Timescale Separation in Sparse Dictionary Learning: Reconstruction Converges Before Reproducibility

Anonymous Authors[1]

Abstract

Sparse autoencoders (SAEs) trained on neural network activations reach low reconstruction loss within tens of epochs, yet cross-seed decoder reproducibility—measured by mean maximum cosine similarity (MMCS) calibrated against an extreme-value-theory random-dictionary null—remains indistinguishable from random dictionaries on the same timescale. We document this timescale separation on a mod-113 algorithmic transformer (d=128) and validate it across Pythia-70M, 160M, and 410M (d=512–1024). On mod-113, reconstruction loss saturates by epoch 20 while TopK decoder MMCS continues climbing for 1,000+ epochs, asymptoting at 0.73 (2.44× random). On Pythia-410M layer 21, a d_{sae} sweep shows that all overparameterization ratios (d_{sae}/erank from 1.3 to 10.0) converge to 4.3–4.7× random given sufficient training, dissolving the apparent d_{sae}/erank stability threshold. We derive this null from the order statistics of inner products between random unit vectors, obtaining predictions within 0.03% of empirical baselines across dimensions 128–1024. Dead-neuron resampling produces a transient MMCS spike to 0.888 at epoch 10 that decays below the unregularized baseline—a *phantom consistency* artifact arising from training dynamics rather than learned structure.

1. Introduction

Sparse autoencoders decompose neural network activations into overcomplete dictionaries of interpretable features (Bricken et al., 2023; Templeton et al., 2024; Gao et al., 2024). Whether the learned features are reproducible across training runs—rather than arbitrary decompositions of the same reconstruction objective—is an open question with practical consequences for interpretability (Paulo & Belrose, 2025; Song et al., 2025; Leask et al., 2025).

We study this question through the lens of training dynamics and high-dimensional geometry. Our central observation is a *timescale separation*: reconstruction loss and decoder reproducibility evolve on different timescales, with reconstruction converging orders of magnitude faster than decoder-column alignment. This separation holds from a 128-dimensional algorithmic setting through 1024-dimensional language model activations, suggesting it reflects a structural property of overcomplete dictionary learning rather than a quirk of any single experimental setup.

Three aspects of this work connect to the themes of the HiLD workshop. First, the timescale separation itself is a learning dynamics result: the reconstruction objective has many near-degenerate minima, and the slow decoder alignment trajectory traces how optimization navigates this flat region. Second, our null calibration uses extreme value theory (EVT) to characterize the geometry of random overcomplete dictionaries in high dimensions, providing an analytic baseline against which learned structure can be measured. Third, the d_{sae}/effective-rank relationship—previously reported as a fixed stability threshold—turns out to be a training budget artifact, with all overparameterization ratios converging given sufficient optimization time.

Our contributions:

1. We quantify the timescale separation between reconstruction convergence (∼20 epochs) and decoder reproducibility (1,000+ epochs) on mod-113, and show the same qualitative pattern on Pythia-70M/160M/410M.

2. We derive an EVT null for MMCS that matches empirical random-dictionary baselines to within 0.03% across $d \in \{128, 512, 768, 1024\}$, providing a dimension-aware calibration for reproducibility claims.

3. We show the d_{sae}/erank stability threshold dissolves with extended training, reinterpreting it as a budget-dependent phenomenon rather than a geometric law.

4. We identify phantom consistency—a transient

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

resampling-induced MMCS spike—as a training dynamics artifact.

2. Background and Related Work

SAE stability. Paulo & Belrose (2025) found that only 30–42% of SAE features are shared across seeds on Pythia-160M and Llama 3 8B. Song et al. (2025) showed that higher consistency is achievable with TopK SAEs and sufficient training. We study the transition between these regimes as a dynamics question: what determines how long the optimizer takes to reach reproducible solutions?

