DICTIONARY SIZE AND DEAD LATENTS HANDLING JOINTLY SET THE RECOVERY-OPTIMAL SPARSITY OF TOP-K SPARSE AUTOENCODERS

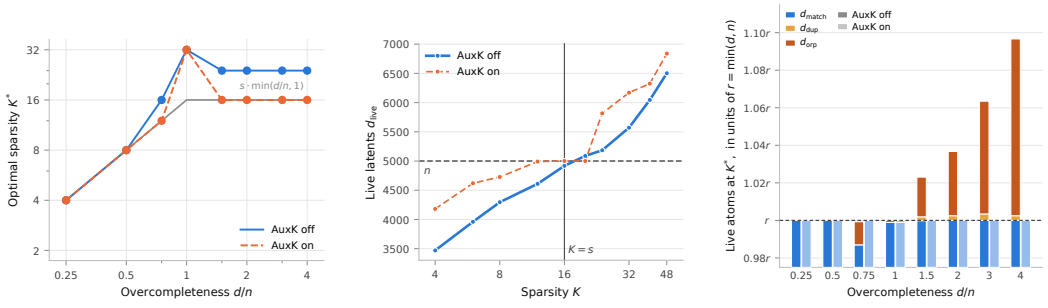
Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse autoencoders (SAEs) are trained to reconstruct activations under a sparsity constraint, but what interpretability requires is the recovery of the features underlying them. In a Top-K SAE toy model with a known ground-truth dictionary, we judge an SAE by the atoms it actually uses rather than by its nominal width, and partition its live atoms against the ground truth and against a second training run. Against the ground truth, the recovery-optimal sparsity is set jointly by dictionary size and dead-latent handling: below the ground-truth dictionary size it rises with the ratio of dictionary size to feature count, whereas above it the dead-latent handling sets its level, and when dead latents are appropriately handled, every feature at the optimum is carried by a single live atom. Against a second run, the same structure appears at the optimum, agreement and recovery come apart elsewhere—a low shared fraction arises even when every atom is correct—and the overlap between the runs, read as a capture–recapture experiment, estimates the number of features to within its sampling error at every feature count and dictionary size we tested.

1 INTRODUCTION

Sparse autoencoders (SAEs) decompose neural network activations into sparse, potentially interpretable components, on the view that networks carry more features than their dimensionality allows through superposition (Elhage et al., 2022). SAEs are used to interpret and steer models (Templeton et al., 2024; Marks et al., 2025) by assigning meaning to individual latents (Meissen et al., 2026; Pach et al., 2025), and such use has to be read through one of two views of what an SAE latent is. On one view, an SAE merely supplies a convenient sparse coordinate system for describing a representation; on the other, SAE latents recover the underlying sparse features themselves. Interpretation in the stronger sense aligns with the second view, and it requires checking directly whether the learned



(a) $K^\star(d/n)$ and $k_{\text{ref}}(d)$. (b) Live latent count at $d/n = 2$. (c) Active dictionary at $K^\star$.

Figure 1: $n = 5000$, $k_{\text{aux}} = 500$, four-seed mean. (a) $K^\star$ against the reference curve $s{\cdot}\min(d/n, 1)$. (b) $d_{\text{live}}(K)$ at $d/n = 2$; the dashed line is n and the vertical line $K = s$. (c) Live atoms at $K^\star$ split into d_{match}, d_{dup}, and d_{orp} ($\tau = 0.9$).

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.

dictionary has in fact recovered the features. We make this check in a toy model where the ground-truth feature dictionary is known: an s-sparse feature vector x is compressed as $h = Wx$, and the SAE is trained on h alone.

The training objective does not guarantee recovery: a low reconstruction error means that the latents provide a useful sparse decomposition of the activation, not that the learned dictionary corresponds to the features the activation encodes. Recent work has therefore begun to adopt, alongside reconstruction quality, consistency across independent training runs and recovery of ground-truth features in controlled settings (Sharkey et al., 2022; Song et al., 2026; Chen et al., 2026; Cui et al., 2026), and an incorrectly chosen sparsity can lead an SAE to learn mixtures of features (Chanin & Garriga-Alonso, 2026). The criteria used in practice—reconstruction error, the dead-latent rate, and the similarity between decoder atoms—were neither designed around nor validated against recovery (Appendix E). The compression itself does not rule out recovery: given W, orthogonal matching pursuit (OMP) (Foucart & Rauhut, 2013) recovers the sparse representation essentially exactly once the recovery sparsity reaches the true sparsity s (Appendix E).

We judge an SAE by the atoms it actually uses rather than by its nominal width. Once training ends not every one of the d atoms need be used, so the nominal dictionary size and the number of atoms in use can diverge, and metrics that compare only dictionary directions give no account of the atoms outside their matching. We therefore partition the live atoms twice: against the ground truth, into atoms matched to a distinct feature, duplicates of a feature already taken, and orphans that correspond to no feature; and against a second training run, into atoms shared with it, redundant copies, and strays (Section 3.2).

The two partitions let us ask two pairs of questions. With the ground truth: what sets the recovery-optimal sparsity, and what does the set of live atoms look like at that optimum? Without it: can the same structure be recognized from a second training run, and what does agreement between runs tell us about recovery? Against the ground truth, the live atoms show that the optimum is set jointly by dictionary size and dead-latent handling, and that above the ground-truth size each feature at the optimum is carried by a single live atom (Section 4). Against a second run, the same structure appears at the overcomplete recovery-optimal sparsity, agreement and recovery come apart elsewhere, and the overlap between the runs counts the features (Section 5). Our contributions are as follows.

- **Whether an SAE has recovered the features is decided by the atoms it actually uses, not by how many it has.** Partitioning the live atoms against the ground truth and against a second run separates correct atoms from duplicated and unmatched ones, a distinction that the nominal width and metrics comparing only dictionary directions do not see (Figures 1b and 1c).
- **Partitioning the live atoms against the ground truth, we find that the recovery-optimal sparsity is set jointly by dictionary size and dead-latent handling, and that above the ground-truth size every feature at that optimum is carried by a single live atom.** Below the ground-truth size the optimal sparsity rises with d/n, unchanged over a fourfold range of feature counts, while above it the dead-latent handling sets the optimum, with a non-monotone response to the revival budget; when dead latents are appropriately handled, the live count equals the feature count however many times the capacity exceeds it, with no duplicates or orphans (Figures 1a and 1c).
- **Partitioning them against a second training run, we recover the same picture at the optimum, find that agreement between runs is a different thing from recovery, and estimate how many features there are.** The pairwise partition agrees with the ground-truth one atom for atom at the optimum; a low shared fraction arises whenever the live atoms are fewer than the features, even when every atom is correct; and read as a capture–recapture experiment, the overlap between the runs recovers the number of features (Figures 2 and 3).

2 Background and Related Work

An SAE maps an activation h to a latent code z with an encoder and reconstructs $\hat{h} = Dz$; the columns of D, called atoms, form the learned dictionary. Sparsity is imposed by ℓ_1 regularization (Bricken et al., 2023; Huben et al., 2024), JumpReLU (Rajamanoharan et al., 2024), BatchTopK

(Bussmann et al., 2024), or Top-K (Makhzani & Frey, 2014); a Top-K SAE keeps the k largest latents for each input, so that $\|z\|_0 = k$ (Gao et al., 2025), and we use it because sparsity is then a single integer hyperparameter.

2.1 Recovery Metrics

Sharkey et al. (2022) swept the L_1 coefficient and the dictionary size together, measured ground-truth recovery with mean max cosine similarity (MMCS), and found the two hyperparameters not independent; they also noted that MMCS does not reflect duplicated copies. Song et al. (2026) measure recovery with GT-MCC, the mean of $|\cos|$ under the optimal one-to-one matching between learned and ground-truth atoms. Chen et al. (2026) give a training procedure with recovery guarantees under statistical assumptions, and Cui et al. (2026) show that reconstruction-based SAEs do not in general recover features completely but can when the ground-truth sparsity is strong enough. MMCS and GT-MCC compare only dictionary directions, and atoms outside the matching contribute nothing; Section 3.2 closes this blind spot by partitioning the live atoms into three groups.

2.2 Sparsity Selection

SAEs are usually tuned on the trade-off between reconstruction error and L_0: Gao et al. (2025) give scaling laws with reconstruction as the criterion, and Ayonrinde et al. (2024) select width and L_0 jointly with information theory. Chanin & Garriga-Alonso (2026) show that the wrong L_0 changes the learned dictionary qualitatively: too low an L_0 absorbs correlated features into a single latent, and too high an L_0 yields a different degenerate solution.

2.3 Dead-Latent Handling

Overcomplete SAEs develop dead latents that are rarely or never selected (Zaigrajew et al., 2025), so the nominal dictionary size and the number of atoms actually used diverge.

AuxK (Gao et al., 2025) treats a latent as dead when it has not fired for a fixed number of training steps and adds an auxiliary loss in which the top k_{aux} dead latents reconstruct the residual $e = h - \hat{h}$ of the main reconstruction:

$$\mathcal{L} = \|h - \hat{h}\|_2^2 + \alpha \frac{\|e - \hat{e}\|_2^2}{\text{Var}(e)}, \qquad \hat{e} = Dz_{\text{aux}}, \tag{1}$$

where z_{aux} keeps the k_{aux} largest encoder pre-activations among the dead latents. Because the target is the residual, revived atoms are pulled toward directions not yet represented. We treat the handling procedure and its budget k_{aux} as one axis of the experiment.

2.4 Consistency Across Seeds

Agreement between independently trained dictionaries has been used as a criterion. Sharkey et al. (2022) compared each dictionary with larger ones trained separately and took the size at which their mean max cosine similarity peaks as the right dictionary size, and Song et al. (2026) proposed the pairwise mean correlation coefficient between two runs (PW-MCC), the mean absolute cosine similarity under the optimal one-to-one matching of their atoms, as a proxy for ground-truth recovery. Disagreement, in turn, has been read as evidence against the features: Paulo & Belrose (2026) found that SAEs trained on the same data with different seeds share only a minority of their features and read the learned features as one useful decomposition among many, and Leask et al. (2025) found that the latents an SAE learns change with its width, larger SAEs learning novel latents and compositions of smaller ones, so that no single width yields a canonical set. Disagreement has also been treated as a defect to remove: mutual feature regularization encourages SAEs trained in parallel to learn similar features (Marks et al., 2024), archetypal SAEs constrain the atoms to the convex hull of the data to stabilize the dictionary across runs (Fel et al., 2025), and ordered SAEs, which extend Matryoshka SAEs (Bussmann et al., 2025), impose a strict ordering on the latents to improve consistency across seeds (Wang et al., 2025). None of these asks how much agreement two runs should show when every atom is correct but there are more features than atoms; without that baseline, the same score can mean perfect recovery at one width and poor recovery at another (Section 5.2).

