CONCEPT-CONDITIONED SCORES CAN RANK SPARSE DICTIONARIES WITHOUT USING THE CONCEPTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse dictionaries are increasingly compared by scores computed from concept labels: how compactly a labeled concept is represented, how well it matches a feature, how much information a feature carries about it. Concept labels give such a score an apparent semantic grounding. Whether the ranking of dictionaries it induces depends on the labels is a separate question, and existing sanity checks, which randomize the dictionary, do not ask it. On dictionaries we train for chest radiographs and on public sparse autoencoders for language models, under three label systems, replacing each concept's support with a size-matched random set of documents leaves the ranking under most scores unchanged. The invariance reaches model selection: on a controlled sweep, compactness scores rank collapsed dictionaries first and give their best value to a dictionary trained on permuted concept targets, whereas probes favor real targets on every one of 20 paired seeds. We therefore propose TWIN (two-way identity null), a protocol for auditing a concept-conditioned score before it is used to compare dictionaries: it destroys concept identity once at evaluation and once at training, controls occupancy, and anchors every score to probing. Applied to fourteen scores, TWIN separates failures of different kinds: rankings that do not move when the concepts are randomized, preferences for real targets that vanish once the feature budget is matched, and a matching score whose ranking follows the live count. The mutual-information gap is the most label-sensitive score, and its sensitivity to discretization keeps it a diagnostic. For compactness the failure can be explained: the score depends only on the sorted shape of a concept's activation profile, that shape stays close to the label-free marginal one on all our pools, and for the primary score we prove that this closeness confines its value to an interval around the value it takes with no labels at all.

1 INTRODUCTION

Sparse dictionaries (sparse autoencoders, transcoders, and their supervised variants) decompose a model's representations into features that a person can name (Bricken et al., 2023; Cunningham et al., 2024; Templeton et al., 2024; Gao et al., 2025). Reconstruction and sparsity are proxies for their quality, and evaluation has moved toward measures grounded in labeled concepts: probing for a concept with a few latents, feature absorption and splitting, matching latents to human annotations, and supervised dictionaries trained to bind a concept to a feature (Karvonen et al., 2025; Chanin et al., 2025; Leask et al., 2025; Klotz et al., 2026; Cassano et al., 2026). Many of these measures are scores computed from a concept's labeled support: how few features carry the concept (Yang et al., 2025), how well its best or matched feature overlaps the support (Klotz et al., 2026), how informative a single feature is about it (Eastwood & Williams, 2018; Chen et al., 2018). We call them *concept-conditioned scores*. They are reported when dictionaries are compared.

Such scores are most attractive where concepts are many and probes are expensive. Medical vision–language modeling is such a setting: report-derived ontologies supply more than a thousand concepts (Johnson et al., 2019; Bodenreider, 2004), which motivates scalar summaries for comparing dictionaries. Because the labels enter the score, the ranking it induces is read as a statement about concepts, without a test of whether the ranking responds to the labels. Existing sanity checks for dictionary evaluation randomize the model side, asking a metric to separate trained dictionaries from untrained or randomly initialized ones (Klotz et al., 2026; Korznikov et al., 2026; Heap et al., 2026),

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.

or audit its reliability under reseeding (Chanin, 2026); none randomizes the concepts. A score can be stable across seeds, and the latents it matches to a concept can respond to perturbations of that concept, while the score still ranks dictionaries by a property that does not depend on the labels.

We test this by destroying concept identity and asking whether the ranking moves. On 39 dictionaries we train for chest radiographs and on 26 public Gemma Scope autoencoders (Lieberum et al., 2024), under three unrelated label systems, replacing each concept's support with a random set of documents of the same size leaves the ranking in place for twelve of fourteen scores on the medical dictionaries and for every compactness score under all three label systems. The problem carries over to model selection: on a controlled sweep, the dictionaries that compactness ranks highest are the collapsed ones, and the best compactness score belongs to a dictionary trained on permuted concept targets, whereas probes favor real targets on all 20 paired seeds.