Dictionary learning theory. Olshausen & Field (1996) observed that sparse codes for natural images converge to Gabor-like filters regardless of initialization, but only after extensive training on large datasets. Formal recovery guarantees (Spielman et al., 2012; Gribonval & Schnass, 2010; Arora et al., 2015) require sparsity conditions stronger than typical SAE settings. Cui et al. (2026) derive identifiability conditions for SAEs—extreme sparsity of the ground truth, sparse SAE activation, and sufficient hidden dimensions—that are violated in our dense-activation regime ($\text{erank}/d_{\text{model}} \approx 63\%$). Our empirical observations are consistent with their predictions: short-budget MMCS matches the random null, and extended training raises decoder agreement without formal identifiability guarantees.

SAE pathologies. O'Neill et al. (2024) identify within-run amortization gaps where reconstruction appears converged but the encoder remains suboptimal; this provides a plausible mechanism for the reconstruction-reproducibility dissociation we measure across runs. Bussmann et al. (2024b) document feature absorption. Heap et al. (2025) show SAEs trained on random transformers produce similar interpretability scores to those from trained models. Korznikov et al. (2026) independently advocate random-baseline controls for SAE evaluation, complementing our EVT calibration approach.

3. Methods

3.1. Experimental Setup

Mod-113 transformer. A 2-layer transformer (d_{model}=128, n_{heads}=4, d_{mlp}=512) trained on modular addition (mod 113) through grokking (Power et al., 2022; Nanda et al., 2023), reaching 100% accuracy. Effective rank of layer-1 residual stream activations: 80.5.

SAE training. We trained TopK (k=32), ReLU (λ=10^{-3}), JumpReLU (Rajamanoharan et al., 2024), and BatchTopK (Bussmann et al., 2024a) SAEs with d_{sae}=1,024 (8× expansion). Each architecture used 5 seeds with identical data and hyperparameters; only random initialization differed. Mod-113 SAEs trained with Adam (lr $= 3\times10^{-4}$, batch size 256). Extended runs reached 2,000 epochs. A 20-seed robustness study ($\binom{20}{2} = 190$ pairs) confirmed the short-budget result.

Pythia models. SAEs trained on cached activations from Pythia-70M (d=512), 160M (d=768), and 410M (d=1024) (Biderman et al., 2023), using 200K activation vectors from wikitext-103. The d_{sae} sweep on Pythia-410M layer 21 used $d_{\text{sae}} \in \{512, 1024, 2048, 4096\}$ with 3 seeds each over 200 epochs (batch size 4,096). Pythia-70M multilayer runs swept layers $\{0, 2, 4, 5\}$ with $d_{\text{sae}} \in \{256, 512, 1024, 2048, 4096\}$.

3.2. Metrics

MMCS. For each decoder column in SAE_1, we compute its maximum absolute cosine similarity to any column in SAE_2, averaged symmetrically. MMCS is a nearest-neighbor decoder geometry statistic; it permits many-to-one matching and does not establish one-to-one feature identity, activation-level agreement, or causal equivalence.

EVT null calibration. The inner product between two independent random unit vectors in $\mathbb{R}^d$ follows a scaled Beta distribution: $(z+1)/2 \sim \text{Beta}((d-1)/2, (d-1)/2)$. For m dictionary elements, the expected maximum absolute inner product is the m-th order statistic of this distribution, computable by numerical integration. This gives dimension-specific analytic predictions:

Table 1. EVT null predictions vs. empirical random-dictionary MMCS. The analytic null matches within 0.03% across four orders of magnitude in dimension.

Setting	d	EVT pred.	Observed
Mod-113 (m=1024)	128	0.2989	0.2990
Pythia-70M (m=1024)	512	0.1510	0.1511
Pythia-410M (m=1024)	1024	0.1073	0.1074

Effective rank. Following Roy & Vetterli (2007): $\text{erank}(A) = \exp(-\sum_i \bar{\sigma}_i \log \bar{\sigma}_i)$, where $\bar{\sigma}_i$ are normalized singular values of the mean-centered activation matrix.