2.5 Conjectured Reference Curve

Given a dictionary size d and a dead-latent handling procedure, we study

$$K^\star(d) = \arg\max_k \ \text{GT-MCC}(d, k), \tag{2}$$

with GT-MCC as in Section 3.2. For $d < n$, Chanin & Garriga-Alonso (2026) propose that a narrower SAE, which should learn the d features that fire most often, should use as its L_0 the expected number of those features active per input; with uniform firing frequencies, as in our model, this is $s \cdot (d/n)$. At $d = n$, Song et al. (2026) report that the optimal sparsity is s. For $d > n$ there is no prior observation, and we conjecture that the optimum stops at s, since added capacity does not change the number of features active per sample. Together these give

$$k_{\text{ref}}(d) = s \cdot \min\left(\frac{d}{n}, 1\right), \tag{3}$$

which we use as a reference curve; since Chanin & Garriga-Alonso (2026) tested a single width, whether either branch holds across widths is for the experiment to decide.

3 Setup

3.1 Toy Model

Each sample activates s of n ground-truth features, $x \in \mathbb{R}^n$ with $\|x\|_0 = s$, and is compressed by $W \in \mathbb{R}^{m \times n}$; the SAE, with $E \in \mathbb{R}^{d \times m}$ and $D \in \mathbb{R}^{m \times d}$, receives only h:

$$h = Wx, \qquad z = \text{TopK}_k(Eh), \qquad \hat{h} = Dz.$$

The columns W_i are the ground-truth atoms, standard Gaussian vectors normalized to unit norm. The s active coordinates are drawn uniformly without replacement, so firing frequencies are uniform, and the amplitudes are drawn from $\text{Uniform}(0.5, 1)$ (Appendix A). We fix $m = 1000$ and $s = 16$, and the main configuration uses $n = 5000$. The SAE has no biases, untied encoder and decoder (E initialized to $D^\top$), and no ReLU before Top-K, and its decoder columns are renormalized to unit norm after every step.

3.2 Measuring Recovery

GT-MCC. Our primary recovery metric is the ground-truth mean correlation coefficient (GT-MCC) of Song et al. (2026): with C_{ji} the cosine between atom D_j and feature W_i, the Hungarian algorithm (Kuhn, 1955) finds the one-to-one matching that maximizes the sum of $|C_{ji}|$, and GT-MCC is that sum divided by $\min(d, n)$ (Appendix B).

Partition against the ground truth. The atoms that enter the Top-K set for at least one evaluation sample are live, and their number is the active dictionary size d_{live}. A live atom D_j is *aligned* to W_i if $|C_{ji}| \geq \tau = 0.9$, and an atom aligned to no feature is an *orphan*; the same Hungarian matching on the aligned pairs splits the rest into *matched* atoms and *duplicates*, which arise exactly when two or more atoms are aligned to the same feature (Appendix B). Thus

$$d_{\text{live}} \ = \ d_{\text{match}} + d_{\text{dup}} + d_{\text{orp}}, \tag{4}$$

and the learned dictionary corresponds one-to-one with the ground truth when $d_{\text{match}} = \min(d, n)$ and $d_{\text{dup}} = d_{\text{orp}} = 0$. The results are insensitive to the choice of τ (Appendix C.5).

Partition against a second run. A real model provides no W, so we also partition the live atoms against a second run that differs only in the training seed. The live atoms of the second run stand in for the columns of W, and those of the first are classified as above: aligned to a partner if their $|\cos|$ is at least δ, set aside if aligned to no partner, and otherwise assigned by Hungarian matching on the aligned pairs. We name the pairwise classes separately,

$$d_{\text{live}} \ = \ d_{\text{shared}} + d_{\text{redundant}} + d_{\text{stray}}, \tag{5}$$

where *shared*, *redundant*, and *stray* play the roles of matched, duplicate, and orphan, and use $\delta = \tau = 0.9$ unless stated otherwise. Each pair is classified in both directions and the two labellings are averaged. The partition does not depend on which run serves as the reference: d_{shared}, the number of edges of a single bipartite matching, is identical in both directions, the live counts differ by at most 4%, and the class fractions by a total-variation distance of at most 0.031; lowering δ to 0.6 changes the agreement with the ground-truth partition by less than 0.007 (Appendix D).

3.3 Experimental Setup

The main configuration sweeps d over eight points from $0.25n$ to $4n$ and K over ten points from 4 to 48, crossed with AuxK off and on, and adds $K = 2$ at $d/n = 0.25$, where the reference value falls on the edge of the grid. We repeat the protocol with AuxK on over $0.5 \leq d/n \leq 2$ at $n \in \{2500, 10000\}$, and retrain the overcomplete cells at $n = 10000$ with k_{aux} from 500 to 8000. For AuxK we use $\alpha = 1/32$, count a latent as dead after $2{,}000$ steps without firing, and set $k_{\text{aux}} = m/2 = 500$ unless stated otherwise. All SAEs are trained for $40{,}000$ steps with Adam (Kingma & Ba, 2015) (learning rate 3×10^{-4}, batch size 4096), every configuration is repeated as two dictionary seeds $\times$ two training seeds, and evaluation uses a held-out set fixed per dictionary seed (Appendix A). For each d, $K^\star$ is the smallest K whose GT-MCC lies within one standard error of the four-seed maximum (Appendix A).

4 Results with Ground Truth

4.1 $K^\star$ Against the Reference Curve

All eight dictionary sizes have an interior optimum under both AuxK conditions (Figure 4, Appendix C): when K is too small a sample's s features cannot be accommodated, and when it is too large the alignment breaks down. With AuxK on, $K^\star$ is exactly 4, 8, and 12 at $d/n = 0.25$, 0.5, and 0.75, matching the reference values, and $16 = s$ throughout $d/n \in \{1.5, 2, 3, 4\}$ (Figure 1a), so more than doubling the dictionary size does not pull the optimal sparsity up with it. $K^\star$ agrees with the reference curve at seven of the eight sizes, although three points on the rising branch and four on the flat one show only that the observations are not inconsistent with it. With AuxK off, the overcomplete optimum settles at $K^\star = 24 = 1.5s$ (Section 4.2).

Repeating the protocol at $n = 2500$ and 10000, all three feature counts give $K^\star = 8$ and 12 at $d/n = 0.5$ and 0.75, so a fourfold change in the feature count leaves the rising branch on the reference curve (Figure 5, Appendix C). The flat branch departs at $n = 10000$, rising to 20 and 24 for $d/n \geq 1.5$. Section 4.2 shows that this departure moves with the revival budget; because n, interference, and s/n change together, we cannot rule out a contribution from the feature count itself. The exception is $d = n$, where $K^\star$ lies well above s at every feature count: $32 = 2s$ at $n = 5000$ under both AuxK conditions, 40 at $n = 2500$, and at least 48, the edge of the grid, at $n = 10000$. At $n = 5000$, $K = 16$ leaves 8–11% of the atoms dead (Table 2), so at least as many features lack a live atom, and GT-MCC is 0.933; at $K = 20$ nearly all atoms are live with AuxK on (Table 5) and GT-MCC is 0.997, within 0.001 of its value at every larger K up to 48. The shortfall at $K = s$ is thus a cliff at small K, with the optimum on the flat top above it rather than at a sharp peak. Song et al. (2026) report $K^\star = s$ at $d = n$ with $m = 8$, $n = d = 40$, $s = 8$, and no auxiliary loss; whether the cliff is a property of matched capacity in general or of this scale is open.

4.2 The Optimum Depends on Dead-Latent Handling

Leaving dead latents alone moves the optimum. With AuxK off the overcomplete optimum is $K^\star = 24 = 1.5s$, and although d_{match} reaches 5000, the dictionary there retains between 106 and 471 orphans, depending on the width, and 10–17 duplicates (Figure 1c; Table 4, Appendix C.4). This shift is not confined to the overcomplete regime: at $d/n = 0.75$ the off optimum is 16, above the reference value of 12, with 46 orphans.

Which directions are learned matters, not how many atoms survive. Under both conditions the live fraction d_{live}/d rises with K (Figure 6, Appendix C.2). With AuxK off, GT-MCC rises along with it, because on the main path only atoms that pass Top-K receive gradient, so when K

is small the remaining atoms are left without a learning signal. AuxK circumvents this constraint without changing K: the top k_{aux} dead latents also receive gradient in every step, so the number of atoms trained grows from K to $K + k_{\text{aux}}$, and because their target is the residual, they are pulled toward feature directions the main path has not yet captured. With AuxK on, the live fraction still spreads over the same range while GT-MCC concentrates in a narrow band of 0.969–1.000, although AuxK lowers the K-sweep mean dead rate by only 3–14% (Appendix C). What separates recovery is therefore not how many atoms survive but which directions were learned.

The revival budget moves the optimum. How far AuxK moves the optimum depends on its budget k_{aux}. At $n = 10000$, where the flat branch lies above s with $k_{\text{aux}} = 500$ (Section 4.1), raising the budget to 1000, the same 10% of the feature count as at $n = 5000$, brings $K^\star$ back to the reference value 16 at $d/n = 1.5$ but only from 24 to 20 at $d/n = 2$. The response is not monotone: at $d/n = 2$, $k_{\text{aux}} = 2000$ brings $K^\star$, the NMSE(Appendix B) minimum, and $d_{\text{live}} = n$ together to $K = s$, while at 4000 $K^\star$ returns to 20 and d_{live} no longer equals n (Appendix C.3). This is inconsistent with the reading that k_{aux} need only be scaled with the feature count. With every nominal setting held fixed, the revival budget alone moves $K^\star$ and the number of live atoms together over a narrow, non-monotone working range, so the optimal sparsity is a function of dead-latent handling as well as of the dictionary size.

4.3 Structure at the Optimum: One Live Atom per Feature

The active size is pinned to n. With AuxK on at $k = s = 16$, $d_{\text{live}} = n = 5000$ at all four overcomplete widths, in all four seeds to the integer (Figure 1b; Table 4, Appendix C.4). This is not forced by the construction: in the same SAEs d_{live} exceeds n at $k = 24$ (about 5300–6100) and falls slightly short at $k = 12$ (4948–5000), and with AuxK off it varies continuously with d and k and is not pinned at $k = s$ (about 4770–5150 at $k = 16$).