These observations call for a check that complements reliability and correspondence: before a concept-conditioned score is used to compare dictionaries, its ranking should be shown to depend on the concepts. We formulate the check as a protocol, TWIN (two-way identity null; Figure 1(b)): an evaluation-time null that randomizes the supports at fixed dictionary, a training-time null that retrains on permuted targets, an occupancy control that equalizes the number of live features, and probing anchors as positive control. On fourteen scores from the compactness, disentanglement, and matching literatures, TWIN exposes failures of different kinds, and for compactness the failure can be explained (Figure 1(c)). Our contributions are as follows.

- **Rankings that do not depend on the concepts.** Across 65 sparse dictionaries and three label systems, destroying concept identity leaves the ranking under most concept-conditioned scores unchanged. Selecting by compactness then picks collapsed dictionaries and cannot tell a dictionary trained on real concept targets from one trained on permuted targets, a difference on which occupancy-controlled probes favor real targets on 20 of 20 paired seeds.
- **TWIN, a validation protocol for concept-conditioned scores.** It combines an evaluation-time null, a training-time null, occupancy control, and probing anchors. Under TWIN, probing anchors pass; compactness and importance scores keep their ranking under random labels; max-over-features scores lose their preference for real targets once the feature budget is matched, and the ranking of a matching score follows the live count; the mutual-information gap is the most label-sensitive score, and its sensitivity to discretization limits it to a diagnostic.
- **An explanation of the clearest failure.** Compactness scores depend only on the sorted shape of a concept's profile, which stays close to the marginal one on our pools; for the primary score we prove that this closeness confines the score to an interval around its label-free null, and we use the gap δ_v to delimit the regime in which the result holds. Subtracting a random-support baseline recovers sensitivity to the training signal without yielding a selection score.

2 Related Work

From proxies to concept-grounded evaluation. Evaluation of sparse dictionaries began with reconstruction and sparsity (Bricken et al., 2023; Cunningham et al., 2024; Gao et al., 2025; Rajamanoharan et al., 2024), moved to the features themselves through automated explanation (Bills et al., 2023; Paulo et al., 2025), and increasingly adds task- and concept-grounded measures. SAEBench measures concept detection, how well individual latents map to concepts, by sparse probing, and feature disentanglement, whether independent concepts are separated, by targeted interventions (Karvonen et al., 2025); absorption and splitting studies measure how a concept's feature fractures across widths (Chanin et al., 2025; Leask et al., 2025); Klotz et al. (2026) match annotated concepts to latents and validate the match under targeted perturbations (see also Pach et al., 2025); and supervised dictionaries are trained to centralize a concept on a single latent (Cassano et al., 2026; Yang et al., 2025). The requirement these works share, few features per concept, also has static scalarizations, and they are already reported as evidence: Yang et al. (2025) support a supervised dictionary by the effective number of features and the top-1 concentration of a concept's activation profile, and Lim et al. (2025) characterize latents by the entropy of their activation mass over labels, the converse conditioning. Effective support, entropy, Gini impurity, and top-t mass belong to this

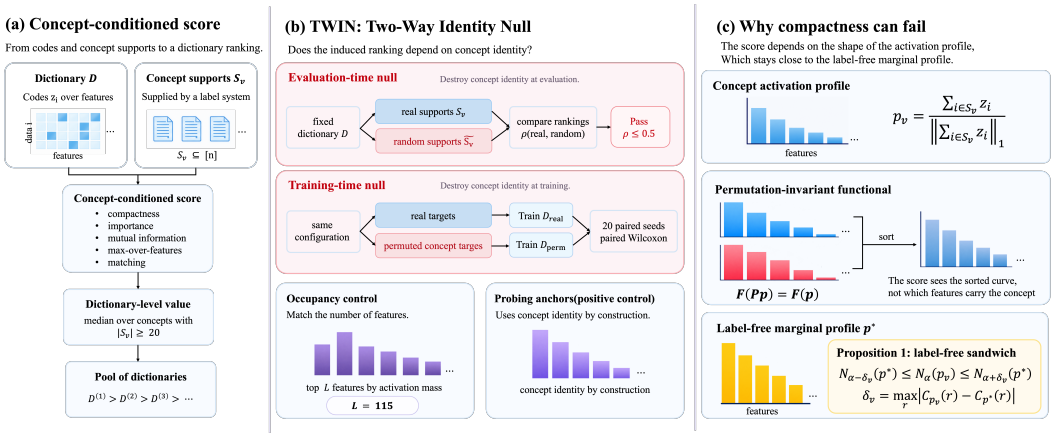

Figure 1: **Overview of the TWIN (two-way identity null) protocol.** (a) A concept-conditioned score is computed from a dictionary's codes and the concepts' supports, and ranks dictionaries. (b) TWIN tests whether that ranking depends on the concepts. (c) The compactness case, in which the score sees only the sorted shape of a concept's activation profile and Proposition 1 confines it to an interval around its value on the marginal profile p^*.