4. Results

4.1. Timescale Separation on Mod-113

At 30 epochs (∼1,500 gradient steps), all four SAE architectures achieve low reconstruction loss while decoder MMCS remains within 1.5% of the random null (Table 2). The 20-seed study confirms this: the trained-minus-random gap is 0.0040 (seed-resampled bootstrap 95% CI $[0.0037, 0.0043]$; permutation p=0.0001), statistically detectable but within our pre-specified equivalence margin ε=0.005.

Extended training reveals the full dynamics (Figure 1). Reconstruction loss saturates within ∼20 epochs, but TopK

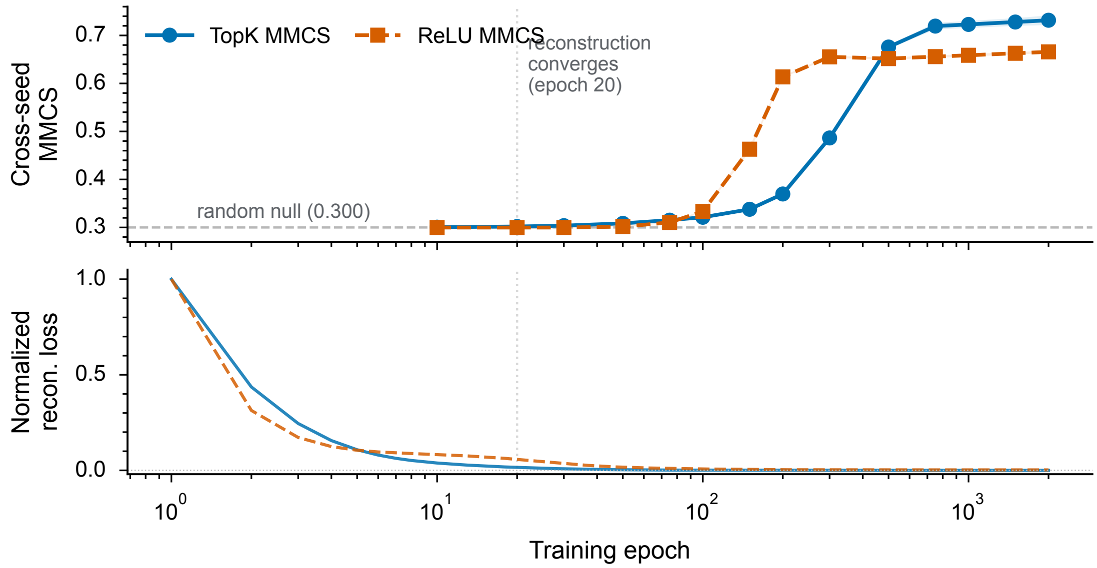

Figure 1. **Reconstruction converges before decoder reproducibility.** (Left) Normalized reconstruction loss saturates within ∼20 epochs. (Right) Cross-seed MMCS starts near the random-dictionary null and continues climbing for 1,000+ epochs. TopK asymptotes at ∼0.73 (2.44× random); ReLU plateaus at ∼0.66 (2.22× random). The gap between reconstruction convergence and decoder alignment spans two orders of magnitude in training time.

Table 2. Four architectures at 30 epochs on mod-113 (d_{sae}=1024). All remain near the MMCS random null despite low reconstruction loss.

Architecture	MMCS	Over Random	L0
TopK	0.304 ± 0.001	1.015×	32
BatchTopK	0.302 ± 0.001	1.008×	32
ReLU	0.300 ± 0.001	1.002×	∼328
JumpReLU	0.300 ± 0.001	1.002×	∼365

MMCS continues climbing for 1,000+ epochs, reaching 0.723 by epoch 1,000 and gaining only +0.009 over the final 1,000 epochs (2.44× random at asymptote). ReLU plateaus at ∼0.66 (2.22× random), confirmed from epoch 300 through 2,000. This represents a ∼50× timescale ratio between reconstruction convergence and decoder stabilization.

A held-out multi-metric audit corroborates this picture: at epoch 2,000, decoder Hungarian matching reaches 0.678 (vs. 0.289 random), activation-pattern Hungarian reaches 0.776 (vs. 0.547 random), and firing-overlap Hungarian reaches 0.540 (vs. 0.361 random). The metrics form their own ordering—decoder geometry aligns first, activation patterns follow, and firing overlap rises last—suggesting a cascade of increasingly specific forms of agreement.