The live atoms correspond one-to-one with the features. $d_{\text{live}} = n$ alone does not guarantee this, since some live atoms may be duplicates or orphans; Figure 1c and Table 4 (Appendix C.4) give the decomposition at $K^\star$. With AuxK on and $d > n$, $d_{\text{match}} = n$ and $d_{\text{dup}} = d_{\text{orp}} = 0$ at all four widths, so every live atom corresponds to a different feature; for $d < n$, $d_{\text{match}} = d$ and the rest are zero, so the learned dictionary corresponds to a size-d subset of W. At $d = n$, d_{match} falls short of n under both conditions at 4995 with 5 orphans, the only width at which the correspondence fails with AuxK on.

The pinning follows the optimum, not s. Across every four-seed configuration with $d > n$, wherever some K has $d_{\text{live}} = n$ in all four seeds, $K^\star$ is one of those K (Table 3, Appendix C.3). Where the two separate, at $n = 10000$ with $k_{\text{aux}} = 500$, the exact pinning occurs at $K^\star = 20$ rather than at $s = 16$: the agreement is attached to the recovery optimum, not to the sparsity equalling s. The converse does not hold: at $n = 10000$, $d/n = 2$ with $k_{\text{aux}} = 8000$, $K^\star = 16 = s$, yet d_{live} at $K = 16$ is 9995–9998 (Appendix C.3).

5 Results without Ground Truth

Section 4 used the ground-truth dictionary W, which a real model does not provide. Here we partition the live atoms against a second training run instead (Section 3.2) and ask what the pairwise partition says about the ground truth, where and why it departs from it, and how that departure yields an estimate of the number of features.

5.1 Agreement at the Optimum and Two Directions of Departure

For $d > n$ at $K^\star$, the pairwise label equals the ground-truth label for every atom with AuxK on (agreement 1.0000) and for all but 0.09% with AuxK off (0.9991); at $d = n$ the agreement is 0.998. This is consistent with the use of consistency as a proxy for recovery by Song et al. (2026). Away from the optimum the two depart in a direction fixed by the class (Figure 2).

Stray over-counts orphans but misses none. The two stray panels are red or neutral throughout: $d_{\text{stray}} \geq d_{\text{orp}}$ in 159 of the 160 cells, and the remaining cell falls short by less than one atom. For $d > n$ the red is confined to $K < K^\star$, where only 8% of stray atoms are orphans, against 96% at or

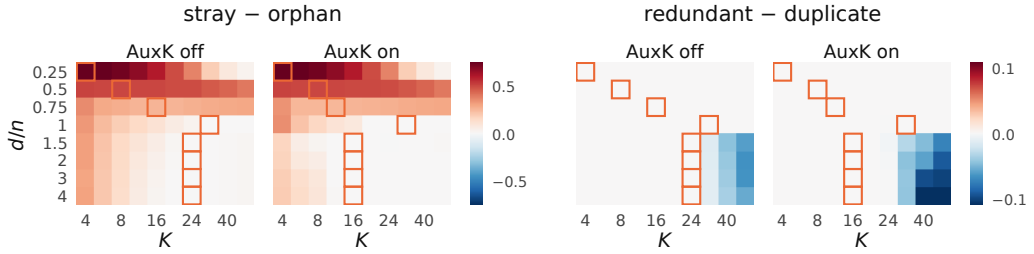


Figure 2: Departure of the pairwise partition from the ground-truth one ($n = 5000$, $\tau = \delta = 0.9$, four-seed mean): $(d_{\text{stray}} - d_{\text{orp}})/d_{\text{live}}$ and $(d_{\text{redundant}} - d_{\text{dup}})/d_{\text{live}}$, each with AuxK off and on. Red: the pair assigns more atoms to the class than the ground truth; blue: fewer. Orange boxes mark $K^\star$.

Table 1: Recovery and run-to-run consistency for $d < n$ ($n = 5000$, four seeds). The shared fraction is $d_{\text{shared}}/d_{\text{live}}$ at $K^\star$. PW-MCC averages over all d atoms; nearly all atoms are live at these K, and restricting it to live atoms changes it at $K^\star$ by at most 0.0005.

		at $K^\star$				at $\arg\max_K$ PW-MCC	
d/n	AuxK	GT-MCC	shared frac.	PW-MCC	$K^\star$	K	GT-MCC
0.25	off	0.9962	0.258	0.3346	4	32	0.7749
0.5	off	0.9977	0.503	0.5526	8	48	0.8933
0.75	off	0.9948	0.735	0.7741	16	20	0.9915
0.25	on	0.9962	0.258	0.3346	4	32	0.7749
0.5	on	0.9977	0.503	0.5526	8	48	0.8927
0.75	on	0.9985	0.753	0.7754	12	24	0.9875

above $K^\star$; for $d < n$ it extends across every K. Conversely, for $d > n$ more than 99.9% of orphans are labelled stray.

Redundant under-counts duplicates. The two redundant panels are blue or neutral throughout: $d_{\text{redundant}} \leq d_{\text{dup}}$ in all 160 cells, and the blue appears only above $K^\star$ for $d > n$, where duplicates exist. Duplicates are reproduced across runs: when both runs place extra copies on the same feature, the copies pair with each other and are counted as shared, which happens to 17–67% of duplicates above $K^\star$.

With AuxK on, $d_{\text{stray}} + d_{\text{redundant}}$ is zero at $K^\star = 16$ at all four overcomplete widths and at most 2.5 at $K = 20$, so the pairwise partition is cleanest at the optimum; with AuxK off, orphans remain at every K, the minimum is not zero, and at $d/n \geq 3$ it falls below $K^\star = 24$ (Table 8, Appendix D.1).

5.2 Chance Agreement and Consistency as a Proxy

Let $F^{(1)}$ and $F^{(2)}$ be the sets of features the two runs cover, of sizes $d^{(1)}_{\text{match}}$ and $d^{(2)}_{\text{match}}$. If each run covers a uniformly random subset of the n features, independently of the other, the overlap is hypergeometric with

$$\mathbb{E}\left|F^{(1)} \cap F^{(2)}\right| = \frac{d^{(1)}_{\text{match}}\, d^{(2)}_{\text{match}}}{n}. \tag{6}$$

Over the 309 labellings with no duplicates and orphans below 1% of d_{live}, the observed d_{shared} is 0.994–1.037 times this prediction. For $d < n$ every atom at $K^\star$ is a correct feature (Figure 1c), yet only a fraction close to d/n has a partner in the other run, because the two runs represent different subsets of the features. The same mechanism produces the stray excess below $K^\star$ for $d \geq n$, where dead latents leave fewer than n live atoms.

Low consistency need not mean wrong atoms. The same baseline sets the level of PW-MCC.[1] For $d < n$ no atom is dead, so PW-MCC is approximately $f + (1 - f)\, c_0$, where $f = d_{\text{shared}}/d$ is the shared fraction and $c_0 \approx 0.1$ the mean cosine of the pairs the matching is forced to form among unshared atoms: at $K^\star$, GT-MCC is at least 0.995 while PW-MCC is 0.33, 0.55, and 0.77 (Table 1). In this regime the level of PW-MCC is set by d/n, not by how well the features were recovered.

High consistency need not mean right atoms. For $d < n$, raising K beyond $K^\star$ moves the two metrics in opposite directions: at $d/n = 0.25$, going from $K = 4$ to $K = 32$ lowers GT-MCC from 0.996 to 0.775 while PW-MCC rises from 0.335 to 0.427, so the K that maximizes PW-MCC lies far above $K^\star$ at all three widths (Table 1). What the runs come to share are mixtures of features that recur below the alignment threshold: at $\delta = 0.9$ only 3.5% of the orphans for $d < n$ find a partner in the other run, but at $\delta = 0.6$ the fraction is 45%. For $d > n$ a different effect dominates: because PW-MCC averages over all d atoms, dead atoms dilute it (Appendix D).

Paulo & Belrose (2026) read a minority of shared features as evidence that the learned features are one of many possible decompositions (Section 2.4). In our setting the ground-truth set is the same for both runs, yet a low shared fraction arises whenever the live atoms are fewer than the features, even when every atom is a correct feature; agreement alone therefore cannot tell learned features that are not true features from a shortfall of atoms.

5.3 Estimating the Number of Features

The chance overlap that limits consistency as a proxy also carries information about n. Reading the two runs as a capture–recapture experiment, the first run marks $d^{(1)}_{\text{live}}$ atoms, the second captures $d^{(2)}_{\text{live}}$, and d_{shared} of them are recaptures; we use the Lincoln–Petersen estimator (Petersen, 1896; Lincoln, 1930) with the bias correction of Chapman (1951),

$$\hat{n} = \frac{\left(d^{(1)}_{\text{live}} + 1\right)\left(d^{(2)}_{\text{live}} + 1\right)}{d_{\text{shared}} + 1} - 1. \tag{7}$$

The estimator counts features, whereas $d^{(1)}_{\text{live}}$ and $d^{(2)}_{\text{live}}$ count atoms, so it is exact only when each live atom is a distinct feature; we call a pair *contaminated* when either run has a live atom that is an orphan or a duplicate. Orphans can only inflate the estimate, and duplicates do not reverse this in practice (Appendix D). At small K there is no room for duplicates: an atom that does not win a feature dies rather than copying one, and at $K = 4$ no cell has a duplicate and no cell has more than 18 orphans. When fewer than n atoms are live the runs sample different subsets and equation 6 applies; when nearly n atoms are live, which happens only for $d \geq n$, the estimate is close to a census.

The estimate across feature counts. We read the estimate at the smallest K, where no cell has a duplicate and none has more than 18 orphans. Across all 26 combinations of $n \in 2500, 5000, 10000$, d/n from 0.25 to 4, and the AuxK arm, $\hat{n}/n$ at K = 4 lies between 0.969 and 1.013 (Table 11), within the estimator's own sampling error: the relative standard error is 4% at $d/n = 0.25$ and under 0.5% at $d/n \geq 0.75$, and on clean pairs the observed error is 0.70 of the predicted coefficient of variation against 0.67 expected (Table 12).

6 Discussion

At a fixed width, the number of live atoms and their partition change with the sparsity and the dead-latent handling (Figures 1b and 1c), and recovery is separated by which directions the live atoms take rather than by how many survive (Section 4.2); the nominal width says little about what an SAE has learned. Dead latents, proposed as a sign that the dictionary exceeds the feature count (Sharkey et al., 2022), also appear below it (11–14% at $d/n = 0.75$, $K = 4$), so a dead-latent count alone does not show that the width is too large. Because the optimum moves with the dead-latent handling and the criteria used in practice do not point at it reliably (Appendix E), a rule for choosing

[1] We compute PW-MCC with the reference implementation of Song et al. (2026); Appendix B compares it with their definition.