family, and the disentanglement metrics DCI (Eastwood & Williams, 2018), the mutual-information gap (Chen et al., 2018), SAP (Kumar et al., 2018), and modularity (Ridgeway & Mozer, 2018) are inherited into a regime of thousands of concepts and mostly inactive features, where they already disagree with one another (Locatello et al., 2019; Carbonneau et al., 2022). To our knowledge, whether the dictionary rankings induced by these scores depend on the concepts has not been tested.

Sanity checks, reliability audits, and falsification. Our tests follow the sanity checks of Adebayo et al. (2018), which randomize what a method claims to depend on, the model in one test and the labels in another, and ask whether the output changes. Audits of dictionary evaluation have adopted the model side. Automated interpretability scores do not distinguish dictionaries of trained and randomly initialized transformers (Heap et al., 2026); dictionaries with frozen random components match trained ones on interpretability, sparse probing, and causal editing (Korznikov et al., 2026); Klotz et al. (2026) require a concept-matching metric to separate trained from untrained dictionaries and subtract an untrained baseline when dictionary sizes differ; and Chanin (2026) asks whether benchmark metrics are stable under reseeding and faithful to a synthetic ground truth. Feature-level claims have been falsified in the same spirit (Leask et al., 2025; Wu et al., 2025; Ma et al., 2026). The label side has a precedent in probing, where control tasks assign random labels to test whether a probe's accuracy reflects the representation or the probe (Hewitt & Liang, 2019), and in the comparison of clusterings, where scores are corrected for agreement expected by chance (Vinh et al., 2010). TWIN (§3.2) carries the label side to dictionary evaluation, with two differences. The unit of the test is the ranking of dictionaries, which can stay fixed while individual scores respond to the labels, and the number of live features is a confound that survives label randomization and has to be controlled.

3 METHODS

We fix the scores under test (§3.1), define TWIN (§3.2), its recoverability anchors and decision rules (§3.3), and the instrument for the training-time null (§3.4), and then analyze the compactness family, for which the outcome of TWIN can be explained (§3.5). Section 4 mirrors this structure: TWIN on the battery (§4.2), the consequences for model selection (§4.3), the bound (§4.4), and the label-sensitive scores (§4.5).

3.1 CONCEPT-CONDITIONED SCORES

A dictionary D maps an input to a code of which the 32 entries of largest magnitude are kept and the rest set to zero; negative entries among the kept ones are then set to zero as well, which leaves a nonnegative code $\boldsymbol{z} \in \mathbb{R}^m_{\geq 0}$ with at most 32 nonzero entries (a median of 16 on the medical pool; Appendix C). For a corpus of n documents with codes $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$, a concept v is a support set $S_v \subseteq [n]$ supplied by a label system. A *concept-conditioned score* is a function of the codes and the supports; its dictionary-level value is the median over concepts with support at least 20, and a pool of dictionaries is ranked by that value. Several scores are functionals of a concept's *profile*, its normalized activation mass over features,

$$p_v \;=\; \frac{\sum_{i \in S_v} \boldsymbol{z}_i}{\left\| \sum_{i \in S_v} \boldsymbol{z}_i \right\|_1} \;\in\; \Delta^{m-1}, \tag{1}$$

and the *marginal profile* p^* is the same quantity over all n documents; the *label-free null* of a score is the same functional applied to p^*. We evaluate fourteen scores, grouped into five families by what they are functionals of (Table 1): six compactness scores, which summarize how concentrated p_v is (§3.5); an importance score, DCI-completeness; the mutual-information scores MIG and modularity, with which we group SAP; the max-over-features scores max-activation purity, best-feature AUROC, and max IoU; and one matching score, matched IoU.