4.2. Cross-Scale Validation on Pythia

The timescale separation is not specific to d=128. On Pythia-410M layer 21 (200 epochs, 5 seeds), both TopK (4.33× random, MMCS = 0.465 ± 0.016) and ReLU (4.41× random, MMCS = 0.473 ± 0.005) are still climbing at the end of training. A 500-epoch held-out audit on the same layer reaches decoder MMCS 0.521 (4.85× random) and decoder Hungarian 0.453 (4.37×). On Pythia-160M layer 10, ReLU reaches 5.11× random (MMCS = 0.632) at 200 epochs. These runs use roughly 40M token-equivalent SAE update exposures, well below production-scale training budgets (Gao et al., 2024; Templeton et al., 2024).

Across model scales, the same qualitative pattern holds at short training budgets (Table 3): deeper layers with lower effective rank show higher stability ratios, but absolute MMCS remains low (< 0.25) at 10 epochs everywhere. Layers with lower effective rank (higher spectral concentration) show earlier stability gains, consistent with the optimization having fewer effective symmetries to break when the activation distribution is more structured.

4.3. The d_{sae}/erank Threshold Dissolves with Training

At short training budgets, overparameterized SAEs ($d_{sae} \gg$ erank) appear fundamentally less stable than matched ones. We tested whether this reflects the geometry of the problem or just the training budget by sweeping $d_{sae} \in \{512, 1024, 2048, 4096\}$ on Pythia-410M layer 21 with 200 epochs (Table 4).

Table 3. Cross-scale scope checks at 10 epochs. Lower effective rank relative to d_{model} corresponds to higher stability. All conditions remain near-random in absolute terms.

Model	Layer	d_{model}	Eff. Rank	Best Ratio
Pythia-70M	0	512	391	1.60×
Pythia-160M	1	768	573	1.84×
Pythia-160M	10	768	512	1.66×
Pythia-410M	2	1024	757	2.09×
Pythia-410M	21	1024	409	2.45×

Table 4. The overparameterization-stability tradeoff dissolves with training. All d_{sae} values converge to 4.3–4.7× random at 200 epochs despite 10-epoch ratios ranging from 1.55× to 2.45×.

d_{sae}	d_{sae}/erank	10-ep ratio	200-ep MMCS	200-ep ratio
512	1.3	2.45×	0.479	4.73×
1024	2.5	2.14×	0.458	4.27×
2048	5.0	1.88×	0.523	4.63×
4096	10.0	1.55×	0.537	4.54×

The d_{sae}=4096 condition (d_{sae}/erank=10.0), which showed the weakest 10-epoch stability, reaches 4.54× random at 200 epochs—comparable to d_{sae}=512. The short-budget threshold is not a fixed property of the loss landscape; it reflects how quickly optimization breaks the additional symmetries introduced by overparameterization.

On Pythia-70M, a multilayer analysis at short training (10 epochs) shows the complementary pattern: at d_{sae}=4096, all layers collapse to ∼1.1× random regardless of effective rank. This matches the mod-113 short-budget behavior and suggests the geometry of overcomplete random dictionaries dominates before sufficient training breaks the degeneracy.

4.4. Phantom Consistency

Under cached dead-neuron resampling, MMCS spikes to 0.888 at epoch 10, then decays monotonically to 0.562 by epoch 2,000—below the unregularized TopK baseline for the remainder of training (Figure 2). This *phantom consistency* is a training dynamics artifact, not learned structure: resampled neurons are initialized from high-loss data points, creating temporary cross-seed alignment that optimization subsequently erodes.

We tested the simplest mechanism—shared resampling pools across seeds—by replacing shared pools with seed-specific pools. The spike persisted (epoch-10 MMCS: 0.524 shared, 0.543 seed-specific, 0.301 no resampling), ruling out shared-pool anchoring as the sole explanation. A Pythia-70M streaming-resampling stress test confirmed the harder failure mode: resampling increased decoder MMCS while driving held-out explained variance to -5.94×10^{6}, demonstrating that resampling-induced consistency can coexist with catastrophic reconstruction failure.