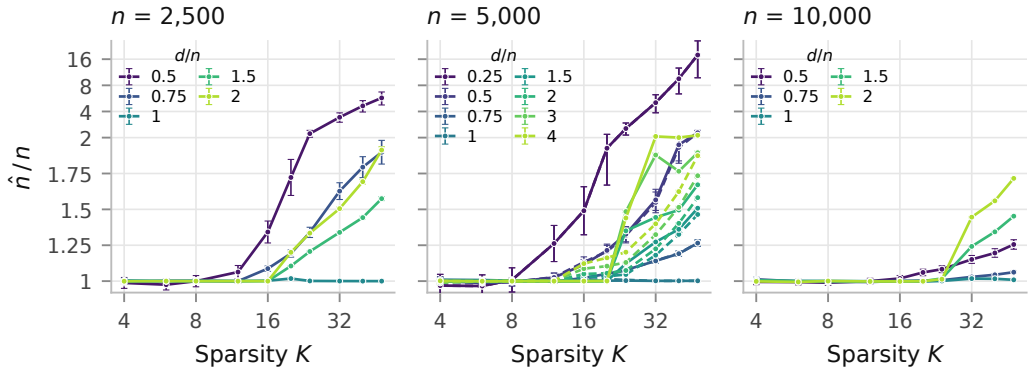


Figure 3: The Chapman estimate $\hat{n}/n$ against K ($\delta = 0.9$; 95% intervals, Appendix D.4). Solid lines: AuxK on; dashed: AuxK off ($n = 5000$ only). The y axis is linear up to 2 and logarithmic above.

the sparsity would have to account for the relative relationship between width and sparsity, for the active rather than the nominal dictionary, and for the dead-latent handling.

At the optimum, agreement between runs tracks recovery, consistent with Song et al. (2026); away from it the two come apart. When the live atoms are fewer than the features, correct runs share only the chance overlap of equation 6, so a low shared fraction, read by Paulo & Belrose (2026) as evidence against the features, does not by itself distinguish wrong features from a shortfall of atoms; conversely, maximizing agreement would pick a K far above the optimum for $d < n$ (Table 1). Whether agreement is used as a criterion or disagreement as evidence against the features, the observed overlap should be read against equation 6. The same overlap counts the features: the capture-recapture estimate at the smallest K lies within sampling error of n in every combination we tested (Section 5.3), supplying one of the two unknowns of the reference curve in the toy model.

7 Limitations and Open Questions

The estimate of the feature count and the selection of the sparsity have been shown under uniform firing frequencies and a constant L_0, with exactly s features active per sample, and not for real models. The values of m and s were fixed, and the feature-count comparison covers $d/n \leq 2$; because n, interference, and s/n move together, we could not separate the effect of the feature count into its causes. At $d = n$, $K^\star$ is not s under any condition and the one-to-one correspondence fails with AuxK on at all three feature counts; the dead atoms below $K = 20$ account for this at $n = 5000$, but whether they do at other scales is open; and we found no rule that predicts the narrow, non-monotone working range of k_{aux}.

8 Conclusion

We judged an SAE by the atoms it actually uses rather than by its nominal width, and partitioned those atoms against the ground truth and against a second training run. Against the ground truth, the recovery-optimal sparsity is set jointly by dictionary size and dead-latent handling, and above the ground-truth size every feature at the optimum is carried by a single live atom when dead latents are appropriately handled. Against a second run, the same structure appears at the optimum, agreement and recovery come apart elsewhere, and the overlap between runs counts the features. These results hold under uniform firing frequencies and a constant L_0; what they say a rule for real activations must account for is the relative relationship between width and sparsity, the active rather than the nominal dictionary, and the dead-latent handling.

AI USE STATEMENT

In this work, we used generative AI tools to assist with implementation and debugging code for experiments, and compiling of experimental results. We additionally used generative AI tools for literature retrieval and discovery, writing and polishing portions of the manuscript, and improving the clarity and readability of the writing. All AI-assisted code was used only after inspection by the authors, and all AI-assisted manuscript text was reviewed and verified by the authors. We take responsibility for the final content of this work, including text, claims, code, and other artifacts produced with the aid of generative AI.

REFERENCES

Kola Ayonrinde, Michael T. Pearce, and Lee Sharkey. Interpretability as compression: Reconsidering SAE explanations of neural activations with MDL-SAEs. *arXiv preprint arXiv:2410.11179*, 2024. URL https://arxiv.org/abs/2410.11179.

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

Bart Bussmann, Patrick Leask, and Neel Nanda. BatchTopK sparse autoencoders. In *NeurIPS 2024 Workshop on Scientific Methods for Understanding Deep Learning*, 2024. URL https://arxiv.org/abs/2412.06410.

Bart Bussmann, Noa Nabeshima, Adam Karvonen, and Neel Nanda. Learning multi-level features with Matryoshka sparse autoencoders. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*. PMLR, 2025. URL https://proceedings.mlr.press/v267/bussmann25a.html.

David Chanin and Adrià Garriga-Alonso. Sparse but wrong: Incorrect L_0 leads to incorrect features in sparse autoencoders. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026. URL https://arxiv.org/abs/2508.16560.

Douglas G. Chapman. Some properties of the hypergeometric distribution with applications to zoological sample censuses. *University of California Publications in Statistics*, 1:131–160, 1951.

Siyu Chen, Heejune Sheen, Xuyuan Xiong, Tianhao Wang, and Zhuoran Yang. Taming polysemanticity in LLMs: Theory-grounded feature recovery via sparse autoencoders. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=VtWkPIbAQ8.

Jingyi Cui, Qi Zhang, Yifei Wang, and Yisen Wang. On the limits of sparse autoencoders: A theoretical framework and reweighted remedy. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=DSOTgzeH3w.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022. arXiv:2209.10652.

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. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*. PMLR, 2025.

Simon Foucart and Holger Rauhut. *A Mathematical Introduction to Compressive Sensing*. Applied and Numerical Harmonic Analysis. Birkhäuser, New York, NY, 2013. ISBN 978-0-8176-4948-7. doi: 10.1007/978-0-8176-4948-7.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=tcsZt9ZNKD.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2309.08600.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *The Third International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

Harold W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97, 1955. doi: 10.1002/nav.3800020109.

Patrick Leask, Bart Bussmann, Michael Pearce, Joseph Bloom, Curt Tigges, Noura Al Moubayed, Lee Sharkey, and Neel Nanda. Sparse autoencoders do not find canonical units of analysis. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=9ca9eHNrdH.

Frederick C. Lincoln. Calculating waterfowl abundance on the basis of banding returns. Circular 118, United States Department of Agriculture, 1930.

Alireza Makhzani and Brendan Frey. k-sparse autoencoders. In *The Second International Conference on Learning Representations*, 2014. URL https://arxiv.org/abs/1312.5663.

Luke Marks, Alasdair Paren, David Krueger, and Fazl Barez. Enhancing neural network interpretability with feature-aligned sparse autoencoders. *arXiv preprint arXiv:2411.01220*, 2024. URL https://arxiv.org/abs/2411.01220.

Samuel Marks, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2403.19647.

Felix Meissen, Kenza Bouzid, Tristan Lazard, Shruthi Bannur, Daniel C. Castro, Javier Alvarez-Valle, and Stephanie Hyland. Automatic interpretation of visual concepts. In *ICLR 2026 Workshop on Unifying Concept Representation Learning (UCRL)*, 2026.

Mateusz Pach, Shyamgopal Karthik, Quentin Bouniot, Serge Belongie, and Zeynep Akata. Sparse autoencoders learn monosemantic features in vision-language models. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*, 2025. URL https://arxiv.org/abs/2504.02821.

Gonçalo Paulo and Nora Belrose. Sparse autoencoders trained on the same data learn different features. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2501.16615.

C. G. Johannes Petersen. The yearly immigration of young plaice into the Limfjord from the German Sea. *Report of the Danish Biological Station to the Home Department*, 6:1–48, 1896.

Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. Jumping ahead: Improving reconstruction fidelity with JumpReLU sparse autoencoders. *arXiv preprint arXiv:2407.14435*, 2024. URL https://arxiv.org/abs/2407.14435.

George A. F. Seber. *The Estimation of Animal Abundance and Related Parameters*. Griffin, London, second edition, 1982.

Lee Sharkey, Dan Braun, and Beren Millidge. Taking features out of superposition with sparse autoencoders. AI Alignment Forum. Interim research report, 2022. URL https://www.alignmentforum.org/posts/z6QQJbtpkEAX3Aojj/interim-research-report-taking-features-out-of-superposition.

Xiangchen Song, Aashiq Muhamed, Yujia Zheng, Lingjing Kong, Zeyu Tang, Mona T. Diab, Virginia Smith, and Kun Zhang. Mechanistic interpretability should prioritize feature consistency in sparse autoencoders. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pp. 2172–2210, 2026.

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

Sophie L. Wang, Alex Quach, Nithin Parsan, and John J. Yang. Enforcing orderedness in SAEs to improve feature consistency. In *NeurIPS 2025 Workshop on Mechanistic Interpretability*, 2025. URL https://arxiv.org/abs/2512.02194.

Vladimir Zaigrajew, Hubert Baniecki, and Przemyslaw Biecek. Interpreting CLIP with hierarchical sparse autoencoders. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*. PMLR, 2025.

A Experimental Details

This appendix gives the details of the setup.

Data. Each sample activates exactly $s = 16$ of the n features, drawn uniformly without replacement, with amplitudes drawn from $\mathrm{Uniform}(0.5, 1)$. The ground-truth atoms are i.i.d. standard Gaussian vectors normalized to unit norm.

Amplitude range. The lower bound of 0.5 removes a failure mode unrelated to the two axes we study. The score of an inactive latent is a sum of the interference shed by the co-active features, with standard deviation $\sigma = \sqrt{s\,\mathbb{E}[A^2]/m} = 0.097$ at our settings, and the largest of the d such scores is about $\sigma\,\Phi^{-1}(1 - 1/d)$, or 0.36 at $d = 10000$. Under $\mathrm{Uniform}(0, 1)$ the weakest of the s active coefficients has expectation $1/(s+1) = 0.06$, and 27% of active coefficients fall below the interference floor; those features cannot be recovered at any K or d, which would confound the sweeps of Section 4. With $\mathrm{Uniform}(0.5, 1)$ the weakest active coefficient has expectation 0.53 and none falls below. The range is held fixed across all settings; since $\mathbb{E}\|h\|^2 = s\,\mathbb{E}[A^2]$ does not depend on n, the loss scale and hence the learning rate transfer across the three feature counts. This calculation assumes an ideal dictionary and ignores the interference on the active scores themselves, so we use it to fix the direction of the dependence, $\sigma \propto \sqrt{s/m}$, rather than as a quantitative prediction.