3.2 TWIN: THE TWO-WAY IDENTITY NULL

TWIN (Figure 1(b)) audits whether the ranking of dictionaries under a concept-conditioned score can be read as a statement about concepts. It destroys concept identity in two ways, at evaluation and at training, and controls a confound that survives both. **Evaluation-time null.** With the dictionary fixed, every support S_v is replaced by a uniformly random subset $\tilde{S}_v$ of documents of size $|S_v|$ (*random labels*, for short). We report the fraction of dictionaries on which real labels beat random ones, the median relative effect, and the Spearman correlation $\rho(\text{real}, \text{random})$ between the two rankings; a score whose ranking is unchanged ranks without the labels, and we use $\rho(\text{real}, \text{random}) \leq 0.5$ as the pre-specified operational threshold for a pass, which indicates label dependence and does not by itself establish validity (§3.3). **Training-time null.** The same configuration is retrained on real versus permuted concept targets over 20 paired seeds (the two *conditions*), and we report the number of paired seeds on which real-target training scores better (*real-target wins*) with a paired Wilcoxon test. This null requires a dictionary trained with concept supervision, which §3.4 supplies; for unsupervised dictionaries only the evaluation-time null is available. **Occupancy control.** The training-time comparison is repeated after restricting each dictionary to its top L features by activation mass ($L = 115$, the smallest live count among the conditions of §3.4), a matched feature budget; a score that favors real targets before the control and not after is sensitive to that budget.

3.3 RECOVERABILITY ANCHORS AND DECISION RULES

The nulls establish label dependence and leave correctness open, so every score is also anchored to recoverability. For each concept a ridge probe on the kept codes yields an AUROC (Appendix C); dictionary-level anchors are the median probe AUROC, the *decodable fraction* (concepts with AUROC > 0.60), and k-sparse probing after SAEBench (Karvonen et al., 2025), a logistic probe refit on the $k \in \{1, 2, 5\}$ latents with the largest mean-activation difference between documents in and out of S_v (Appendix C). We report the partial correlation of a score with an anchor given the score's label-free null, which measures what the score adds beyond what the marginal profile already predicts; for scores that are not functionals of the profile, the partial correlation conditions on the label-free null of mass coverage, $N_{0.8}(p^*)$. The anchors use concept identity by construction, so they also serve as TWIN's positive control. Because $\rho(\text{real}, \text{random}) \approx 0$ can arise from noise alone, a score that passes both nulls is not yet recommended. For the mutual-information gap (MIG), a functional of the mutual information $I(z_j; c_v)$ between feature firing and concept membership, we pre-registered four further conditions: alignment with the anchors beyond the null (M1), correct ordering of the training conditions (M2), seed reproducibility (M3), and stability under a change of the discretization from which the mutual information is computed (M4), with the rule that MIG would be recommended as a selection scalar only if M1, M2, and M4 all held.

Table 1: **The battery of fourteen concept-conditioned scores.** Direction gives the sign read as better. Notation: $p_{v,(l)}$ is the l-th largest entry of p_v; c_v is the indicator of S_v; z_j is the code of feature j and fires_j the set of documents on which it is positive after truncation; $W_{\cdot v}$ is the weight vector of the ridge probe for v. Binary 2×2: mutual information from the contingency table of firing and membership; firing-rate AUROC: AUROC of a feature's binary firing indicator. SAP is grouped with the mutual-information scores although its per-feature quantity is an AUROC.