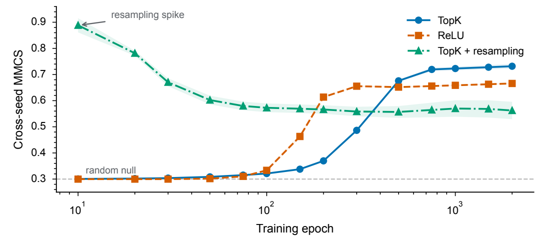

Figure 2. **Phantom consistency.** Dead-neuron resampling (green) produces a transient MMCS spike to 0.888 at epoch 10 that decays monotonically to 0.562 by epoch 2,000—below standard TopK (blue). The spike reflects a training dynamics artifact: resampled neurons briefly align across seeds before optimization disperses them.

4.5. Subspace Stability Precedes Feature Stability

The metric cascade—decoder geometry before activation patterns before firing overlap—suggests that coarse subspace structure aligns before individual features do. To test this directly, we measured the overlap of principal decoder subspaces across seeds at epoch 30 on mod-113. The top-8 principal decoder subspace has 2.98× more overlap than random (0.187 vs. 0.063), even when individual feature MMCS matches the random baseline. This ratio decays with subspace rank: 2.03× at rank 16, 1.43× at rank 32, 1.11× at rank 64. The decay is consistent with a picture where the leading singular modes of the decoder are partially shared across seeds from early in training, while higher-frequency modes remain random until later.

5. Discussion

Timescale separation as a high-dimensional phenomenon. The reconstruction objective for overcomplete dictionaries in $\mathbb{R}^d$ admits a large family of near-equivalent solutions related by rotations within the overcomplete subspace. Our EVT null characterizes the geometry of random points within this solution family: at d=128 with m=1024 dictionary elements, the expected maximum cosine similarity between random unit vectors is already 0.299. Optimization must navigate from this random starting configuration to a reproducible decomposition, and this symmetry-breaking process requires far more gradient steps than achieving low reconstruction loss. The subspace stability result (Section 4.5) provides direct evidence for this hierarchy: low-rank decoder structure emerges first, followed by feature-level alignment.

The cascade of metric convergence—decoder geometry, then activation patterns, then firing overlap—parallels known learning dynamics in deep networks, where low-frequency components are learned before high-frequency ones. In the dictionary learning setting, the "low-frequency" component is the subspace spanned by the decoder columns, while "high-frequency" corresponds to the precise arrangement of individual columns within that subspace. This ordering implies that short-budget stability measurements are dominated by the geometry of random vectors in the yet-unbroken symmetry group, not by any meaningful property of the learned representation.

The d_{sae}/erank result. The dissolution of the overparameterization threshold with extended training has a geometric interpretation. Higher $d_{\text{sae}}/\text{erank}$ ratios introduce more redundant directions in the dictionary, creating a larger continuous symmetry group that optimization must break. At short budgets, this manifests as a stability barrier; with sufficient training, the barrier is traversed and all conditions converge to similar reproducibility levels. This reframes what looked like a geometric constraint as a dynamics one. The practical implication is that practitioners cannot determine whether an overparameterized SAE "lacks reproducible structure" or "has not yet been trained long enough" from a single-budget evaluation.

Phantom consistency. The resampling spike illustrates how training dynamics can manufacture metrics that look like learned structure. The spike's decay profile—monotonically decreasing from epoch 10 through 2,000—is consistent with optimization eroding an initialization artifact rather than losing learned features. The co-occurrence of high decoder MMCS and catastrophic reconstruction failure (EV $= -5.94\times10^6$) in the Pythia streaming stress test makes clear that decoder agreement is not a sufficient condition for useful feature recovery.

