

The above discussion motivates our statement of the linear representation hypothesis: that there exists a well-behaved matrix $A \in \mathbb{R}^{d \times M}$ and a K-sparse function $z : X \to \mathbb{R}^M_+$ such that for all $x \in X$,

$$f(x) = Az(x). \tag{3}$$

We formalize this as Postulate 1.

Deriving atomicity and a rigidity property. From the postulate, we can derive the atomic property. Theorem 1 considers a "competing" dictionary (B, u): a well-behaved matrix $B \in \mathbb{R}^{d \times M'}$ and K'-sparse function $u : X \to \mathbb{R}^{M'}_+$ satisfying

$$f(x) = Bu(x) \tag{4}$$

for all $x \in X$. Then the result shows if X is sufficiently rich (exploring each combination of features), then if B, u is Pareto-efficient in dictionary size and description length, then either it is equal to A, z or $K' \geq 2K$ (it sometimes uses *twice* as many features to describe an input). Theorem 1 shows that if a sparse system of representation exists, and the space being described is sufficiently rich, then this sparse system is "rigid." To get an intuitive sense of this rigidity, we draw on the analogy of dictionaries in natural language. We can state the following informal analog of Theorem 1. Consider a dictionary of M words and a set of situations such that each situation can be exactly described by K words, and suppose that there exist situations spanning each combination of words. Suppose that you were tasked with coming up with a new dictionary with one fewer total word that was still capable of exactly describing each situation (you are allowed to create new words with new definitions). Theorem 1 says that in this "vocabulary construction game," for every such dictionary, there must exist some situation such that to exactly describe it you now need twice as many words.

From theory to practice. Up to this point, we have established a pretty but (thus far) hypothetical picture of the underlying structure of language model representations. How can we know if this picture is a useful description of reality? A natural approach is to try and find the hypothesized atoms. To do this, we can try to apply sparse dictionary learning: to find a matrix $\hat{A}$ and sparse function $\hat{z}$ such that

$$f(x) \approx \hat{A}\hat{z}(x). \tag{5}$$

But major challenges arise in practice. First, the true number of atoms M is unknown (and likely very large). Second, the learned dictionary is shaped by the specific training data. Finally, given a learned dictionary, how are we to know that we have recovered the true atoms?

[1] Of course, we are drawing from a long line of work on dictionary learning here and throughout (Olshausen & Field, 1996).

[2] It suffices for A to satisfy the $(\gamma, 2K)$-lower restricted isometry property (LRIP). This property, which we will simply call $(\gamma, 2K)$-stability, occurs with high probability for random matrices as long as $d = \Omega(\log M)$.

The recovery principle. There is a powerful mathematical consequence of Postulate 1 that helps resolve these issues. Theorem 2 shows that even sparse dictionaries of size $m < M$ that are trained on idiosyncratic (but "well-behaved") distributions $\mathcal{D}$ approximately recover a subset of atoms. In particular, as we increase the size of a sparse dictionary, we recover a growing prefix of atoms comprising the most prevalent atoms in the distribution.[3]

Notice how the recovery principle implies testable implications. First, training large SAEs on a fixed distribution, we should see a growing set of "persistent features"—i.e., features in smaller SAEs that appear in every larger SAE. Second, even if we train SAEs on different data distributions, we should expect to see a set of shared features—the atomic features that are most prevalent in both distributions. The recovery principle and its implications are further illustrations of what seems to be a rigidity property of sparse systems of representation. Returning to the vocabulary construction game, suppose your goal were to construct a limited vocabulary of m words (say, 200) to best describe situations arising in your daily life, using at most K words per situation.[4] Again, you are allowed to invent new words (say, combining concepts that appear especially frequently for you). The recovery principle suggests that you would not pick entirely new, invented words; rather you would adopt many existing words—and in particular, the existing words that are most relevant for your own daily life. In a sense, humanity has been playing the vocabulary game for millenia. Different societies have come up with languages of different lengths to describe different distributions of situations. Quite strikingly, these languages have an enormous amount of overlap in the concepts given words. In linguistics, a "Swadesh list" give concepts that appear in essentially *all* languages, including concepts like I, you, five, short, knee, to drink, to fly, to count, river, sky, red, green, year, and round (among many others).

Recovery of hierarchical features. The recovery principle shown in Theorem 2 makes the assumption that the training distribution is well-behaved in the sense that for every pair of features z_i, z_j, there exist inputs for which z_i is active but z_j is not (and vice-versa). In Theorem 3, we show that the recovery principle continues to hold (with an adjustment) for distributions where activations are hierarchically correlated: that is there exist parent features that are always active when a child feature is active. The result implies that in sufficiently large SAEs, we should expect to simultaneously recover parent features and child features. This prediction is especially interesting when compared to existing empirical observations in which a broad parent feature that is recovered in a smaller SAE, "splits" as SAE size increases. Under our theory, the expected behavior is that as the size of the SAE increases even further, we should expect the parent feature to once again be recovered.

3 Theoretical Results: Deriving Testable Predictions

We now mathematically present the atomic theory, starting from a single postulate, and deriving the testable predictions P1-P3 as consequences.

Postulate 1 (The Linear Representation Hypothesis)**.** *There exists a $(\gamma, 2K)$-stable matrix $A \in \mathbb{R}^{d\times M}$ with unit-norm columns and a K-sparse function $z : X \to \mathbb{R}^M_+$ such that for all $x \in X$,*

$$f(x) = Az(x). \tag{6}$$

Here, $(\gamma, 2K)$-stability means that for all $2K$-sparse vectors $v \in \mathbb{R}^M$, $\|Av\|_2 \geq \gamma\|v\|_2$. In each of our results, it will also be necessary to specify some properties of X. For example, if we considered a set X where $z(x)$ was identical for all $x \in X$, there would not be much to say.

The first thing we show is that when $z(X)$ is sufficiently rich, the decomposition $f = Az$ is unique (or, perhaps more precisely, "rigid"). We say that $z(X)$ is K-rich if it is dense in some open subset of every K-coordinate subspace of $[0, 1]^M$. Intuitively, X explores each combination of K features.

Then for an alternate dictionary B, u, where $B \in \mathbb{R}^{d\times M'}$ is $(\gamma', 2K')$-stable and $u : X \to \mathbb{R}^{M'}_+$ is K'-sparse, we say that it is Pareto-efficient if (M', K') is not Pareto-dominated by (M, K), where lower is better. We write $(B, u) \equiv (A, z)$ if they are equal up to permutation and scaling.

[3]To mix metaphors, the recovery principle suggests that training larger SAEs is something like building larger particle accelerators: by scaling, we expect to be able to recover more and more fundamental particles (here, atoms).

[4]Perhaps, you were confined to tweeting.

Theorem 1 (Atomicness of Features)**.** *Suppose that $z(X)$ is K-rich. Then for a Pareto-efficient competing dictionary (B, u) such that*

$$f(x) = Bu(x) \tag{7}$$

for all $x \in X$, then either $(B, u) \equiv (A, z)$ or $K' \geq 2K$.

Previous results have shown that when $(M', K') = (M, K)$ we must have $(B, u) \equiv (A, z)$ (Aharon et al., 2006; Hillar & Sommer, 2015; Garfinkle & Hillar, 2019). Theorem 1 shows (A, z) is unique in a much broader sense: no alternate dictionary has description length less than K, and any dictionary with fewer than M features has description length at least double.[5]

We now turn to the recovery principle, showing that undersized dictionaries optimized to specific distributions can recover atoms. For $m < M$, we consider dictionaries $\hat{A}, \hat{z}$ that minimize the dictionary learning objective

$$\mathcal{L}_{\mathcal{D}}(B, u) := \mathbb{E}_{x \sim \mathcal{D}} \|f(x) - Bu(x)\|_2^2 \tag{8}$$

over the reference class of $(\gamma, 2K)$-stable matrices in $\mathbb{R}^{d \times m}$ with unit-norm columns and K-sparse functions $X \to \mathbb{R}_+^m$. In fact, in what follows, we let $\hat{A}^{(\mathcal{D},m)}, \hat{z}^{(\mathcal{D},m)}$ be any dictionary that achieves a fixed constant factor C_0 of the optimal loss in this class. The question is if and when $\hat{A}, \hat{z} = \hat{A}^{(\mathcal{D},m)}, \hat{z}^{(\mathcal{D},m)}$ recovers the atomic features and directions. We say that $\hat{A}$ η-recovers a_i if there exists $j \in [m]$ such that $\langle \hat{a}_j, a_i \rangle > \eta$. We say that $\hat{z}$ $(\eta, \mathcal{D})$-recovers z_i if there exists $j \in [m]$ and a threshold τ such that $\mathrm{F1}_{\mathcal{D}}(\mathbf{1}\{z_i > 0\}, \mathbf{1}\{\hat{z}_j > \tau\}) > \eta$.

We show such recovery for well-behaved $\mathcal{D}$. (For example, we cannot expect recovery if features are exactly correlated.) Fix constants $\rho, c_-, c_+ > 0$. We say that $\mathcal{D}$ is regular if $\mathrm{supp}(z(x)) = K$ with probability one and the following two conditions hold. First, for all pairs $i, j \in [M]$, $\Pr[j \notin \mathrm{supp}(z(x)) \,|\, i \in \mathrm{supp}(z(x))] \geq \rho$. This means that feature i activates without feature j some of the time. Second, letting $h_S : (0,1)^K \to \mathbb{R}_+$ denote the density of $z_S(x)$ conditional on $\mathrm{supp}(z(x)) = S$, $c_- \leq h_S(w) \leq c_+$ for all $w \in (0,1)^K$. This says that no region of feature activations occurs too rarely or too often. Given $\mathcal{D}$, let $p_i := \Pr[i \in \mathrm{supp}(z(x))]$ and $\sigma : [M] \to [M]$ sort these the atomic features in monotonically decreasing order of prevalence ($\sigma(1)$ is the most prevalent). Define $T(m) = \sum_{j=m+1}^{M} p_{\sigma(j)}$.

Theorem 2 (The Recovery Principle)**.** *Let $\mathcal{D}$ be regular. Then for every $\eta < 1$, there exists a constant C not depending on m or M such that $\hat{A}^{\mathcal{D},m}$ η-recovers $a_{\sigma(i)}$ and $\hat{z}^{\mathcal{D},m}$ $(\eta, \mathcal{D})$-recovers $z_{\sigma(i)}$ for all $i \in [M]$ satisfying*

$$p_{\sigma(i)} \geq CT(m). \tag{9}$$

The result guarantees recovery of atomic features that appear more frequently than the tail mass beyond the m-th most frequent atom. For example, if the probabilities p_i follow a power-law distribution with coefficient $\alpha > 1$, the result guarantees that at least the $\Omega(m^{1-1/\alpha})$ most prevalent atoms are recovered. In Section G, we show that in practice (under simulated data), a much larger fraction of atoms can be recovered.

We derive predictions P1 and P2 as corollaries. For matrices U and V, we define $\mu_t(U, V)$ to be the size of the largest one-to-one matching between columns of U and V such that matches have cosine similarity at least t. Let $\mu_t(U; V_1, V_2, \cdots, V_s)$ be the size of the largest subset of columns of U that can be matched in this way into each $V_1, V_2, \cdots, V_s$. Then for $t < 1$, there exists a constant C_t such that for $R_{\mathcal{D},t}(m) := \max\{r \in [M] : p_{\sigma(r)} \geq C_t \sum_{j=m+1}^{M} p_{\sigma(j)}\}$ the following results hold.

Corollary 1 (P1: Stability Across Size)**.** *For a regular distribution $\mathcal{D}$ and dictionary sizes $m_1, \ldots, m_s \geq m$,*

$$\mu_t\left(\hat{A}^{(\mathcal{D},m)}; \hat{A}^{(\mathcal{D},m_1)}, \ldots, \hat{A}^{(\mathcal{D},m_s)}\right) \geq R_{\mathcal{D},t}(m).$$

Corollary 2 (P2: Stability Across Data)**.** *For any regular distributions $\mathcal{D}$ and $\mathcal{D}'$ on the same atomic representation,*

$$\mu_t\left(\hat{A}^{(\mathcal{D},m)}; \hat{A}^{(\mathcal{D}',m)}\right) \geq \left|\sigma([R_{\mathcal{D},t}(m)]) \cap \sigma'([R_{\mathcal{D}',t}(m)])\right|.$$

[5] A related rigidity result shows that one cannot achieve a constant fraction of dictionary size without description length blowing up to be on the order of embedding dimension (Sitharam et al., 2014). We state this result as Theorem 4 and give an alternate proof.

Here, $\sigma([R_{\mathcal{D},t}(m)]) \cap \sigma'([R_{\mathcal{D}',t}(m)])$ is the set of atoms in both distributions' guaranteed prefixes.

The hierarchical recovery principle. We now establish a version of the recovery principle for distributions $\mathcal{D}$ with hierarchical feature correlation, violating regularity. We let $M = 3N$ and $K = 2L$, and re-index the atoms as $i, i1, i2$ for $i \in [N]$. Here, the parent feature z_i is always active when the child features z_{i1} or z_{i2} are active. (For simplicity, we consider families of size three, but it is possible to generalize.) We use a notion of hierarchical regular distributions in which parent features satisfy the standard regularity assumption (details in Section B). For a given $\mathcal{D}$ and $i \in [N]$, let $p_i = \Pr[i \in \text{supp}(z(x))]$ and $\sigma : [N] \to [N]$ sort the the atomic parent features in monotonically decreasing order of prevalence (so $\sigma(1)$ is the most prevalent parent feature). Define $T(n) = \sum_{j=n+1}^{N} p_{\sigma(j)}$.

Theorem 3 (The Hierarchical Recovery Principle)**.** *Let $\mathcal{D}$ be hierarchical regular. Then for every $\eta < 1$, there exists a constant C not depending on m or M such that $\hat{z}^{\mathcal{D},m}$ $(\eta, \mathcal{D})$-recovers the family of features $z_{\sigma(i)}, z_{\sigma(i)1}, z_{\sigma(i)2}$ for all $i \in [N]$ satisfying*

$$p_{\sigma(i)} \geq CT\left(\left\lfloor \frac{m}{3} \right\rfloor\right). \tag{10}$$

Therefore, feature recovery carries over. Direction recovery does not carry over in the strict sense. The intuition is that if $z_j(x) > 0 \implies z_i(x) > 0$, then for all $\lambda \geq 0$ we can re-write $z_i(x)a_i + z_j(x)a_j = \hat{z}_i(x)\hat{a}_i + \hat{z}_j(x)\hat{a}_j$ where $\hat{z}_i(x) = z_i(x) + \lambda z_j(x), \hat{a}_i = a_i, \hat{z}_j(x) = z_j(x)$, and $\hat{a}_j = a_j - \lambda a_i$.[6] This preserves sparsity and non-negativity of u. More importantly, both $\hat{z}_i$ and $\hat{z}_j$ activate on exactly the same examples as z_i and z_j.

The prediction of simultaneous recovery of both parent and child features at large SAE sizes seems in tension with empirical results that have demonstrated feature splitting and absorption (Chanin et al., 2026; Bussmann et al., 2025; Templeton et al., 2024). P3 suggests that as SAE size continues to increase, we should expect lost parent features to return. We give a "molecular" model of this phenomoenon in Section H.

4 Experiments: Testing P1, P2, and P3

We now test the predictions P1, P2, and P3 derived from Postulate 1. We train TopK SAEs on two large embedding models, gemini-embedding-2 and llama-embed-nemotron-8b, which give representations of dimension $d = 3072$ and $d = 4096$, respectively. We train SAEs of sizes m in powers of two ranging from 512 to 131,072 with sparsity parameter $k = 32, 32, 32, 32, 64, 64, 64, 128, 128$.[7] We train on embeddings derived from 90M text inputs (the majority of non-synthetic training data used to fine-tune llama-embed-nemotron-8b). This training mix includes 13 datasets. We also train SAEs of size 16,384 on four additional splits of this data, each including six datasets. Two of these splits separate sources that include Wikipedia data from those that do not. The other two are a random partition of 12 datasets. Details of our experimental setup are given in Section C. Extended analyses for P1, P2, and P3 are collected in Sections D to F, respectively. Figure 2 displays our main results.[8]

P1: Persistent features across size. Following Corollary 1, for each trained width $m < 131{,}072$, we measure the maximum number of feature directions in $\hat{A}^{(m)}$ that can be matched to directions in *every* larger SAE dictionary, where we take the cosine similarity threshold $t = 0.7$. Figure 2a shows that at width 4,096, 73% of Gemini features directions and 66% of Nemotron directions can be matched simultaneously into dictionaries of width 8K, 16K, 32K, 64K, *and* 131K. Persistent stability also holds when choosing matches using feature activations: Figure 7 shows that at width 4,096, 60% of Gemini features and 66.4% of Nemotron features have matches in every larger SAE using Pearson correlation at least 0.7. Figure 10 shows that larger SAEs learn a similar number of coarse features as smaller SAEs, while learning a longer tail of more granular features, while Figure 11 shows that

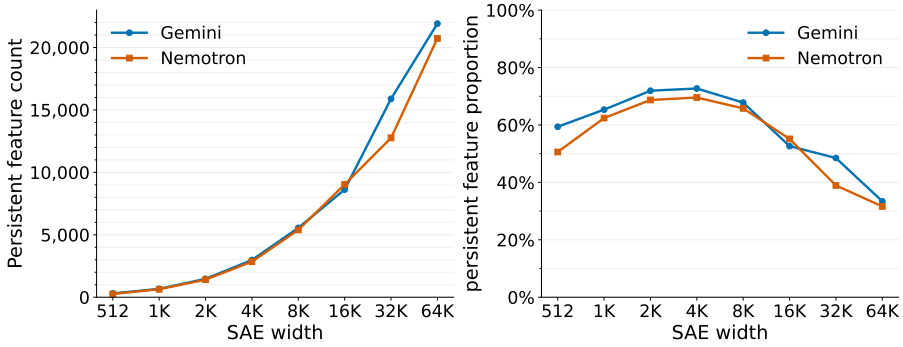
[6] This identity is studied by Chanin et al. (2026) as a potential source of feature absorption; an important distinction in our assumptions is that we allow child feature activations to be larger than the parent, meaning that $\hat{z}_i = z_i(x) + \lambda z_j(x) \geq 0 \implies \lambda \geq 0$.

[7] In Section D, we also analyze results with a fixed sparsity $k = 128$.

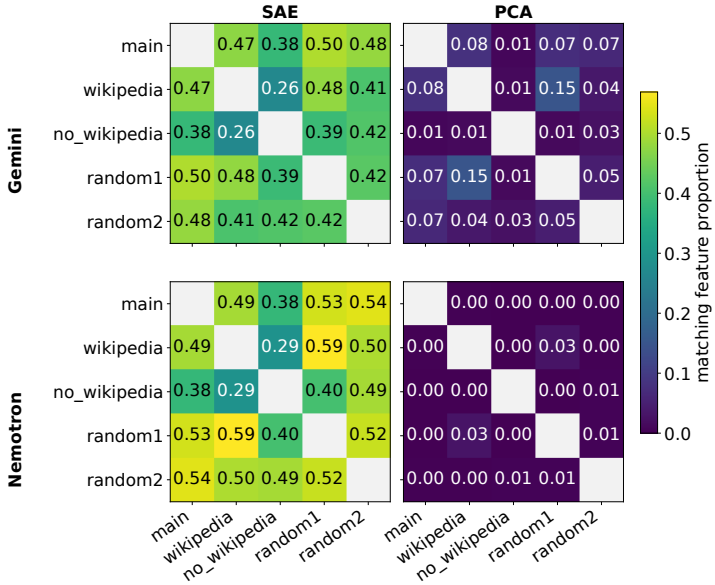
[8] Code used to generate our main results is available at this anonymous repository: https://anonymous.4open.science/r/atomic_features_iclr-B28B/README.md.

Figure 2: Testing P1, P2, and P3—the predictions of the atomic theory.

(a) **P1: Persistence across size.** The maximum number of feature directions at each width that can be simultaneously matched to all larger widths with cosine similarity at least $t = 0.7$. About 70% of features recovered by size-4K SAEs simultaneously have matches in SAEs of size 8K, 16K, 32K, 65K, and 131K.



(b) **P2: Stability over data distributions.** Proportion of features shared between dictionaries trained on different distributions with cosine similarity at least $t = 0.7$. We use absolute cosine similarity for PCA directions. 26%-59% of SAE feature directions have matches, compared to mostly 0% and at most 15% for PCA.



(c) **P3: Hierarchical Recovery.** Mean test F1 by hierarchy level and SAE width. Feature recovery at all levels are highest in the largest dictionaries. **Blue** marks the highest mean F1; **Purple** marks second-highest. N counts category-model pairs for each level.

Domain	Level	N	512	1K	2K	4K	8K	16K	32K	65K	131K
Prefix	1-letter	52	0.864	0.853	0.846	0.833	0.899	0.900	0.888	**0.922**	**0.918**
Prefix	2-letter	432	0.405	0.519	0.604	0.667	0.709	0.719	0.716	**0.727**	**0.719**
Cities	Continent	12	0.642	0.646	0.681	0.676	0.681	0.664	0.669	**0.698**	**0.686**
Cities	Country	174	0.426	0.556	0.657	0.739	0.770	0.778	0.775	**0.779**	**0.779**
GBIF	Kingdom	8	0.543	0.605	0.614	0.626	0.594	0.595	0.581	**0.639**	**0.629**
GBIF	Phylum	24	0.329	0.439	0.516	0.537	0.590	0.616	0.614	**0.673**	**0.637**
GBIF	Class	66	0.234	0.325	0.433	0.505	0.551	0.589	0.617	**0.640**	**0.646**
GBIF	Order	262	0.154	0.212	0.293	0.391	0.446	0.492	0.529	**0.542**	**0.552**
GBIF	Family	552	0.122	0.174	0.235	0.325	0.397	0.448	0.499	**0.523**	**0.538**
GBIF	Genus	86	0.173	0.238	0.299	0.401	0.473	0.543	0.600	**0.621**	**0.653**

more prevalent features are more likely to be persistent. These are consistent with the recovery of a growing prefix of atomic features, where the order is determined by the prevalence of the feature in the distribution. In Section D, we also consider an alternate *exclusive* matching criteria (where a match must be unique), finding similar results. We also show how results vary by the threshold t.

P2: Shared features across training data. We follow Corollary 2 and determine the number of matching directions shared across pairs of size-16K SAEs trained on the five distributions `full`, `wiki`, `no_wiki`, `rand1`, and `rand2`. We also compare baseline PCA dictionaries, consisting of the $d \in \{3072, 4096\}$ principal components fitted to each distribution. For PCA, we use absolute cosine similarity, since the signs of principal components are arbitrary. Figure 2b shows the proportion of features shared across pairs of different dictionaries trained on different datasets for $t = 0.7$. SAE dictionaries have many cross-dataset matches (between 26% and 59%), while PCA dictionaries have substantially fewer (at most 15%, median 1%). The lower matching rates in the PCA dictionaries suggest that the high amount of overlap in SAE dictionaries is not explained just by similarity in the underlying data distributions. Figure 12 shows that SAE dictionaries have many matching directions (25-57%) even under the exclusive matching criteria, whereas the PCA match proportion is at most 7%. Figure 13 also shows that SAEs have many cross-distribution matches based on activation similarity. Figure 14 shows that features with greater prevalence in both distributions are more likely to be selected in direction matching, consistent with the recovery of common prevalent atoms.

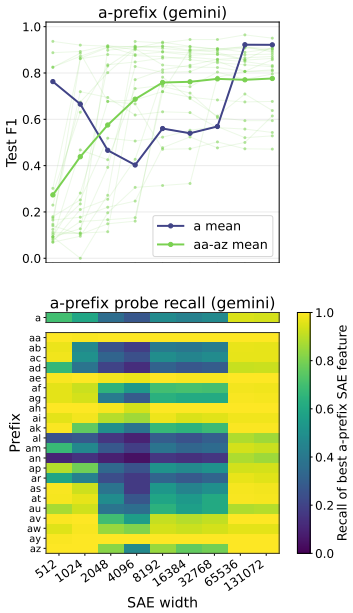
P3: Hierarchical feature recovery. For ground truth binary features z^*_ℓ in three hierarchical datasets (`prefix`, `cities`, `gbif`) and for each SAE size m, we report the test F1 score $\mathrm{F1}_{\mathcal{E}_{\text{test}}}(z^*_\ell, \mathbf{1}\{\hat{z}^{(m)}_j \geq \tau\})$, where $j \in [m]$ and $\tau > 0$ are selected to maximize the same score under $\mathcal{E}_{\text{train}}$. Here $\mathcal{E}_{\text{train}}$ and $\mathcal{E}_{\text{test}}$ are the empirical distributions on the train and held-out test splits of the labeled dataset. Dataset and experiment details are given in Section C.5. Figure 2c shows that across levels of the hierarchy, mean test F1 generally improves with size and is highest at width 65K or 131K. We also find examples of mesoscale feature splitting in our experiments: Figure 3a shows that F1 recovery of the parent 'a' prefix feature decreases as SAE size increases initially (splitting), but increases again at even larger sizes. (Analogs for all 25 remaining Gemini letter prefixes appear in Figure 17.) We also extract hierarchical structure directly (Section F.3), finding 114,762 parent-child pairs in the 131K gemini SAE. Figure 3b shows examples of extracted three-level hierarchies, with top-level features "Food" and "Music." Example lineages include "Food" $\rightarrow$ "Potatoes" $\rightarrow$ "Hash browns / rosti" and "Music" $\rightarrow$ "Jazz" $\rightarrow$ "John Coltrane."

5 Discussion and Additional Results

We developed an "atomic" theory of language model representations, deriving three testable predictions from the LRH and testing these predictions on SAEs trained on two large embedding models. A next step is to test these predictions on other representations (e.g., residual streams). Here, we discuss applications of the atomic theory and connections to related work, alongside additional experiments.

Scaling SAEs and Simulations of the Recovery Principle. The recovery principle suggests the promise of scaling standard SAEs. In particular, large SAEs should be able to simultaneously recover stable coarse and granular features, reducing the need for incorporating hierarchical structure directly (c.f., Muchane et al. (2025); Bussmann et al. (2025); Costa et al. (2026); Zaigrajew et al. (2025); Cao et al. (2026)). In simulations, we study two potential challenges: First, the recovery principle only guarantees a growing number of recovered atoms—in large SAEs, the proportion of features that are atoms may decrease. Second, the recovery principle assumes correct sparsity. In practice, sparsity needs to be tuned (Chanin & Garriga-Alonso, 2025). In Section G, we find that SAEs trained on simulated data can recover a much larger proportion of atomic features than guaranteed by Theorems 2 and 3, and that there is some robustness to the choice of sparsity.

A molecular view of the recovery principle, mesoscale feature splitting, and polysemanticity. In Section H, we suggest a molecular view of the recovery principle: that some feature directions in smaller SAEs can be viewed as molecules, which—as size increases—split into smaller molecules and eventually atoms. Tables 7 and 8 give a molecular explanation of mesoscale feature splitting (a large family molecule splits into parent-child molecules into atoms) that highlights asymmetry between



(a) Mesoscale feature splitting.

Level 1	Level 2	Level 3
Food	Cheese	Cheddar; Mozzarella / related cheeses; Blue cheeses; Gruyère / Emmental; Brie / Camembert
	Potatoes	Potato cultivars; Mashed potatoes; Baked potatoes; Hash browns / rösti; Tater tots
	Fast food	Burger King; Wendy's; Jack in the Box; Hardee's; McDonald's burgers
	Rice dishes	Biryani; Nasi goreng; Fried rice
	Sausages	Chorizo; Salami; German sausages; Hot dogs
	Nutritional information	Fast-food calorie questions; Calorie-breakdown template; Percent Daily Value text; Vitamins / minerals
Music	Jazz	Charlie Parker / bebop; John Coltrane; Miles Davis; Mahavishnu / John McLaughlin; Marsalis family
	Hip-hop / rap	N.W.A; Eminem; Snoop Dogg; Tupac Shakur; Lil Wayne
	Country music	George Jones; Rascal Flatts; Patty Loveless / Vince Gill; Hank Williams
	Rock / metal	Hard-rock musicians / bands; Nu / alternative metal; Power / symphonic metal; Doom / stoner metal
	Electronic / dance	Detroit techno cluster; Trance mixes / compilations; Basement Jaxx; deadmau5
	The Beatles	Please Please Me / early Beatles; Hey Jude; Let It Be / Get Back; Abbey Road / late Beatles

(b) Learned feature hierarchies.

Figure 3: **Hierarchical feature recovery. (a)** Top: parent and child F1 for a-prefixes. Bottom: recall of the selected a-prefix feature, broken down by prefix. **(b)** Extracted Food and Music hierarchies in the Gemini 131K SAE. Activations for each feature given in Tables 5 and 6

directions and activations. Tables 9 and 10 show that molecules can explain polysemanticity: a polysemantic feature in a smaller SAE can be understood as a molecule that splits into monosemantic atoms in larger SAEs. Molecules also explain clusters of learned feature directions observed in practice (e.g., Templeton et al. (2024)): these may be molecules that share an underlying atom.