Evaluation set. The live count and all partitions are computed on a held-out evaluation set fixed per dictionary seed, of $N = 65{,}536$ samples for $n \in \{2500, 5000\}$ and $131{,}072$ for $n = 10000$. Each feature therefore fires about 210 times in expectation at $n \in \{5000, 10000\}$ and 420 times at $n = 2500$, above 200 in every case; GT-MCC depends only on D and W and is unaffected by this choice.

Initialization and training. The decoder is initialized with i.i.d. standard Gaussian entries and unit-norm columns, and the encoder with its transpose; the two are separate parameters thereafter, and the initialization depends on both the dictionary seed and the training seed. All SAEs are trained for $40{,}000$ steps with Adam at learning rate 3×10^{-4} and batch size 4096, and the decoder columns are renormalized after every step. A preliminary gate at $d = n = 5000$, $K = 16$, AuxK off, over learning rates $\{10^{-4}, 3 \times 10^{-4}, 10^{-3}\}$ at $40{,}000$ steps ranked 10^{-4} first, but the fraction of features matched at $\tau = 0.9$ differed by only 0.7% from 3×10^{-4} (0.8996 against 0.8932; 0.8596 at 10^{-3}).

AuxK. A latent is a dead candidate when it has not entered the Top-K set in the last $2{,}000$ steps, a condition recomputed at every step. In our implementation the auxiliary term is normalized by the batch variance of the residual, $\|e-\hat{e}\|_2^2/\mathrm{Var}(e)$, with the residual detached, and is added with weight $\alpha = 1/32$. The number of latents AuxK actually trains is $\min(k_{\mathrm{aux}}, \text{number of dead candidates})$, with $k_{\mathrm{aux}} = m/2 = 500$ unless stated otherwise. The training-time window of $2{,}000$ steps covers 62–125 times as many samples as the evaluation set on which d_{live} is counted, so the two notions of a dead latent differ.

Grids. The main configuration uses $n = 5000$, $d/n \in \{0.25, 0.5, 0.75, 1, 1.5, 2, 3, 4\}$, and $K \in \{4, 6, 8, 12, 16, 20, 24, 32, 40, 48\}$, with AuxK off and on, plus $K = 2$ at $d/n = 0.25$. At $n \in \{2500, 10000\}$ the protocol is repeated with AuxK on for $d/n \in \{0.5, 0.75, 1, 1.5, 2\}$. The revival-budget runs use $K \in \{8, 12, 16, 20, 24, 32\}$, with $k_{\mathrm{aux}} = 1000$ at $n \in \{5000, 10000\}$ and $d/n \in \{1.5, 2\}$ ($K \geq 12$ at $n = 5000$), and $k_{\mathrm{aux}} \in \{1500, 2000, 4000, 8000\}$ at $n = 10000$, $d/n = 2$. Every configuration is repeated as two dictionary seeds $\times$ two training seeds.

Selecting $K^\star$. For each d we select the optimum of GT-MCC over K with a one-standard-error rule, since taking the maximum outright would let seed variation alone flip neighboring cells: every K within one standard error of the four seeds below the maximum is treated as indistinguishable from it, and the smallest such K is $K^\star$. We compare only the location of the optimum across d, not GT-MCC values themselves, because the level GT-MCC can attain depends on (m, d).

Hardware. Runs were distributed over RTX 3090, RTX 4090, and RTX 4070 Ti SUPER GPUs. Between the 3090 and the 4090 the data stream and the Top-K selections are bit-identical and the final loss differs by 3×10^{-6}, well below the variation across seeds.

B Metric Definitions

This appendix details the two partitions of Section 3.2, the PW-MCC used in Section 5.2, and the reconstruction error.

GT-MCC. With the columns of $D \in \mathbb{R}^{m \times d}$ and $W \in \mathbb{R}^{m \times n}$ normalized to unit norm, we form

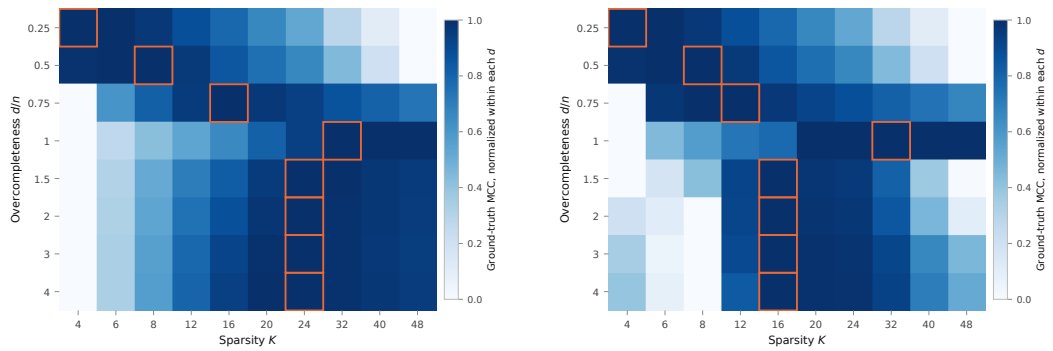
$$C_{ji} \;=\; \frac{\langle D_j,\, W_i \rangle}{\|D_j\|_2 \, \|W_i\|_2}, \tag{8}$$

find with the Hungarian algorithm the one-to-one matching M that maximizes the sum of $|C_{ji}|$, and average under it:

$$\text{GT-MCC}(D, W) \;=\; \frac{1}{r} \max_{M} \sum_{(j,i) \in M} \big|C_{ji}\big|, \qquad r = \min(d, n). \tag{9}$$

Partition against the ground truth. Only some of the d atoms are used once training ends; the number that enter the Top-K set for at least one evaluation sample is the active dictionary size d_{live}. We partition the live atoms in two stages. An atom D_j is *aligned* to a feature W_i if $|C_{ji}| \geq \tau$, with $\tau = 0.9$, and an atom aligned to no feature is an *orphan*. On the bipartite graph of aligned pairs we then solve the same Hungarian matching, so an atom can only be assigned to a feature it is aligned to; assigned atoms are *matched* and the rest *duplicates*. When $d_{\mathrm{match}} = \min(d, n)$ and $d_{\mathrm{dup}} = d_{\mathrm{orp}} = 0$, for $d > n$ every feature has a dedicated atom with no duplicated copies and no unmatched live atoms, and for $d < n$ the d atoms each correspond to a different feature, which is not the same as covering every feature.

Partition against a second run. The pairwise partition uses the same two-stage procedure with the live atoms of the second run in place of the columns of W and the threshold δ in place of τ. Both reference sets are restricted to live atoms: W has no dead columns, and if all d atoms of the second run were used, its dead atoms, which learned nothing, could be matched to live atoms of the first run and counted as agreement. The number of shared atoms is the number of edges of a single bipartite matching and is therefore the same whichever run serves as the reference; each pair is classified in both directions and the two labellings are averaged (Appendix D.2).



(a) AuxK off.

(b) AuxK on.

Figure 4: GT-MCC over the grid ($n = 5000$, four-seed mean), min–max normalized within each row; orange boxes mark $K^\star$. Color is comparable only within a row.

PW-MCC. Song et al. (2026) define PW-MCC as the mean of $|\cos|$ under the optimal one-to-one matching between the atoms of two runs. Their reference implementation, which we follow in Section 5.2, matches signed cosines over all d atoms of both decoders and divides by d. Across all 320 pairs the two conventions differ by at most 0.042 when all d atoms are matched and by at most 0.003 when only live atoms are, and the difference comes from dead atoms rather than from sign-flipped features: the fraction of pairs matched with a negative cosine has median 0 and maximum 0.0008, and the cell with the largest difference (AuxK on, $d = 7500$, $K = 16$; 0.8505 against 0.8925) has none. Under $|\cos|$, a dead atom can collect a large negative cosine with any atom of the other run as a positive contribution; under signed cosines it cannot. Where we restrict PW-MCC to live atoms, the matching is over the live atoms of both runs and the sum is divided by $\min(d^{(1)}_{\text{live}}, d^{(2)}_{\text{live}})$.

Reconstruction. We measure reconstruction with $\text{NMSE}_h = \mathbb{E}\,\|h - \hat{h}\|_2^2 / \mathbb{E}\,\|h\|_2^2$.

C Supplementary Results with Ground Truth

This appendix collects the results behind Section 4: the location of the optimum (C.1), dead-latent handling (C.2), the revival budget (C.3), the partition at the optimum and the live counts behind the pinning (C.4), and the threshold.

C.1 Location of the Optimum

Figure 4 shows GT-MCC over the full grid at $n = 5000$; every width has an interior optimum under both AuxK conditions (Section 4.1).

Interior optimum. At the left edge of the grid, $d/n = 0.25$, the optimum falls on $K = 4$ and could therefore be a boundary value, so we additionally trained $K = 2$: GT-MCC is 0.9932 at $K = 2$ against 0.9962 at $K = 4$ (1-SE $\approx 3 \times 10^{-5}$), confirming $K = 4$ as an interior optimum.

With AuxK off, $d/n \in \{1.5, 2, 3, 4\}$ all reach GT-MCC 0.9992–0.9993 at $K = 24$ and then fall to about 0.988 at $K = 48$. Even with latent capacity to spare, raising the number of active latents beyond the optimum degrades the dictionary alignment. The same structure holds with AuxK on; only the location of the optimum changes.

In the overcomplete regime the differences around the optimum are small. At $d = 7500$ with AuxK on, GT-MCC moves from 0.988 at $K = 4$ to 0.99971 at $K = 16$ and back to 0.988 at $K = 48$, so a twelvefold change in K moves the value within a band of only 0.012. The optimum is nonetheless resolved. In the same row the gap between $K = 16$ and $K = 12$ (0.99876) is under 0.001, but that is 21.7 times the spread of the four seeds, so it is not explained by seed variation. Direction

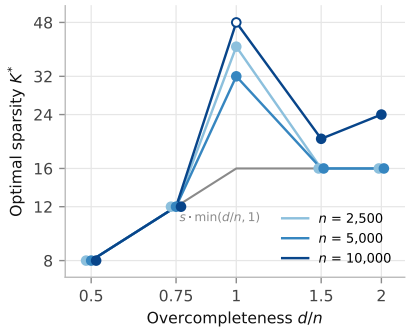


Figure 5: $K^\star(d/n)$ at three feature counts (AuxK on, $k_{\text{aux}} = 500$, four seeds, one-standard-error rule). A hollow marker is a lower bound at the edge of the grid; markers are nudged horizontally only where they would overlap.