Family	Score	Definition (per concept)	Direction	Estimator
Compactness (symmetric functional of p_v; §3.5)	mass coverage $N_{0.8}$	fewest features covering 80% of p_v	low	–
	coverage $N_{0.7}$	fewest features covering 70% of p_v	low	–
	exp-entropy	$\exp H(p_v)$, H the Shannon entropy	low	–
	inverse Simpson	$1/\sum_j p_{v,j}^2$	low	–
	Gini impurity	$1 - \sum_j p_{v,j}^2$	low	–
	top-10 mass	$\sum_{l \le 10} p_{v,(l)}$	high	–
Importance	DCI-completeness	$1 - H(\lvert W_{\cdot v}\rvert / \lVert W_{\cdot v}\rVert_1) / \log m$	high	ridge probe, penalty 1
Mutual information	MIG	gap between the two largest $I(z_j; c_v)$, normalized by $H(c_v)$	high	binary 2×2
	SAP	gap between the two largest per-feature predictive scores	high	firing-rate AUROC
	modularity	deviation from a one-concept-per-feature ideal	high	binary 2×2
Max-over-features	max-activation purity	$\max_j$ lift of v among the top-activating documents of j	high	–
	best-feature AUROC	$\max_j \text{AUROC}(\mathbf{1}[z_j > 0]; c_v)$	high	firing-rate AUROC
	max IoU	$\max_j \text{IoU}(\text{fires}_j, S_v)$	high	–
Matching	matched IoU	mean IoU under a one-to-one Hungarian assignment (a variant of Klotz et al., 2026)	high	–

3.4 Report-Aligned Dictionary for the Training-Time Null

The training-time null needs a dictionary whose supervision can be replaced by noise without changing anything else: a report-aligned dictionary, an encoder that maps 3,072-dimensional image features to a sparse code of $m = 2{,}048$ features (two hidden layers of width 2,048, post-hoc top-32), trained with a reconstruction loss and an alignment loss, weighted by λ_{align}, that pulls the code of each study (one radiology exam with its report) toward a target derived from the concepts in its report. Four conditions, each trained with 20 seeds (the *seed grid*, 80 runs), differ in three of 44 configuration keys (Table A9, Appendix K): `poincare` uses the Poincaré distance to the targets; `shuffled` uses the same loss with targets permuted across studies; `euclid_matched` uses the Euclidean distance with λ_{align} chosen so that the alignment term contributes the same amount to the loss; `noalign` sets $\lambda_{\text{align}} = 0$. The training-time null compares `poincare` with `shuffled`, and the other two conditions are controls.

3.5 Compactness Scores and the Label-Free Bound

A compactness score is a functional $F : \Delta^{m-1} \to \mathbb{R}$ of the profile p_v in equation 1 (Table 1); the primary one is mass coverage, $N_\alpha(p) = \min\{j : \sum_{l \le j} p_{(l)} \ge \alpha\}$ with $p_{(1)} \ge p_{(2)} \ge \cdots$ and $\alpha = 0.8$. Every such score satisfies $F(Pp) = F(p)$ for any permutation matrix P. Two concepts whose profiles have the same sorted shape receive the same score whether they occupy the same features or disjoint ones (Figure 1(c)), so $F(p_v)$ can depend only on the sorted curve of p_v. Invariance to feature permutations is shared by most scores, MIG included, and does not by itself make a score insensitive to the labels; the insensitivity arises when the sorted curve of a concept's profile stays close to that of the marginal profile, and for mass coverage this closeness bounds the score around a quantity that never looks at the labels, the score of the marginal profile. Write $C_p(r) = \sum_{l \le r} p_{(l)}$ for the sorted cumulative mass and let $\delta_v = \max_r |C_{p_v}(r) - C_{p^*}(r)|$ be the largest gap between the sorted cumulative curve of concept v and that of the marginal.

Proposition 1 (Label-free sandwich)**.** *For every concept v and every $\alpha \in (\delta_v,\, 1 - \delta_v)$,*

$$N_{\alpha - \delta_v}(p^*) \;\le\; N_\alpha(p_v) \;\le\; N_{\alpha + \delta_v}(p^*). \tag{2}$$

Moreover $\delta_v \le \frac{1}{2}\varepsilon_v$ with $\varepsilon_v := \lVert p_v - p^* \rVert_1$. The statement makes no assumption about how the labels were produced or how ε_v scales with $|S_v|$.

The labels enter the bounds only through the gap δ_v: every support whose sorted cumulative curve lies within δ_v of the marginal one is confined to the same interval around the label-free null $N_\alpha(p^*)$, whichever documents it contains. The gap therefore delimits the regime of the result: the smaller δ_v, the closer the score is held to its label-free null, whereas in a dictionary that binds each concept to a dedicated supervised feature (Yang et al., 2025) δ_v can approach one, where the bound is vacuous. §4.4 reports δ_v and ε_v, the distance after and before sorting. The other compactness scores are

likewise functions of the sorted curve alone and are evaluated empirically (Table A6). A synthetic check with planted concept features shows the same split (Appendix B).