Multi-dimensional, non-linear, and manifold hypotheses. Some work has noted that representations can exhibit higher-dimensional non-linear structure (Engels et al., 2025; Bhalla et al., 2026; Fel et al., 2026; 2025b). In Section I, we suggest that these observations are consistent with the atomic theory. We show that correlations in feature activations alone (absent any correlation in their associated directions) is enough to produce manifold structure. Figure 23 shows that small sets of features recover manifolds on dates and hex colors.[9]

The rigidity property, platonic representations, and the puzzle of human-interpretability. A growing literature documents similarity in representations learned between different AI models as well as similarity in representations learned by between humans and AI (Huh et al., 2024; Mahner et al., 2026). We suggest a candidate explanation based on the rigidity property illustrated by Theorems 1 to 3: that sparse systems of representation tend to involve the same concepts (recall, again, the vocabulary construction game). If AI models all use sparse representations, then these should be the same.[10] In Appendix Section J, we identify matches between features recovered in SAEs trained on gemini and nemotron. We find, for example, the 131K gemini and nemotron SAEs have 11K matching features at Pearson similarity threshold 0.7 and 27K matching features at similarity 0.5. Example shared features (shown in Tables 12 to 23) include "mechanical bearings," "comments in source code," "trade tariffs and protectionism," "trigonometric half-angle identities," and "astragalus and milkvetch species." Note that the fact that SAE features are *human*-interpretable is not obvious a priori: we could have imagined AI models to learn different, inscrutable representations. Rigidity gives one candidate explanation: that if both human and machine intelligence rely on sparse systems of representation, we should expect shared concepts to emerge.

[9] As a bonus, Table 11 shows that these color features generalize across hex codes, text, images.

[10] Formally, we could imagine that different representations factor as $f = Az$ and $f = A'z$ for different choices of $A \in \mathbb{R}^{d \times M}$ and $A \in \mathbb{R}^{d' \times M}$, where the atomic feature representation $z : X \to \mathbb{R}^M$ is shared.

AI USE STATEMENT

We used AI to prove theorem statements and help draft appendix sections. We reviewed and edited proofs (and re-wrote proofs of central claims for clarity). Lean verifications for theoretical results are given in this anonymous repository: https://anonymous.4open.science/r/atomic-features-formalization-31F0/README.md.

REFERENCES

Michal Aharon, Michael Elad, and Alfred M Bruckstein. On the uniqueness of overcomplete dictionaries, and a practical way to retrieve them. *Linear algebra and its applications*, 416(1): 48–67, 2006.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. MS MARCO: A human generated machine reading comprehension dataset. *CoRR*, abs/1611.09268, 2016. URL http://arxiv.org/abs/1611.09268.

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.

Vera Boteva, Demian Gholipour, Artem Sokolov, and Stefan Riezler. A full-text learning to rank dataset for medical information retrieval. 2016. URL http://www.cl.uni-heidelberg.de/~riezler/publications/papers/ECIR2016.pdf.

Bart Bussmann, Noa Nabeshima, Adam Karvonen, and Neel Nanda. Learning multi-level features with matryoshka sparse autoencoders. *arXiv preprint arXiv:2503.17547*, 2025.

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 and Adrià Garriga-Alonso. Sparse but wrong: Incorrect l0 leads to incorrect features in sparse autoencoders. *arXiv preprint arXiv:2508.16560*, 2025.

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.

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.

Thomas Fel, Ekdeep Singh Lubana, Jacob S Prince, Matthew Kowal, Victor Boutin, Isabel Papadimitriou, Binxu Wang, Martin Wattenberg, Demba Ba, and Talia Konkle. Archetypal sae: Adaptive and stable dictionary learning for concept extraction in large vision models. *arXiv preprint arXiv:2502.12892*, 2025a.

Thomas Fel, Binxu Wang, Michael A Lepori, Matthew Kowal, Andrew Lee, Randall Balestriero, Sonia Joseph, Ekdeep S Lubana, Talia Konkle, Demba Ba, et al. Into the rabbit hull: From task-relevant concepts in dino to minkowski geometry. *arXiv preprint arXiv:2510.08638*, 2025b.

Thomas Fel, Matthew Kowal, Mozes Jacobs, Dron Hazra, Usha Bhalla, Lee Sharkey, Lucius Bushnaq, Satchel Grant, Tal Haklay, Thomas Icard, et al. Structuring sparsity: Block-sparse featurizers capture visual concept manifolds. *arXiv preprint arXiv:2606.25234*, 2026.

Charles J Garfinkle and Christopher J Hillar. On the uniqueness and stability of dictionaries for sparse representation of noisy signals. *IEEE Transactions on Signal Processing*, 67(23):5884–5892, 2019.

Christopher J Hillar and Friedrich T Sommer. When can dictionary learning uniquely recover sparse data from subsamples? *IEEE Transactions on Information Theory*, 61(11):6290–6297, 2015.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. triviaqa: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. *arXiv e-prints*, art. arXiv:1705.03551, 2017.

Daniel Khashabi, Amos Ng, Tushar Khot, Ashish Sabharwal, Hannaneh Hajishirzi, and Chris Callison-Burch. Gooaq: Open question answering with diverse answer types. *arXiv preprint*, 2021.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026/.

Patrick Lewis, Yuxiang Wu, Linqing Liu, Pasquale Minervini, Heinrich Küttler, Aleksandra Piktus, Pontus Stenetorp, and Sebastian Riedel. Paq: 65 million probably-asked questions and what you can do with them. 2021.

Florian P Mahner, Johannes Roth, Ka Chun Lam, Michael F Bonner, Francisco Pereira, and Martin N Hebart. Characterizing universal object representations across vision models. *arXiv preprint arXiv:2605.13675*, 2026.

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

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

Gonçalo Paulo and Nora Belrose. Sparse autoencoders trained on the same data learn different features. *arXiv preprint arXiv:2501.16615*, 2025.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In Jian Su, Kevin Duh, and Xavier Carreras (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL https://aclanthology.org/D16-1264.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. CARER: Contextualized affect representations for emotion recognition. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3687–3697, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1404. URL https://aclanthology.org/D18-1404.

Meera Sitharam, Mohamad Tarifi, and Menghan Wang. An incidence geometry approach to dictionary learning. *arXiv preprint arXiv:1402.7344*, 2014.

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. transformer circuits thread, 2024.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL https://openreview.net/forum?id=wCu6T5xFjeJ.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In *NAACL-HLT*, 2018.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7534–7550. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.609. URL https://aclanthology.org/2020.emnlp-main.609/.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259.

Xiang Yue, Tuney Zheng, Ge Zhang, and Wenhu Chen. Mammoth2: Scaling instructions from the web. *Advances in Neural Information Processing Systems*, 2024.

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

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. MIRACL: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131, 2023. URL https://aclanthology.org/2023.tacl-1.63/.

APPENDIX CONTENTS

A Deriving the Atomic Property **14**

B Proof of the Recovery Principle **17**

C Experimental Setup **23**

C.1 Data and SAE training details 23

C.2 Matching criteria and computation 24

C.3 P1: Persistence across SAE widths 26

C.4 P2: Shared features across training distributions 26

C.5 P3: Hierarchical datasets and feature-recovery protocol 27

D Extended Experiments for P1: Persistence across Size **29**

D.1 Replication with fixed sparsity 29

D.2 Replication with an alternate training seed 29

D.3 Alternative stability comparison 30

D.4 Persistence by activation similarity 31

D.5 Sensitivity to the matching threshold 32

D.6 Feature activation rates 33

D.7 Feature prevalence and direction persistence 34

E Extended Experiments for P2: Stability across Distributions **35**

E.1 Cross-distribution stability under exclusive matching 35

E.2 Cross-distribution stability by activation similarity 36

E.3 Feature prevalence and cross-distribution matching 37

F Extended Experiments for P3: Hierarchical Feature Recovery **38**

F.1 Counts and proportions of recovered categories and adjacent pairs 38

F.2 Recovery of letter-prefix hierarchies 38

F.3 Empirical extraction of feature hierarchies 41

G Simulation experiments **45**

G.1 Recovery under width and sparsity variation 45

G.2 Persistent features across simulated SAE widths 48

H A Molecular Model of SAE Scaling **50**

I An Atomic View of Feature Manifolds **54**

J An Atomic View of the Platonic Representation Hypothesis **57**

A Deriving the Atomic Property

For a matrix $A \in \mathbb{R}^{d\times M}$, define

$$U_K(A) := \bigcup_{\substack{S\subseteq[M]\\|S|=K}} \operatorname{span}(A_S), \tag{11}$$

where $A_S := \{a_i\}_{i\in S}$.

Lemma 3. *Suppose that (A, z) is an (M, K)-featurization of f and z is K-rich. Then if (B, z') is an (M', K')-featurization of f, then*

$$U_K(A) \subseteq U_{K'}(B) \tag{12}$$

Proof. Let $E_S := \{e_i\}_{i\in S}$. By the definition of K-richness, for each $S \subseteq [M]$ with $|S| = K$, there is some non-empty open subset $O_S \subseteq \operatorname{span}(E_S)$ such that $O_S \subseteq \overline{z(X)}$. Then since $\operatorname{spark}(A) > 2K$, A, when restricted to have domain $\operatorname{span}(E_S)$, is injective. Therefore, $A(O_S)$ is an open subset of $\operatorname{span}(A_S)$ that is contained in $\overline{f(X)} = A\overline{z(X)} \subseteq U_{K'}(B)$.

A nonempty open subset of a finite-dimensional subspace cannot be covered by finitely many proper linear subspaces. Therefore, there exists $T \subseteq [M']$ with $|T| = K'$ such that $A(O_S) \subseteq \operatorname{span}(B_T)$. If a nonempty open subset of a finite-dimensional subspace is included in another linear subspace, this implies that the first subspace is a subset of the second. Therefore, $\operatorname{span}(A_S) \subseteq \operatorname{span}(B_T)$. Since this holds for every $S \subseteq [M]$ with $|S| = K$, we have that $U_K(A) \subseteq U_{K'}(B)$, as desired. □

The inclusion (12) is the primary consequence of richness that we invoke.

We first prove that $K' \geq K$. Invoking Lemma 3, we have that $U_K(A) \subseteq U_{K'}(B)$. Consider $S \subseteq [M]$ with $|S| = K$. Then $\operatorname{span}(A_S) \subseteq U_K(A) \subseteq U_{K'}(B)$. Since a finite-dimensional subspace cannot be covered by finitely many proper linear subspaces, there exists some $T \subseteq [M']$ with $|T| = K'$ such that $\operatorname{span}(A_S) \subseteq \operatorname{span}(B_T)$. Since $|S| = K < \operatorname{spark}(A)$, $\operatorname{rank}(A_S) = K$. Therefore, $K' = \operatorname{rank}(B_T) \geq K$.

Next we prove that if $M', K' = M, K$, then $(A, z) \equiv (B, u)$. Suppose that $M' = M$ and $K' = K$. Invoking Lemma 3, we have that $U_K(A) \subseteq U_{K'}(B)$. As in the proof above, for each $S \subseteq [M]$ with $|S| = K$, there must exist some $T \subseteq [M]$ with $|T| = K$ such that $\operatorname{span}(A_S) \subseteq \operatorname{span}(B_T)$. Since both are K-dimensional (since $\operatorname{spark}(A), \operatorname{spark}(B) > 2K$), we must have $\operatorname{span}(A_S) = \operatorname{span}(B_T)$. Define $g : \binom{[M]}{K} \to \binom{[M]}{K}$, $S \mapsto T$. Since $\operatorname{span}(A_S) \neq \operatorname{span}(A_{S'})$ when $S \neq S'$ (again from the spark condition), g is a bijection.

Moreover,

$$|S \cap S'| = \dim(\operatorname{span}(A_S) \cap \operatorname{span}(A_{S'})) = \dim(\operatorname{span}(B_{g(S)}) \cap \operatorname{span}(B_{g(S')})) = |g(S) \cap g(S')|. \tag{13}$$

Now for each $i \in [M]$, we can choose $S, S' \in \binom{[M]}{K}$ such that $S \cap S' = \{i\}$, so $g(S) \cap g(S') = \{j\}$ for some $j \in [M]$. This implies that

$$\operatorname{span}(a_i) = \operatorname{span}(A_S) \cap \operatorname{span}(A_{S'}) = \operatorname{span}(B_{g(S)}) \cap \operatorname{span}(B_{g(S')}) = \operatorname{span}(b_j). \tag{14}$$

So $b_j = \lambda_i a_i$ for some $\lambda_i \neq 0$. Since columns of A are pairwise linearly independent, each i maps to a distinct j and an invertible diagonal matrix D such that $B = APD$. Then for each $x \in X$, $Bz'(x) = Az(x)$. Since $\operatorname{spark}(B) > 2K$, there is a unique K-sparse choice of $z'(x)$. In particular, $z'(x) = D^{-1}P^T z(x)$.

Finally, we show that if $M' < M$, then $K' \geq 2K$.

The proof rests on the following new characterization of when Hall's condition holds: if Hall's condition is satisfied for subsets of size at most $2K$ and subsets of size at most K have neighborhoods of size at most $2K - 1$, then Hall's condition is satisfied globally.

Lemma 4 (Local-to-global Hall lemma)**.** *Let $G = (X, Y, E)$ be a bipartite graph, and let $K \geq 1$. For $S \subseteq X$, let $N(S) := \{y \in Y : (x, y) \in E \text{ for some } x \in S\}$. Suppose that for every $S \subseteq X$,*

(i) if $|S| \leq 2K$, then $|N(S)| \geq |S|$;

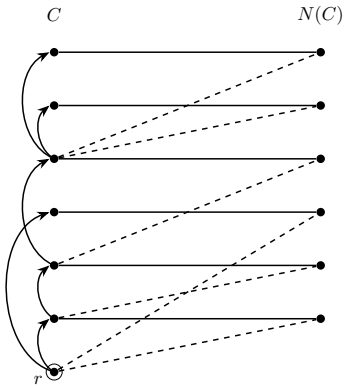


Figure 4: A sketch of how the tree construction completes the proof. Removing one vertex r from C, there exists a matching denoted by the solid lines from $C \setminus \{r\}$ to $N(C)$ (by the "Hall minimality" of C). We can then construct a tree on C where the existence of an edge ij implies that $m(j) \in F_i$ (so $F_i \cap F_j$ is non-empty). Splitting the tree into its two parts, this gives us a subset of C that has a large neighborhood, providing the desired contradiction.

(ii) if $|S| \leq K$*, then* $|N(S)| \leq 2K - 1$.

Then G satisfies Hall's condition: $|N(S)| \geq |S|$ for every $S \subseteq X$. Consequently, there exists a matching that saturates X, and in particular $|Y| \geq |X|$.

Proof of Lemma 4. Assume for sake of contradiction that (i) and (ii) are satisfied, but that there exists $C \subseteq X$ such that $|N(C)| < |C|$. Take C to be inclusion minimal, in the sense that $|N(C')| \geq |C'|$ for all $C' \subsetneq C$. Then $|N(C)| = |C| - 1$, since otherwise we can freely remove any vertex from C to obtain C' violating Hall's condition. By (i), we have that $|C| \geq 2K + 1$. We now show that (ii) is contradicted.

Choose any $r \in C$. By the minimality of C, every strict subset of C satisfies Hall's condition. Therefore, we can construct a perfect matching $m : C \setminus \{r\} \to N(C)$.

We now iteratively construct a tree $T = (C, E_T)$ alongside a bijection $g : E_T \to N(C)$. Start with the set of vertices $R = \{r\}$ and the empty set of edges $E_T = \emptyset$. At each step, choose a vertex $j \in C \setminus R$ such that $m(j) \in N(\{i\})$ for some $i \in R$. Then add j to R, the edge ij to E_T, and set $g(ij) := m(j)$. Such a vertex j exists as long as $R \neq C$. Indeed, suppose that there were not such a vertex. This would mean that for every $u \in N(R)$, $m^{-1}(u) \in R \setminus \{r\}$. Since m is a bijection, this implies that $|N(R)| \leq |R| - 1$. Therefore, if $R \neq C$, this would contradict the minimality of C.

A key property of T is that if $g(ij) = m(j)$, then $m(j) \in N(\{i\})$ and $m(j) \in N(\{j\})$.

Now take any subtree of T with $2K + 1$ vertices (equivalently, $2K$ edges). Let the two sides of the tree (which is bipartite) have vertices S_1 and S_2. Without loss of generality, suppose $|S_1| \leq K$. Every edge ij in the subtree has one endpoint in S_1. Therefore, for each such edge ij, $g(ij)$ is in $N(S_1)$. Since g is a bijection, this implies that $|N(S_1)| \geq 2K$. This contradicts (ii), as desired.

Finishing the proof. We now complete the proof. Invoking Lemma 3, we have that $U_K(A) \subseteq U_{K'}(B)$. Therefore, for all $i \in [M]$, since $a_i \in U_K(A) \subseteq U_{K'}(B)$, a_i has K'-sparse representation in B. Since $\mathrm{spark}(B) > 2K'$, this representation is unique. Let $F_i \subseteq [M']$ denote the support of this representation.

Now consider a bipartite graph $G = ([M], [M'], E)$ where $E = \{(i, j) : i \in [M], j \in F_i\}$. Then we have $N(S) = \cup_{i \in S} F_i \subseteq [M']$.

Our approach is to show that if $K' \leq 2K - 1$, then G satisfies the two conditions in Lemma 4, implying that $M' \geq M$.

To show (i) holds, observe that since $|S| \leq 2K < \text{spark}(A)$, then $\text{rank}(A_S) = |S|$. Also, $\text{span}(A_S) \subseteq \text{span}(B_{\cup_{i\in S} F_i})$. Now observe that $|N(S)| = |\cup_{i\in S} F_i| \geq \text{rank}(B_{\cup_{i\in S} F_i}) \geq \text{rank}(A_S) = |S|$. (Notice that this property holds regardless of K'.)

To show (ii) holds, observe that if $|S| \leq K$, $\text{span}(A_S) \subseteq U_K(A) \subseteq U_{K'}(B)$. Since a subspace can't be covered by the union of a finite number of strictly lower-dimensional subspaces, there exists $T_S \subseteq [M']$ with $|T_S| = K'$ such that $\text{span}(A_S) \subseteq \text{span}(B_{T_S})$. For each $i \in S$, since $a_i \in \text{span}(A_S)$ and F_i is the support of the unique K'-sparse representation of a_i in B, then $F_i \subseteq T_S$. Therefore, $|N(S)| = |\cup_{i\in S} F_i| \leq |T_S| \leq K' \leq 2K - 1$, as desired.

Theorem 4 (Generic rigidity)**.** *Fix d, M, K with $M > d \geq 2$. For almost every $A = (a_1, \ldots, a_M) \in (\mathbb{S}^{d-1})^M$, the following holds. Suppose that (A, z) is a K-rich (M, K)-featurization of $f : X \to \mathbb{R}^d$. Then every (M', K')-featurization (A', z') of f, for $M' > d$, satisfies*

$$K' \geq 1 + (d-1)\frac{M - M'}{M}. \tag{15}$$

Since $d + 1 \geq \text{spark}(A') > 2K'$, (15) implies that $M' > \frac{M}{2}$.

Proof of Theorem 4. Suppose that (A, z) is a K-rich (M, K)-featurization of $f : X \to X \to \mathbb{R}^d$. Then by Lemma 3, we have that $a_i \in U_{K'}(B)$. We show that if

$$K' < 1 + (d-1)\frac{M - M'}{M}, \tag{16}$$

then for almost every $A = (a_1, a_2, \cdots, a_M) \in (\mathbb{S}^{d-1})^M$, there does not exist $B \in \mathbb{R}^{d\times M'}$ such that $a_1, a_2, \cdots, a_M$ all have K'-sparse representations in B. (Notice that this argument does not require that $U_{K'}(B)$ include all of $U_K(A)$, but rather just $a_1, a_2, \cdots, a_M$, so it seems natural that the bound given here could be further improved.)

The idea is a dimension counting argument on parameter spaces. In particular, we first define a smooth manifold that parameterizes A. We then define a smooth manifold that parameterizes (M', K')-featurizations B, z' that can represent $a_1, a_2, \cdots, a_M$. Finally, we define a smooth map between the B-manifold and the A-manifold, with the property that every representable A-featurization must be in the image of this map. Then invoking Sard's theorem, if the dimension of the B-manifold is smaller than the A-manifold, then the image of the map has measure zero. So for a measure one subset of A, no (M, K)-featurization A, z has an (M', K')-featurization.

Let $\mathcal{A}_M = (\mathbb{S}^{d-1})^M$. Let $\mathcal{B}_{M',K',T} = (\mathbb{S}^{d-1})^{M'} \times \prod_{i=1}^{M} S^{|T_i|-1}$. Let $T = (T_1, \cdots, T_M)$ where $T_i \subseteq [M']$ with $1 \leq |T_i| \leq K'$. Define the smooth function $\Phi : \mathcal{B}_{M',K',T} \to (\mathbb{R}^d)^M$ such that

$$(B, \gamma_1, \gamma_2, \cdots, \gamma_M) \mapsto \left(\sum_{j\in T_1} \gamma_{1j} b_j, \sum_{j\in T_2} \gamma_{2j} b_j, \cdots, \sum_{j\in T_M} \gamma_{Mj} b_j \right). \tag{17}$$

Now define $\Theta : (\mathbb{R}^d \setminus \{0\})^M \to \mathcal{A}_M : (a'_1, a'_2, \cdots, a'_M) \mapsto (\frac{a'_1}{\|a'_1\|}, \frac{a'_2}{\|a'_2\|}, \cdots, \frac{a'_M}{\|a'_M\|})$, which is also smooth. Let $\mathcal{B}^*_{M',K',T} = \Phi^{-1}((\mathbb{R}^d \setminus \{0\})^M)$. Since $(\mathbb{R}^d \setminus \{0\})^M$ is open, $\mathcal{B}^*_{M',K',T}$ is itself a smooth manifold. $\mathcal{B}^*_{M',K',T}$ has dimension at most $M'(d-1) + M(K'-1)$. $\mathcal{A}_M$ has dimension $M(d-1)$. So if

$$M'(d-1) + M(K'-1) < M(d-1), \tag{18}$$

by Sard's theorem, $(\Theta \circ \Phi)(\mathcal{B}^*_{M',K',T}) \subseteq \mathcal{A}_M$ has measure zero. Now observe that every $(a_1, a_2, \cdots, a_M)$ that has a K'-sparse decomposition in $(b_1, b_2, \cdots, b_{M'})$ must lie in the image of $(\Theta \circ \Phi)(\mathcal{B}^*_{M',K',T})$ for some T. Since there are a finite number of supports T, and the union of a finite number of measure zero sets has measure zero, this implies that with probability one, $(a_1, a_2, \cdots, a_M)$ does not have a K'-sparse decomposition in any $(b_1, b_2, \cdots, b_{M'})$. Similarly, we can repeat the argument for all M', K' satisfying (18).

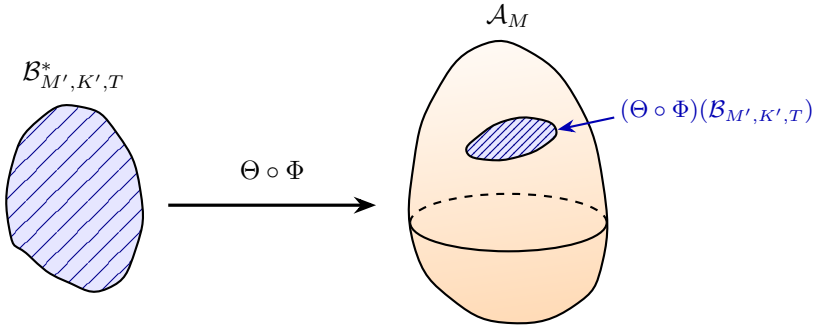


Figure 5: When $\dim(\mathcal{B}_{M',K',S}) < \dim(\mathcal{A}_M)$, the set of (M, K)-factorizations that also have an (M', K')-factorization is measure zero under a generic distribution over (M, K)-factorizations.

Finally, observe that since $d + 1 \geq \text{spark}(A') > 2K'$, then $K' - 1 < \frac{d-1}{2}$, so

$$\frac{d-1}{2} > (d-1)\frac{M - M'}{M} \tag{19}$$

$$\frac{M}{2} > M - M' \tag{20}$$

$$M' > \frac{M}{2}, \tag{21}$$

as desired.

B Proof of the Recovery Principle

We assume throughout that A and $\hat{A}$ have unit-norm columns and are $(\gamma, 2K)$-stable, that $z(x)$ and $\hat{z}(x)$ are nonnegative, and that $\hat{z}(x)$ is at most K-sparse. The atomic support $S(x) := \text{supp}(z(x))$ has size K almost surely. Conditional on any positive-probability support S, the coefficients $z_S(x)$ lie in $(0, 1)^K$ and have density h_S satisfying $c_- \leq h_S \leq c_+$. The densities may differ across supports; only their bounds are shared. All probabilities are with respect to $\mathcal{D}$. We abbreviate $\hat{A}^{(\mathcal{D},m)}, \hat{z}^{(\mathcal{D},m)}$ by $\hat{A}, \hat{z}$ and write

$$E := L_{\mathcal{D}}(\hat{A}, \hat{z}), \qquad E_S := \mathbb{E}_{x\sim\mathcal{D}}\left[\|Az(x) - \hat{A}\hat{z}(x)\|_2^2 \,\middle|\, S(x) = S\right].$$

For equal-dimensional subspaces $U, V \subseteq \mathbb{R}^d$, let $\text{gap}(U, V) := \sup_{u\in U:\, \|u\|_2=1} \text{dist}(u, V)$; the same value is obtained after exchanging U, V. We set $\text{gap}(\{0\}, \{0\}) = 0$. The first three lemmas do not require regularity, so we can use them for both the ordinary and hierarchical settings.

The core idea is the following lemma, which shows that if a full atomic support span is accurately reconstructed by the learned dictionary, then there must be a single learned support whose span matches the atomic span (in a way that is independent of m and M). Then, by intersecting these spans, we can show the recovery of individual atoms. A key challenge in showing the result is that the atomic span can, in theory, be "tiled" by many different learned spans (the number of which depends on m).

Lemma 5. *For every $\epsilon > 0$, there exists $\delta = \delta(\epsilon, K, \gamma, c_-) > 0$ such that any positive-probability support S with $E_S \leq \delta$ has a learned support $T \subseteq [m]$ of size K satisfying* $\text{gap}(\text{span}(A_S), \text{span}(\hat{A}_T)) \leq \epsilon$.

Proof. Fix $\epsilon > 0$. Set $r_0 := \gamma/4$ and $r_{j+1} := \gamma r_j/(16K)$ for $j = 0, \ldots, K-1$. Write $R := r_K$ and choose $\alpha := R\min\{\epsilon, 1/8\}$. All these quantities are independent of m, M.

Define $U_K(\hat{A}) := \bigcup_{T\subseteq[m]:\,|T|\leq K} \text{span}(\hat{A}_T)$ and $g(z) := \text{dist}(A_S z, U_K(\hat{A}))$. The conditional density lower bound gives $E_S \geq c_- \int_{(0,1)^K} g(z)^2\, dz$. Moreover, since distance to a fixed set is 1-Lipschitz and $\|A_S\|_{\text{op}} \leq \sqrt{K}$, we have $|g(z) - g(z')| \leq \sqrt{K}\|z - z'\|_2$.

Suppose $g(z_0) > \alpha$ at some $z_0 \in [1/4, 3/4]^K$. Then $g(z) > \alpha/2$ throughout the coefficient ball of radius $\alpha/(2\sqrt{K})$ centered at z_0. This ball lies inside $(0,1)^K$: unit norms and stability imply $\gamma \le 1$, and our choice of α makes its radius smaller than $1/4$. Writing ω_K for the volume of the unit ball in $\mathbb{R}^K$, choose

$$\delta := \frac{c_- \omega_K}{2^{K+3} K^{K/2}} \alpha^{K+2}.$$

The contribution of this ball would give $E_S \ge c_-(\alpha^2/4)\omega_K(\alpha/(2\sqrt{K}))^K = 2\delta$, a contradiction. Thus $g(z) \le \alpha$ throughout $[1/4, 3/4]^K$.

Write $V := \operatorname{span}(A_S)$ and $y_0 := A_S(1/2, \dots, 1/2)$. Since $\sigma_{\min}(A_S) \ge \gamma$, the image $A_S[1/4, 3/4]^K$ contains the ball $\mathcal{B}_0 := B_V(y_0, r_0)$. Indeed, any displacement in V of norm at most $\gamma/4$ equals $A_S w$ for some $\|w\|_2 \le 1/4$. Here and below, B_V denotes a closed ball taken within V.