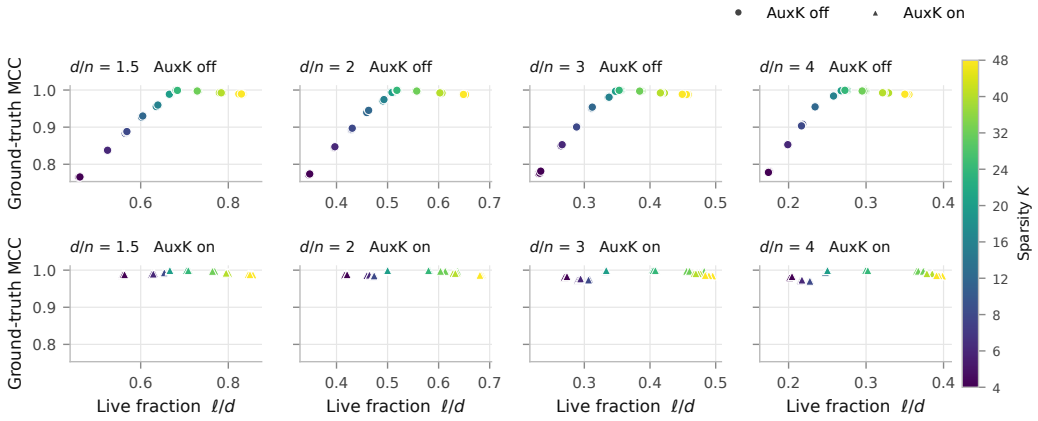


Figure 6: GT-MCC against the live fraction d_{live}/d for every cell with $d/n \geq 1.5$ ($n = 5000$; seeds plotted separately; color encodes K). With AuxK off (top) the points trace a single curve; with AuxK on (bottom) they spread over the same live fractions while GT-MCC stays near the top.

recovery simply works well over a wide range, which is why the range of values is narrow; within it, the location of the optimum is read off stably.

Three feature counts. Keeping the same protocol, we re-measured the rising branch with n changed to 2500 and 10000 (m and s fixed). This comparison examines only the location of $K^\star$; the absolute levels of the metric are not compared, because the range of d differs. The departure of the flat branch at $n = 10000$ is not noise: at $d/n = 2$, GT-MCC is 0.9662, 0.9735, 0.9922, 0.9988, and 0.9984 at $K = 12$, 16, 20, 24, and 32, so $K = 16$ is clearly not optimal.

The $d = n$ outlier. $d = n$ is the point at which spare capacity disappears. For $d/n \leq 0.5$ no atoms die, and at $d/n = 0.75$ almost all survive at $K^\star$ (3747–3750 of 3750), whereas at $d = n$ the live fraction is already 0.98 at $K = 24$, so there is no room to replace an atom that dies. AuxK is also at its weakest here, lowering the K-sweep mean dead rate only from 0.108 to 0.089.

C.2 Dead-Latent Handling

Section 4.2 attributes the effect of AuxK to the atoms it trains beyond the Top-K set. With $K = 4$ and $k_{\text{aux}} = 500$, five hundred atoms are trained on the auxiliary path while the main path trains four. With AuxK off at $d/n = 1.5$, $K = 4$ starts at a live fraction of 0.46 with GT-MCC 0.766, and by the time the live fraction reaches 0.7 GT-MCC has reached its ceiling (Figure 6).

Table 2: Fraction of atoms that are not live, $1 - d_{\text{live}}/d$, for $d \leq n$ ($n = 5000$, four-seed mean), AuxK off / on. At $d/n \in \{0.25, 0.5\}$ it is 0 at every K under both conditions.

d/n	$K = 4$	8	12	16	20	24	32	48
0.75	.136/.113	.028/.000	.010/.000	.001/.000	0/0	0/0	0/0	0/0
1	.326/.343	.184/.153	.146/.100	.106/.077	.057/.002	.019/.002	0/0	0/0

Table 3: $k_{\text{aux}} = 500$ against 1000 (AuxK on, four seeds). All values are computed on the common grid $K \in \{8, 12, 16, 20, 24, 32\}$. The "live $= n$" column lists the K at which d_{live} equals n exactly in all four seeds. The two $n = 5000$ rows have no $K = 8$ cell at $k_{\text{aux}} = 1000$ and are therefore computed on $K \in \{12, 16, 20, 24, 32\}$.

		$k_{\text{aux}} = 500$			$k_{\text{aux}} = 1000$		
n	d/n	$K^\star$	live $= n$	NMSE	$K^\star$	live $= n$	NMSE
5000	1.5	16	16, 20	16	16	12, 16, 20	16
5000	2	16	16	16	16	16	16
10000	1.5	20	20	20	16	16, 20	16
10000	2	24	none	20	20	none	20

With AuxK on, the K-sweep mean dead rate goes from 0.352 to 0.304 at $d/n = 1.5$ and from 0.740 to 0.716 at $d/n = 4$, reductions of 14% and 3%, and the reduction shrinks as the dictionary grows. Recovery nonetheless rises sharply (at $d/n = 3$, from 0.9811 to 0.9997 at $K = 16$ and from 0.9972 to 0.9994 at $K = 20$).

Dead latents are not confined to $d > n$ (Table 2). At $d/n = 0.75$ and $d = n$ they appear at small K under both conditions, and AuxK removes them at $d/n = 0.75$ from $K = 8$ on; this is the observation used in Section 6.

C.3 Revival Budget

At $n = 5000$, $k_{\text{aux}} = 500$ is 10% of the feature count. Retraining at the matched ratio $k_{\text{aux}} = 1000$ leaves $K^\star$ and the NMSE minimum unchanged at $n = 5000$, whereas at $n = 10000$ $K^\star$ drops from 20 to 16 at $d/n = 1.5$, matching the reference value, and from 24 to 20 at $d/n = 2$ (Table 3).

At $n = 10000$, $d/n = 2$, $k_{\text{aux}} = 2000$ brings all three quantities to $K = s$ (Figure 7): $K^\star = 16$, d_{live} is exactly n at $K = 16$ in all four seeds, and the NMSE minimum is at $K = 16$. At 4000, $K^\star$ returns to 20, the NMSE minimum grows from 0.0181 to 0.0292, and the exact pinning disappears; at 8000, $K^\star$ is 16 again but without pinning. The ratio-matched budget of 1000 did not restore $d/n = 2$, and 4000 was worse than 2000.

The mean d_{live} at $K = 16$ at $n = 10000$, $d/n = 2$ moves with the budget as $9149 \to 9181 \to 9604 \to 10000 \to 9899 \to 9996$ for $k_{\text{aux}} = 500$, 1000, 1500, 2000, 4000, and 8000, with seed-to-seed ranges of $68 \to 75 \to 234 \to 0 \to 18 \to 3$. It is hard to read a rule for setting the budget out of this.

C.4 Partition at the Optimum and Live Counts Behind the Pinning

Table 4 gives $K^\star$ and the partition of the live atoms at $K^\star$ for every width at $n = 5000$ (Section 4.3).

Tables 5 and 6 give the spread of the live count over the four runs behind the pinning of Section 4.3.

In pinned cells the four runs agree to the integer, while in the other cells they scatter by tens to hundreds of atoms.

C.5 Threshold

With AuxK off at $d/n = 2$, the fraction of features matched at threshold τ is the same to four decimals for $\tau \in \{0.7, 0.8, 0.9\}$ at every K, because recovered atoms clear 0.9 by a wide margin

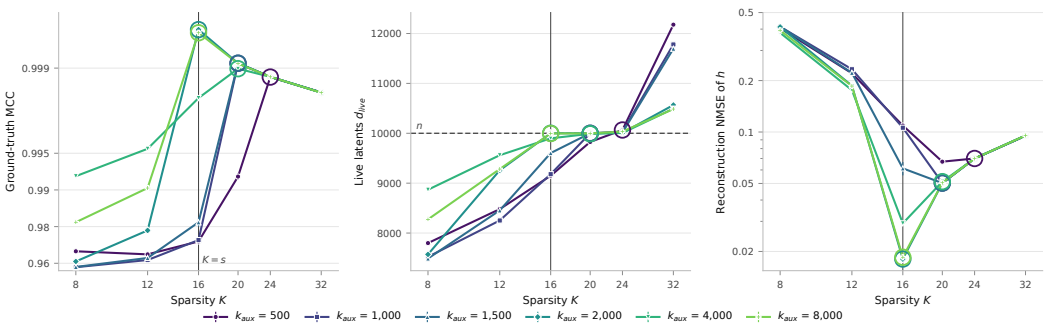


Figure 7: Six revival budgets at $n = 10000$, $d/n = 2$, four-seed mean $\pm$ 1 SE: (a) GT-MCC, (b) d_{live} (dashed: n), (c) NMSE_h (log scale). Hollow circles mark each budget's $K^\star$: 24, 20, 20, 16, 20, and 16 for $k_{\text{aux}} = 500$ to 8000.

Table 4: $K^\star$ and the partition of the live atoms at $K^\star$ ($n = 5000$, $k_{\text{aux}} = 500$, four-seed mean, $\tau = 0.9$). The $d/n = 0.25$ row uses the grid that includes $K = 2$.

	AuxK off					AuxK on				
d/n	$K^\star$	d_{live}	d_{match}	d_{dup}	d_{orp}	$K^\star$	d_{live}	d_{match}	d_{dup}	d_{orp}
0.25	4	1250	1250	0	0	4	1250	1250	0	0
0.5	8	2500	2500	0	0	8	2500	2500	0	0
0.75	16	3747.25	3701.25	0	46	12	3750	3750	0	0
1	32	5000	4994.75	0	5.25	32	5000	4994.50	0	5.50
1.5	24	5115.25	4999.25	10.25	105.75	16	5000	5000	0	0
2	24	5183.50	4999.75	12.75	171	16	5000	5000	0	0
3	24	5317.50	5000	17.25	300.25	16	5000	5000	0	0
4	24	5483.25	5000	12.50	470.75	16	5000	5000	0	0

(the mean matched cosine is 0.9993 at $K = 24$). Only $\tau = 0.95$ separates, at $K \geq 32$ (1.0000 against 0.9428 at $K = 48$), where the dictionary degrades and some atoms clear 0.9 but not 0.95. Lowering τ counts looser copies as duplicates, but at every τ duplicates grow sharply from $K \geq 32$. Table 7 shows the decomposition of one cell with many orphans.