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Dictionaries. The medical pool is 39 report-aligned dictionary variants trained on CLIP ViT-B/16 features (Radford et al., 2021) of MIMIC-CXR images (Johnson et al., 2019), spanning 12 configurations of encoder, sparsification, alignment granularity, target geometry, and alignment weight, five of them collapsed with 9–16 live features, plus the 80 runs of the seed grid (§3.4); every score, anchor, and test uses the held-out split. The public pool is 26 Gemma Scope SAEs on the residual stream of Gemma-2-2B (Lieberum et al., 2024; Gemma Team et al., 2024), layers 5, 12, and 20 at widths 16k–65k, applied to token activations and mean-pooled into document codes. Every code is truncated as in §3.1, and a feature is *live* if it is positive on at least one code (live@k32). **Label systems.** UMLS concepts extracted from MIMIC-CXR reports with QuickUMLS (Soldaini & Goharian, 2016; Bodenreider, 2004) (767 with support ≥ 20), EuroVoc descriptors on EURLEX (Chalkidis et al., 2019) (100), and GoEmotions labels (Demszky et al., 2020) (28; a stress case, Appendix H). **Statistics.** Confidence intervals are cluster bootstraps over the 12 configurations of the medical pool, which are of unequal size, and over the five layer–width groups of the public pool (Appendix D); p-values are nominal and unadjusted for the number of scores. Every confirmatory test was pre-registered, with decision rules recorded before results were computed, and post hoc analyses are marked as such; deviations from the pre-registered predictions are in Appendix O, and data, splits, and compute are in Appendix C.

4.2 RESULTS OF THE TWIN PROTOCOL

This section applies TWIN (§3.2) to the fourteen scores of Table 1. Figure 2 summarizes the verdicts; Appendix G lists every value and the full training-time null (Table A7). A second evaluation-time null, which permutes the rows of the label matrix and so preserves label co-occurrence, gives the same verdict in every cell (Table A8).

Positive control. Under the evaluation-time null on the 39 medical dictionaries the median probe AUROC falls to 0.497–0.505 on every dictionary, the median decodable fraction from 0.37 to 0.07, and the median sparse-probe AUROC ($k = 5$) from 0.571 to 0.499, with real labels winning on 38 of 39 for all three anchors; on the 26 public SAEs the probe AUROC wins 26 of 26 under both label systems (Appendix M). Under the training-time null the anchors favor real targets on 20 of 20 paired seeds (§4.3). The sparse probe is near chance only on the five collapsed dictionaries (0.50–0.54) and agrees with the decodable fraction at $\rho = 0.88$.

Compactness and importance-based scores. On the medical dictionaries twelve of fourteen scores keep their ranking of the dictionaries under random labels ($\rho(\text{real}, \text{random}) \geq 0.97$). The six compactness scores do so under all three label systems (0.936–1.000; Table A6), and none favors real targets under the training-time null (5–7 wins of 20; exp-entropy nominally favors permuted targets, $p = 0.044$). DCI-completeness keeps its ranking under all three label systems (0.985, 0.905, 0.971) although it beats random labels on 89.7% of medical dictionaries: it responds to the labels, but the response does not reorder dictionaries. SAP prefers random labels (10/39 wins, -7%; 0/26 on Gemma Scope, -80%) and permuted targets under the training-time null (2/20, $p = 1.7 \times 10^{-4}$). Both summarize the sorted shape of a label-derived importance profile, as the compactness scores do for the activation profile (§3.5).

Max-over-features and matching scores. The max-over-features scores favor real targets under the raw training-time null (17–18/20) and no longer do so under the matched feature budget (5–12/20), where max-activation purity and best-feature AUROC nominally favor permuted targets ($p = 0.030$ and 0.022). Matched IoU, a one-to-one variant of the matching of Klotz et al. (2026) without their untrained-baseline subtraction, passes the training-time null with and without control (19/20, 20/20) yet largely keeps its ranking under random labels (0.98, 0.87, 0.77) and correlates with live@k32 at $\rho = 0.95$: it sees the training signal, but its ranking follows the live count, in part by construction: the mean runs over all 767 concepts, and a concept assigned no live feature contributes zero (Appendix C).