For each $y \in \mathcal{B}_0$, choose a best signed K-sparse coefficient vector $v(y)$. Such a vector exists because $U_K(\hat{A})$ is a finite union of subspaces. The preceding bound gives $\|y - \hat{A}v(y)\|_2 \le \alpha$. Since $v(y) - v(y')$ is at most $2K$-sparse, stability and the triangle inequality imply

$$\|v(y) - v(y')\|_2 \le \frac{\|\hat{A}(v(y) - v(y'))\|_2}{\gamma} \le \frac{\|y - y'\|_2 + 2\alpha}{\gamma} \qquad (y, y' \in \mathcal{B}_0). \tag{22}$$

Uniqueness of the best coefficient vector is not needed.

We now construct nested balls $\mathcal{B}_j := B_V(y_j, r_j)$ and sets $T_j \subseteq [m]$ of size j such that $T_j \subseteq \operatorname{supp}(v(y))$ throughout $\mathcal{B}_j$. Initially, take $T_0 = \varnothing$.

Suppose $\mathcal{B}_j, T_j$ have been constructed, with $j < K$. Since $\dim V = K > j$, choose a unit vector in V orthogonal to $\operatorname{span}(\hat{A}_{T_j})$. At least one of the two points obtained by displacing y_j by $\pm r_j/2$ in this direction has distance at least $r_j/2$ from $\operatorname{span}(\hat{A}_{T_j})$. Call this point y_{j+1}. Using the residual bound and the unit norms of the learned columns gives

$$\frac{r_j}{2} \le \operatorname{dist}\big(y_{j+1}, \operatorname{span}(\hat{A}_{T_j})\big) \le \alpha + \sum_{\ell \notin T_j} |v_\ell(y_{j+1})|.$$

Since $\alpha \le r_j/4$ and $v(y_{j+1})$ has at most K nonzero entries, some $i \notin T_j$ satisfies $|v_i(y_{j+1})| \ge r_j/(4K)$.

Set $T_{j+1} := T_j \cup \{i\}$ and $\mathcal{B}_{j+1} := B_V(y_{j+1}, r_{j+1})$. This ball lies inside $\mathcal{B}_j$, since its center is displaced by $r_j/2$ and $r_{j+1} \le r_j/2$. For every $y \in \mathcal{B}_{j+1}$, (22) and $2\alpha \le r_{j+1}$ give $|v_i(y) - v_i(y_{j+1})| \le 2r_{j+1}/\gamma = r_j/(8K)$. Hence $|v_i(y)| \ge r_j/(8K) > 0$ throughout the smaller ball. The coordinates in T_j remain nonzero by the induction hypothesis, completing the construction.

After K steps, $T := T_K$ has size K, so sparsity forces $\operatorname{supp}(v(y)) = T$ throughout $\mathcal{B}_K$. Consequently, $\operatorname{dist}(y, \operatorname{span}(\hat{A}_T)) \le \alpha$ for every $y \in B_V(y_K, R)$.

Finally, let P_T be the orthogonal projection onto $\operatorname{span}(\hat{A}_T)$. For any unit vector $w \in V$, both $y_K + Rw$ and $y_K - Rw$ belong to the final ball. Therefore,

$$2R\|(I - P_T)w\|_2 \le \|(I - P_T)(y_K + Rw)\|_2 + \|(I - P_T)(y_K - Rw)\|_2 \le 2\alpha.$$

Both subspaces have dimension K, so taking the supremum over unit vectors $w \in V$ bounds their gap by $\alpha/R \le \epsilon$. □

The next lemma shows that approximate matching of support spans preserves their intersections. Importantly, its bound does not depend on the number of supports being intersected.

Lemma 6. *Suppose a nonempty finite collection of pairs S_ℓ, T_ℓ, with $|S_\ell| = |T_\ell| = K$, satisfies*

$$\operatorname{gap}(\operatorname{span}(A_{S_\ell}), \operatorname{span}(\hat{A}_{T_\ell})) \le r \qquad \textit{for every } \ell.$$

Write $I := \bigcap_\ell S_\ell$ and $J := \bigcap_\ell T_\ell$. If $r(1 + 2K/\gamma) < 1$, then

$$|I| = |J|, \qquad \operatorname{gap}\big(\operatorname{span}(A_I), \operatorname{span}(\hat{A}_J)\big) \le r(1 + 2K/\gamma).$$

Proof. Take a unit vector $u \in \operatorname{span}(A_I)$. For each ℓ, choose $v_\ell \in \mathbb{R}^m$, supported on T_ℓ, such that $\|u - \hat{A}v_\ell\|_2 \le r$. Fix one pair, indexed by 1. For every $j \in T_1 \setminus J$, some T_ℓ omits j. Stability then gives $|(v_1)_j| \le \|v_1 - v_\ell\|_2 \le 2r/\gamma$. Deleting these at most K coordinates from v_1, and using the unit norms of the learned columns, gives

$$\operatorname{dist}(u, \operatorname{span}(\hat{A}_J)) \le \|u - \hat{A}v_1\|_2 + \sum_{j \in T_1 \setminus J} |(v_1)_j| \le r + 2Kr/\gamma.$$

Exchanging A, S_ℓ with $\hat{A}, T_\ell$ gives the same bound in the reverse direction. If the two intersection spans had different dimensions, the larger would contain a unit vector orthogonal to the smaller, contradicting the bound being less than 1. Their dimensions are $|I|$ and $|J|$ by stability, proving both claims. This argument also covers the case in which an intersection is empty. □

We next connect these geometric matches with the actual learned activations. This step uses both nonnegativity of $\hat{z}$ and the upper density bound. It does not assume that $\hat{z}(x)$ minimizes the reconstruction error for each x.

Lemma 7 (Activation recovery)**.** *For every $\eta \in (0,1)$, there exist $r, c, \tau > 0$, depending only on η, K, γ, c_+, with the following property.*

Let $\mathcal{G}$ be a collection of positive-probability atomic supports. Suppose each $S \in \mathcal{G}$ has a learned support T_S of size K satisfying $\operatorname{gap}(\operatorname{span}(A_S), \operatorname{span}(\hat{A}_{T_S})) \le r$. Suppose also that an atomic coordinate i of prevalence $p_i := \Pr(z_i > 0) > 0$ and a learned coordinate j satisfy $j \in T_S$ if and only if $i \in S$, for every $S \in \mathcal{G}$.

If $E + \Pr(S(x) \notin \mathcal{G}) \le cp_i$, then

$$\mathrm{F1}_{\mathcal{D}}\big(\mathbf{1}\{z_i > 0\}, \mathbf{1}\{\hat{z}_j > \tau\}\big) > \eta.$$

Proof. We first derive a bound for any $0 \le r \le 1/2$ and $\tau > 0$ satisfying $2Kr \le \gamma\tau$. Write $\beta := \Pr(S(x) \notin \mathcal{G})$.

Fix $S \in \mathcal{G}$, and let P_{T_S} denote orthogonal projection onto $\operatorname{span}(\hat{A}_{T_S})$. Let H_S be the $K \times K$ matrix satisfying $\hat{A}_{T_S} H_S = P_{T_S} A_S$. It exists and is unique because $\hat{A}_{T_S}$ has full column rank. For an input with support S, let $v_S(x) \in \mathbb{R}^m$ have coordinates $H_S z_S(x)$ on T_S and zero elsewhere. These auxiliary coefficients may have either sign.

Since the atomic coefficients lie in $(0,1)^K$ and the columns have unit norm, $\|Az(x)\|_2 \le K$. The gap bound therefore gives $\|Az(x) - \hat{A}v_S(x)\|_2 \le Kr$. Writing $e(x) := \|Az(x) - \hat{A}\hat{z}(x)\|_2$, stability and the triangle inequality imply

$$\|\hat{z}(x) - v_S(x)\|_2 \le \frac{\|\hat{A}(\hat{z}(x) - v_S(x))\|_2}{\gamma} \le \frac{e(x) + Kr}{\gamma}. \tag{23}$$

Here stability applies because $\hat{z}(x) - v_S(x)$ is at most $2K$-sparse.

If $j \in T_S$, the coordinate $(v_S(x))_j$ is a linear function of $z_S(x)$ with at least one coefficient of magnitude $\gamma/(2K)$ or larger. To see this, for every $w \in \mathbb{R}^K$,

$$\sqrt{K}\|H_S w\|_2 \ge \|P_{T_S} A_S w\|_2 \ge \sqrt{1 - r^2}\|A_S w\|_2 \ge \frac{\gamma}{2}\|w\|_2.$$

Thus $\sigma_{\min}(H_S) \ge \gamma/(2\sqrt{K})$. Since H_S is square, each row has at least this norm, and hence has an entry of magnitude at least $\gamma/(2K)$.

Holding the other $K - 1$ atomic coefficients fixed, the values of the corresponding coefficient for which $|(v_S(x))_j| \le 2\tau$ form an interval of length at most $8K\tau/\gamma$. Integrating over the remaining coordinates and using $h_S \le c_+$ gives

$$\Pr\big(|(v_S(x))_j| \le 2\tau \,\big|\, S(x) = S\big) \le \frac{8Kc_+\tau}{\gamma}. \tag{24}$$

No independence of the coefficients is needed.

By Markov's inequality, outside an event of probability at most $\beta + 4E/(\gamma^2\tau^2)$, we have $S(x) \in \mathcal{G}$ and $e(x) \le \gamma\tau/2$. On this complement, (23) and $Kr \le \gamma\tau/2$ give $\|\hat{z}(x) - v_S(x)\|_2 \le \tau$, where $S = S(x)$.

If $i \notin S$, then $j \notin T_S$, so $(v_S(x))_j = 0$ and $\hat{z}_j(x) \le \tau$. Thus there is no false positive on this complement. If $i \in S$ but $\hat{z}_j(x) \le \tau$, nonnegativity of $\hat{z}_j(x)$ implies $-\tau \le (v_S(x))_j \le 2\tau$. Summing (24) over supports in $\mathcal{G}$ containing i, the probability of these false negatives is at most $(8Kc_+\tau/\gamma)p_i$.

Let $Y := \mathbf{1}\{z_i > 0\}$ and $\widehat{Y} := \mathbf{1}\{\hat{z}_j > \tau\}$. The total disagreement probability is therefore at most $\beta + 4E/(\gamma^2\tau^2) + (8Kc_+\tau/\gamma)p_i$. Since $\Pr(Y = 1) = p_i > 0$, the definition of the F1 score gives

$$\begin{aligned} \mathrm{F1}_{\mathcal{D}}(Y, \widehat{Y}) &= 1 - \frac{\Pr_{\mathcal{D}}(Y \neq \widehat{Y})}{\Pr_{\mathcal{D}}(Y = 1) + \Pr_{\mathcal{D}}(\widehat{Y} = 1)} \\ &\ge 1 - \frac{\beta}{p_i} - \frac{4E}{\gamma^2\tau^2 p_i} - \frac{8Kc_+\tau}{\gamma}. \end{aligned} \tag{25}$$

Finally, choose $\tau > 0$ sufficiently small that $8Kc_+\tau/\gamma < (1-\eta)/2$, and choose $0 < r \le \min\{1/2, \gamma\tau/(2K)\}$. Choose $c > 0$ sufficiently small that $c(1 + 4/(\gamma^2\tau^2)) < (1-\eta)/2$. These choices depend only on η, K, γ, c_+. If $E + \beta \le cp_i$, then the sum of the first two error terms in (25) is at most $c(1 + 4/(\gamma^2\tau^2))$, so the F1 score exceeds η. □

The ordinary recovery principle. We now impose regularity of the support law: for a fixed $\rho \in (0, 1]$ and distinct atoms i, k with $p_i > 0$, assume $\Pr(k \notin S(x) \mid i \in S(x)) \ge \rho$. Let σ order the atom prevalences nonincreasingly and let $T(m) := \sum_{j=m+1}^{M} p_{\sigma(j)}$. Write $L^*_{\mathcal{D}}(m)$ for the infimum of the dictionary learning loss over the unit-norm, $(\gamma, 2K)$-stable reference class with nonnegative, at most K-sparse codes. We do not need the infimum to be attained.

Proof of Theorem 2. Retain the m most prevalent columns of A, together with their true coefficients. This gives an admissible dictionary. For every input, the squared norm of the omitted sum is at most $K \sum_{k \notin \sigma([m])} z_k(x)^2$, which is at most K times the number of omitted active atoms. Therefore,

$$E \le C_0 L^*_{\mathcal{D}}(m) \le C_0 K T(m). \tag{26}$$

We will show that an atom of prevalence $p_i > 0$ is recovered whenever $E \le cp_i$, for a sufficiently small constant $c > 0$.

It suffices to take $0 < \eta < 1$. Let r_*, c_*, τ be the constants supplied by Lemma 7. Choose $0 < r \le r_*$ small enough that $r(1 + 2K/\gamma) < \min\{\gamma/(4\sqrt{2}), \sqrt{1-\eta^2}\}$. Let δ be the threshold supplied by Lemma 5 with $\epsilon = r$, decreasing it if necessary so that $\delta < c_-\gamma^2(\gamma/(8K))^{K-1}/32$. Finally, choose $c > 0$ such that $c < \rho\delta$ and $c(1 + \delta^{-1}) \le c_*$. These choices depend only on the parameters in the theorem.

Call a positive-probability support S good if $E_S \le \delta$, and write $\mathcal{G}$ for the good supports. Their complement has probability $\beta \le E/\delta$. For each good support, choose T_S as in Lemma 5, so its span has gap at most r from $\mathrm{span}(A_S)$.

Fix an atom i with $p_i > 0$ and $E \le cp_i$, and let $\mathcal{G}_i := \{S \in \mathcal{G} : i \in S\}$. Since $\beta < \rho p_i \le p_i$, this collection is nonempty. For every $k \neq i$, regularity gives $\Pr(i \in S(x), k \notin S(x)) \ge \rho p_i > \beta$. Hence some good support contains i and omits k, implying $\bigcap_{S \in \mathcal{G}_i} S = \{i\}$. Lemma 6 now gives a single learned coordinate j such that

$$\bigcap_{S \in \mathcal{G}_i} T_S = \{j\}, \qquad \mathrm{gap}\big(\mathrm{span}(a_i), \mathrm{span}(\hat{a}_j)\big) \le r(1 + 2K/\gamma). \tag{27}$$

Thus $j \in T_S$ whenever a good support contains i. If a good support S omits i, adding S to the preceding intersection makes the atomic intersection empty. Lemma 6 makes the learned intersection empty as well, so $j \notin T_S$. We have proved $j \in T_S$ if and only if $i \in S$ on all good supports, not just on those containing i.

The line bound in (27) gives $|\langle a_i, \hat{a}_j \rangle| > \eta$. We must also fix its sign. If the inner product were negative, the same bound and our choice of r would give $\|a_i + \hat{a}_j\|_2 < \gamma/4$. Choose any $S \in \mathcal{G}_i$

and consider coefficients with $z_i \in (1/2, 1)$ and $z_k \in (0, \gamma/(8K))$ for $k \in S \setminus \{i\}$. Since $\hat{z}$ is nonnegative and at most K-sparse, $\hat{z} + z_i e_j$ is at most $2K$-sparse and has norm at least z_i. Stability and the triangle inequality therefore give, on this coefficient region,

$$\|Az - \hat{A}\hat{z}\|_2 \geq (\gamma - \|a_i + \hat{a}_j\|_2)z_i - \sum_{k \in S \setminus \{i\}} z_k \geq \gamma/4.$$

The density lower bound gives this region conditional probability at least $(c_-/2)(\gamma/(8K))^{K-1}$. Thus $E_S \geq c_- \gamma^2 (\gamma/(8K))^{K-1}/32$, contradicting goodness and our choice of δ. Consequently, $\langle a_i, \hat{a}_j \rangle > \eta$, as required for signed direction recovery.

For activation recovery, observe that $E + \beta \leq E(1 + \delta^{-1}) \leq c_* p_i$. Lemma 7 therefore applies to the same coordinate j and gives an F1 score greater than η at threshold τ. This threshold is shared by all recovered coordinates. Finally, take $C := C_0 K/c$. By (26), the condition $p_i \geq CT(m)$ implies $E \leq cp_i$ and completes the proof. □

The hierarchical recovery principle. We make precise the hierarchical regularity invoked in the main text. Let $M = 3N$ and $K = 2L$, with families i, i_1, i_2 for $i \in [N]$. An input chooses a set $F(x) \subseteq [N]$ of exactly L active families. Each active family contributes its parent and exactly one child, so $S(x) = \bigcup_{i \in F(x)} \{i, i_{b_i(x)}\}$ with $b_i(x) \in \{1, 2\}$. The law of $F(x)$ satisfies $\Pr(k \notin F(x) \mid i \in F(x)) \geq \rho$ for distinct parents i, k whenever $\Pr(i \in F(x)) > 0$. In addition, for a fixed $\kappa \in (0, 1/2]$, we require

$$\Pr\big(b_i(x) = b \mid F(x) = F\big) \geq \kappa \qquad (i \in F,\ b \in \{1, 2\}) \tag{28}$$

for every positive-probability family set F. Conditional on each resulting atomic support, all K coefficients have a joint density between c_- and c_+ on $(0, 1)^K$, as above. We call such a distribution hierarchical regular.

The child choices need not be independent, and neither do the active coefficients. Independent uniform child choices give the special case $\kappa = 1/2$. The child-choice bound and the full conditional density bounds are part of this definition; they do not follow from the child-implies-parent condition alone.

As in the main text, now let $p_i := \Pr(i \in F(x))$ be the parent prevalences, let σ sort these prevalences, and let $T(n) := \sum_{j=n+1}^{N} p_{\sigma(j)}$. For a learned dictionary of width m, we set $n := \lfloor m/3 \rfloor$, with $T(0) := \sum_{i=1}^{N} p_i = L$.

Proof of Theorem 3. Retain all three atoms in each of the $n = \lfloor m/3 \rfloor$ most prevalent families. If $3n < m$, fill the remaining slots with unused columns of A and set their codes to zero. This gives an admissible dictionary of exactly m unit-norm columns. Each omitted active family contributes two omitted atoms, so the same truncation calculation as before gives

$$E \leq C_0 L^*_{\mathcal{D}}(m) \leq 2C_0 K T(n). \tag{29}$$

We show that a family of prevalence $p_i > 0$ is recovered whenever $E \leq cp_i$, for a sufficiently small constant $c > 0$.

Again it suffices to take $0 < \eta < 1$. Let r_*, c_*, τ be the constants from Lemma 7. Choose $0 < r \leq r_*$ with $r(1 + 2K/\gamma) < 1$, and let δ be the threshold from Lemma 5 with $\epsilon = r$. Choose $c > 0$ so that $c < \kappa\rho\delta$ and $c(1 + \delta^{-1}) \leq \kappa c_*$. As before, let $\mathcal{G}$ contain the positive-probability supports with $E_S \leq \delta$, choose their learned supports T_S, and write $\beta \leq E/\delta$ for the probability of the remaining supports.

Fix a family i with $p_i > 0$ and $E \leq cp_i$. Each child has prevalence at least κp_i. Moreover, for a different family k, (28) and parent regularity give

$$\Pr\big(i_b \in S(x),\ k \notin F(x)\big) \geq \kappa \Pr\big(i \in F(x),\ k \notin F(x)\big) \geq \kappa\rho p_i > \beta.$$

Thus, even after discarding the bad supports, either child can occur with i, and each chosen child can occur without any specified other family. It follows that the collections below are nonempty and

$$\bigcap_{S \in \mathcal{G}:\, i \in S} S = \{i\}, \qquad \bigcap_{S \in \mathcal{G}:\, i_b \in S} S = \{i, i_b\} \quad (b \in \{1, 2\}). \tag{30}$$

Indeed, the other child is excluded whenever i_b occurs, and the displayed probability bound excludes every atom belonging to any other family.

Applying Lemma 6 to these intersections gives learned sets

$$J_i := \bigcap_{S \in \mathcal{G}:\, i \in S} T_S = \{j_i\}, \qquad J_{i_b} := \bigcap_{S \in \mathcal{G}:\, i_b \in S} T_S = \{j_i, j_{i_b}\}.$$

Here $J_i \subseteq J_{i_b}$ because every support containing i_b also contains i. The good supports containing i are exactly the union of those containing i_1 and those containing i_2. Hence $J_{i_1} \cap J_{i_2} = J_i$, so j_i, j_{i_1}, j_{i_2} are distinct.

We claim that each of these learned coordinates has the same membership pattern as its corresponding atom across all good supports. For the parent, this is the same argument as in Theorem 2: adding a good support omitting i to its intersection forces that support's learned set to omit j_i. For a child i_b, every good support containing i_b contains J_{i_b} on the learned side. If a good support omits the family, its atomic intersection with $\{i, i_b\}$ is empty, so its learned intersection with J_{i_b} is empty as well. If it instead contains i and the other child, that atomic intersection is $\{i\}$. Lemma 6 makes the learned intersection a singleton, which must be $\{j_i\}$. In particular it omits j_{i_b}. This proves the claimed membership patterns.

Finally, $E + \beta \le E(1 + \delta^{-1}) \le \kappa c_* p_i$, and each of the three target coordinates has prevalence at least κp_i. Lemma 7 therefore gives an F1 score greater than η for each target, using the distinct learned coordinates j_i, j_{i_1}, j_{i_2}. The same threshold τ works for all three and for all recovered families. Take $C := 2C_0K/c$ and invoke (29). This proves the theorem. □

C EXPERIMENTAL SETUP

This section describes the shared data, training setup, and matching metrics, followed by the setup for the main P1, P2, and P3 experiments in Figures 2a to 2c, respectively. Extended experiments (including results and setup) are discussed in Sections D to F.

C.1 DATA AND SAE TRAINING DETAILS

Training corpus. The full training mix, `full`, combines 13 source datasets used to fine-tune the Nemotron embedding model. Table 1 counts the stored *embedding vectors* available to SAE training: a question and its answer, for example, are embedded separately and contribute two vectors. These counts differ from source-record counts and do not deduplicate text across datasets or text columns.

Table 1: SAE training datasets, with number of embeddings computed for each model.

Source collection	Gemini	Nemotron
Emotion (Saravia et al., 2018)	19,930	19,930
FEVER (Thorne et al., 2018; Thakur et al., 2021)	11,079,420	11,079,420
GooAQ (Khashabi et al., 2021)	6,024,992	6,024,992
HotpotQA (Yang et al., 2018)	10,564,510	10,564,510
MS MARCO (Bajaj et al., 2016)	9,351,785	9,351,785
Natural Questions (Kwiatkowski et al., 2019)	200,462	200,462
NFCorpus (Boteva et al., 2016)	10,503	10,503
PAQ, sampled 10M pairs (Lewis et al., 2021)	20,000,000	20,000,000
SciFact (Wadden et al., 2020)	11,475	11,475
SQuAD (Rajpurkar et al., 2016)	175,198	175,198
TriviaQA (Joshi et al., 2017)	146,692	146,692
WebInstructSub (Yue et al., 2024)	4,670,440	4,670,440
Non-MIRACL subtotal	62,255,407	62,255,407
MIRACL, sampled passages	7,978,232	7,378,232
MIRACL, unique titles per language	19,593,919	19,593,919
Full mix (`full`)	89,827,558	89,227,558

Corpus preparation. Each dataset gives a query and an answer. We embed both separately. We sample 10M question-answer pairs from PAQ. For MIRACL (Zhang et al., 2023), we sample up to 500,000 passages per language. All available passages are used when a language has fewer than 500,000. (Due to an error, our nemotron embeddings are missing 500,000 Thai passages and 100,000 sampled Telugu passages present for Gemini. Thus the full corpora contain 89.8M Gemini vectors and 89.2M Nemotron vectors.) When comparing activations, we use the 89.2M shared embeddings.

Embedding representations. For gemini embeddings, we use `gemini-embedding-2-preview` with the task type "RETRIEVAL_DOCUMENT" and dimension $d = 3072$. Nemotron representations use `nvidia/llama-embed-nemotron-8b` with $d = 4096$. We truncate at 10,000 characters for gemini and 256 tokens for Nemotron.

Additional data splits for P2 experiments. We train a separate width-16,384, $k = 64$ SAE for each of the following mixtures and each embedding model:

- `rand1`: emotion, fever, gooaq, natural-questions, sampled PAQ, and scifact; 37,336,279 vectors.
- `rand2`: hotpotqa, msmarco, nfcorpus, squad, trivia-qa, and WebInstructSub; 24,919,128 vectors.
- `wiki`: fever, hotpotqa, natural-questions, sampled PAQ, squad, and trivia-qa; 42,166,282 vectors.
- `no_wiki`: emotion, gooaq, msmarco, nfcorpus, scifact, WebInstructSub; 20,089,125 vectors.

Each mixture contains six complete prepared datasets. The `wiki/no_wiki`and `rand1/rand2`pairs are two partitions of the same 12 non-MIRACL datasets; both MIRACL collections occur only in `full`. `wiki/no_wiki`splits based on if the dataset includes Wikipedia data or not.

SAE architecture and objective. We use the ordinary per-example (non-batch) TopK SAE implementation from https://github.com/bartbussmann/BatchTopK/tree/main. For an embedding $x \in \mathbb{R}^d$ the forward pass is

$$h(x) = \mathrm{ReLU}(W_{\mathrm{enc}}(x - b_{\mathrm{dec}}))\,, \tag{31}$$

$$z(x) = \mathrm{TopK}_k(h(x))\,, \tag{32}$$

$$\hat{x} = W_{\mathrm{dec}} z(x) + b_{\mathrm{dec}}, \tag{33}$$

where $W_{\mathrm{enc}} \in \mathbb{R}^{m\times d}$, $W_{\mathrm{dec}} \in \mathbb{R}^{d\times m}$, and $b_{\mathrm{dec}} \in \mathbb{R}^d$. There is no encoder bias. TopK_k retains the k largest nonnegative activations in each example.

Let D be the set of dead neurons: neurons features with no positive post-TopK activation in the last 512 consecutive training batches. For each example, $z^{\mathrm{AuxK}}(x)$ retains the largest $\min\{32, |D|\}$ entries of the *pre-TopK* activation vector $h(x)$ restricted to D, with all other entries set to zero. For a batch of B embeddings, the implemented loss is

$$\mathcal{L} = \frac{1}{Bd}\sum_{i=1}^{B}\left[\|x_i - \hat{x}_i\|_2^2 + w_{\mathrm{aux}}\left\|W_{\mathrm{dec}} z^{\mathrm{AuxK}}(x_i) - (x_i - \hat{x}_i)\right\|_2^2\right]. \tag{34}$$

The auxiliary term is zero when D is empty. Its residual $x_i - \hat{x}_i$ remains in the gradient computation. We use no ℓ_1 penalty. The auxiliary coefficient is $w_{\mathrm{aux}} = 0.25$ except for the Gemini width-131,072 SAE, which uses $w_{\mathrm{aux}} = 1.0$ (we found that at 0.25 in this setting, the number of dead neurons would remain large).

Optimization and training budget. The main width sweep contains nine SAEs per embedding model, with widths, sparsities, and initialization seeds in Table 2. The four additional distribution mixtures give 26 SAEs in the main width/split experiments. We use float32 parameters, batch size $B = 2048$, and Adam with learning rate 5×10^{-4}, $(\beta_1, \beta_2) = (0.9, 0.99)$, $\epsilon = 10^{-8}$, and no weight decay or learning-rate schedule. Gradients are clipped to global norm 1. The encoder is initialized with Kaiming-uniform weights, the decoder with the normalized transpose of the encoder, and the bias with zeros. Before each optimizer step, decoder directions are normalized and their radial gradient components are removed.

All of these SAEs train for 1,000,000 optimizer steps except the Gemini width-131,072 SAE, which trains for 2,000,000 steps. These budgets correspond to 2.048 billion and 4.096 billion sampled embedding presentations, respectively, with repeated sampling from the finite corpora above. A shard is selected with probability proportional to its row count and supplies four batches before the next shard selection. Each batch is a randomly positioned contiguous block with its rows shuffled; shards smaller than one batch are sampled with replacement. Training uses NVIDIA A100 (80GB) or A6000 GPUs. Seeds are fixed separately for each run, rather than paired between embedding models. The synthetic experiments use a separate training recipe described in Section G.

C.2 Matching criteria and computation

The P1, P2, and cross-model comparisons combine a pairwise similarity score with a rule for selecting matches. We define the shared metrics and procedures here; the experiment-specific sections specify the dictionaries and evaluation corpora. The default threshold is $t = 0.7$.

C.2.1 Direction similarity

For nonzero dictionary directions u and v in the same embedding space, we use signed cosine similarity,

$$s_{\mathrm{dir}}(u, v) = \frac{\langle u, v\rangle}{\|u\|_2\|v\|_2}.$$

For SAEs, u and v are decoder columns. The main PCA baseline uses $|s_{\mathrm{dir}}(u, v)|$, because the signs of principal components are arbitrary. The exclusive criterion below uses signed cosine for both SAEs and PCA.