D Supplementary Results without Ground Truth

This appendix collects the results behind Section 5: how the pairwise labels relate to the ground-truth ones (D.1), how the two directions of a pair compare (D.2), why PW-MCC is diluted for $d > n$ (D.3), and the capture–recapture estimate (D.4).

D.1 Agreement with the Ground-Truth Partition

Table 9 crosses the two labellings over all 320 pairs of the main configuration, both directions summed, and Figure 8 splits the same counts by region. For $d < n$, 43% of the matched atoms are labelled stray, which is the chance overlap of Section 5.2: every atom is correct, but the other run represents a different subset of the features. For $d > n$, more than 99.9% of orphans are labelled stray under both conditions, while duplicates split between redundant (67% with AuxK off, 52% with AuxK on) and shared (29% and 47%), because copies placed on the same feature by both runs pair with each other. Lowering δ from 0.9 to 0.6 changes the agreement in every region by less than 0.007 (Table 10).

D.2 The Two Directions of a Pair

d_{shared} is identical in both directions in all 1,280 labelling pairs across the four values of δ, since it is the number of edges of one bipartite matching; the two directions can differ only in the redundant

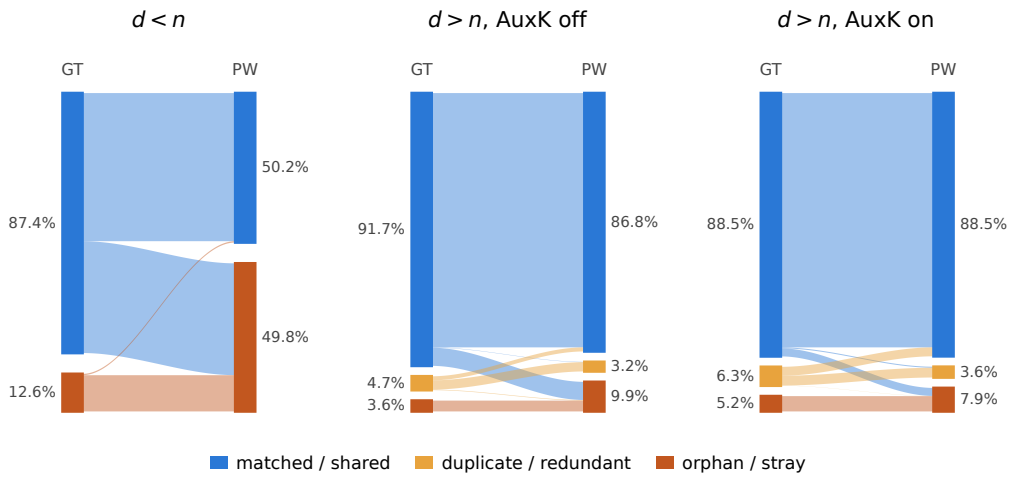


Figure 8: Ground-truth label (left) to pairwise label (right) of the live atoms at $n = 5000$, $\delta = \tau = 0.9$, both directions and four seeds summed, for $d < n$, $d > n$ with AuxK off, and $d > n$ with AuxK on. Shares are of the region.

and stray counts and in d_{live}. The total-variation distance between their class fractions averages 0.0023 and is at most 0.0313, reached at $d = n$; for $d < n$ it is exactly zero in 52 of the 60 cells.

D.3 Dilution of PW-MCC

For $d > n$ many atoms are dead, and because PW-MCC averages over all d atoms, the pairs the matching is forced to form with dead atoms contribute cosines near zero and dilute it (Section 5.2). With AuxK on at $d/n = 4$ and $K^\star = 16$, PW-MCC is 0.44, against 0.9994 when the matching is restricted to live atoms. The live-restricted version peaks at $K^\star$ at all four overcomplete widths with AuxK on.

D.4 Capture–Recapture Estimate

Why orphans inflate the estimate. Where there are no duplicates, $d^{(i)}_{\text{live}} = d^{(i)}_{\text{match}} + d^{(i)}_{\text{orp}}$, only matched atoms can pair across runs, and $d_{\text{shared}} = |F^{(1)} \cap F^{(2)}|$. Substituting these two facts and then the mean in equation 6,

$$\begin{aligned} \hat{n} &\approx \frac{d^{(1)}_{\text{live}}\, d^{(2)}_{\text{live}}}{d_{\text{shared}}} = \frac{\left(d^{(1)}_{\text{match}} + d^{(1)}_{\text{orp}}\right)\left(d^{(2)}_{\text{match}} + d^{(2)}_{\text{orp}}\right)}{\left|F^{(1)} \cap F^{(2)}\right|} \\ &\approx n \cdot \frac{\left(d^{(1)}_{\text{match}} + d^{(1)}_{\text{orp}}\right)\left(d^{(2)}_{\text{match}} + d^{(2)}_{\text{orp}}\right)}{d^{(1)}_{\text{match}}\, d^{(2)}_{\text{match}}} \;\geq\; n, \end{aligned} \tag{10}$$

with equality exactly when there are no orphans. Duplicates do not reverse this in practice: they recur across runs and add to d_{shared} as well as to d_{live}, and over the 377 contaminated pairs the bias correlates with the orphan count (0.545) but not with the duplicate count (0.006), with the largest underestimate at 3.1%.

Sampling and census. At $K = 4$ the at most 18 orphans are 0.6% of d_{live}. When fewer than n atoms are live—always for $d < n$, and for $d \geq n$ when dead latents leave the live count below n—equation 6 applies: at $d/n = 0.25$ and $K = 4$, $d_{\text{shared}} = 322$ against a predicted 313, and with AuxK off at $d/n = 2$, 3470 atoms are live and $d_{\text{shared}} = 2410$ against a predicted 2408. When nearly n atoms are live, $d^{(1)}_{\text{live}} \approx d^{(2)}_{\text{live}} \approx d_{\text{shared}} \approx n$.



(a) Reconstruction NMSE of h.

(b) Ground-truth dictionary recovery (GT-MCC).

(c) Feature recovery with a known-dictionary OMP oracle.

Figure 9: Reconstruction quality does not generally identify the recovery-optimal SAE ($n = 5000$, AuxK on, four-seed mean). The minimum of (a) and the maximum of (b) disagree for $d/n \leq 1$; the OMP oracle (c) reaches 1.000 once $K \geq s$, and its rows repeat because it uses no learned dictionary.

Estimates by combination. Table 11 gives the estimate behind Figure 3 for all 26 combinations. The minimum over K lies at $K = 4$ or 6 in 20 of them, and the minimum and the estimate at the smallest K lie within $[0.966, 1.000]$ and $[0.969, 1.013]$.

Variance and interval. The Chapman estimator of equation 7 has the variance (Seber, 1982)

$$\widehat{\mathrm{Var}}(\hat{n}) = \frac{\left(d_{\mathrm{live}}^{(1)} + 1\right)\left(d_{\mathrm{live}}^{(2)} + 1\right)\left(d_{\mathrm{live}}^{(1)} - d_{\mathrm{shared}}\right)\left(d_{\mathrm{live}}^{(2)} - d_{\mathrm{shared}}\right)}{\left(d_{\mathrm{shared}} + 1\right)^2\left(d_{\mathrm{shared}} + 2\right)}. \tag{11}$$

and the 95% intervals of Figure 3 are $\hat{n} \pm 1.96\,\widehat{\mathrm{SE}}$ on the raw scale. Table 12 checks them on the pairs free of orphans and duplicates. For a correct variance, the median ratio of the observed relative error to the predicted coefficient of variation should be 0.674, the median of $|z|$ for a standard normal; the pooled ratio is 0.700. Near a census the interval is too narrow. As $d_{\mathrm{live}}^{(i)}/n$ approaches 1 the variance in equation 11 vanishes, while atoms that miss the alignment threshold still leave a residual of a few atoms: at $n = 10000$, $d/n = 2$, $K = 20$, $\hat{n}$ falls 11 short of n with a standard error of 1.4. Accordingly, at $n = 10000$ the ratio rises to 1.395 and the coverage of the 95% interval drops to 0.69.

E Reconstruction and Standard Criteria

This appendix gives the evidence behind the statements in the introduction and in Section 6 that the criteria used in practice were not validated against recovery and do not point at the recovery optimum reliably.

Figure 9 makes the distinction between reconstruction and recovery concrete. Comparing reconstruction quality, ground-truth dictionary recovery, and oracle sparse coding over the same design grid shows that the two landscapes do not generally agree. A configuration that reconstructs h better is not thereby one that recovers the ground-truth dictionary better, and the gap is pronounced in the undercomplete and matched-size regimes. This mismatch is not caused by information being irreversibly lost in compression. Given the ground-truth dictionary W, orthogonal matching pursuit (OMP) recovers the sparse representation essentially exactly as soon as the recovery sparsity reaches the true sparsity s. In other words, h retains enough information for exact recovery, and yet an SAE trained to reconstruct h can still fail to recover the corresponding dictionary.

We return to the question raised in the introduction: do the criteria used in practice point at the recovery optimum? The three are reconstruction error, the dead-latent rate, and the similarity between decoder atoms. For each criterion we select K with the same one-standard-error rule used for GT-MCC, pointed at the minimum in all three cases, and compare the result with $K^\star$ (Figure 10).

Reconstruction. When $d > n$ and dead latents are controlled, the NMSE minimum equals $K^\star$: both are 16 at the four overcomplete widths with AuxK on (Figure 11; four of eight points). Outside

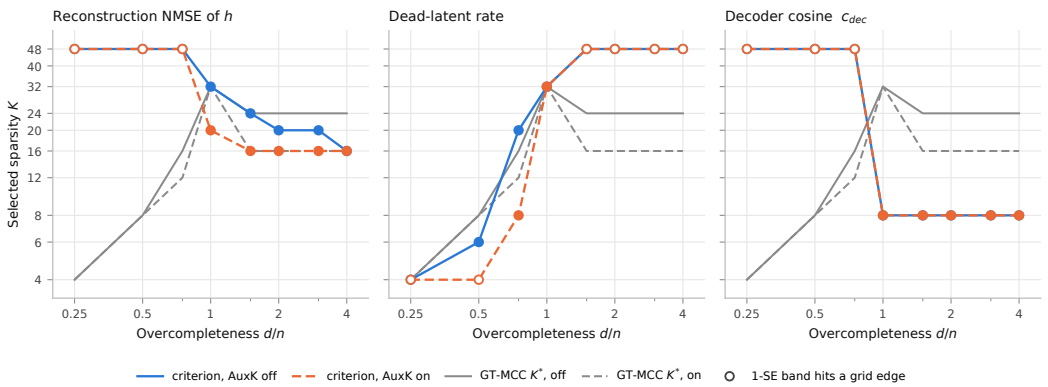


Figure 10: K selected by each practical criterion (colored) against $K^\star$ (gray) ($n = 5000$, $k_{\text{aux}} = 500$, one-standard-error rule, $\log_2 y$ axis). Hollow markers are bounds at the edge of the grid; c_{dec} uses live atoms only.