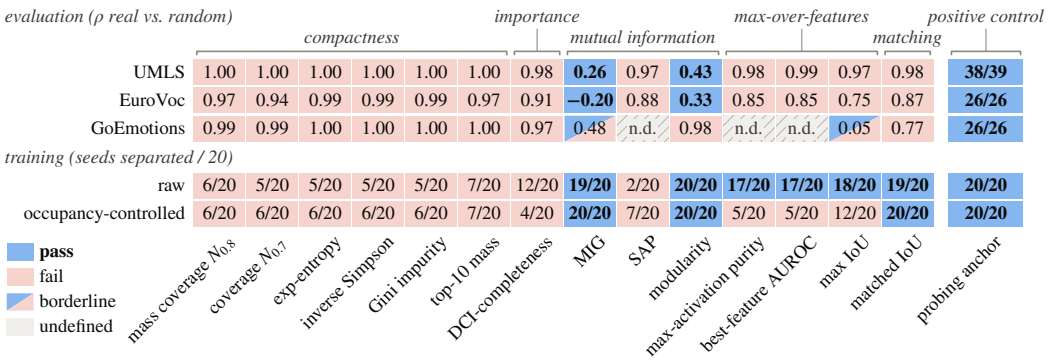

Figure 2: **Verdicts of the fourteen scores and the probing anchor under TWIN.** Columns are the scores of Table 1 and the probing anchor (positive control); rows are the evaluation-time null under UMLS, EuroVoc, and GoEmotions (ρ between the rankings under real and random labels) and the training-time null before and after occupancy control (real-target wins out of 20 paired seeds). Cells mark pass or fail under the rules of §3.2; hatched cells are undefined, and split cells are borderline (draw-dependent; Appendix G). The anchor (decodable fraction on UMLS, median probe AUROC on the public corpora) reports real-label wins in its evaluation-time cells. Values are in Appendix G.

Mutual-information scores. MIG's ranking changes under random labels on all ten draws of the random supports under UMLS (0.09–0.34) and EuroVoc (−0.42–0.07) and on eight of ten under GoEmotions (0.17–0.61; borderline in Figure 2), and it passes the training-time null with and without control; modularity does so under two systems and fails on GoEmotions (0.98); §4.5 examines whether MIG can be recommended.

4.3 Consequences for Model Selection

The instrument of §3.4 measures what selecting by such a ranking costs.

Selection by compactness. Ranked by mass coverage, the five most compact dictionaries are the five collapsed ones (9–16 live features); ranked by the decodable fraction, their median position is 37 of 39. Selecting on the validation split and evaluating on held-out data, five of the six compactness scores pick the same collapsed dictionary as the label-free null (regret 0.30, against 0.32 for a random choice), DCI-completeness picks the worst dictionary in the pool, MIG the fifth best (0.12), and the validation anchor the best (Appendix L). Among healthy dictionaries alone (live > 100, $n = 32$) mass coverage does track the decodable fraction ($\rho = -0.89$), but so does its null: the partial correlation given the null is -0.18, whose magnitude is below the 0.30 pre-registered for condition M1 (Appendix J), so compactness adds little beyond its label-free null there.

Detection of the training signal. In Table 2, the decodable fraction favors real-target training on 20 of 20 paired seeds ($p = 2 \times 10^{-6}$), whereas mass coverage does so on 6 of 20 ($p = 0.232$) and gives its best mean score to `shuffled`. DCI-completeness favors real targets on 12/20 seeds before occupancy control and permuted targets after it (4/20, $p = 0.011$), the clearest instance of the two nulls measuring different things.

4.4 Analysis of Compactness Scores

Compactness is the clearest failure of §4.2, and §3.5 predicts its form.