Table 2: Sparsities and initialization seeds for the 26 SAEs in the main width and distribution experiments. Subset runs all use $m = 16,384$ and $k = 64$. The Gemini width-131,072 run uses 2M steps and auxiliary coefficient 1.0; every other listed run uses 1M steps and coefficient 0.25.

Width or subset	k	Gemini seed	Nemotron seed
512	32	1138983768	1588734130
1,024	32	1643307215	515450099
2,048	32	1903409744	1277871182
4,096	32	296799549	578010941
8,192	64	1244342741	1357330459
16,384	64	707885418	1610719394
32,768	64	1480725346	1923885051
65,536	128	1317536270	150416361
131,072	128	172912	1728433352
`wiki`	64	752675140	1058356814
`no_wiki`	64	65683026	223132717
`rand1`	64	1028554591	222067858
`rand2`	64	669991270	616594595

C.2.2 Activation similarity

For two SAE features evaluated on the same examples $x_1, \ldots, x_N$, let a_n and b_n be their continuous post-TopK activations. With $\bar{a} = N^{-1} \sum_n a_n$ and $\bar{b} = N^{-1} \sum_n b_n$, their signed Pearson correlation is

$$s_{\text{act}}(a, b) = \frac{\sum_{n=1}^{N} (a_n - \bar{a})(b_n - \bar{b})}{\sqrt{\sum_{n=1}^{N} (a_n - \bar{a})^2} \sqrt{\sum_{n=1}^{N} (b_n - \bar{b})^2}}.$$

Every evaluation example receives equal weight, including examples with zero activation. Constant activation vectors have undefined correlation and cannot form eligible pairs. The evaluation corpus is part of this metric: changing it can change a pair's correlation.

C.2.3 One-to-one and common-subset matching

For either similarity score, eligible pairs have similarity at least t and form the edges of a bipartite graph. A maximum-cardinality matching is the largest set of eligible pairs in which each feature occurs at most once. For directions, this is the main-text quantity $\mu_t(U, V)$, with the absolute-cosine modification for PCA. Reported proportions divide the matching count by the source dictionary width, including inactive and unmatched features in the denominator.

For persistence across widths, we maximize the size of a *common source subset* independently matchable into every target dictionary: $\mu_t(U; V_1, \ldots, V_s)$ for directions, with an analogous definition for activations. Destination features need not match one another across targets. This is neither the minimum of the pairwise matching counts nor the intersection of independently optimized pairwise matchings.

Let $x_u \in \{0, 1\}$ indicate whether source feature u is selected, and let $y_{uv}^{(\ell)} \in [0, 1]$ indicate its assignment to feature v in target ℓ, with assignment variables only for eligible edges. We maximize $\sum_u x_u$ subject to

$$\sum_{v:(u,v)\text{ eligible}} y_{uv}^{(\ell)} = x_u \quad \text{for every } u, \ell, \qquad \sum_{u:(u,v)\text{ eligible}} y_{uv}^{(\ell)} \leq 1 \quad \text{for every } v, \ell. \tag{35}$$

Once the source subset is fixed, each target's bipartite matching problem admits an integral solution whenever feasible. We discard sources lacking a neighbor in at least one target, decompose independent groups of sources sharing possible destination partners, and use exact bipartite-matching bounds or mixed-integer optimization to find each group's optimum. We reconstruct and verify a matching for the selected subset in every target.

C.2.4 CANDIDATE GRAPH CONSTRUCTION

Direction comparisons evaluate all pairs. For activation comparisons, we use 512-dimensional CountSketch representations to propose 32 candidates per feature in each search direction, with the same evaluation rows and sketch hashes for both dictionaries. We recompute candidate correlations over the full evaluation corpus and retain the five highest-scoring candidates per feature in each direction. Thresholded matching uses the union of retained pairs with correlation at least t.

Pairwise and common-subset matching are exact on these retained graphs. Because candidate search and retention can omit qualifying pairs, the counts are lower bounds on exhaustive Pearson matching. The mutual-nearest-neighbor comparison in Section J uses the same retrieval and rescoring procedure but a different selection rule, specified there.

C.2.5 ALTERNATIVE EXCLUSIVE MATCHING CRITERION

The exclusive direction comparisons in Figures 6 and 12 count a pair (u, v) only when its signed cosine exceeds t and neither endpoint has another above-threshold partner. Writing $s(u, v) = s_{\mathrm{dir}}(u, v)$, the requirements are

$$s(u, v) > t, \qquad s(u, v') \le t \text{ for all } v' \neq v, \qquad s(u', v) \le t \text{ for all } u' \neq u. \tag{36}$$

Thus only isolated edges of the signed-cosine threshold graph count. We use $t = 0.7$ both to admit pairs and to identify competitors, checking every direction in the opposite dictionary. For example, if u has similarities 0.90 and 0.80 with v_1, v_2 and there are no other eligible pairs, maximum-cardinality matching counts one pair but exclusive matching counts none. A PCA pair with cosine -0.90 is also excluded, although eligible under the main absolute-cosine baseline.

The exclusive pairs themselves form a feasible one-to-one matching, so their count is bounded above by μ_t. Lowering t can reduce the exclusive count by introducing competitors, whereas maximum-cardinality counts cannot decrease. For persistence, intersecting the exclusive source-feature sets across targets gives the exact count, since the pairs are already injective.

C.3 P1: PERSISTENCE ACROSS SAE WIDTHS

We apply the common-subset procedure in Section C.2.3 separately within Gemini and Nemotron, using each source SAE below width 131,072 and every larger trained SAE as targets. We report counts and source-width-normalized proportions. The largest SAE is a target only, since it has no larger comparison dictionary. Figure 2a uses the signed decoder-cosine metric in Section C.2.1.

C.4 P2: SHARED FEATURES ACROSS TRAINING DISTRIBUTIONS

Figure 2b compares dictionaries trained on `full`, `wiki`, `no_wiki`, `rand1`, and `rand2`, separately for each embedding model and dictionary family. SAE dictionaries contain 16,384 directions; PCA uses all d principal components, with $d = 3072$ for Gemini and $d = 4096$ for Nemotron.

PCA dictionaries. We fit PCA separately for each embedding model and training mixture, using every embedding in that mixture. We stream the data to accumulate the count N, sum $s = \sum_{i=1}^{N} x_i$, and second-moment matrix $Q = \sum_{i=1}^{N} x_i x_i^\top$ in float64, then compute

$$\bar{x} = s/N, \qquad \Sigma = \frac{Q - N\bar{x}\bar{x}^\top}{N - 1}.$$

We diagonalize this centered covariance matrix and order its orthonormal eigenvectors by decreasing eigenvalue. The PCA dictionary consists of these directions in the original embedding space; P2 uses the full d-component basis. PCA fitting uses the stored embeddings without additional length normalization or coordinate standardization.

Dictionary matching. For each pair of training distributions, Figure 2b applies the direction metric, pairwise matching rule, and source-width normalization in Section C.2. Equal-width direction comparisons give symmetric heatmaps; self-comparisons are omitted. Exclusive direction matching and distribution-specific activation comparisons are reported in Section E.

C.5 P3: HIERARCHICAL DATASETS AND FEATURE-RECOVERY PROTOCOL

For Figure 2c, we evaluate individual SAE features against binary category labels in three domains: word prefixes, geography, and organism taxonomy. The labels define hierarchical relationships, such as a two-letter prefix implying its first letter or a city's country determining its continent (the organism taxonomy has more levels).

English prefixes (`prefix`). We use the English frequency list from the `wordfreq` package. We retain the top 200,000 words in frequency order. Each normalized word is embedded on its own. The first character defines a parent label, and the first two characters define a child label for words of length at least two. We retain categories with at least 100 positive word rows across the full evaluation collection, giving 26 one-letter and 216 two-letter categories per embedding model.

Geography (`cities`). We use all 69,182 records in the experiment's GeoNames `cities5000` snapshot, with country labels from each record's ISO country code and continent labels obtained from `countryInfo.txt`. We retain 87 country and six continent categories with at least 100 city rows each. For the reported results, we embed two question variants of every city name: `what country is [city] in` for country prediction and `what continent is [city] in` for continent prediction, following the use of elicitation in Bhalla et al. (2026). (We found that without this elicitation, SAEs at no widths reliably could detect the continent.)

Organism taxonomy (`gbif`). We use the GBIF Backbone Taxonomy. From its `Taxon.tsv` and `VernacularName.tsv` tables, we select accepted species having at least one nonempty English vernacular name. This yields 123,495 species. We embed one scientific-name row per species and 259,286 deduplicated English common-name strings, leading to 382,781 name rows in total. Binary labels follow the species' kingdom, phylum, class, order, family, and genus. We retain categories with at least 100 distinct species. This gives 4 kingdoms, 12 phyla, 33 classes, 131 orders, 276 families, and 43 genera. Once labels are constructed, F1 is evaluated over name rows, so species with more names can contribute more evaluation examples.

Selection and test splits. We partition the data into selection and test sets. Word rows are divided equally within each two-letter-prefix stratum, with one-character words forming their own strata. City rows are divided equally within each country, and both question variants inherit this partition. GBIF partitions the underlying species into 61,748 selection species and 61,747 test species, with scientific names inheriting the species assignment. Common-name rows are assigned a single split, even when their species associations span both partitions.

Every probe reads these saved split labels unchanged, and the same assignments are used across both embedding models and all widths. Exact reruns require these row tables and their source snapshots: the provided preparation scripts supply deterministic seed-1729 reconstructions for the word and city splits, but do not reproduce the historical row-by-row assignments. Fresh source downloads can also change the evaluation collections.

Embedding and SAE evaluation. We compute gemini and nemotron embeddings for each input and compute activations for the nine full-mix SAEs for each model, with widths 512 through 131,072 and the corresponding TopK sparsities in Table 2.

Selecting a single-feature probe. For a binary target z_i^* and SAE activations $\hat{z}$, define

$$g_{j,\tau}(x) = \mathbf{1}\{\hat{z}_j(x) \geq \tau\}, \qquad \tau > 0.$$

For each coordinate active on selection rows, we sort its positive selection activations and evaluate every distinct observed value as an inclusive threshold. We choose the coordinate and threshold maximizing selection F1, requiring at least one selection true positive.

For $E = \{z_i^* = 1\}$ and $F = \{g_{j,\tau} = 1\}$, we use

$$\mathrm{F1}_{\mathcal{D}}(E, F) = \frac{2\,\Pr_{\mathcal{D}}(E \cap F)}{\Pr_{\mathcal{D}}(E) + \Pr_{\mathcal{D}}(F)} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

on the relevant split. This evaluates how well a single thresholded feature identifies a category.

Figure 2c details. Each cell in Figure 2c is the unweighted mean of held-out test F1 over all retained categories at that level and both embedding models. Thus N is twice the number of categories listed above, and the same category–model pairs contribute at every width. We describe main text a-prefix example and empirical Food and Music hierarchies are described in Section F, alongside other additional hierarchical experiments.

D Extended Experiments for P1: Persistence across Size

We extend the main persistent feature experiment in Figure 2a with controls for sparsity and training seed, activation-based matching, alternative matching criteria and similarity thresholds, and analyses of feature prevalence.

D.1 Replication with fixed sparsity

At widths 512, 4,096, and 32,768, we replace each original source SAE by its $k = 128$ counterpart, retaining every original larger-width target and the direction-matching protocol in Section C.3.

Table 3: Fixed-sparsity source replacement, using the same full-sweep target sets as the seed control. Only the source changes from its original k to $k = 128$; all original larger-width targets are retained. Same-width counts match the original and replacement source dictionaries. All comparisons use signed decoder cosine similarity ≥ 0.7; percentages divide by source width.

Model	Original width/k	Same-width cross-sparsity matches	Original source persistence	$k = 128$ source persistence
Gemini	512/32	389 (76.0%)	304 (59.4%)	221 (43.2%)
	4,096/32	2,949 (72.0%)	2,978 (72.7%)	2,409 (58.8%)
	32,768/64	16,466 (50.3%)	15,890 (48.5%)	12,906 (39.4%)
Nemotron	512/32	410 (80.1%)	259 (50.6%)	210 (41.0%)
	4,096/32	3,286 (80.2%)	2,849 (69.6%)	2,614 (63.8%)
	32,768/64	18,198 (55.5%)	12,761 (38.9%)	13,633 (41.6%)

Table 3 shows persistence remains substantial, but is somewhat lower in general, especially for smaller widths. This seems to suggest that the conventional wisdom to increase k alongside m is reasonable (and also is somewhat in alignment with our simulations Section G).

D.2 Replication with an alternate training seed

We repeat the direction comparisons for gemini SAEs 512/32, 4,096/32, and 32,768/64 (width/k) with different training seeds. Table 4 shows a substantial number of matches across seeds, and that persistent remains similar (suggesting that persistence doesn't vary heavily between SAEs). The same-width and persistence columns use the respective matching rules in Section C.2.3. Consistent with Paulo & Belrose (2025), there are a substantial number of different features learned when varying the random seed (especially for larger widths); simultaneously, there are many features that have matches even in all larger dictionaries.

Table 4: Gemini direction stability across training seeds.

Width/k	Same-width cross-seed matches	Original source persistence	Alternate source persistence
512/32	470 (91.8%)	304 (59.4%)	293 (57.2%)
4,096/32	3,564 (87.0%)	2,978 (72.7%)	2,993 (73.1%)
32,768/64	17,856 (54.5%)	15,890 (48.5%)	16,014 (48.9%)

D.3 ALTERNATIVE STABILITY COMPARISON

Figure 6 repeats the source–target comparisons in Section C.3 using the exclusive persistence criterion in Section C.2.5.

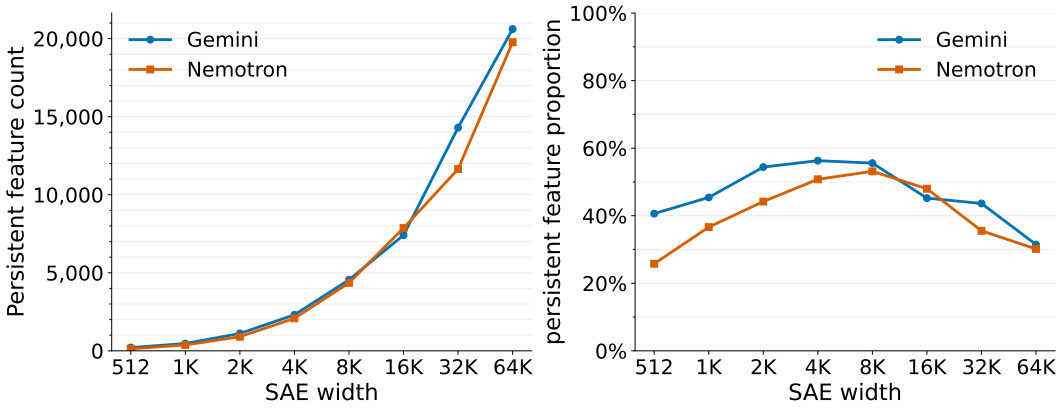


Figure 6: Persistent stability under strict exclusive matching at 0.7. A source direction counts only if it has a signed-cosine match above 0.7 in every larger trained dictionary, with no competing partner above 0.7 at either endpoint in each comparison.

D.4 PERSISTENCE BY ACTIVATION SIMILARITY

Figure 7 repeats the source–target comparisons in Section C.3 using signed Pearson correlation at threshold 0.7, with the metric and candidate-graph procedure defined in Sections C.2.2 and C.2.4.

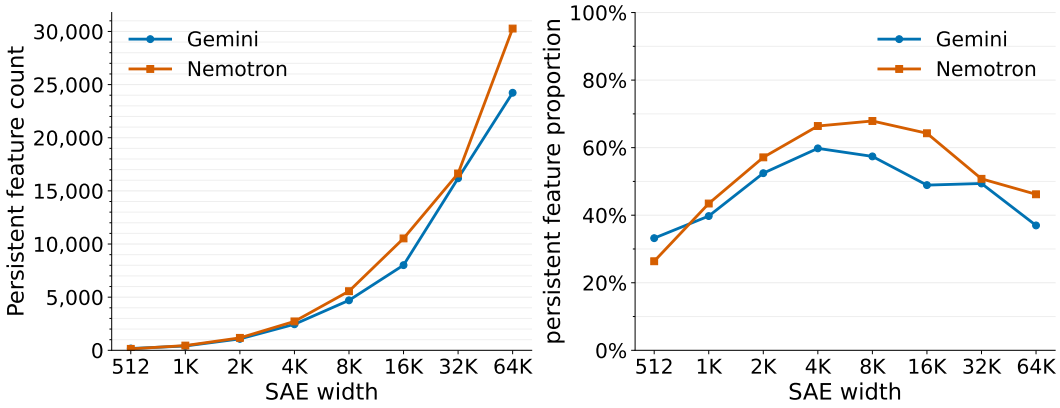


Figure 7: Persistent stability across SAE widths using activation similarity at signed Pearson correlation at least 0.7. Left: the largest common source subset injectively matchable into every larger trained SAE on the retained candidate graphs. Right: this count divided by source width. Both panels compare Gemini and Nemotron; counts are lower bounds on exhaustive Pearson persistence. The 131K source width is omitted because no larger trained dictionary is available.

D.5 SENSITIVITY TO THE MATCHING THRESHOLD

We vary the matching threshold over $t \in \{0.5, 0.6, 0.7, 0.8\}$ used to compute persistent matches. Figure 8 uses signed decoder cosine similarity, and Figure 9 uses signed activation Pearson correlation.

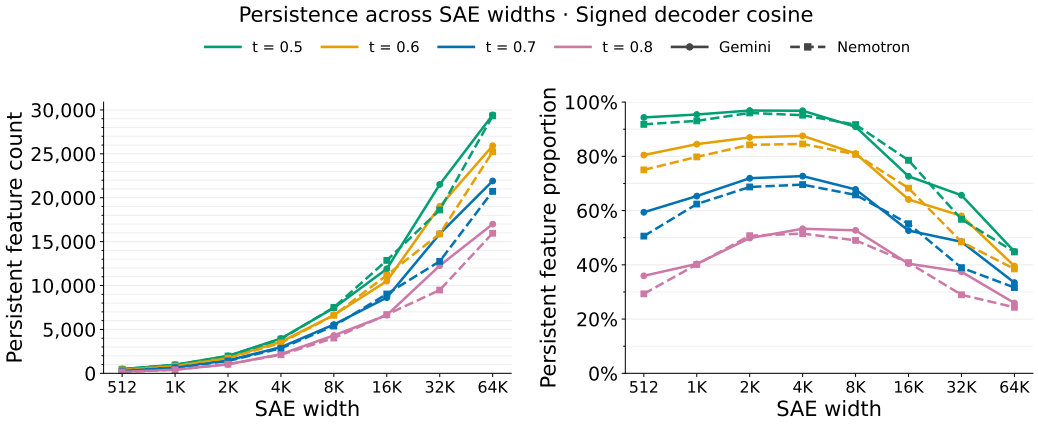


Figure 8: Threshold sensitivity of the persistent-stability experiment in Figure 2a.

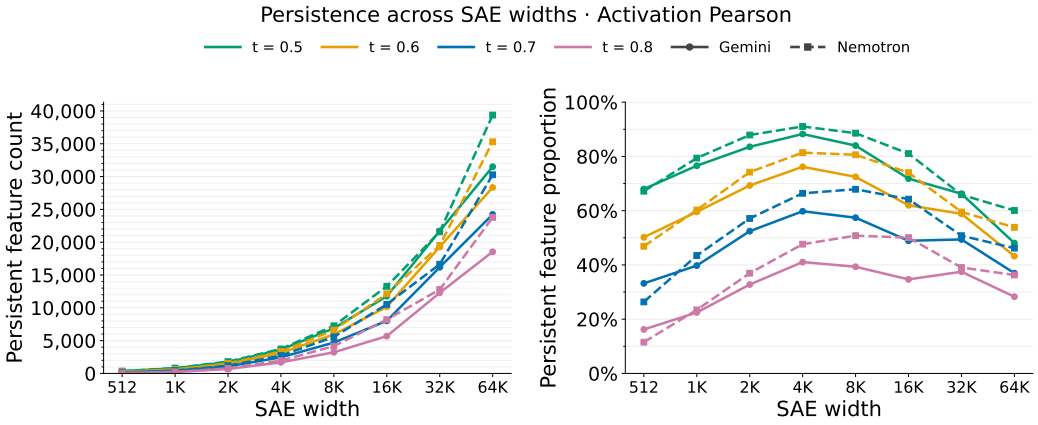


Figure 9: Threshold sensitivity of persistence measured by feature activations.

D.6 Feature activation rates

Figure 10 complements the persistence comparisons by showing the prevalence of learned features across SAE widths. For every feature, we compute the fraction of the model's full evaluation corpus on which its continuous post-TopK activation exceeds either zero or the SAE's pooled median positive activation. The latter is one threshold per SAE, computed over all positive activation events across features and examples.

We plot kernel density estimates of $\log_{10}$ activation rates, using a bandwidth of 0.12 and scaling each curve to integrate to the number of features with positive rates. Zero-rate features are omitted from the logarithmic axis. The vertical axis uses a square-root scale with tick labels in the original density units.

We observe a "nested" behavior, in which larger SAEs tend to include a similar number of coarse features as smaller SAEs, while also learning a longer tail of more granular features, consistent with the recovery principle.

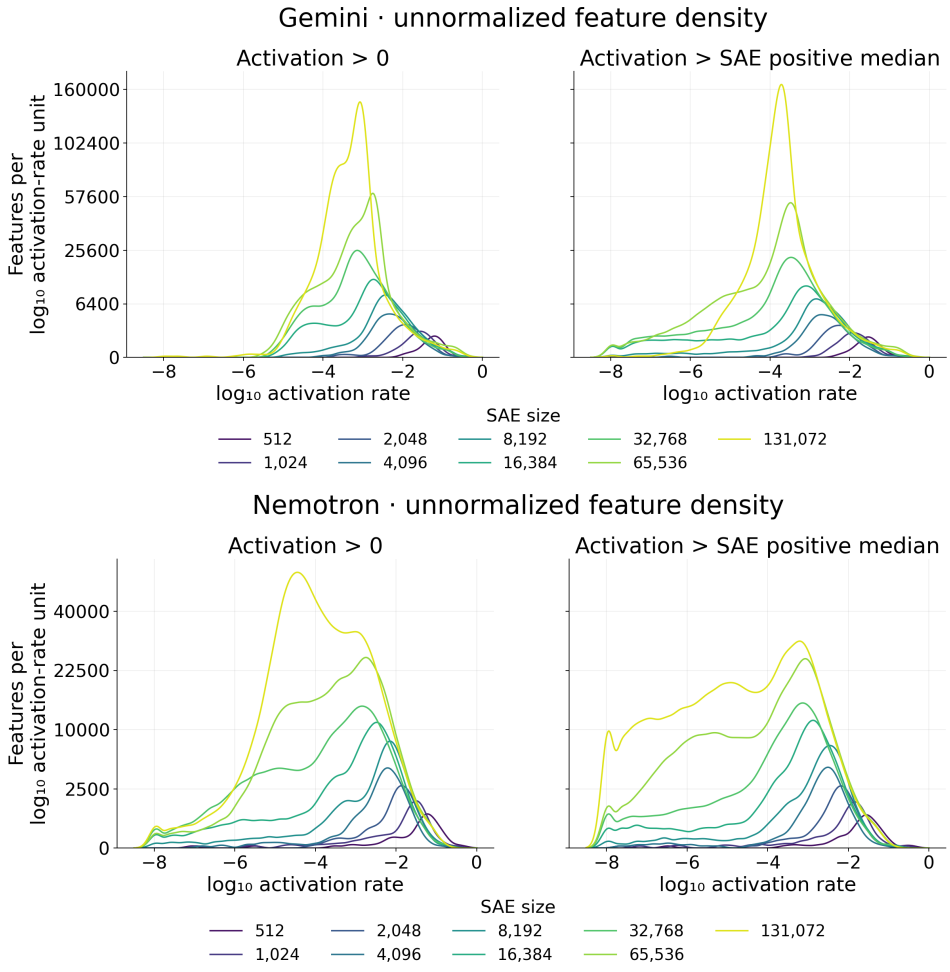


Figure 10: Feature activation rates across SAE widths for Gemini (top) and Nemotron (bottom). Left columns use activation strictly greater than zero; right columns use activation strictly greater than the SAE's pooled median positive activation. Colors indicate SAE width. Curves show feature-count-scaled densities of $\log_{10}$ activation rates, with a square-root vertical scale.

D.7 Feature prevalence and direction persistence

We directly compare feature prevalence with direction persistence using the median activation threshold from Figure 10. For an SAE with continuous post-TopK activations $\hat{z}_i(x)$, let τ_{SAE} be the pooled median of all strictly positive activation events, across features and examples in its embedding model's full evaluation corpus. We define

$$p_i(D) = \frac{1}{|D|} \sum_{x \in D} \mathbf{1}\{\hat{z}_i(x) > \tau_{\text{SAE}}\}. \tag{37}$$

There is one threshold per SAE, rather than a separate median for each feature; the denominator includes every example. The full corpora contain 89,827,558 Gemini and 89,227,558 Nemotron examples. These are descriptive full-corpus measurements.

For each source width in the original sweep, we compare p_i between features selected in one optimal direction matching under the persistence protocol in Section C.3 and the remaining source features.

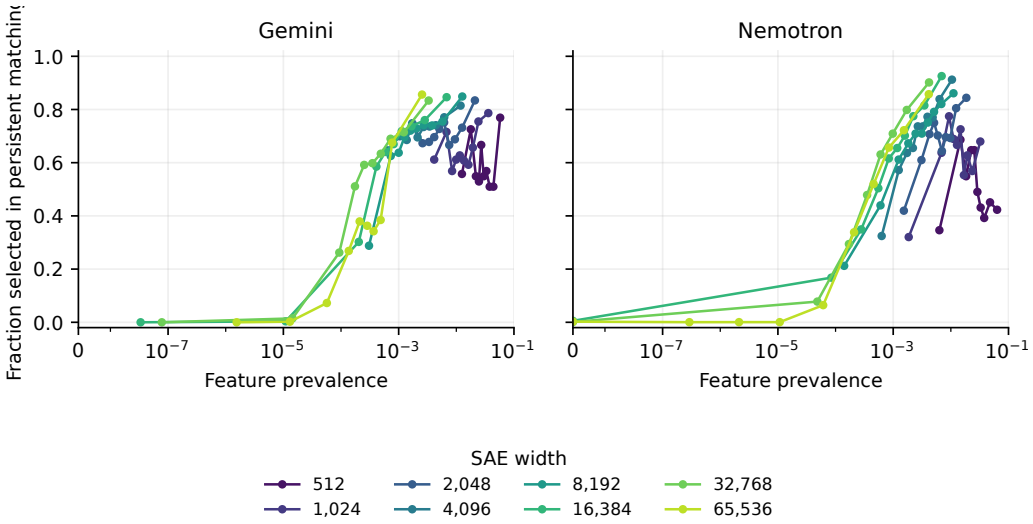


Figure 11: Median-threshold feature prevalence and direction persistence in Gemini (left) and Nemotron (right). Each curve represents a source width in the original sparsity schedule. Points group source features into up to ten prevalence-quantile bins, keeping ties together, and show the median prevalence and fraction selected in the persistent matching within each bin. All source features are included. Bin sizes can differ; lines connect descriptive bin summaries. The horizontal axis is logarithmic above 10^{-8} with a linear segment retaining zero-prevalence bins.

We measure the AUC of feature prevalence as a predictor of persistence. At widths 16,384, 32,768, and 65,536, AUCs are 0.815, 0.802, and 0.822 for Gemini, and 0.819, 0.905, and 0.917 for Nemotron. Restricting to features with nonzero prevalence gives respective ranges of 0.787–0.821 and 0.793–0.905. The association is weaker at the smallest widths.

Overall, these results are consistent with the recovery principle's prediction that the most prevalent atomic features are recovered.

E Extended Experiments for P2: Stability across Distributions

We extend the direction-based comparison in Figure 2b using the exclusive matching criterion, activation-based matching, and the relationship between cross-distribution stability and cross-distribution prevalence.

E.1 Cross-distribution stability under exclusive matching

Figure 12 applies the exclusive signed-cosine criterion from Section C.2.5 to SAE and PCA dictionaries trained on different distributions, at $t = 0.7$. This tests whether cross-distribution agreement persists when each match must have an unambiguous, same-sign partner.