Limitations. Our primary experiments use d=128; Pythia audits are selected-layer checks through 410M parameters, not a harmonized scaling law. MMCS is a geometric proxy for decoder-column agreement; high MMCS does not establish activation-level, causal, or semantic feature agreement. Our held-out Hungarian audits partially address this gap but do not include downstream-task or auto-interpretability validation. The Pythia MMCS curves are still climbing at the end of training, so we do not know their asymptotes. Dense mod-113 activations ($\text{erank}/d_{\text{model}} = 63\%$) may exaggerate the timescale separation relative to sparser activation distributions.

6. Conclusion

Reconstruction loss and decoder reproducibility in sparse autoencoders evolve on separated timescales. This separation persists across four SAE architectures, dimensions 128–1024, and model scales from algorithmic transformers through Pythia-410M. The apparent $d_{\text{sae}}/\text{erank}$ stability threshold is a training budget artifact. Dead-neuron resampling can produce phantom consistency that decays below unregularized baselines. These results suggest that SAE training dynamics should be studied jointly with reconstruction quality, and that reproducibility claims require calibration against dimension-appropriate random-dictionary nulls.

References

Arora, S., Bhaskara, A., Ge, R., and Ma, T. Simple, efficient, and neural algorithms for sparse coding. In *Conference on Learning Theory*, pp. 113–149, 2015.

Biderman, S., Schoelkopf, H., Anthony, Q. G., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., et al. Pythia: A suite for analyzing large language models across training and scaling. In *ICML*, 2023.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

Bussmann, B., Leask, P., and Nanda, N. BatchTopK sparse autoencoders. *arXiv preprint arXiv:2412.06410*, 2024a.

Bussmann, B. et al. A is for absorption: Studying feature splitting and absorption in sparse autoencoders. *NeurIPS*, 2024b.

Cui, J., Zhang, Q., Wang, Y., and Wang, Y. On the limits of sparse autoencoders: A theoretical framework and reweighted remedy. *ICLR*, 2026.

Gao, L., la Tour, T. D., Tillman, H., Goh, G., Troll, R., Radford, A., Sutskever, I., Leike, J., and Wu, J. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*, 2024.

Gribonval, R. and Schnass, K. Dictionary identification: Sparse matrix-factorization via ℓ_1-minimization. *IEEE Transactions on Information Theory*, 56(7):3523–3539, 2010.

Heap, T., Lawson, T., Farnik, L., and Aitchison, L. Sparse autoencoders can interpret randomly initialized transformers. *arXiv preprint arXiv:2501.17727*, 2025.

Korznikov, A. et al. Sanity checks for SAEs: Do SAEs beat random baselines? *arXiv preprint arXiv:2602.14111*, 2026.

Leask, P., Bussmann, B., et al. Sparse autoencoders do not find canonical units of analysis. *ICLR*, 2025.

Nanda, N., Chan, L., Lieberum, T., Smith, J., and Steinhardt, J. Progress measures for grokking via mechanistic interpretability. In *ICLR*, 2023.

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

O'Neill, C. et al. Compute optimal inference and provable amortisation gap in sparse autoencoders. *arXiv preprint arXiv:2411.13117*, 2024.

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

Power, A., Burda, Y., Edwards, H., Babuschkin, I., and Misra, V. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022.

Rajamanoharan, S., Conmy, A., Smith, L., Lieberum, T., Varma, V., Kramár, J., Shah, R., and Nanda, N. Jumping ahead: Improving reconstruction fidelity with JumpReLU sparse autoencoders. *arXiv preprint arXiv:2407.14435*, 2024.

Roy, O. and Vetterli, M. The effective rank: A measure of effective dimensionality. In *15th European Signal Processing Conference (EUSIPCO)*, pp. 606–610, 2007.

Song, X., Muhamed, A., Zheng, Y., Kong, L., Tang, Z., Diab, M. T., Smith, V., and Zhang, K. Position: Mechanistic interpretability should prioritize feature consistency in sparse autoencoders. *arXiv preprint arXiv:2505.20254*, 2025.

Spielman, D. A., Wang, H., and Wright, J. Exact recovery of sparsely-used dictionaries. *arXiv preprint arXiv:1206.5882*, 2012.

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., et al. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024.