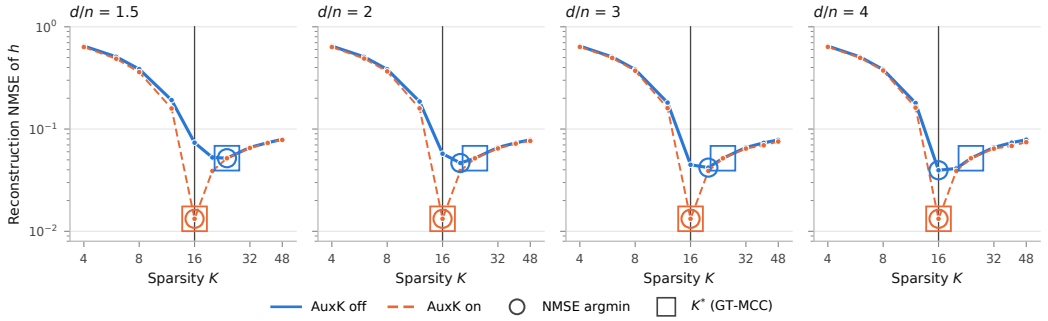


Figure 11: $\text{NMSE}_h(K)$ at $d/n = 1.5, 2, 3, 4$ ($n = 5000$, $k_{\text{aux}} = 500$, log scale). Circles mark the NMSE minimum and squares $K^\star$: with AuxK on they coincide at $K = 16$; with AuxK off the minimum is at 24, 20, 20, and 16 while $K^\star$ is 24.

that condition they diverge. With AuxK off the minimum scatters between 16 and 24 over the same range and matches $K^\star = 24$ only at $d/n = 1.5$, and for $d < n$ both conditions point at the grid edge of 48, far from the $K^\star$ values of 4, 8, and 12. Reconstruction therefore points at the recovery optimum only under specific conditions, and whether those conditions hold cannot be determined from reconstruction alone.

Dead-latent rate. The K at which the dead rate is minimized rises to the grid edge of 48 for $d \geq n$. This follows directly from the fact that a larger K lets more latents pass Top-K and survive, and that direction is unrelated to what recovery requires. Only two of eight points agree, and those are at $d = n$ and at the narrowest width $d/n = 0.25$, where the two meet at the lower grid edge $K = 4$. As Section 4.2 showed, moreover, AuxK changes recovery substantially while reducing the dead rate by only 3–14%, which means two SAEs with the same dead rate can be in entirely different recovery states. This criterion measures how much of the width is used, not which directions the surviving atoms occupy.

Decoder orthogonality. Chanin & Garriga-Alonso (2026) report that the further L_0 is from the correct value, the more latents mix several features and the less orthogonal they become, so that the mean $|\cos|$ between decoder columns is minimized at the true L_0. Measuring the same quantity over the live atoms in our setting, it agrees with $K^\star$ at none of the eight widths (0/8 under both AuxK conditions). For $d < n$, c_{dec} decreases monotonically in K and the grid edge of 48 is selected; for $d \geq n$ a minimum appears at $K = 8$, diverging from $K^\star$.

Table 5: Range of the live atom count over the four runs (two dictionary seeds × two training seeds; evaluation set, AuxK on, $k_{\text{aux}} = 500$). A single number means all four runs are equal. Bold marks the cells with $d > n$ in which all four runs have $d_{\text{live}} = n$ exactly; at $d = n$, $d_{\text{live}} = n$ means only that every atom is live.

n	d/n	$K = 4$	6	8	12	16	20	24	32
2500	1	1816–1831	2081–2094	2252–2266	2472–2478	2494–2499	2472–2480	2499–2500	2500
2500	1.5	2459–2488	2492–2500	**2500**	**2500**	**2500**	2614–2643	2732–2763	2862–2904
2500	2	2385–2432	2492–2500	2499–2500	**2500**	**2500**	2709–2768	2852–2921	3039–3070
5000	1	3281–3296	3828–4181	4218–4245	4500–4501	4601–4630	4977–4999	4985–4995	5000
5000	1.5	4184–4224	4687–4736	4891–4905	4994–5000	**5000**	**5000**	5284–5329	5723–5768
5000	2	4155–4208	4591–4648	4711–4742	4988–4995	**5000**	5000–5001	5802–5826	6037–6370
10000	1	6983–7197	9274–9308	9371–9395	9470–9472	9471–9474	9173–9198	9820–9855	9897–9920
10000	1.5	7014–7035	8006–8042	8473–8599	9140–9286	9567–9732	**10000**	10079–10099	11199–11292
10000	2	6881–6933	7474–7560	7733–7876	8467–8502	9110–9178	9785–9851	10053–10081	12161–12191

Table 6: Range of the live atom count over the four runs across the revival budgets at $n = 10000$, $d/n = 2$ ($d = 20000$), in the format of Table 5. All four runs equal n exactly only at $k_{\text{aux}} = 1500$, $K = 20$ and $k_{\text{aux}} = 2000$, $K = 16$.

k_{aux}	$K = 8$	12	16	20	24	32
500	7733–7876	8467–8502	9110–9178	9785–9851	10053–10081	12161–12191
1000	7510–7555	8234–8277	9137–9212	9999–10000	10028–10038	11711–11853
1500	7475–7515	8422–8488	9511–9745	**10000**	10012–10021	11673–11724
2000	7545–7594	9221–9282	**10000**	9999–10000	10004–10013	10514–10664
4000	8859–8874	9553–9567	9889–9907	9988–9992	10003–10010	10477–10505
8000	8251–8296	9255–9320	9995–9998	9996–9999	10022–10059	10471–10498

Table 7: Decomposition of the live atoms as τ varies at $d/n = 4$, AuxK off, $K = K^\star = 24$ ($n = 5000$, four-seed mean, $d_{\text{live}} = 5483.25$).

τ	d_{match}	d_{dup}	d_{orp}
0.7	5000	34	449.25
0.8	5000	32	451.25
0.9	5000	12.5	470.75
0.95	5000	0.25	483

Table 8: Pairwise departures $d_{\text{stray}} + d_{\text{redundant}}$ for $d > n$ ($n = 5000$, $\delta = 0.9$, mean over both directions and both pairs). Bold marks $K^\star$, which is 16 with AuxK on and 24 with AuxK off.

d/n	$K = 8$	12	16	20	24	32
AuxK on						
1.5	100.0	2.5	**0.0**	0.0	290.5	562.5
2	258.8	8.5	**0.0**	0.2	806.5	897.0
3	380.8	24.0	**0.0**	0.8	1091.0	1755.8
4	417.0	49.8	**0.0**	2.5	998.0	2092.2
AuxK off						
1.5	636.0	428.2	335.2	163.5	**113.2**	402.5
2	608.2	361.8	310.8	179.5	**180.0**	503.8
3	584.2	313.2	346.2	272.0	**313.0**	727.2
4	572.5	296.2	407.0	404.8	**479.8**	888.5

Table 9: Ground-truth label against pairwise label for every live atom, over all 320 pairs of the main configuration with both directions summed ($\tau = \delta = 0.9$; 2,633,810 labels, overall agreement 0.8584).

GT \ pairwise	shared	redundant	stray	recall
matched	2,059,604	3,829	327,189	0.862
duplicate	37,200	53,826	1,983	0.579
orphan	2,702	7	147,470	0.982

Table 10: Agreement between the ground-truth and pairwise labels by region as the pairwise threshold δ varies ($\tau = 0.9$ fixed).

region	$\delta = 0.9$	0.8	0.7	0.6
$d < n$	0.6196	0.6175	0.6161	0.6132
$d = n$	0.9142	0.9142	0.9141	0.9141
$d > n$, AuxK off	0.9223	0.9226	0.9218	0.9214
$d > n$, AuxK on	0.9391	0.9387	0.9377	0.9372
$d > n$, AuxK on, $K = K^\star$	1.0000	1.0000	1.0000	1.0000

Table 11: The Chapman estimate $\hat{n}/n$ for every combination of feature count, AuxK arm, and d/n ($\delta = 0.9$, mean over the two pairs of each cell, K over the full ten-point grid): the minimum over K, the K at which it occurs, and the estimate at the smallest $K = 4$.

n	AuxK	d/n	$\min_K$	K	K=4	n	AuxK	d/n	$\min_K$	K	K=4
2500	on	0.5	0.9750	6	0.9873	5000	on	0.25	0.9659	6	0.9689
2500	on	0.75	0.9939	4	0.9939	5000	on	0.5	0.9868	6	1.0018
2500	on	1	0.9999	12	1.0044	5000	on	0.75	0.9933	6	1.0029
2500	on	1.5	0.9997	4	0.9997	5000	on	1	1.0001	16	1.0098
2500	on	2	0.9991	4	0.9991	5000	on	1.5	0.9970	4	0.9970
5000	off	0.25	0.9659	6	0.9689	5000	on	2	0.9995	4	0.9995
5000	off	0.5	0.9868	6	1.0018	5000	on	3	0.9978	6	0.9984
5000	off	0.75	0.9898	6	0.9975	5000	on	4	0.9993	6	1.0019
5000	off	1	0.9926	4	0.9926	10000	on	0.5	0.9890	8	0.9947
5000	off	1.5	0.9955	4	0.9955	10000	on	0.75	0.9939	12	1.0126
5000	off	2	0.9992	4	0.9992	10000	on	1	0.9970	20	1.0054
5000	off	3	0.9883	4	0.9883	10000	on	1.5	0.9985	12	1.0056
5000	off	4	0.9988	6	1.0000	10000	on	2	0.9943	6	0.9982

Table 12: Calibration of the Chapman interval on pairs free of orphans and duplicates ($\delta = 0.9$). Pairs with $d^{(1)}_{\text{live}} = d^{(2)}_{\text{live}} = d_{\text{shared}}$ have zero estimated variance; coverage is given over the pairs with a positive standard error and over all clean pairs. The ratio is the median of the observed relative error over the predicted coefficient of variation.

n	clean pairs	SE > 0	coverage (SE > 0)	coverage (all)	ratio
2500	42	23	0.870	0.929	0.554
5000	68	55	0.927	0.941	0.672
10000	31	29	0.690	0.710	1.395
all	141	107	0.850	—	0.700