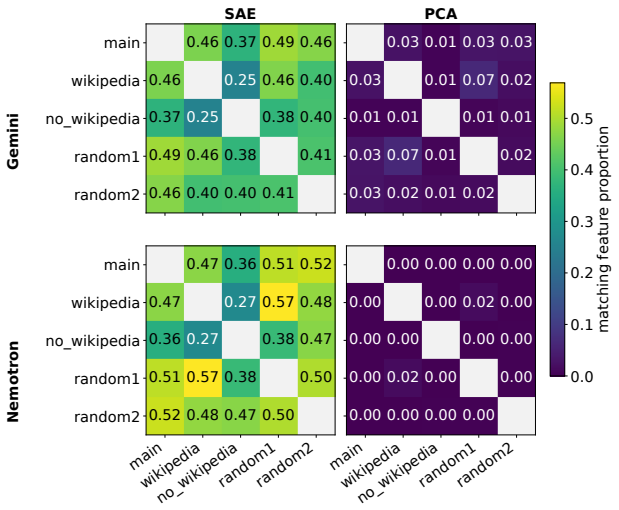


Figure 12: Cross-distribution SAE and PCA matches under the alternative exclusive signed-cosine criterion at $t = 0.7$. Each counted pair exceeds the threshold and has no other above-threshold partner at either endpoint. Entries divide the pair count by the dictionary width. Compared with the maximum-cardinality matching in Figure 2b, this criterion discards all pairs involving competing above-threshold partners. It also excludes opposite-sign PCA directions, which remain eligible in the main PCA baseline. It therefore gives a stricter measure of unambiguous agreement between dictionaries.

E.2 Cross-distribution stability by activation similarity

Figure 13 applies the activation metric, pairwise matching rule, and normalization in Section C.2 to the width-16,384 SAEs trained on `full`, `wiki`, `no_wiki`, `rand1`, and `rand2`. Within each embedding model, the cell at row D and column D' evaluates both SAEs on all examples in D. Reversing the row and column changes the evaluation distribution, so the heatmap need not be symmetric.

We construct sketches and candidate graphs for each row distribution using Section C.2.4, with moments computed over that distribution.

Matched proportions range from 22.0% to 48.9% for Gemini and from 29.3% to 60.7% for Nemotron.

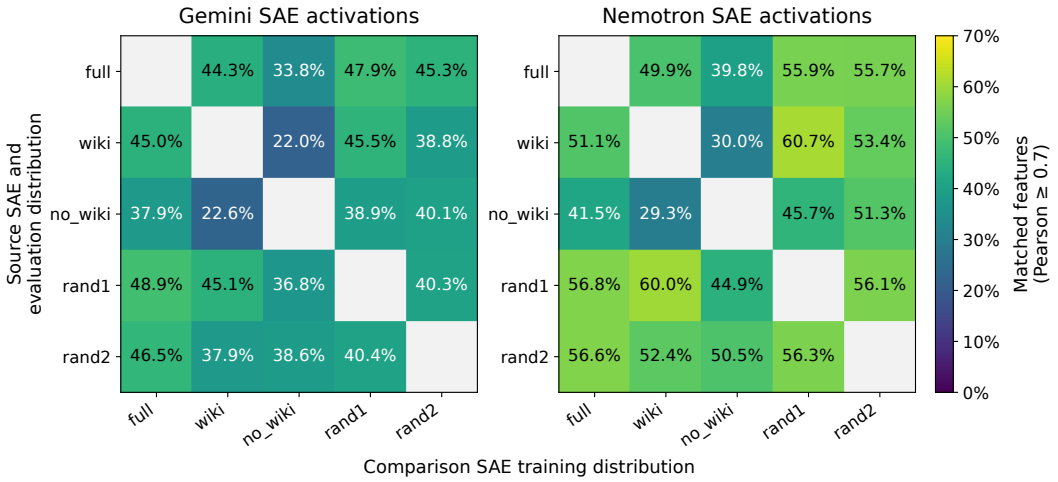


Figure 13: Cross-distribution SAE activation matching for Gemini (left) and Nemotron (right). Each cell reports the maximum number of one-to-one matches at signed Pearson correlation ≥ 0.7, divided by the SAE width of 16,384. Both SAEs are evaluated on the row's distribution; columns identify the comparison SAE's training distribution.

E.3 Feature prevalence and cross-distribution matching

We test whether features matched between distribution-specific SAEs are more prevalent in both distributions, using the median-threshold definition in Equation (37). Corollary 2 suggests that the atomic features that are prevalent in both data distributions are recovered.

For each width-16,384 source SAE trained on D, we compute its pooled median positive activation on the embedding model's full evaluation corpus, including MIRACL. We freeze that single threshold when evaluating the source SAE on both D and the target training distribution D', and summarize prevalence by

$$b_i(D, D') = \min\{p_i(D), p_i(D')\}. \tag{38}$$

Both frequencies refer to the same feature in the same source SAE, including when it has no match in the target SAE. This avoids assigning a partner to unmatched features. Reversing source and target evaluates a different SAE, so the prevalence analysis is directional.

We use the disjoint `wiki`/`no_wiki`and `rand1`/`rand2`dataset partitions from Section C. These evaluation pairs exclude MIRACL, although threshold calibration uses the full corpus. They contain respectively 42,166,282/20,089,125 and 37,336,279/24,919,128 examples; they partition datasets rather than equal numbers of rows. The outcome is membership in one optimal SAE direction matching under the protocol in Section C.4. All 16,384 source features enter the analysis, with direction matches fixed independently of the prevalence threshold.

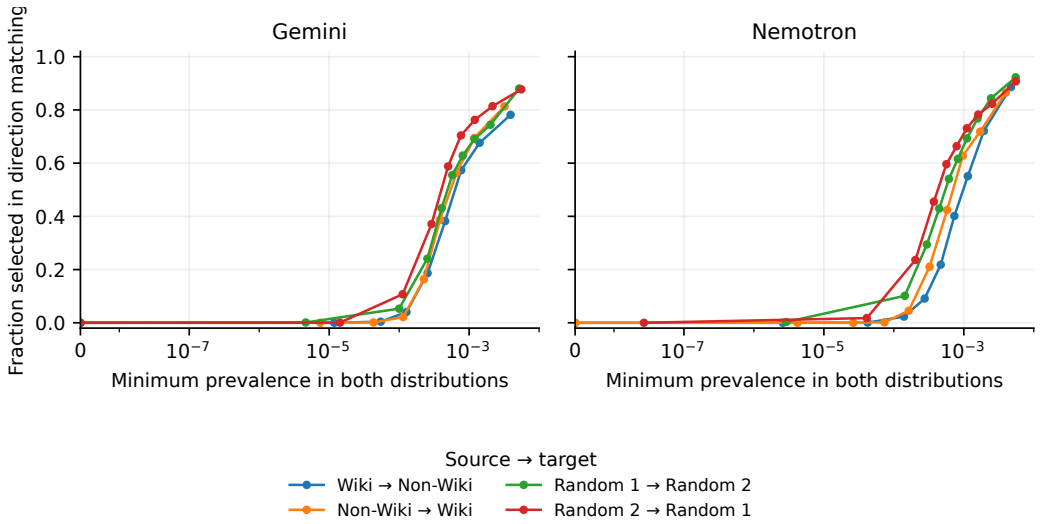


Figure 14: Median-threshold prevalence in both distributions and cross-distribution direction matching for Gemini (left) and Nemotron (right). Curves identify the source and target training distributions. The horizontal coordinate is the bin median of the minimum prevalence of a source feature across the two distributions; the vertical coordinate is the fraction selected in a one-to-one direction matching. Quantile binning preserves ties and includes all source features, with the same axis convention as Figure 11.

Higher prevalence in both distributions is associated with matching in all eight directional comparisons. AUCs, defined as in Section D.7, range from 0.864 to 0.923 for Gemini and from 0.839 to 0.927 for Nemotron. They remain positive after restricting to features with nonzero prevalence in both distributions (ranges 0.844–0.905 and 0.828–0.921, respectively).

F Extended Experiments for P3: Hierarchical Feature Recovery

We complement the aggregate hierarchy results in Figure 2c with the a-prefix example in Figure 3a and the empirical Food and Music hierarchies in Figures 3b and 18.

F.1 Counts and proportions of recovered categories and adjacent pairs

To complement mean F1, Figures 15 and 16 count categories and parent–child pairs (in gbif, this means adjacent levels of the hierarchy) whose held-out test F1 is strictly greater than 0.6.

F.2 Recovery of letter-prefix hierarchies

The upper panel of Figure 3a shows Gemini test F1 for the `a` parent and its eligible two-letter children, selecting a separate detector for each. The lower recall heatmap instead fixes the selection-chosen `a` detector at each width and measures its recall within each prefix group. That descriptive heatmap uses the full 200,000-word collection, including selection and test rows, rather than only the held-out split.

Figure 17 extends the upper F1 panel to all 25 remaining one-letter prefixes, `b`–`z`, using the same saved probes, nine widths, and category eligibility rules. These curves report continuous held-out test F1 without applying the recovery cutoff above. The `q` prefix has one eligible two-letter child, while `x` has none and therefore shows only its parent curve. We see more instances of mesoscale splitting (a U-shaped parent F1 curve).

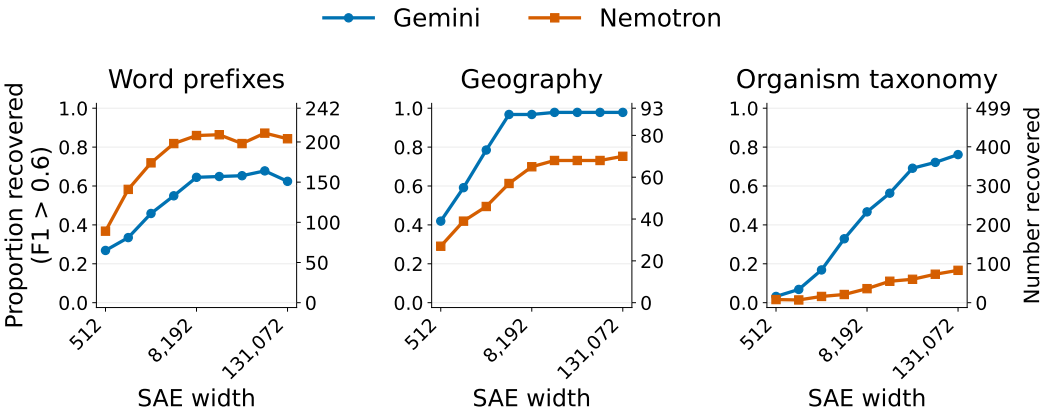


Figure 15: **Recovery of individual categories across SAE widths.** A category is recovered when its selection-chosen detector achieves held-out test F1 > 0.6. Left axes show proportions; right axes show raw counts among 242 word prefixes, 93 geographic categories, and 499 organism categories. Geography uses different continent and country prompts.

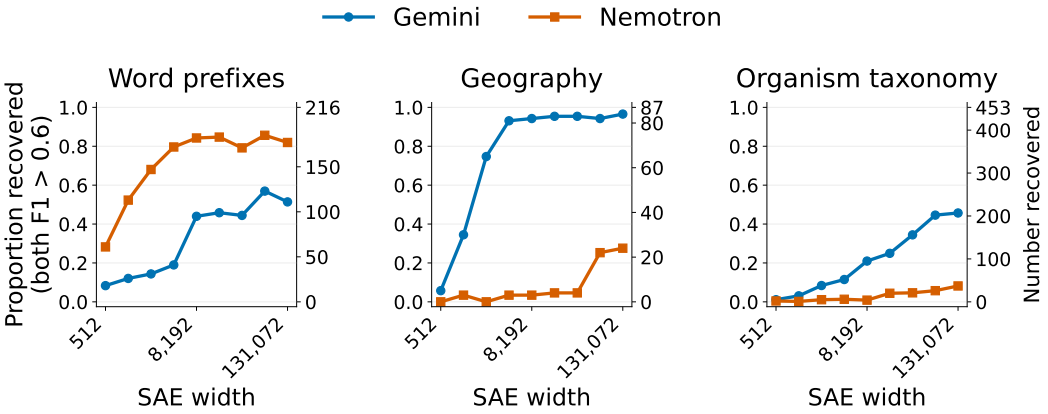


Figure 16: **Simultaneous recovery of adjacent parent–child pairs.** Both selection-chosen detectors must achieve held-out test F1 > 0.6 and use distinct SAE coordinates within the pair. Left axes show proportions; right axes show raw counts among 216 prefix pairs, 87 continent–country pairs, and 453 adjacent-rank organism pairs.

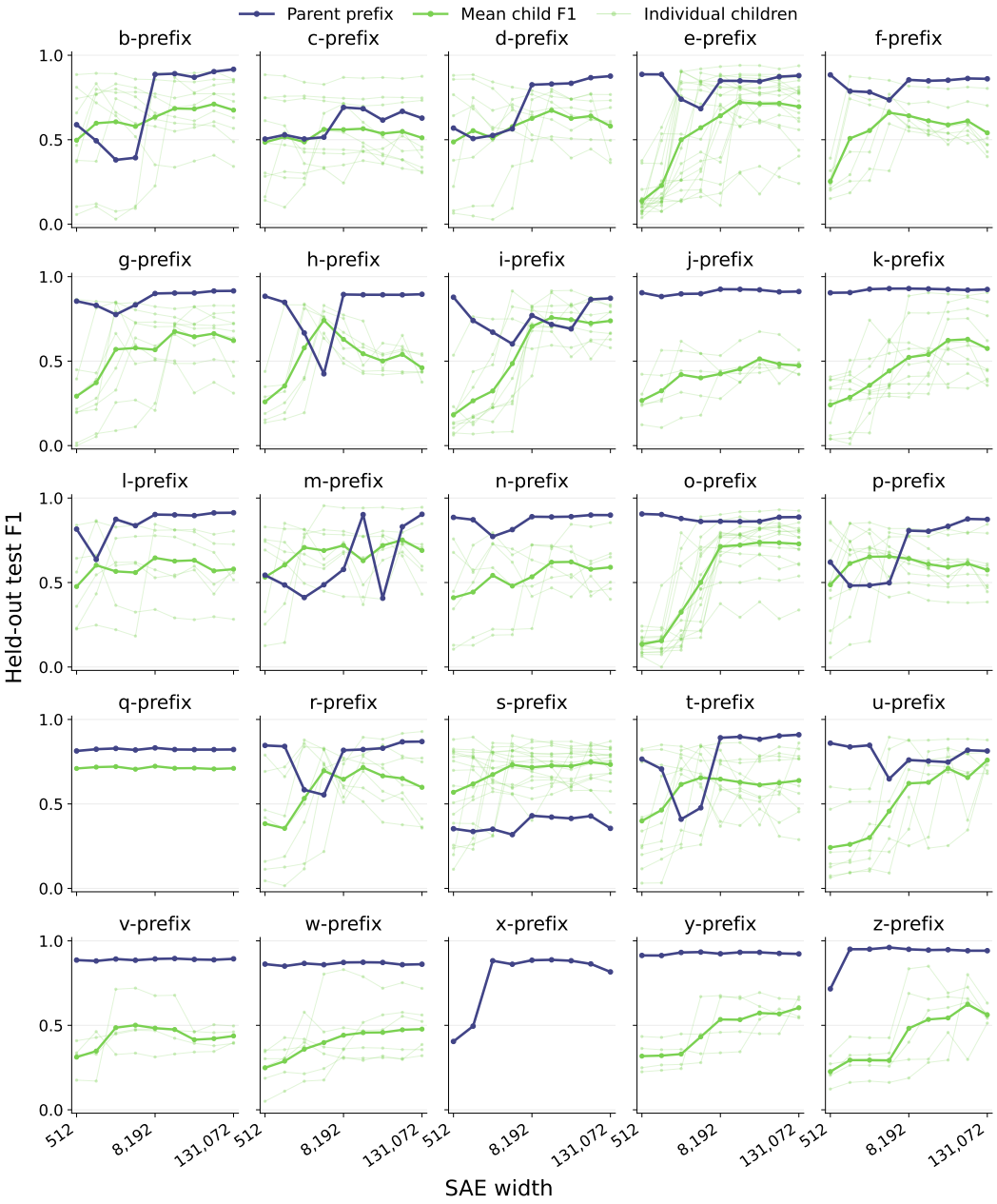


Figure 17: **Gemini F1 recovery for the remaining 25 letter prefixes.** Each panel is the analog of the upper panel of Figure 3a for one prefix from b to z. Blue curves show parent F1; bold green curves show the unweighted mean F1 of eligible two-letter children; faint green curves show individual children.

F.3 EMPIRICAL EXTRACTION OF FEATURE HIERARCHIES

The food and music hierarchies in Figure 18 are curated subsets of a graph extracted from activation co-occurrences in the Gemini SAE with 131,072 features and $k = 128$. We use cached post-TopK activations on $N = 89{,}827{,}558$ corpus rows. A feature is active when its activation is strictly greater than 0.05; this threshold is the same for candidate parents and children. For a candidate parent P and child C, let n_P and n_C denote their activation counts and n_{PC} their co-activation count. We estimate forward and reverse containment as

$$\widehat{p}_{P|C} = \frac{n_{PC}}{n_C}, \qquad \widehat{p}_{C|P} = \frac{n_{PC}}{n_P}. \tag{39}$$

High forward containment indicates that the parent usually activates when the child does; low reverse containment indicates that the parent also covers examples outside the child.

Candidate edges and directionality. We retain a candidate edge $P \to C$ only if $n_C \geq 100$, $n_{PC} \geq 100$, $n_P > n_C$, and

$$L_{P|C} \geq 0.70, \qquad \widehat{p}_{C|P} \leq 0.50, \qquad \frac{\widehat{p}_{P|C}}{n_P/N} \geq 5. \tag{40}$$

The last condition is a lift threshold: containment must be at least five times the parent's marginal activation rate. The reverse-containment and support conditions orient edges from broader to narrower activation sets; the strict support ordering also rules out directed cycles.

Here $L_{P|C}$ is the lower endpoint of the two-sided 95% Wilson interval for a binomial proportion. Writing $\widehat{p} = \widehat{p}_{P|C}$, $n = n_C$, and $z = 1.96$, it is

$$L_{P|C} = \frac{\widehat{p} + z^2/(2n) - z\sqrt{\widehat{p}(1-\widehat{p})/n + z^2/(4n^2)}}{1 + z^2/n}. \tag{41}$$

This penalizes estimates based on fewer observations: for example, $90/100$ and $900/1000$ both give 90% containment, but their lower bounds are approximately 82.6% and 88.0%. We use the bound as a screening heuristic, since corpus rows can be duplicated or dependent and the interval does not adjust for testing many feature pairs.

Removing ancestor shortcuts. After collecting all qualifying edges, we remove $P \to C$ whenever a feature M has both qualifying edges $P \to M$ and $M \to C$. Thus, Food $\to$ Cheddar is removed if Food $\to$ Cheese and Cheese $\to$ Cheddar both qualify. This is a single reduction using two-edge paths in the candidate graph, rather than a full transitive reduction over paths of arbitrary length. Multiple parents can remain, and a broader ancestor edge can remain if an intermediate edge fails the thresholds. The procedure yields 154,316 qualifying ancestor pairs and 114,762 retained parent–child pairs. For inspection, we sort pairs by

$$\max\left\{0, \frac{L_{P|C} - n_P/N}{1 - n_P/N}\right\}\left(1 - \widehat{p}_{C|P}\right), \tag{42}$$

but this ranking does not select a unique parent or determine which shortcuts are removed.

Selection and interpretation of the figures. Each hierarchy in Figure 18 displays 33 selected features and 32 retained edges. The compact main-text table in Figure 3b represents all 66 features and 64 parent–child relationships in three hierarchy-level columns, omitting only the feature IDs. We inspected cached activating texts to choose interpretable branches and assign descriptive labels; Tables 5 and 6 provide example texts. All edge statistics use the full corpus counts, rather than the sampled texts. The diagrams illustrate selected activation-containment relationships, not an exhaustive taxonomy or disjoint partition, and the labels are interpretations rather than independent ground truth.

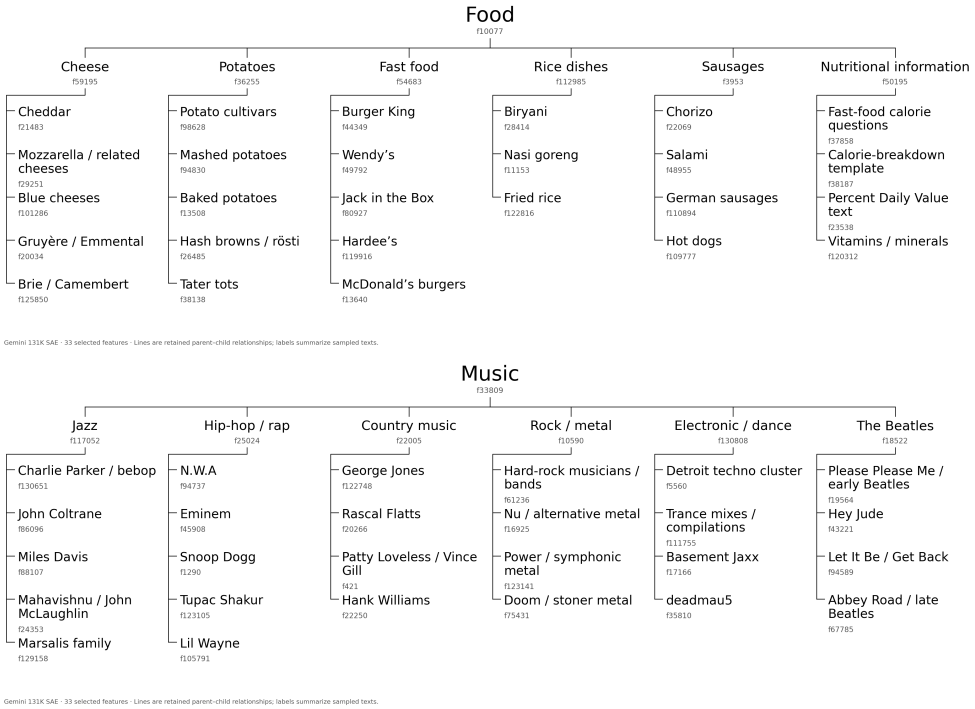


Figure 18: Learned hierarchies for Music and Food. The labels here are automatically generated using Codex; activating examples for each feature in the hierarchy are given in Tables 5 and 6.

Table 5: Food hierarchy with five unique cached text examples per feature.

Food

Food f10077	Food render • Food in England • Food Race • Food and Magic • Culinary diplomacy
Cheese f59195	チーズはいただき • Cheese (album) • Liste de mets à base de fromage • Premium Cheese! • Fromage à gogo
Cheddar f21483	Cheddar • CheddarU • Hombre de Cheddar • how cheddar cheese is made? • Cheddar (Käse)
Mozzarella / related cheeses f29251	what is mozzarella • Mozzarella • 모차렐라 • موتزارلا • モッツァレッラ
Blue cheeses f101286	Bleue d'Auvergne • Roquefort • Bleu d'Auvergne • Gorgonzola • Fourme d'Ambert
Gruyère / Emmental f20034	what is gruyere • Emmental de Savoie • what is gruyere cheese • 3. Jarlsberg. Jarlsberg is Norwegian variant of the cheese having holes and is a Swiss type cheese. This mellow and… • Emmental cheese
Brie / Camembert f125850	Camembert • Coulommiers (Käse) • what does camembert cheese smell like • what's in brie cheese • what gives brie unique taste
Potatoes f36255	Potato Project • Potatoes & You • when is potato day • Похороны картофеля • Jus de pomme de terre
Potato cultivars f98628	Lapin Puikula • Golden Wonder potato • Umatilla Russet • Lenape potato • Golden Wonder (potato)
Mashed potatoes f94830	how long will mashed potatoes stay good in the refrigerator? • do mashed potatoes go bad? • how long can mashed potatoes last? • how long are leftover mashed potatoes good for in the fridge? • do mashed potatoes go bad in the fridge?
Baked potatoes f13508	Check out this recipe for the Perfect Baked Potato: http://www.recipe4living.com/Recipes/Rec... It says to bake it for… • Jacket potato • how to check if baked potato done • Medium Plain Baked Potato: 150 Calories Size matters. A small baked potato (about 2 inches in diameter) has about 130… • how long is a baked potato good in the fridge
Hash browns / rösti f26485	how to oven cook hash browns • Hash Browns • what is rosti potatoes • هاش براونز • 뢰슈티
Tater tots f38138	Tater Tots • With the lack of animal products not bearing any weight on making a decent batch of tots, tater tots are usually… • are tater tots bad for your diet? • are tater tots unhealthy for you? • are tater tots bad for your health?
Fast food f54683	which fast food chain launched a "value" campaign in 1988 that included drive-through • Wendy's • gillian clare mcdonald is the ceo of which fast food chain • which fast food chain did paramount partner with to release movies at participating restaurants • Burger King
Burger King f44349	Burger King • Burger King advertising • Burger King Specialty Sandwiches • who started burger king • burger king hours for breakfast
Wendy's f49792	The Wendy's Company • Wendy's • calories in wendys medium fries • وندیز • how many wendy's are in florida
Jack in the Box f80927	Jack in the Box • what are in jack in the box tacos? • are jack in the box tacos made with meat? • what are jack in the box tacos made with? • are jack in the box tacos made with real meat?
Hardee's f119916	Hardee's • Hardee's • The Carl's Jr near me has this also. Since Carl's Jr and Hardee's are the same company they are probably exactly the… • hardee's breakfast hours • are carl's jr and hardee's the same?
McDonald's burgers f13640	720 Cal. 720 Cal. Each Double Quarter Pounder with Cheese features two quarter pound* 100% fresh beef burger patties… • 510 Cal. 510 Cal. Each Quarter Pounder with Cheese burger features a ¼ lb. * of 100% fresh beef that's hot,… • 510 Cal. 510 Cal. Each Quarter Pounder with Cheese burger features a ¼ lb. • 1250 Cal. 1250 Cal. Get double the fresh beef** flavor with a Double Quarter Pounder® with Cheese made with fresh beef… • Double Quarter Pounder®* with Cheese Meal. 1260 Cal. 1260 Cal. Get double the fresh beef** flavor with a Double…
Rice dishes f112985	Locrio • دمیختک • Galinhada • Pulao • Pelau
Biryani f28414	what vessel is biriyani cooked in • What is the main ingredient in biriani? • where is the movie biriyani being performed • Each community and region of India has its own traditional style of making biryani, some of the most famous styles of… • what is ulavacharu veg biryani?
Nasi goreng f11153	how long is the yangzhou fried rice • Nasi goreng pattaya • Yangzhou fried rice • 1 Mix and cook/fry the rice and egg together (in a wok is best). 2 Add the long green onion, chashaorou/bacon, carrots… • Yangzhou fried rice Yangzhou fried rice or Yeung Chow fried rice (Traditional Chinese: 揚州炒飯; Simplified Chinese :…
Fried rice f122816	Arroz frito Yangzhou • 揚州チャーハン • 揚州炒飯 • Nasi goreng pattaya • Nasi goreng tionghoa
Sausages f3953	Port sausage • Список колбас • sausage • Chinese sausage • Garlic sausage
Chorizo f22069	what is chorizo used for • what is chorizo • Spanish chorizo and Mexican chorizo are very different from each other in looks, texture and taste. Since the Spanish… • Mexican Chorizo. 1 How it's made and sold: Mexican chorizo is a spicy ground meat sausage that is most commonly sold… • Chorizo
Salami f48955	Salami Genoa • calories in hard salami lunch meat • how is salami processed • Yes, salami is very keto-friendly and low-carb. Salami has a very high-fat content and almost no carbs. Just be aware… • what is the difference between hard salami and dry salami?
German sausages f110894	Wollwurst • Bratwurst • Regensburger Würstchen • what is bratwurst • Appenzeller Siedwurst
Hot dogs f109777	Cooking Instructions. Grill: Place over medium heat and turn until browned. Boil: Bring water to a boil, reduce heat,… • For just two or three hot dogs, put them in a shallow bowl with about 1/3 cup of water. Cover with plastic wrap. Be… • To cook hot dogs on the stove top, add up to 10 hot dogs to saucepan of boiling water. Return to boil; cover. Remove… • Ingredients: Mechanically Separated Turkey, Water, Salt, Contains 2% or Less Modified Food Starch, Potassium Lactate,… • How long to cook hot dogs. OSCAR MAYER Franks or Wieners can be cooked in just minutes on the grill. Heat the grill to…
Nutritional information f50195	what is the nutritional value of flounder? • what is the nutritional value of eggplant? • Carbs: 45.6 grams. Fiber: 0.7 grams. Sodium: 399 mg. Folate: 24% of the Daily Value (DV) • The Basics. A 1-cup serving of fresh hearts of palm contains 4 grams of protein, 0.3 grams of total fat, 37 grams of… • what is the nutritional value of corn?
Fast-food calorie questions f37858	how many calories in subway bacon egg and cheese flatbread? • how many calories in chicken noodle soup from panera? • how many calories in a burger king vanilla cone? • how many calories in a kfc donut chicken sandwich? • how many calories are in a mcdonald's egg white?
Calorie-breakdown template f38187	There are 111 calories in 1 cup of Cheerios. Calorie breakdown: 13% fat, 75% carbs, 12% protein. • There are 74 calories in 1 serving of Nescafe 3 in 1 Coffee. Calorie breakdown: 26% fat, 72% carbs, 2% protein. • There are 270 calories in a 1 link serving of Johnsonville Italian Sausage. Calorie breakdown: 73% fat, 4% carbs, 22%… • There are 80 calories in a 1/3 cup serving of Carnation Nonfat Dry Milk Powder. Calorie breakdown: 0% fat, 60% carbs,… • There are 320 calories in 1 serving of Taco Bell Nacho Fries. Calorie breakdown: 51% fat, 44% carbs, 5% protein.
Percent Daily Value text f23538	* Percent Daily Values are based on a 2,000 calorie diet. Your daily values may be higher or lower depending on your… • Percent Daily Values are based on a 2000 calorie diet. Your daily values may be higher or lower depending on your… • Percent Daily Values are based on a 2,000 calorie diet. Your daily values may be higher or lower depending on your… • * Percent Daily Values are based on a 2,000 calorie diet. Your daily values may be higher or lower depending on your… • Percent Daily Values are based on a 2000 calorie diet. Your daily values may be higher or lower depending on your…
Vitamins / minerals f120312	what vitamins and minerals are in red peppers? • which vitamins and minerals are in bananas? • what are the vitamins and minerals in cucumbers? • what vitamins and minerals does cabbage have? • what are the vitamins of mango?

Table 6: Music hierarchy with five unique cached text examples per feature.

Music

Music f33809	مجله موسیقی • موسیقی إسبانیا • Diamond Music • Unduhan musik • BlueBeat Music
Jazz f117052	Kids: Live at Dizzy's Club Coca-Cola • Sunset-Sunside • Sunset/Sunside • Robbins' Nest: Live at the Jazz Showcase • Love and Peace: A Tribute to Horace Silver
Charlie Parker / bebop f130651	Bird and Diz • Charlie Parker Memorial, Vol. 2 • Fats Navarro • Charlie Parker at Storyville • Groovin' High (Dizzy Gillespie album)
John Coltrane f86096	Coltrane for Lovers • コルトレーン • Coltrane (Album) • Coltrane (1962 album) • John Coltrane
Miles Davis f88107	Miles Davis: The Complete Columbia Album Collection • ไมล์ส เดวิส • マイルス・デイヴィス • Дэвис, Майлз • Forever Miles Davis
Mahavishnu / John McLaughlin f24353	マハヴィシュヌ・オーケストラ • Mahavishnu Orchestra • Mahavishnu Orchestra: The Complete Columbia Albums Collection • mahavishnu orchestra vision of the emerald beyond album • John McLaughlin
Marsalis family f129158	Branford Marsalis • Марсалис, Уинтон • Wynton Marsalis • Branford Marsalis Quartet • 윈튼 마살리스
Hip-hop / rap f25024	Hip-hop indonésien • Liste von Hip-Hop-Musikern • Hip-hop espagnol • Something from Nothing: The Art of Rap • Hip hop Tuga
N.W.A f94737	NWA: Straight Outta Compton • N.W.A • N.W.A. • NWA • Straight Outta Compton: N.W.A 10th Anniversary Tribute
Eminem f45908	Eminem • what is the name of the movie that features eminem • Eminem Is Back • Discographie d'Eminem • My Son Marshall, My Son Eminem
Snoop Dogg f1290	Snoop Dogg has worked with Death Row Records. • Snoop Dogg released an album on Death Row Records. • Snoop Dogg • Snoop Dogg comes from California. • Snoop Dogg had a soundtrack released under Death Row Records.
Tupac Shakur f123105	Tupac Shakur • Tupac Shakur Legacy • توباك شاكور • what movie is about tupac shakur • who wrote the book about tupac shakur
Lil Wayne f105791	what is lil wayne's record label called • Tha Carter III • what is lil wayne's label called • what is lil wayne's latest album called • lil wayne sorry 4 the wait album
Country music f22005	Now That's What I Call Country Volume 2 • This is a list of notable events in country music that took place in 2012. • 1988 in country music • My Life's Been a Country Song • List of number-one country singles of 1982 (U.S.)
George Jones f122748	Discographie de George Jones • George Jones Sings His Greatest Hits • 조지 존스 • how old was george jones when he died • George Jones
Rascal Flatts f20266	Discographie de Rascal Flatts • Rascal Flatts • who created rascal flatts music • what is the highest selling rascal flatts album • rascal flatts here comes goodbye album
Patty Loveless / Vince Gill f421	Pam Tillis • Patty Loveless • Suzy Bogguss • Kathy Mattea • pam tillis all the good ones are gone album
Hank Williams f22250	hank wiiliams where did bocephus come from • هانك ويليامز • Hank Williams discography • what was hank williams first hit song • ハンク・ウィリアムズ
Rock / metal f10590	Phil Campbell and the Bastard Sons • Rhino Bucket • The Wildhearts • Taking Dawn • Ugly Kid Joe
Hard-rock musicians / bands f61236	Dokken • Dave Meniketti • Jeff Pilson • Bobby Blotzer • ドン・ドッケン
Nu / alternative metal f16925	Nonpoint • Spineshank • 40 Below Summer • Ill Niño • Stereomud
Power / symphonic metal f123141	Ecliptica – Revisited: 15th Anniversary Edition • Sonata Arctica • Liste der Lieder von Sonata Arctica • The Metal Opera Part II • A Tribute to Sonata Arctica
Doom / stoner metal f75431	Sleep's Holy Mountain • Spirit Caravan • Electric Wizard • The Atomic Bitchwax • Scott Weinrich
Electronic / dance f130808	Antoine Clamaran • Robbie Rivera • Sander Kleinenberg • Mauro Picotto • Ivan Gough
Detroit techno cluster f5560	The Belleville Three • Juan Atkins • Kevin Saunderson • where did the belleville three come from • where did the belleville three first perform
Trance mixes / compilations f111755	A State of Trance 2013 • A State of Trance • A State of Trance 2009 • A State of Trance 2012 • A State of Trance 2010
Basement Jaxx f17166	Basement Jaxx • Discographie de Basement Jaxx • basement jaxx never say never album • when did basement jaxx album come out • ベースメント・ジャックス
deadmau5 f35810	what is the title of deadmau5's fourth album • Deadmau5 Circa 1998-2002 • Full Circle (Deadmau5 album) • when does deadmau5 upload stuff i used to do • Mau5trap
The Beatles f18522	The Beatles • Beatles Day • ビートルズ! • 비틀즈 • Beatles Stories
Please Please Me / early Beatles f19564	The Beatles' Please Please Me album was rush-released by Parlophone on 22nd March, 1963 to capitalise on the enormous… • Please Please Me (альбом) • Please Please Me • Please Please Me (album) • when was please please me released in uk
Hey Jude f43221	Hey Jude • where did the beatles film hey jude • what tv show featured the beatles singing hey jude • how many weeks did hey jude stay on the billboard hot 100 • Hey Jude (Wilson Pickett album)
Let It Be / Get Back f94589	Let It Be • Let It Be (song) "Let It Be" is a song by the English rock band the Beatles, released in March 1970 as a single, and… • Let It Be (песня) • Let It Be (song) • Let It Be (2004 film)
Abbey Road / late Beatles f67785	Abbey Road (album) • Abbey Road (álbum) • The Best of Abbey Road • what is the title of abbey road's last album • who sings lead vocals on best of abbey road

G SIMULATION EXPERIMENTS

G.1 RECOVERY UNDER WIDTH AND SPARSITY VARIATION

We study recovery in finite synthetic distributions with known ground-truth atoms. These experiments assess whether trained TopK SAEs exhibit prefix recovery and how this behavior depends on the frequency distribution and the SAE sparsity. Figures 19 and 20 report approximate recovered prefix lengths, expressed in atoms and divided by SAE width.

Data-generating distributions. Inputs lie in $\mathbb{R}^{512}$ and are nonnegative linear combinations of exactly $K = 8$ atoms. We draw each ground-truth direction independently from a standard Gaussian and normalize it to unit length. In the flat setting, there are $M = 4096$ atoms. An eight-element support S is sampled according to

$$\Pr(S) = \frac{\prod_{i \in S} i^{-\alpha}}{\sum_{T:\,|T|=8} \prod_{j \in T} j^{-\alpha}}, \qquad |S| = 8,$$

and its nonzero coefficients are independent $\mathrm{Uniform}(0, 1)$ draws. This is a product-weight distribution over fixed-size subsets; the power law specifies the support weights, rather than the marginal activation probabilities directly. In the hierarchical setting, $M = 3072$ atoms form 1024 families, each containing a parent and two children. We sample four families using the same product-weight law over family ranks, then activate each selected parent and one of its two children, chosen uniformly. The eight active coefficients are again independent $\mathrm{Uniform}(0, 1)$ draws. Thus a child can only occur with its parent, while either child can accompany the parent.

Training and sweeps. We train nonnegative TopK SAEs at widths $m \in \{128, 256, 512, 1024, 2048, M\}$. Generating sparsity $K = 8$ is fixed throughout; k denotes the SAE's activation budget. The frequency sweep uses $\alpha \in \{0.8, 1.2, 1.6, 2.0, 2.4\}$ at $k = 8$. The sparsity sweep uses $k \in \{1, 2, 4, 6, 8, 10, 12, 16, 32\}$ at $\alpha = 1.6$. The finite distribution at $\alpha = 0.8$ is well-defined, but serves as a control outside the theorem's assumption $\alpha > 1$. We use three paired random seeds per setting, giving 468 distinct training trajectories after deduplicating the shared $\alpha = 1.6, k = 8$ setting.

Each trajectory trains on 1,024M examples using Adam with learning rate 10^{-3}, $\beta_1 = 0.9$, $\beta_2 = 0.99$, and batch size 1024, with no learning-rate schedule. Decoder directions have unit norm and the decoder bias is fixed to zero. An auxiliary reconstruction loss with coefficient 0.25 targets up to 64 neurons inactive for 256 steps. Model and optimizer checkpoints are retained at cumulative budgets of $1, 2, 4, 8, 16, 32, 64, 128, 256, 512$, and 1024 million examples; the figures use the final checkpoint.

Recovery criteria. Geometric recovery requires absolute cosine similarity at least 0.9 with a learned decoder direction. For hierarchical data, child comparisons project both directions onto the orthogonal complement of the true parent; recovery of a complete family requires three distinct decoder directions meeting the parent and two projected-child criteria. Activation recovery uses 200,000 independent validation examples to select coordinate thresholds and a maximum-cardinality injective assignment among atom–coordinate pairs with validation F1 at least 0.8. A matched atom is recovered if its F1 is also at least 0.8 on a separate set of 400,000 test examples. An activation-recovered family must recover all three members.

For either criterion, let q_i indicate recovery of atom i in the flat setting or complete family i in the hierarchical setting, ordered by generating weight. We compute

$$r = \max\left(\{0\} \cup \left\{s : \sum_{i=1}^{s} q_i \geq 0.9s\right\}\right).$$

The plotted quantity is r/m for flat data and $3r/m$ for hierarchical data. It measures the size of an approximate prefix in atom units relative to SAE width, rather than total recovery across all ranks or the fraction of the ground-truth population recovered. The 10% miss allowance means this ratio can exceed one. Curves show means over three seeds and shaded bands show their minima and maxima.

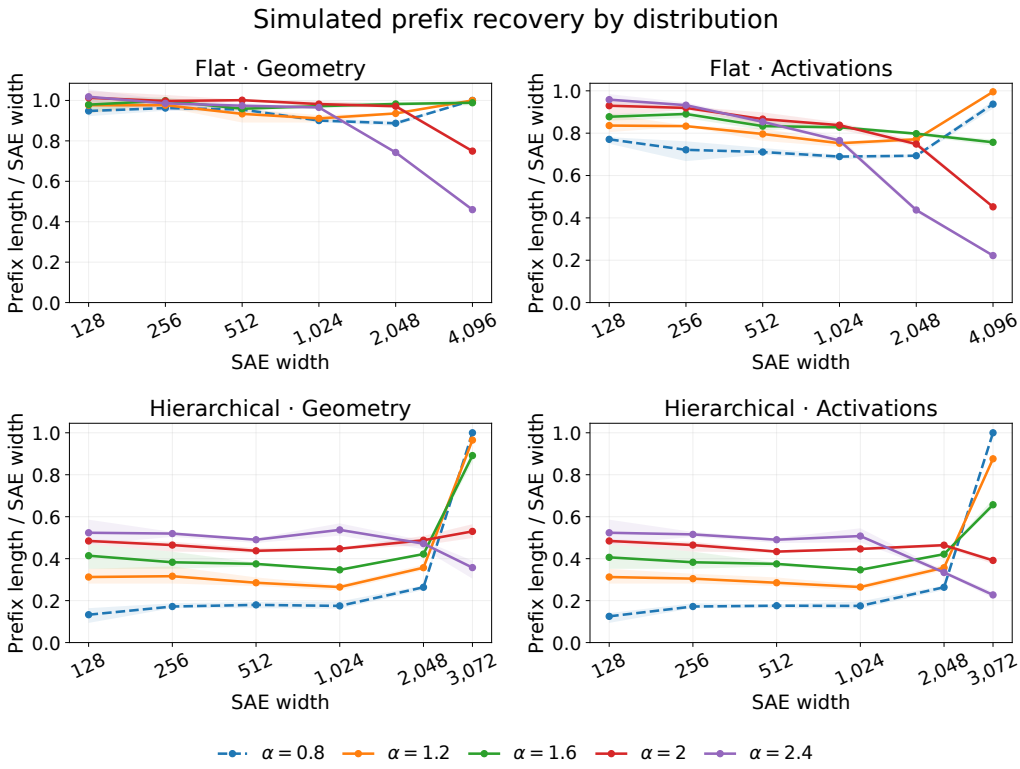


Figure 19: **Width-normalized prefix recovery across frequency distributions.** Generating sparsity $K = 8$ and SAE sparsity $k = 8$ are fixed; each curve corresponds to a different α. All models train on 1,024M examples. Rows show flat and hierarchical data; columns show geometric and activation recovery. The vertical axis is approximate prefix length in atoms divided by SAE width (three atoms per complete family for hierarchical data), with a 10% miss allowance. Lines show three-seed means and bands show seed ranges.

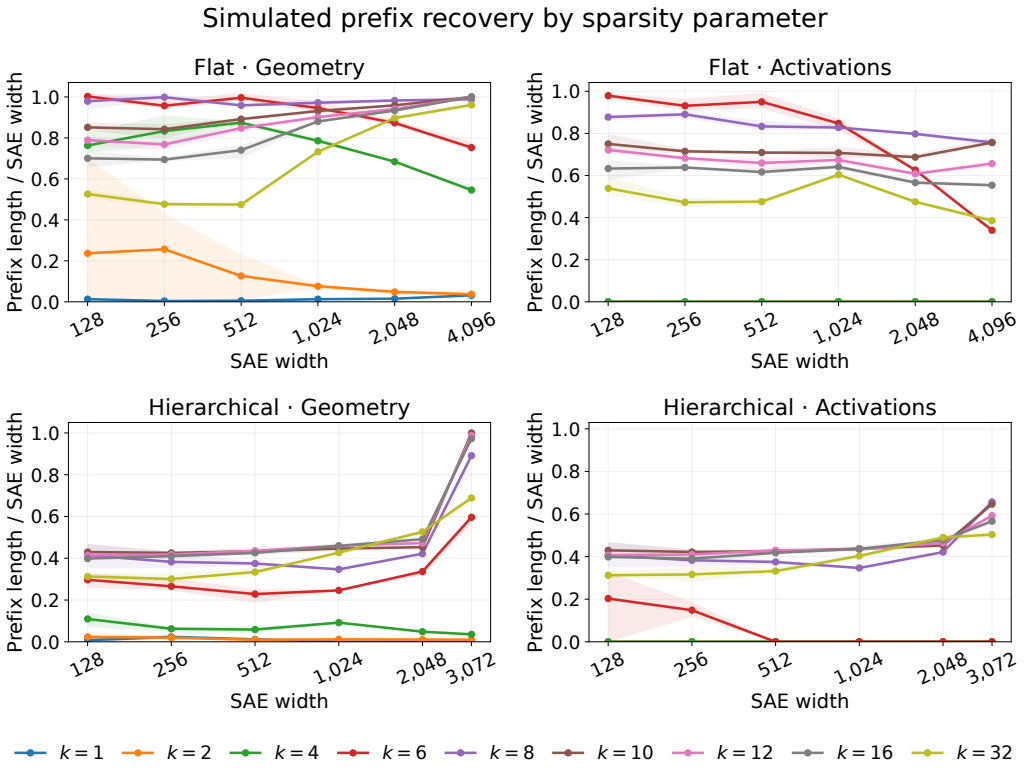


Figure 20: **Width-normalized prefix recovery across SAE sparsities.** Generating sparsity $K = 8$ and frequency exponent $\alpha = 1.6$ are fixed; curves vary SAE sparsity k. All models train on 1,024M examples. Panels, normalization, and the 10% prefix miss allowance follow Figure 19. Lines show three-seed means and bands show seed ranges.

G.2 Persistent features across simulated SAE widths

We apply the persistence criterion of Figure 2a to the final checkpoints above. For each source width m, we find the largest common subset of learned features that admits a separate injective matching into every larger trained SAE, holding the data distribution, SAE sparsity, and paired seed fixed. Matches require signed decoder cosine similarity or signed Pearson correlation of continuous post-TopK activations at least $t = 0.8$. The largest width (4096 for flat data, 3072 for hierarchical data) is a matching target only. We report the persistent-feature proportion, dividing the size of the common subset by all m source features, including inactive features.

We evaluate each group of six widths on the same 400,000 held-out test examples and compute all pairwise similarities without sketching or candidate truncation. Constant activation columns cannot form Pearson edges. The common-subset solver in Section C.2.3 gives exact maxima on these empirical threshold graphs. This measures persistence between learned SAEs.

Figure 21 varies the frequency exponent; Figure 22 varies SAE sparsity. At $\alpha = 1.6, k = 8$, mean activation persistence across source widths ranges from 93.6–96.1% for flat data and 91.9–99.1% for hierarchical data. Hierarchical decoder persistence rises from 61.5% at width 128 to 91.8% at width 2048. Severe sparsity underspecification particularly reduces activation persistence, showing that persistent decoder directions need not imply persistent activation patterns.

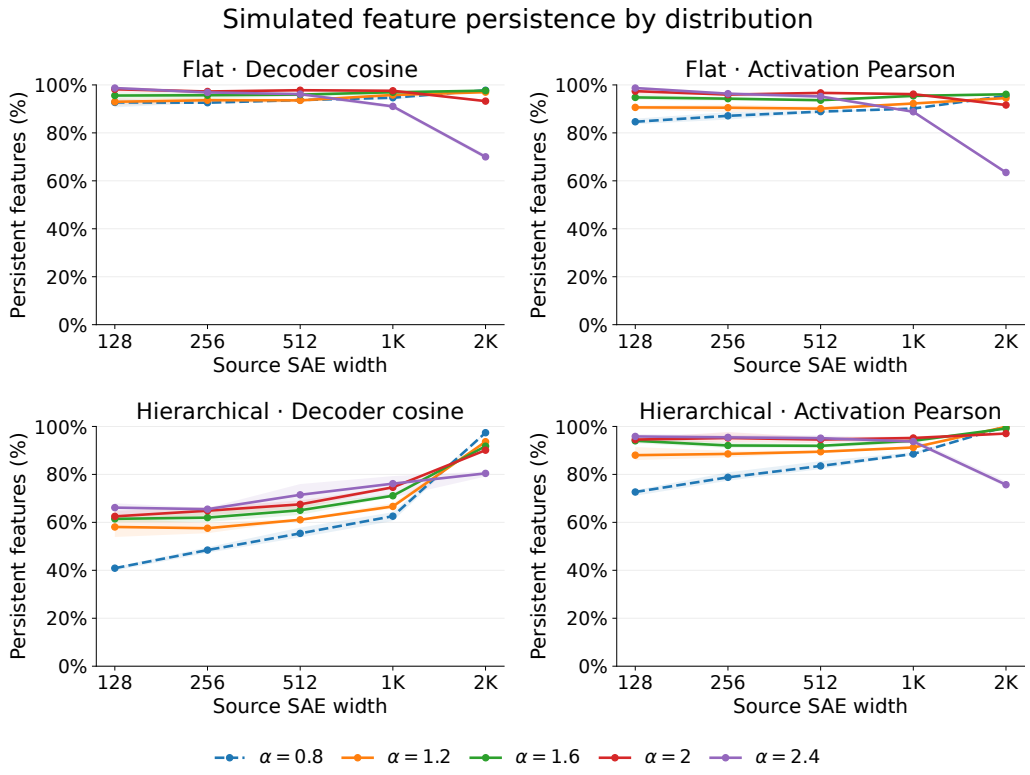


Figure 21: **Persistent-feature proportions across frequency distributions at $t = 0.8$.** The largest common source subset injectively matchable into every larger trained width, divided by source SAE width. Generating sparsity $K = 8$ and SAE sparsity $k = 8$ are fixed; colors vary α. Rows show flat and hierarchical data; columns show signed decoder cosine and activation Pearson similarity. All models train on 1,024M examples. Lines show three-seed means and bands show seed ranges. The dashed $\alpha = 0.8$ curve is the finite-distribution control outside the theorem's assumptions.

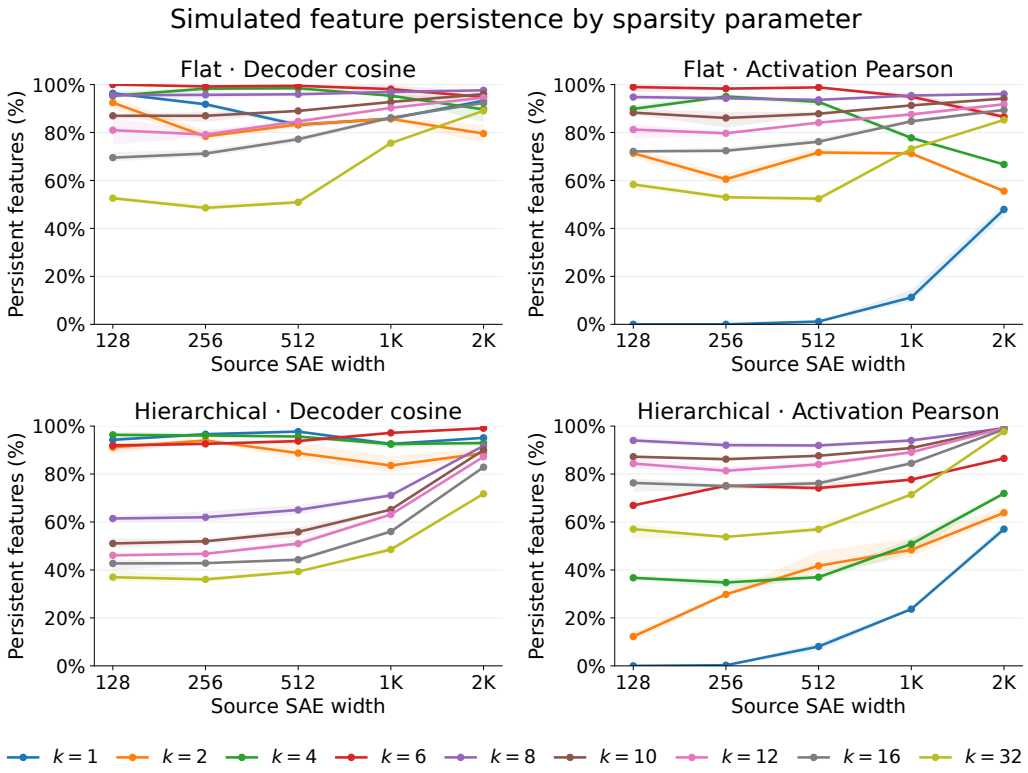


Figure 22: **Persistent-feature proportions across SAE sparsities at $t = 0.8$.** The largest common source subset injectively matchable into every larger trained width, divided by source SAE width. Generating sparsity $K = 8$ and frequency exponent $\alpha = 1.6$ are fixed; colors vary SAE sparsity k. Rows show flat and hierarchical data; columns show signed decoder cosine and activation Pearson similarity. All models train on 1,024M examples. Lines show three-seed means and bands show seed ranges.

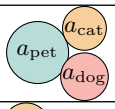
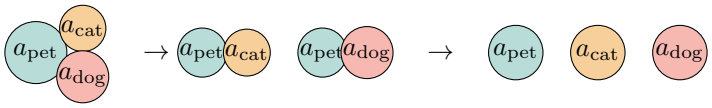
dict. size	i	$\hat{a}_i$ with molecule illustration	$\hat{z}_i(\text{cat})$	$\hat{z}_i(\text{dog})$	$\hat{A}\hat{z}(\text{cat})$	$\hat{A}\hat{z}(\text{dog})$
1	1	$\frac{2}{\sqrt{6}}a_{\text{pet}} + \frac{1}{\sqrt{6}}a_{\text{cat}} + \frac{1}{\sqrt{6}}a_{\text{dog}}$ 	$\frac{3}{\sqrt{6}}$	$\frac{3}{\sqrt{6}}$	$\frac{2a_{\text{pet}} + a_{\text{cat}} + a_{\text{dog}}}{2}$	
2	1	$\frac{1}{\sqrt{2}}a_{\text{pet}} + \frac{1}{\sqrt{2}}a_{\text{cat}}$ 	$\sqrt{2}$	0	$a_{\text{pet}} + a_{\text{cat}}$	$a_{\text{pet}} + a_{\text{dog}}$
	2	$\frac{1}{\sqrt{2}}a_{\text{pet}} + \frac{1}{\sqrt{2}}a_{\text{dog}}$ 	0	$\sqrt{2}$		
3	1	a_{pet} 	1	1	$a_{\text{pet}} + a_{\text{cat}}$	$a_{\text{pet}} + a_{\text{dog}}$
	2	a_{cat} 	1	0		
	3	a_{dog} 	0	1		

Table 7: Learned directions illustrate molecular splitting. Corresponding learned features illustrate mesoscale feature splitting and absorption: at size 2 only, there is no learned feature that is active on both cat and dog.

H A Molecular Model of SAE Scaling

We consider a richer view of the recovery principle, in which non-atomic learned directions can be approximated as molecules: linear combinations of a small number of atomic directions. The theory and results here are intended to be illustrative, rather than fully developed.

The intuition for molecules is that if two atomic features $z_i(x)$ and $z_j(x)$ are heavily correlated over $\mathcal{D}$, then it can be efficient for $\hat{A}$ (trained to minimize $\mathcal{L}_{\mathcal{D}}$) to mix the atomic directions a_i and a_j into a single learned direction, saving budget in terms of dictionary. As budget increases, as per the recovery principle, it becomes optimal to separately represent atoms.



(43)

A three-atom instance of molecule splitting. Consider a stylized example with three atomic directions $a_{\texttt{pet}}, a_{\texttt{cat}}$, and $a_{\texttt{dog}}$ and three atomic features $z_{\texttt{pet}}, z_{\texttt{cat}}$, and $z_{\texttt{dog}}$ (which measure how much an input is about pets, cats, and dogs respectively). So an input is represented by

$$f(x) = Az(x) = z_{\texttt{pet}}(x)a_{\texttt{pet}} + z_{\texttt{cat}}(x)a_{\texttt{cat}} + z_{\texttt{dog}}(x)a_{\texttt{dog}}. \tag{44}$$

Now consider a distribution $\mathcal{D}$ that is heavily concentrated on the inputs "cat" and "dog," which have representations

$$f(\texttt{cat}) = a_{\texttt{pet}} + a_{\texttt{cat}} \qquad f(\texttt{dog}) = a_{\texttt{pet}} + a_{\texttt{dog}}. \tag{45}$$

(Notice also that the atomic direction corresponding to the cat feature is not the same as the representation of the cat input, which also contains the atomic pet direction.)

Table 7 shows optimal choices of learned atomic features for dictionary sizes $1, 2$, and 3, illustrating feature splitting. In particular, it illustrates mesoscale feature splitting, in which the parent pet feature is lost in the middle-budget regime, but returns in the large-budget regime.

Low budget: Full family molecule. At size 1, there is a full-family molecule, that is utilized whenever either cat or dog is present. With only one budgeted feature, the optimal choice is to have a single fully family molecule that is active for both cats and dogs. This means that $\hat{z}_1 \approx z_{\texttt{pet}}$.

Medium budget: Parent-child molecules; feature splitting. At size 2, there are now separate parent-child molecules. Now, notice, that $\hat{z}_1$ is only active for cat; if it were active for dog, then this would unnecessarily write the cat direction into the representation of dog. Likewise, $\hat{z}_2$ is only active for dog. In this setting, $\hat{z}_1 \approx z_{\texttt{cat}}, \hat{z}_2 \approx z_{\texttt{dog}}$, but now there is no feature that is active on both

dict. size	i	$\hat{a}_i$ decoder molecule	recall on ah	recall on ab	recall on ap	recall on as
512	108	$0.91\,a^*_{\mathrm{a}} + 0.11\,a^*_{\mathrm{as}} + 0.08\,a^*_{\mathrm{ab}} + 0.03\,a^*_{\mathrm{ap}} + 0.02\,a^*_{\mathrm{ah}}$	1.000	0.962	0.877	0.993
2048	1457	$0.92\,a^*_{\mathrm{a}} + 0.10\,a^*_{\mathrm{ah}}$	1.000	0.236	0.339	0.388
	1794	$0.85\,a^*_{\mathrm{ab}} + 0.36\,a^*_{\mathrm{a}}$	0.000	0.873	0.003	0.000
	413	$0.84\,a^*_{\mathrm{ap}} + 0.34\,a^*_{\mathrm{a}}$	0.000	0.003	0.887	0.000
	183	$0.79\,a^*_{\mathrm{as}} + 0.36\,a^*_{\mathrm{a}}$	0.029	0.003	0.000	0.783
131K	38618	$1.00\,a^*_{\mathrm{a}}$	1.000	0.950	0.932	0.980
	121657	$1.00\,a^*_{\mathrm{ah}}$	0.871	0.000	0.000	0.000
	100965	$1.00\,a^*_{\mathrm{ab}}$	0.000	0.749	0.000	0.000
	120576	$1.00\,a^*_{\mathrm{ap}}$	0.000	0.003	0.798	0.000
	32170	$1.00\,a^*_{\mathrm{as}}$	0.029	0.000	0.000	0.765

Table 8: Empirical a-prefix feature splitting viewed in terms of molecules.

cat and dog—i.e., there is no feature that approximates the atomic feature $z_{\mathtt{pet}}$.

High budget: Parent and child atoms recovered. Finally, at size 3, it is possible to separately include each of the atomic directions, in which case $z_1 = z_{\mathtt{pet}}, z_2 = z_{\mathtt{cat}}, z_3 = z_{\mathtt{dog}}$, so we recover both the atomic directions and atomic features.

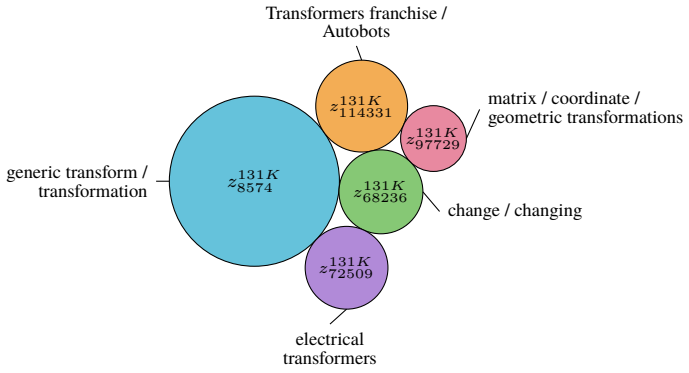
Molecular explanation of feature splitting and absorption. Table 8 shows an empirical analog of Table 7, where we have restricted our attention to a subset of proxy atoms that correspond to one- and two-letter "a" prefixes. A very similar pattern holds where at small dictionary size, there is a large molecule that is active for ah, ab, ap, and as. At intermediate size, there are instead molecules that mix the broad parent atom with the individual ah, ab, ap, and as child atoms. This is the splitting/absorption regime where there is no feature that is active for the whole family. Finally, at largest width, we have recovery of individual atoms.

Molecular explanation of polysemanticity. Oftentimes, SAE features can seem polysemantic, mixing loosely related concepts. This can be understood in the molecular view, where correlated features can be mixed together in small SAEs, but eventually split in larger SAEs. Table 9 gives an example of a direction in the 4K gemini SAE that decomposes into five 131K gemini SAE directions (which we treat as proxy atoms). The table gives example activating texts for the 4K feature, as well as the activations of the corresponding 131K features. The feature mixes different "transform" related features: including mathematical transforms and the Transformers franchise. The molecule also includes a broader molecule corresponding to the concept of "change."

Molecular explanation of correlation in learned feature directions (feature clusters). Various papers report seeing clusters of learned SAE directions that are correlated (Templeton et al., 2024). The model we have presented here suggests that this is evidence that molecules are being learned rather than atoms. In underbudgeted regimes (in either dictionary size or description length) mixing together atomic directions can be efficient. This observation is consistent with results by Chanin et al. (2026), who show that mis-specified k leads to undesirable correlation in learned feature directions.

A broad implication is that many undesirable aspects of learned SAEs are explained by molecules. So problems can be solved by scaling SAEs, so that molecules split into their constituent atoms.

Table 9: Selected feature-3290/4K activating texts and their activations on the five 131K atoms. This feature appears polysemantic, covering both mathematical notions of transforms and mentions of the Transformers franchise (without the word transform).



z^{4K}_{3290}	Feature-3290 activating text	z^{131K}_{8574}	z^{131K}_{68236}	z^{131K}_{114331}	z^{131K}_{72509}	z^{131K}_{97729}
0.387	This is a list of transforms in mathematics.	0.252	0.087	0.000	0.067	0.014
0.330	what is the importance of data transformation?	0.255	0.102	0.000	0.009	0.000
0.318	A transformation group is a group G that acts on a set X by bijections. In other words, for each element g in G, there ...	0.250	0.047	0.000	0.000	0.000
0.285	A non-linear transformation is a mathematical function that doesn't follow the rules of linearity. In simple terms, a linear ...	0.211	0.068	0.000	0.000	0.014
0.257	To transform the given sum, utilize the decomposition dfrac 1(k(k+1)(k+1)!)=left(dfrac 1(kcdot k!)-dfrac 1((k+1)(k+1)!) ...	0.217	0.124	0.000	0.000	0.000
0.163	The Metamorphosis (German: Die Verwandlung) is a novella written by Franz Kafka which was first published in 1915.	0.085	0.078	0.000	0.000	0.000
0.148	what is ctrl t in photoshop?	0.079	0.043	0.000	0.000	0.000
0.204	... voice actors Peter Cullen and Hugo Weaving voice Optimus Prime and Megatron respectively.	0.030	0.000	0.210	0.000	0.000
0.169	are autobots stronger than decepticons?	0.000	0.000	0.242	0.032	0.000
0.338	A transformer is a device that transfers electrical energy from one circuit to another by magnetic coupling.	0.165	0.055	0.036	0.299	0.000
0.186	why crgo core is used?	0.055	0.022	0.000	0.177	0.000
0.168	Translation: 15 units left, 3 units down; Reflection: across x-axis; Scale: along y-axis by a factor of 2.	0.075	0.061	0.000	0.000	0.063
0.152	1. Reflect about the x-axis: - A(1, 5) → A'(1, -5) - B(0, 1) → B'(0, -1) 2. Translate vertically by 3: - A'(1, -5) → A''(1, -2) - B'(0, -1) → B''(0, 2) 3. Rotate about the origin...	0.037	0.075	0.000	0.000	0.097

Table 10: Activating examples for the five 131K atoms. Each column contains twelve examples, sorted by decreasing activation. These appear more monosemantic than the feature in Table 9.

8574	114331	68236	72509	97729
Transform (0.281)	Transformers: Cybertron (0.288)	is gewijzigd in of naar? (0.320)	Transformer (0.273)	* To multiply Row 1 by 3, use the elementary matrix: “‘ [3 0 0 0] [0 1 0 0] [0 0 1 0] [0 0 0 1]... (0.129)
meaning of transformed (0.273)	who does optimus prime fight in transformers autobots (0.270)	تغيير (0.313)	Transformer types (0.265)	The concept lies in the fact that every elementary row operation can be represented as a matrix m... (0.117)
Transformation (Politikwissenschaft) (0.265)	Transformers: Prime (0.250)	Cambios de imagen (0.265)	how does transformer work (0.265)	1. Reflect about the y-axis: The new endpoints become (1, 4) and (-2, 3). <img>https://useruploa... (0.117)
when was the transformation theory introduced in quantum mechanics (0.236)	what is the name of the god in transformers (0.245)	when was the roman liturgical celebration changed (0.247)	Why do transformers exhibit voltage ratios different from their specified values when connected i... (0.239)	Following the steps from the previous example, we first find the reflection matrix for the line y... (0.111)
To transform sqrt(2) into 1+frac(1)(sqrt(2)+1), we can follow these steps: 1. Start with ... (0.232)	Transformers: The Score (0.237)	how do i change my azure subscription? (0.239)	The buzzing noise from the transformer is caused by the rapid fluctuations of magnetic fields, ty... (0.238)	To rotate a vector around a point (2, 3) by 30 degrees, follow these steps: 1. Translate the vect... (0.108)
what are the steps in graph transformation called (0.197)	Transformers (film) voices of Ryan, Cullen, and Dunn, as well as Welker as Megatron. The Target c... (0.197)	when did the grammy award for best traditional pop vocal album change (0.231)	What are the consequences of mistakenly short-circuiting both the primary and secondary sides of... (0.228)	The Hadamard gate represents a 180-degree rotation around the axis connecting the states |0⟩ and... (0.104)
The transformations in the generating function are based on two key concepts: the derivative of a... (0.187)	a unit commanded by Beast Wars Megatron, who ordered them to attack following the deaths of Optim... (0.197)	If you want to change the unit vector vec a to vec a’ with a distance d’ from vec b, n... (0.188)	How should I wire a center-tapped transformer with input voltages of 0V and 230V and output volta... (0.199)	To obtain the graph of y=-2f(x) from the graph of f(x), we can perform the following transfor... (0.102)
what is oncogenic transformation (0.186)	Transformers: Exodus Cybertron where each individual determining their own path in life highly ap... (0.194)	when did the west coast hockey league change to the east coast (0.177)	what type of transformer is needed to achieve the nominal impedance of valve outputs (0.186)	To represent the combined rotation of a 3D point (x, y, z), you can multiply the rotation matri... (0.102)
A Reflection is a transformation in which the figure is the mirror image of the other. 1 Reflecti... (0.142)	Greg “ M Sipher ” Sepelak is an American artist and writer for various Transformers comic ventu... (0.149)	The 2015 expenditure for Army research, development and acquisition changed from 32 billion proj... (0.141)	Commonly, transformers are used to increase or decrease the voltages of alternating current in el... (0.182)	To transform a vector from System (1) to System (2) using intrinsic rotations, follow these s... (0.101)
Is it correct to use Tietze transformations to transform a group into the group langle a, b, c... (0.121)	Sentinel Prime (0.148)	[’Open your Google Account. You might need to sign in.’, ’Under ”Security,” select Signing in to... (0.140)	Control transformers are used to provide single point wiring when the control voltage differs fro... (0.149)	In the context of rigid body transformations preserving the centroid at the origin, the transform... (0.052)
At transform plate boundaries plates grind past each other side by side. This type of boundary se... (0.095)	Barricade is the name of four fictional characters in the various Transformers universes. (0.139)	The integrals span one full period of the DTFT, which means that the x[n] samples are also the co... (0.097)	Earthing transformers are used to create a neutral point in a three-phase system, which provides... (0.140)	Find the transformations applied to the function f(x)=frac1x to obtain the function y=dfr... (0.040)
What is the correct interpretation of a Wick rotation, specifically regarding the transformation... (0.036)	에이수스 이패드 트랜스포머 (0.084)	Constant Change Constant Change is the sixth studio album by Filipino singer-songwriter Jose Mari... (0.086)	Will a step-down transformer work for using my US electronic devices (LED lights, string lights,... (0.137)	How can I effectively visualize Lorentz transformations in the context of Minkowski spacetime? Sp... (0.018)

I AN ATOMIC VIEW OF FEATURE MANIFOLDS

A growing body of work has observed that features do not necessarily exclusively have one-dimensional representations. For example, considering two-dimensional PCA projections of days of the year, one often recovers circular structure. The observance of such "manifolds" is sometimes used to argue against the linear representation hypothesis (which the atomic hypothesis strengthens).

Here, we point out that the atomic hypothesis is not inconsistent with the observation of higher-dimensional non-linear structure. However, the atomic hypothesis can be viewed as implying that it should be possible to reconstruct this non-linear structure using atomic features. In particular, if we believe that the information represented by f is captured entirely by z, we should see the same non-linear structure arise in z. Looking exclusively at z disentangles the information stored in the features themselves from information potentially stored by the directions corresponding to each feature.

We consider subsets of SAE features $S \subseteq [131072]$ and consider PCA projections of $z_S(x) \in \mathbb{R}^M$ for $x \in X^*$. We demonstrate that this procedure reveals geometric structure from atomic features. Notice this procedure only uses the learned features z and not the learned directions A. Importantly, this disentangles geometric structure arising from the embedding space (i.e., if different atoms are represented along correlated directions in $\mathbb{R}^d$) from geometric structure arising purely from correlation in feature activations in data.

We study two datasets: dates and colors. Dates consists of the 366 strings "[Month] [Day]" for each day of the year (e.g., "January 1" or "July 17"). Colors consist of 374 hex codes "#[code]". We plot the PCA of the raw embeddings. We then select a subset S of 131K SAE features and plot the PCA of the feature representations $z_S(x)$. These are shown in Figure 23.

Table 11 plots top activating texts for each of the color features, as well as the top activating artworks from the WikiArt dataset. (The WikiArt dataset is outside of the SAE training corpus, and also consists exclusively of images. The result shows that the features used to construct the color manifold in Figure 23 generalize beyond hex codes to both multilingual text and images.)

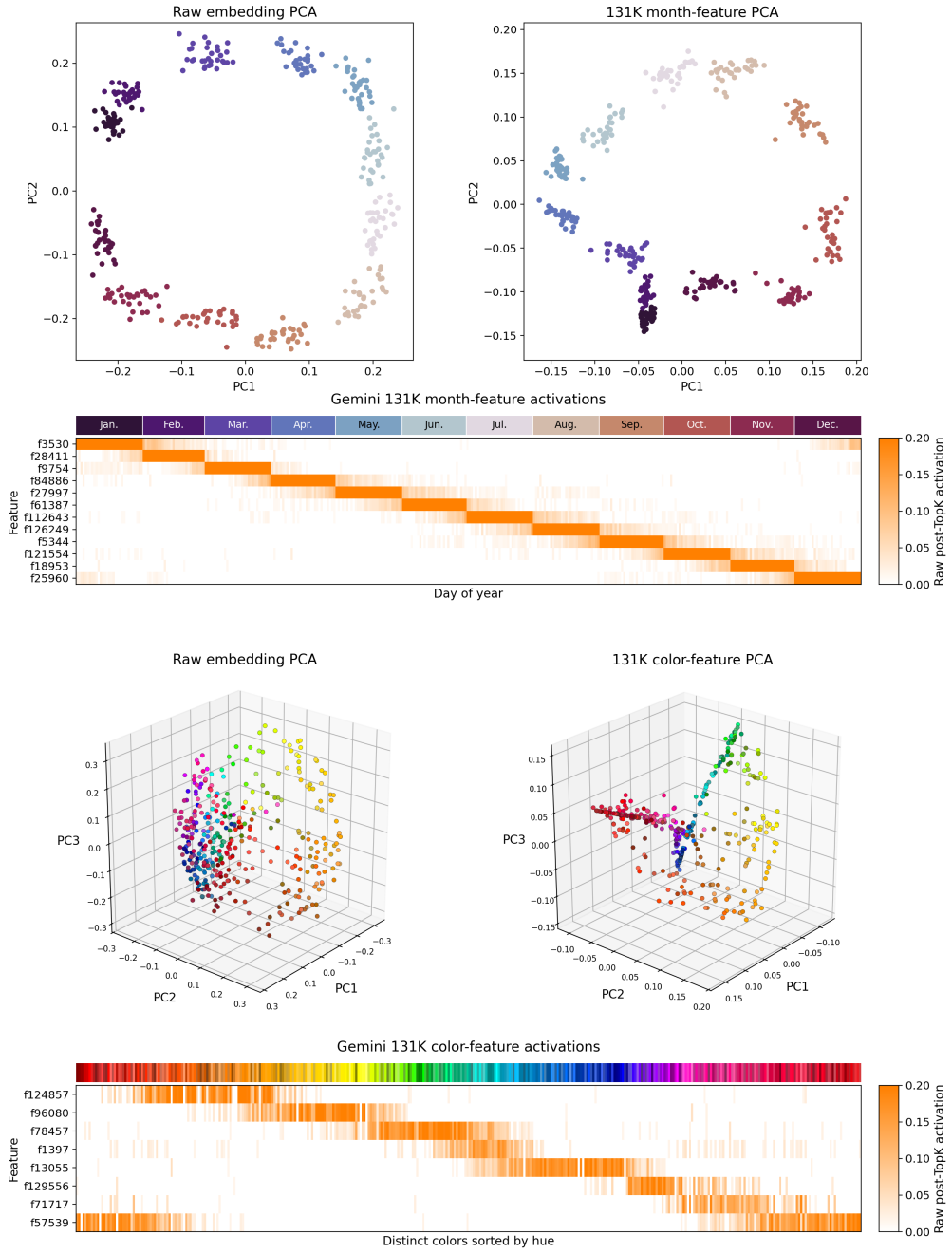


Figure 23: Manifolds arising from small sets of 131K gemini SAE activations. Top left shows PCA of raw embeddings, top right shows PCA of the SAE feature representation restricted to a subset of coordinates, and the bottom shows the activation pattern on these selected features. Importantly, the SAE-induced manifolds arise purely from correlation in activations, not correlation in feature directions.

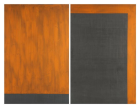
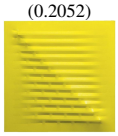
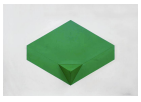
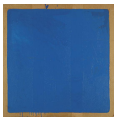
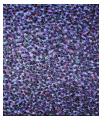
	top activating texts	top activating artwork				
f124857	Оранжевый — хит сезона (0.4119) Orange Record Label (0.3469) Orange Hat (0.3398) Orange Link (0.3359) Orange ribbon (0.3344)	 (0.2292)	 (0.2244)	 (0.2110)	 (0.2063)	 (0.2052)
f96080	Yellow hat (0.3367) Yellow Arrow (0.3251) Yellow monitor (0.3236) Yellow dog (0.3185) Yellow Hair (0.3169)	(0.1967)	 (0.1800)	 (0.1798)	(0.1796)	 (0.1787)
f78457	Green hat (0.3725) Green and Silver (0.3619) Green Rose (0.3583) Green S (0.3556) Green Light (0.3518)	 (0.1837)	 (0.1826)	 (0.1816)	 (0.1791)	 (0.1762)
f1397	Turquoise is not a color choice you make for your car. It chooses you! I've been... (0.2814) Turquoise ribbon (0.2587) 「ターコイズブルー」(C90 Y20) (0.2533) turquoise number (0.2505) Colors That Go With Turquoise. ... Different shades of turquoise don't have to match... (0.2317)	 (0.1757)	 (0.1655)	 (0.1618)	 (0.1565)	 (0.1511)
f13055	Blue Rainbow (0.3702) Blue hat (0.3697) Blue Girl (0.3693) Blue Snowman (0.3691) Blue Monkey (0.3658)	 (0.1800)	 (0.1731)	 (0.1716)	 (0.1708)	 (0.1637)
f129556	Purple Numbers (0.3519) Purple crocodile (0.3409) i feel like rich purple and gold are a match made in heaven and this reinforces that... (0.3329) Purple Love (0.3328) Purple Radio (0.3317)	 (0.2225)	 (0.2112)	 (0.1996)	 (0.1966)	 (0.1950)
f71717	Pink Subaru (0.3593) Pink contract (0.3525) Pink Pineapple (0.3458) Pink and Gray (0.3367) Pink Man (0.3366)	 (0.2032)	 (0.1902)	 (0.1841)	 (0.1840)	 (0.1839)
f57539	Красное на красном (0.4016) Red Elephant (0.3942) Red Bridge (0.3836) Red Post (0.3779) Red jersey (0.3765)	 (0.1869)	 (0.1832)	 (0.1821)	 (0.1800)	 (0.1792)

Table 11: Top-activating text and image examples for eight color-selective features in the Gemini 131K sparse autoencoder. Each row shows the five highest-activating distinct general-corpus texts and five highest-activating WikiArt images for one feature, ordered by decreasing activation within each corpus. Parentheses give raw post-TopK activations. Texts are truncated for display.

J An Atomic View of the Platonic Representation Hypothesis

If we interpret the rigidity property at an even higher level, then we can arrive at an even stronger theory: that atomic features are not atomic only with respect to an individual model, but are actually atomic as descriptions of the world (or at least, of descriptions of human-generated text). We might describe these atomic features as also being "platonic" (Huh et al., 2024).

So while there is no reason for atomic directions to be model specific (one can have an equivalent representation that is a rotation), it seems possible that z could be model independent, in which case different representations f, f' are just different projections of the feature representation: that is, $f = Az$ and $f' = A'z$.

Same-width cross-model matching. We compare activations between models for all nine SAE widths from 512 to 131,072. At each width and threshold $t \in \{0.5, 0.6, 0.7, 0.8\}$, we apply the pairwise matching and candidate-graph procedures in Sections C.2.3 and C.2.4 to Gemini and Nemotron features.

Figure 24 shows that the number of matched pairs increases with SAE width, while the fraction of each dictionary that is matched generally decreases.

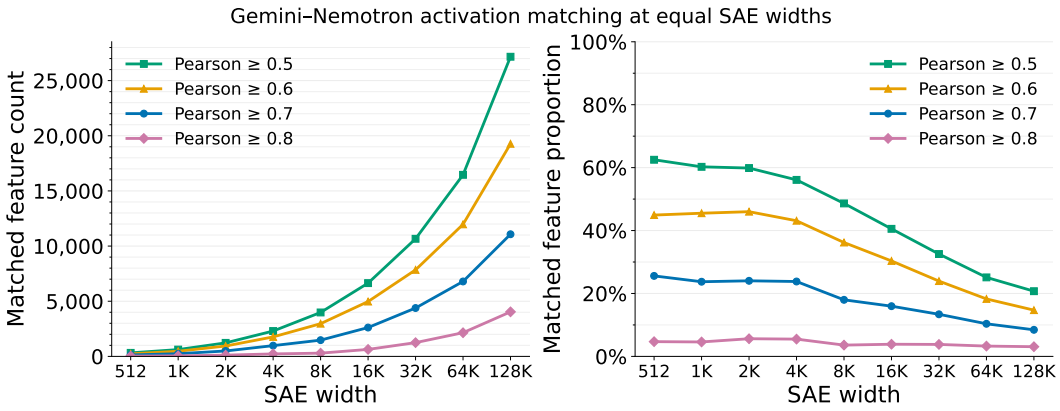


Figure 24: Feature-activation matching between Gemini and Nemotron SAEs of equal width. Curves give maximum-cardinality one-to-one matching at signed Pearson thresholds $t \in \{0.5, 0.6, 0.7, 0.8\}$ on 89,227,558 shared examples. Left: matched-pair count. Right: count divided by the common SAE width, including dead and unmatched features in the denominator. Counts are exact on the retained candidate graphs and lower bounds on exhaustive matching. The 131K comparison is included because matching across models does not require a larger SAE.

Example matches. Tables 12 to 23 show ten correlated pairs of features, with Pearson similarity ranging from 0.3 to 0.9. They cover education, trigonometry, astronomy, trade, programming, mechanics, food, sign languages, music, and botany. These are illustrative selections, and the captions describe qualitative interpretations of the activating texts. Each table contains fifteen sampled texts from each model: eight with source-feature activation in $(0.05, 0.1]$ and seven above 0.1. Gemini feature 71850 is an exception: it has no cached activations above 0.1, so all fifteen of its examples come from $(0.05, 0.1]$.

Table 12: Clarifying the meaning or scope of a term. Pearson $r = 0.352$.

Text	Gemini 71850	Nemotron 119331
Secara khusus, Chang'an Avenue hanya terdiri dari Chang'an Avenue Barat dan Chang'an Avenue T...	0.073	0.067
The term "bile duct" is often used as a shorthand for the "common bile duct," but they are no...	0.064	0.109
In both Great Britain and Ireland, the Continent is widely and generally used to refer to the...	0.064	0.118
Wenn in diesem Text oder in der wissenschaftlichen Literatur "Mevalonsäure" ohne weiteren Nam...	0.064	0.037
While we often use the word "sun" to refer specifically to our star, the Sun, the term "sun"...	0.058	0.128
Typically, "abstraction" is used in the arts as a synonym for abstract art in general.	0.057	0.119
Kangaroo Island most commonly refers to Kangaroo Island in South Australia.	0.057	0.017
Stansted usually refers to London Stansted Airport.	0.056	0.000
The Entombment usually refers to the entombment, or burial, of Jesus.	0.056	0.026
what word is often used casually to refer to both a database and the dbms used	0.056	0.077
Generally yes. When people refer to the company as Coke or as the more formal "Coca-Cola" the...	0.052	0.135
By extension, Himalayas is also the name of the massive mountain system which includes the Hi...	0.052	0.078
North Africa The term "North Africa" has no single accepted definition. It is sometimes defin...	0.051	0.079
Some professional pool players still use the term billiards to describe what's more commonly...	0.051	0.130
Saat ini, istilah "gubernur jenderal" umumnya digunakan lebih ke arah bekas koloni Kerajaan B...	0.050	0.076
Although the term progeria applies strictly speaking to all diseases characterized by prematu...	0.025	0.141
In an older and more restricted sense (not used here), the term chytrids referred just to tho...	0.044	0.121
The term combat (French for fight) typically refers to armed conflict between military forces...	0.023	0.119
These terms are rarely used by the police themselves and instead are used by criminals, priso...	0.000	0.109
In U.S. usage, a "plant condensate" is the equivalent of products classified as "pentanes+" a...	0.021	0.104
While Mochica has been used in place of Moche by people describing this culture, the word Moc...	0.000	0.101
Definition Challenges Since the terms "risk appetite" and "risk tolerance" often are used int...	0.000	0.100
[This agency, the Brain Injury Network, had asked them to change their definition, which form...	0.000	0.085
Blindness is frequently used today to describe severe visual decline in one or both eyes with...	0.000	0.071
Hindi is an Indo-Aryan language or languages descended from Sanskrit , defined with various d...	0.023	0.071
I think the word is used in a electronics in a different way than what the word originally me...	0.000	0.067
The MENA acronym is often used in academia, military planning, disaster relief, and business...	0.000	0.059
As a broad generality, in the United States, the SI metric system of units is predominantly u...	0.021	0.058
My son is 3 and I still consider him a toddler, but I was talking to a friend of mine who has...	0.000	0.058
In computer science, this type of matrix representation is commonly known as a bit array, bit...	0.000	0.057

Table 13: Whether algebraic transformations preserve solutions. Pearson $r = 0.362$.

Text	Gemini 25799	Nemotron 102037
When we cancel out a common factor from both sides of an equation, we are essentially dividin...	0.124	0.128
When proving trigonometric identities, it is crucial to avoid manipulating them as if they we...	0.121	0.098
Manipulating each side of a trigonometric identity separately is advised because it ensures t...	0.116	0.064
Dividing by zero is undefined in mathematics. If we divide both sides of an equation by a var...	0.115	0.132
Dividing both sides of a trigonometric equation by another trigonometric expression, like c...	0.111	0.074
Explain why it is important to consider the possibility of losing solutions when dividing bot...	0.104	0.100
Which axioms or theorems in group theory support the justification of multiplying or adding a...	0.104	0.022
It is not accurate to say that there are solutions to a differential equation that cannot be...	0.091	0.098
When you have the equation AC = B and wish to isolate C by multiplying both sides by A^(-1),...	0.077	0.042
Explain why the method of factorizing a variable to solve a quadratic equation is incorrect,...	0.073	0.084
Unfortunately, removing the floor function from both sides of the equation is not valid becau...	0.070	0.043
If ab>0, then which one of the following must be true? (A) a/b>0 (B) a-b>0 (C)...	0.059	0.000
In the original question, it is mentioned that taking the natural logarithm of both sides of...	0.059	0.042
When we derive a new equation from a system of simultaneous equations, we are creating an equ...	0.056	0.119
Why does the third elementary row operation, adding a row to another one multiplied by a numb...	0.054	0.023
When is it incorrect to square both sides of an inequality to solve it?	0.082	0.149
By substituting the expression for y from the original equation into one of the derivatives a...	0.000	0.134
The zero-product property states that if the product of two factors is zero, then at least on...	0.000	0.122
Move all terms to one side of the inequality to simplify the expression.	0.041	0.108
When integrating a differential equation using separation of variables, why is it incorrect t...	0.034	0.107
Apply the product rule of differentiation to the potential parent term. If the result matches...	0.000	0.106
how to solve equation with completing the square?	0.000	0.101
['Isolate the x2u200b term on one side of the equal sign.', 'Take the square root of both si...	0.025	0.072
In the proof of the solution to a recurrence relation using the characteristic equation, how...	0.000	0.072
how to solve quadratic equations with variables?	0.000	0.067
To solve equations involving absolute values, follow these steps: 1. **Case Analysis**: Recog...	0.013	0.063
['Use the Rational Zero Theorem to list all possible rational zeros of the function.', 'Use s...	0.000	0.061
Suppose I have an equation f(x) = 0, and I know the solution exists. I perform a series of...	0.077	0.054
When calculating the Taylor expansion of a function that has different forms on either side o...	0.000	0.051
As for the second restriction, commas are interpreted as or, so desmos is graphing values tha...	0.000	0.051

Table 14: School activities and educational offerings. Pearson $r = 0.516$.

Text	Gemini 31660	Nemotron 42453
Roanoke Rapids High School's most prestigious club is the Student Council. Other clubs includ...	0.114	0.140
The Manhasset High School offers a variety of clubs and activities for our students.	0.110	0.129
Riverdale Country School Abraham Award, which is an award given annually to high schools that...	0.110	0.108
Untuk menyalurkan bakat para siswa di berbagai serta melatih jiwa kepimpinan mereka, pihak se...	0.104	0.100
Popular extra curricular activities include National Cadet Corps (NCC), National Service Sche...	0.103	0.129
Die Musical-AG führt regelmäßig Musicals auf, unter anderem: Winnfried (2006), Anything Goes...	0.102	0.049
সাধারণ পাঠ্যপুস্তকের শিক্ষার বাইরে সাহিত্য ও সংস্কৃতি অঙ্গনে বিদ্যালয়ের শিক্ষার্থীরা বিভিন্ন...	0.100	0.076
softball, track and field (boys/girls), and cheer. Cave City students participate in competit...	0.089	0.100
what are some types of extracurricular activities?	0.083	0.064
productions, an annual talent show, and the annual ISAC Choir Festival. Schutz American Schoo...	0.068	0.042
why is extracurricular activities important in school?	0.063	0.094
History of Georgetown University or in opposition to it in later years, such as the short liv...	0.060	0.021
and events. It has more than 400 student organizations and 17 intercollegiate sports teams. A...	0.059	0.024
oldest still-running women's college radio station in the country. In the 80s, it earned the...	0.057	0.052
The top 10 countries represented are China, Greece, Hong Kong, India, Turkey, Singapore, Cana...	0.051	0.021
Early Childhood Each class in the Early Childhood section has a designated teacher, fully qua...	0.030	0.149
A few examples of common educational opportunities that may be considered co-curricular inclu...	0.099	0.145
Northland Preparatory Academy is a public charter high school and middle school in Flagstaff...	0.044	0.129
Wah Yan College, Hong Kong of Secondary Education Examination), Aestheic Education (Music and...	0.053	0.125
in the school assembly are: The National Pledge is also recited. Special classes are provided...	0.071	0.121
The Extended Student Services (ESS) Program extends Poway Unified School District's education...	0.027	0.108
Today (4/19) your student will bring home a permission slip for a wonderful opportunity espec...	0.030	0.105
Deer Valley High School offers all Arizona Interscholastic Association-sanctioned sports as o...	0.068	0.092
Outside the classroom, Iona students can join more than 75 campus clubs, including about 10 f...	0.045	0.092
Upon admittance to the school, students are given an academic level based on their proficienc...	0.024	0.091
Marianopolis is the only college that offers five special interest certificates: Environmenta...	0.073	0.083
2.) give first priority to EOF activities over other extra-curricular activities. 3.) make ev...	0.027	0.059
Nord Anglia Education International Baccalaureate. In some locations, A-levels, the French Ba...	0.000	0.058
organization on campus with over 280 members. The band program incorporates a competitive cor...	0.056	0.053
opportunity in which to engage youth and teachers in supporting English language-learning and...	0.019	0.053

Table 15: Trigonometric half-angle identities. Pearson $r = 0.578$.

Text	Gemini 23469	Nemotron 114122
The half-angle identities can be derived from the sum and difference formulas for sine and co...	0.154	0.104
To find the exact value of sin(pi/6) using the half-angle identity, follow these steps: 1. St...	0.131	0.064
How do you use the half-angle formulas to simplify the expression √((1+cos4x)/2)?	0.128	0.069
How do you use the half-angle identity to find the exact value of cos(pi/8) degrees?	0.125	0.077
How do you use a half-angle formula to find the exact value of tan((9π)/8)?	0.116	0.072
Find the exact value of cos(105°) using the half-angle identity.	0.114	0.028
How do you use the half-angle identity to find the exact value of cosecant of 7pi/8?	0.113	0.076
How can we verify the identity #cos 2x = (1-tan^2x)/(1+tan^2x)# using the double angle identity?	0.098	0.000
Given that cos A = frac(8)(17), find: a) sin 2A b) cos 2A Using the Pythagorean ide...	0.084	0.000
How do you find the value of sin(theta/2) given tan(theta) = 2 and 0 < theta < pi/2?	0.080	0.008
Starting with the double angle formula for tangent: tan(2x) = frac(2tan(x))(1-tan^2(x))...	0.076	0.024
We can use the double angle identity for cosine to write: #cos(4x) = cos(2(2x))# #cos(4x) = 2...	0.073	0.013
The exact value of #sec(75°)# can be found as follows: 1. Note that #sec(75°) = frac(1)(cos...	0.066	0.029
To solve the equation #cos 2x = -frac(sqrt(3))(2)#, we can employ the double angle identity...	0.065	0.000
Using the double angle formula for sine, we have: #sin(6theta) = 2sin(3theta)cos(3theta)# Usi...	0.059	0.019
The half-angle identities are trigonometric identities that express the sine, cosine, and tan...	0.142	0.114
To use the half-angle identities to find the exact values of sin 165°, cos 165°, and tan 165°...	0.137	0.113
The half-angle identities can be derived from the sum and difference formulas for sine and co...	0.154	0.104
To derive the half-angle formulas for sine, cosine, and tangent, we can start with the identi...	0.155	0.104
Half-Angle Identities are trigonometric equations that relate the sine, cosine, and tangent o...	0.131	0.103
To find the exact value of trigonometric functions at specific angles using half-angle identi...	0.134	0.103
To derive the half-angle formulas, we can use the following steps: * Use the double-angle for...	0.123	0.102
How do you use the half-angle formulas to find the exact value of tan(3π/8)?	0.136	0.077
How do you use the half angle identity to find the exact value of cos 101.25?	0.121	0.074
The half-angle formula for sine is given by #sin^2(theta/2) = (1 - cos(theta))/2#. To simplif...	0.094	0.062
To simplify #csc ((7pi)/8)# using the half-angle formula, follow these steps: 1. Recognize th...	0.098	0.062
To evaluate tan(-195) using the half-angle formula, we first need to find the value of tan(19...	0.099	0.060
How do you use a half-angle formula to simplify tan[(9π) / 8]?	0.106	0.060
We can use the half-angle formula for cosine to find the value of cos(pi/5) without using a c...	0.063	0.058
For the first question, we can use the half-angle formulas: sinleft(frac(x)(2)right) =...	0.091	0.056

Table 16: Telescopes and astronomical optics. Pearson $r = 0.604$.

Text	Gemini 64628	Nemotron 31241
what are the two key properties of a telescope and why is each important?	0.222	0.127
when was the vega bray telescope built	0.170	0.040
Timeline of telescopes , observatories , and observing technology .	0.145	0.137
He therefore bought a telescope, which enabled him to watch as accurately as did the owner hi...	0.136	0.109
Can you see an American flag on the moon with a telescope? Even the powerful Hubble Space Tel...	0.128	0.127
the version of the track recorded on "Later... with Jools Holland" was included as a "bonus l...	0.110	0.033
sensitivity is about 10 mJy. Steering of the antenna main lobe is accomplished with phase shi...	0.108	0.061
طبق اعلام مجری رصدخانهٔ ملی ایران ساخت بنای ثابت محفظه تلسکوپ ۳٫۴ متری، محوطه‌سازی و ساختمان...	0.100	0.069
Five hundred meter Aperture Spherical Telescope help change the focus to different areas of t...	0.099	0.097
The Square Kilometre Array (SKA) is a large multi radio telescope project aimed to be built...	0.094	0.033
from around 800 K (500 °C / 1000 °F) to 1500 K (1200 °C / 2200 °F) in just 6 hours. An observ...	0.079	0.041
Newsletter. Today's BAFact: The Sun is 400,000 times brighter than the full Moon in the sky....	0.074	0.063
Moon missions. Russian Federation spacecraft is planned to send cosmonauts to the moon orbit...	0.070	0.016
Holders of a bachelor's or a master's degree in astronomy often enter an unrelated field. How...	0.064	0.044
are swfa scopes any good?	0.057	0.057
As a rule you should avoid zoom binoculars, at least for astronomical use. Good zoom eyepiece...	0.023	0.203
I am planning to build a grism spectrometer based on F. Sigernes' work. Can I use single lens...	0.000	0.173
On the other hand, a compound microscope has two sets of lenses, an ocular or eyepiece lens a...	0.000	0.149
What online resources are available for amateur astronomy experiments, particularly those inv...	0.080	0.123
Coronagraph A coronagraph is a telescopic attachment designed to block out the direct light f...	0.025	0.122
1 While looking at the objective lens and the stage from the side, turn the coarse focus knob...	0.000	0.116
از آینه‌های معروف جهان می‌توان به آیینه‌ای که در تلسکوپ رصدخانه مانت پالومار کالیفرنیا کار گذ...	0.112	0.102
what is the name of the spherical lens placed on top of text to magnify the	0.000	0.077
[This root (tel-) means reaching the end (aim).. It is well-illustrated with the old pirate's...	0.091	0.073
Bernard Ferdinand Lyot (27 February 1897 , in Paris – 2 April 1952 , in Cairo) was a Frenc...	0.000	0.067
This was the observation that Venus has phases, much like our Moon does. To the naked eye, Ve...	0.071	0.060
lens system was made obsolete by later developments, primarily advanced by Panavision, Cinema...	0.021	0.059
what year did the hubble space telescope not see neptune's great red spot	0.033	0.058
what does the parabolic reflector of holophones radio produce	0.000	0.054
Lens Creek	0.000	0.052

Table 17: Trade tariffs and protectionism. Pearson $r = 0.651$.

Text	Gemini 61225	Nemotron 124167
Undang-Undang Smoot-Hawley	0.153	0.030
Loi Hawley-Smoot	0.146	0.000
where were the first tariffs introduced in australia	0.145	0.088
what was the name of the tariff introduced by germany in 1925	0.136	0.137
what did the republicans impose to protect themselves from foreign trade	0.135	0.043
A commercial treaty is a formal agreement between states for the purpose of establishing mutu...	0.113	0.194
who was the president when steel tariffs were imposed	0.109	0.096
Customs duties in the United States	0.086	0.097
Canadian-American Reciprocity Treaty primarily to the United States. The treaty was abrogated...	0.083	0.098
define autarky	0.080	0.000
Rückzoll	0.072	0.000
the anti-cotton law league campaigned against the repeal of which laws	0.060	0.015
Aviation in Indonesia airlines by the end of 2015 or early 2016. Indonesia is considering ope...	0.057	0.039
what article did the jay treaty suspend in 1796	0.054	0.013
United Nations Conference on Trade and Development The organisation grew from the view that e...	0.053	0.162
Special 301 Report and enforcement of intellectual property rights". It also states that comp...	0.000	0.177
What are the main drawbacks of adopting a WTO Brexit (also known as a Clean Break)?	0.080	0.149
The so-called Really Good Friends (RGF) club is a sub-group of members of the World Trade Org...	0.027	0.148
Interpreting NAFTA Rules of Origin. A rule of origin might contain. 1) a change in tariff cla...	0.043	0.145
of Canada in the 1920s. Its platform was entitled the "New National Policy", and it advocated...	0.089	0.129
Pollution haven hypothesis estimate the indirect effect of trade liberalization on pollution...	0.036	0.129
Comprehensive and Progressive Agreement for Trans-Pacific Partnership favored that other coun...	0.023	0.118
Post-World War II economic expansion trade, GNP, and general quality of life. The high econom...	0.019	0.076
what act was passed in the kennedy round of trade negotiations	0.107	0.074
__ INDEX __ Arun M. Kumar (born 25 November 1952) has been the Chairman and Chief Executive...	0.000	0.072
Chile-United States Free Trade Agreement	0.067	0.057
Products may come from many different countries – one part of a product may come from China...	0.000	0.053
Both the United States and Japan belong to a variety of global organizations, including: the...	0.000	0.052
The International Jute Study Group (IJSG) is an intergovernmental organisation of states th...	0.027	0.051
which international agreement provides for the protection of sound recordings	0.000	0.050

Table 18: Comments in source code. Pearson $r = 0.667$.

Text	Gemini 96133	Nemotron 76359
In Emacs, the provided examples actually use 2-character comment delimiters, which are extend...	0.167	0.178
Modeling line comments in a CFG can be achieved as follows: 1. If you're not dealing with mul...	0.158	0.171
how to comment out multiple lines in jupyter?	0.151	0.165
are comments passed to the parser?	0.150	0.082
Comment programming	0.150	0.114
how to comment multiple lines in pycharm shortcut?	0.120	0.145
ulysses how to add comment?	0.102	0.022
How can I prevent syntax-propertize in Emacs from treating '#' within SPARQL URI's as the sta...	0.093	0.107
How can I make Yasnippet insert a comment sign on a new line when pressing Return? Q: When us...	0.089	0.108
জাভাস্ক্রিপ্ট এ ভেরিয়েবল ডিক্লায়ার করা হয় var কি ওয়ার্ড দিয়ে : var x; // defines the var...	0.074	0.044
Important As of Internet Explorer 10, conditional comments are no longer supported by standar...	0.068	0.029
Here document	0.058	0.023
A feature that fits into the former category is the extended properties of databases. Introdu...	0.053	0.000
up vote 10 down vote favorite. 4. When I insert a comment or make a revision to some Microsof...	0.053	0.051
How to Query SQL Server 2000 for Stored Procedures Containing Text. For SQL Server 2000, the...	0.051	0.000
The recommended way to comment out multiple lines of code in Python is to use consecutive # s...	0.151	0.234
Multiline comment. A multiline comment starts with /* and can be found at any position in the...	0.160	0.201
Multiline Comments in Bash Unlike most of the programming languages, Bash doesn't support mul...	0.146	0.172
how to comment out multiple lines in mysql workbench?	0.145	0.161
While there isn't a direct way to insert line breaks in comments, they are designed to be con...	0.070	0.145
There is also another way to document Python code using comments that start with ##. These ty...	0.126	0.130
If the PyCharm IDE is used to write Python code - select multiple code rows to comment and pr...	0.090	0.113
Комментарий	0.015	0.094
Looks like you are back with WordPress comments. I have sites with WordPress, DNN Blog, and D...	0.000	0.073
En kommentoi	0.021	0.069
VBdocman allows commenting and the automatic generation of technical documentation from Visua...	0.077	0.065
There isn't a built-in command in c-mode that mimics LaTeX's latex-close-block for #ifdef blo...	0.053	0.064
This topic has been extensively discussed on various Meta sites within the Stack Exchange net...	0.051	0.053
Syntax highlighting for HTML code within a #+BEGIN_EXPORT HTML block in org-mode is affected...	0.020	0.053
Commentaries dealing with the zahir (outward aspects) of the text are called tafsir, and herm...	0.000	0.051

Table 19: Mechanical bearings. Pearson $r = 0.709$.

Text	Gemini 76787	Nemotron 40156
Bearing (mechanical) 1794. His was the first modern ball-bearing design, with the ball runnin...	0.203	0.282
Rolf Hoppler and Reinhold Errath evaluate the types of bearings used in the. cement industry,...	0.191	0.184
A wheel bearing replacement can take anywhere from one to three hours depending on the type o...	0.157	0.195
Fluid bearing generators weighing hundreds of tons. They are also used in very heavy machiner...	0.146	0.218
Driving with 450+ answers. You didn't indicate front or rear, or whether it is a front-wheel...	0.137	0.153
SKF-89976A	0.108	0.097
The noise you're experiencing could be due to a faulty main bearing in the front of the trans...	0.107	0.073
Consider the following diagram: [Image of a diagram showing Jack and Jill's positions and the...	0.098	0.031
Pacific Bearing Corporation guides, linear slides, linear motors and nuts for lead screws. In...	0.096	0.198
ACER Racing is a large OEM manufacturing company specializing in precision high performance r...	0.091	0.067
taken from the shafts of VFD-driven motors throughout the commercial bakery to illustrate the...	0.075	0.085
1. Salvageability: The assembly can likely be salvaged with proper repairs. To verify the con...	0.073	0.128
Bearing an Hourglass is a fantasy novel by Piers Anthony. It is the second of eight books in...	0.064	0.092
A ship is anchored off a long straight shoreline that runs north and south. From two observat...	0.057	0.000
Nicolas Fatio de Duillier (also spelled Faccio or Facio ; 16 February 1664 – 12 May 1753)...	0.052	0.000
Bearing (mechanical) rolling element bearing is a wooden ball bearing supporting a rotating t...	0.214	0.274
608 STANDARD The most common bearing size is the "608." Characterized by an 8mm core, a 22mm...	0.167	0.247
tapered roller bearing in combination with modern lubricants is extremely durable and is used...	0.155	0.239
Slewing bearing. A slewing bearing or slew ring is a rotational rolling-element bearing or pl...	0.153	0.200
ABEC 5 bearings are the norm in skateboarding. You get a reasonable amount of speed, and at a...	0.177	0.174
Ball (bearing) the largest and smallest diameter measurement. For a given lot there is a "lot...	0.085	0.126
Oli pelumas harus selalu melumasi bagian-bagian poros engkol yang bergerak, hal ini tentunya...	0.104	0.117
when was the first roller bearing truck made	0.222	0.079
Tribology Online | Read articles with impact on ResearchGate, the professional network for sc...	0.061	0.069
Cojinete magnético	0.199	0.060
• CHILDBEARING (noun). The noun CHILDBEARING has 1 sense: 1. the parturition process in human...	0.043	0.059
Im August 1993 übernahm die Minebea alle Anteile der PM DM und wurde somit alleiniger Eigentü...	0.039	0.058
Footnote on Crankshaft Rebuilders. Did another search on Crankshaft Rebuilders. They use a ne...	0.077	0.058
Tunnel crankcase there they were used in the more established diesel-electric designs. The fi...	0.074	0.053
The behavior of an object rolling on a surface is influenced by the coefficient of rolling re...	0.029	0.052

Table 20: Butter. Pearson $r = 0.788$.

Text	Gemini 59059	Nemotron 17701
which oil is used in sautéing to make butter	0.240	0.065
Grams in Butter Eight Tablespoon stick of unsalted butter weighs 110 grams. Therefore, 110 di...	0.171	0.235
The rock gunnel or Butterfish(”Pholis gunnellus”) is an eel-like fish found in the intertidal...	0.163	0.132
Basically, cookies made with butter spread more and are flatter and crisper if baked long eno...	0.158	0.213
If a recipe calls for butter in ounces, simply divide the number of ounces by 4 to arrive at...	0.154	0.244
is shea butter good for hair?	0.133	0.096
Hartkäse, Weichkäse und Butter sind relativ ergiebige Quellen für Vitamin K.	0.120	0.063
Fortunately for everyone who loves peanut butter, almond butter, and other nut butters, these...	0.076	0.102
الجبنة القابلة للدهن (الكاسات)، شرائح الجبنة، الزبدة، السمن، كريمة الخفق، كريمة الطبخ، الموزا...	0.070	0.038
otto mønsted aarhus was the founder of which famous margarine	0.069	0.014
There are a number of these cleaners available on the Internet, most of them free to try out...	0.064	0.025
Once melted, remove from heat and whisk in the hot sauce and salt. Coat the wings with the sa...	0.061	0.034
1 Bring a large pot of lightly salted water to a boil. 2 Add pasta and cook for 8 to 10 minut...	0.055	0.055
Directions. For the chocolate sauce: Combine the chocolate and 1/4 cup water in a heavy small...	0.054	0.018
Prize (marketing) prizes from German margarine were molded in a light cream color designed to...	0.051	0.071
what kind of butter is fiddlehead fern pickle	0.169	0.208
when did the buttery become part of connecticut hall	0.036	0.144
Don’t use low fat spread-it might be better for you, but much of the fat is replaced by water...	0.032	0.136
Place the butter cubes on top of the carrots. They will melt all over and do yummy things to...	0.123	0.132
West Buttercreek Airport	0.063	0.121
After 2 hours begin mopping the brisket with a butter/cajun seasoning mix that I call ”mop wa...	0.077	0.107
Pan Frying. 1 1. Shuck the clams. Remove the clams from their shells if they are still intact...	0.081	0.104
Calories in Kroger Peanut Butter based on the calories, fat, protein, carbs and other nutriti...	0.034	0.086
do you have to refrigerate earth balance butter?	0.152	0.085
Ingredients. 1 6-8 ounces salt pork, cut into 1/2 inch chunks. 2 4 Tbsp unsalted butter, divi...	0.070	0.080
A nervous stomach can give someone ”butterflies” in their stomach, or even make a person feel...	0.000	0.077
A Franzbrötchen (]) is a small, sweet pastry, baked with butter and cinnamon. Sometimes othe...	0.042	0.076
how many grams is 1 tablespoon of margarine?	0.066	0.064
You’re better off sticking with skim or low-fat dairy. Participants who ate whole-fat dairy (...	0.046	0.060
Crème fraîche (English pronunciation : -LSB- ˌkɹɛmˈfɹɛʃ -RSB- , -LSB- kʁɛm fʁɛʃ -RSB- , lit...	0.042	0.055

Table 21: Sign languages. Pearson $r = 0.802$.

Text	Gemini 61020	Nemotron 89987
hawaii sign language is an independent language from what other language	0.157	0.117
A free sign language worksheet for kids (a great reference when learning the ASL alphabet) AS...	0.136	0.164
American Sign Language grammar signs: The hand shape is the initial of one of the English nam...	0.129	0.209
You could qualify by registering as a trainee sign language interpreter (TSLI). To register,...	0.118	0.185
Home sign is often the starting point for new deaf sign languages that emerge when deaf peopl...	0.116	0.203
Lithuanian Sign Language (LGK) (Lithuanian : Lietuvių gestų kalba) is the national sign l...	0.109	0.187
who was the founder of indian sign language school in madras	0.103	0.056
Cuban Sign Language is the deaf sign language of Cuba . Little information is available .	0.099	0.156
part. Thirteen of the most widely spoken indigenous languages are considered ”national langua...	0.073	0.132
who did jean massieu’s father play in sign gene	0.072	0.000
Fingerspelling Some writers have suggested that the body and hands were used to represent alp...	0.071	0.112
Naga Hills Sign Language was a village sign language of India. Early in the 20th century, a h...	0.068	0.099
Project Nim (film)	0.058	0.000
Starbucks Malaysia’s Signing Store. Signing Store Workshop; Learn More . Policies; Looking fo...	0.053	0.000
Deaf with the capitalized D represents people that are members of the Deaf Community, a lingu...	0.050	0.048
Black American Sign Language 2004 by Melanie Metzger and Susan Mather found that Black male s...	0.130	0.213
hard to draw. A language ’continuum’ is often described between signing with a strong sign-la...	0.117	0.205
ASLwrite () is a somacheirographic writing system that developed from si5s. It was created to...	0.099	0.180
Many Australian Aboriginal cultures have or traditionally had a manually coded language, a si...	0.082	0.136
Rennellese Sign Language is an extinct form of home sign documented from Rennell Island in 19...	0.081	0.117
estimates for the number of people who use ASL are relatively easy to find in research and pr...	0.108	0.109
can asl be used in other countries?	0.178	0.104
When the 26-year-old signed I’m proud of you, the president quickly signed back a Thank you....	0.081	0.099
Hanabiko ”Koko” (July 4, 1971 - June 19, 2018) was a female western lowland gorilla known for...	0.081	0.086
She is now 13 months old. I started teaching her sign language as a tiny baby (signing as I t...	0.065	0.085
The World Trade Organization (WTO) is an intergovernmental organization that regulates inte...	0.000	0.084
Tom Clerc Tom Clerc is a fictional deaf superhero appearing in Sign Gene, the world’s first f...	0.060	0.071
As in other states, a certified interpreter in Idaho will command a higher fee for legal inte...	0.018	0.059
Centre for Deaf Studies, Bristol The Centre for Deaf Studies was a department of the Universi...	0.046	0.054
The General Agreement on Trade in Services (GATS) is a treaty of the World Trade Organization...	0.000	0.052

Table 22: Eurovision Song Contest entries. Pearson $r = 0.865$.

Text	Gemini 123114	Nemotron 97094
Norway in the Eurovision Song Contest 1966	0.175	0.124
United Kingdom in the Eurovision Song Contest 1968	0.174	0.132
France in the Eurovision Song Contest 1956	0.173	0.139
Denmark in the Eurovision Song Contest 1989	0.169	0.144
Lithuania in the Eurovision Song Contest 2018	0.147	0.142
Greece in the Eurovision Song Contest 2014	0.138	0.157
Macedonia in the Junior Eurovision Song Contest 2013	0.124	0.126
Singapore in the ABU TV Song Festival	0.094	0.090
Bashkortostan in the Turkvision Song Contest 2013	0.087	0.094
Venezuela in the OTI Festival	0.082	0.074
Kazakhstan in the Bala Turkvision Song Contest 2015	0.078	0.093
France au Concours Eurovision de la chanson 1984	0.072	0.026
Ukraine in the Eurovision Young Musicians	0.071	0.108
Eurovision Song Contest 1971	0.056	0.051
Luxembourg au Concours Eurovision de la chanson 1961	0.052	0.036
Switzerland in the Eurovision Song Contest 2009	0.164	0.155
Denmark in the Eurovision Song Contest 2004	0.164	0.144
Belarus in the Eurovision Song Contest 2005	0.154	0.141
Netherlands in the Eurovision Song Contest 1983	0.179	0.141
United Kingdom in the Eurovision Song Contest 1981	0.177	0.135
Belgium in the Junior Eurovision Song Contest	0.116	0.119
Ireland in the Eurovision Young Musicians	0.109	0.118
Wales in Eurovision Choir	0.041	0.088
Switzerland in the Eurovision Song Contest 2019 2018 contest would be selected through a Die...	0.000	0.073
The Republic of Macedonia, participating under the name F.Y.R. Macedonia was represented at t...	0.000	0.065
Ukraine was represented in the Eurovision Song Contest 2005 by GreenJolly with the song "Razo...	0.000	0.060
Montenegro Eurovision laulukilpailussa	0.032	0.058
Spain in the Eurovision Song Contest 1965 Spain took part in the Eurovision Song Contest 1965...	0.000	0.054
Poland in the Early Middle Ages	0.000	0.054
Eurovision Song Contest 1975	0.060	0.052

Table 23: Astragalus and milkvetch species. Pearson $r = 0.871$.

Text	Gemini 104756	Nemotron 51458
Astragalus campylanthus	0.166	0.165
Astragalus brauntonii	0.165	0.151
Astragalus brachystachys	0.164	0.157
Astragalus annularis	0.163	0.160
Astragalus monanthemus	0.160	0.155
Astragalus sprucei is a species of legume in the Fabaceae family. It is found only in Ecuador...	0.120	0.160
Astragalus sabulonum is a species of milkvetch known by the common name gravel milkvetch. It...	0.120	0.226
Astracantha gummifer	0.091	0.075
Astragalus proimanthus is a species of flowering plant in the legume family known by the comm...	0.086	0.185
Astragalus cicer in regards to the inherent durability of the plant. In regards to obtaining...	0.081	0.143
The garage door gasket not only protects the bottom of the garage door from the floor each ti...	0.075	0.000
purtuka or nakhuna (Astragalus Tribuloids),	0.072	0.039
Mesorhizobium sangaii is a bacterium from the genus Mesorhizobium which was isolated from roo...	0.067	0.056
Циклоастрогенол	0.064	0.000
قتاد سمسمي	0.063	0.000
Astragalus bernardinus , known by the common name San Bernardino milkvetch , is a species of...	0.138	0.200
astragalus (a··str ·gae·gă l) , s N latin: Names astragalus, Membranaceus astragalus; gummife...	0.111	0.182
Astragalus cedreti (also called Cedar astragalus ;) is a species of flowering plant in the...	0.103	0.179
Astragalus crotalariae	0.163	0.157
Astragalus newberryi	0.162	0.152
Astragalus pelecinus pelecinus	0.162	0.138
Es una planta herbácea perennifolia originaria de Norteamérica donde se distribuye por Estado...	0.129	0.130
Astragalin is a chemical compound. It can be isolated from "Phytolacca americana" (the Americ...	0.109	0.088
what is the herbage of astragalus molybdenus coated in	0.140	0.086
Astragal	0.207	0.083
why is the seed bank of astragalus nitidiflorus so sensitive	0.104	0.071
Chinese Herbs In Western View - Huang (Bei) qi (Radix astragali) Health Benefits and Side eff...	0.080	0.067
where did the astragaloi perform divination in ancient greece	0.082	0.060
locoweed toxins, pinpointing levels of toxicity in animals once they have ingested locoweed,...	0.064	0.055
Astragalomancy pessomancy (also known as psephomancy) - another act of divination which uses...	0.053	0.054