Label-free null. Table 3 reports the correlation between mass coverage and its null (plotted in Figure A1, Appendix G). On the 39 medical dictionaries, mass coverage correlates with its null at $\rho = 0.998$ and the other compactness scores at 0.986–0.998. The correlation survives when the pool is narrowed to the 80 seed-grid runs (0.990 [0.978, 0.994]) and to each of the four conditions (0.949–0.990), and it is not produced by variation in live counts, since it equals 1.00 within the linear-encoder configurations, whose live counts barely vary ($\rho(\text{score}, \text{live}) = 0.06$). Rebuilding the labels to handle negation, which removed 43% of the concepts, leaves the correlation (0.9983) and the ranking ($\rho = 0.9997$) in place (Appendix E).

Table 2: **Compactness and the probing anchors on the four training conditions of §3.4** (20 seeds; mean ± s.d. for mass coverage, means otherwise; best value per column in bold). The condition trained on permuted targets receives the best compactness score, while the three anchors place both real-target conditions above it. The last row counts the paired seeds on which `poincare` scores better than `shuffled`; paired tests are in Appendix K.

	Compactness		Probing anchors (§3.3)			Diagnostic
Training condition	mass coverage ↓	live@k32	decodable ↑	sparse probe (k=5) ↑	probe AUROC ↑	MIG ↑
`poincare` (real targets)	22.75 ± 4.02	363.6	**0.671**	0.626	**0.644**	0.00307
`euclid_matched` (real targets)	22.10 ± 3.82	226.8	0.663	**0.635**	0.643	**0.00342**
`shuffled` (permuted targets)	**21.60** ± 3.73	255.9	0.502	0.585	0.600	0.00235
`noalign` (no targets)	58.40 ± 15.70	438.5	0.348	0.573	0.564	0.00197
real-target wins over permuted targets / 20	6	–	20	20	20	19

Table 3: **Label-free and evaluation-time nulls across label systems** for mass coverage. $\rho(\cdot, p^*)$: correlation with the label-free null; wins: dictionaries on which real labels beat random ones; effect: median over dictionaries of $(\text{random} - \text{real})/\text{random}$, positive when real labels give the more compact score; $\rho(\text{real}, \text{random})$: correlation of the two rankings. CIs are cluster bootstraps ($B = 4{,}000$).

Label system	Dictionaries	n	$\rho(\cdot, p^*)$ [95% CI]	$\rho(\cdot, \text{live@k32})$	wins / effect	$\rho(\text{real}, \text{random})$
UMLS (MIMIC-CXR)	39 trained	39	0.998 [0.991, 1.000]	0.925	21/39 / +4.5%	0.999
EuroVoc (EURLEX)	Gemma Scope	26	0.975 [0.935, 0.985]	0.652	0/26 / 0.000%	0.975
GoEmotions	Gemma Scope	26	0.987 [0.940, 1.000]	0.703	1/26 / 0.000%	0.987

Label-free bound. The interval of equation 2 holds on all 91 dictionary–label-system pairs, as the proposition requires; the informative quantities are its width and its midpoint, which predicts the measured score at $\rho = 0.995$ on the medical pool and at 0.977 and 0.972 on the public one (Table A1). The labels do move the profile, since ε_v is 41.5% larger under real supports than under random ones, but little survives sorting (median $\delta_v/\varepsilon_v = 0.18$): the top $N_\alpha(p^*)$ features of p^*, which hold at least a share α of the marginal mass by definition, hold 0.797 of a concept's mass under real supports and 0.801 under random ones. The score consequently changes by 4.5% between real and random supports against a $778\times$ range of $N_\alpha(p^*)$ across the medical dictionaries. On the public pool $N_\alpha(p^*)$ ranges over only $1.6\times$ and $1.4\times$, so the ranking result there rests on the measured correlations of Table 3.

Recoverability. Compactness shows no detectable addition to its null for predicting the decodable fraction, which spans 0 to 0.68 across dictionaries: the partial correlation of mass coverage given the null is -0.07 ($[-0.42, 0.13]$), and -0.12 and -0.05 against the two other anchors (Table 4).

4.5 Label-Sensitive Scores and Model Selection

Label dependence does not suffice for a selection score: modularity passes both nulls and selects a collapsed dictionary on the validation split (Appendix L), and the null-corrected compactness D and MIG, which both respond to the labels, do not qualify under the pre-registered rules.

Null-corrected compactness. We pre-registered two corrections (Appendix N): the difference D between a score under real and under size-matched random supports, and the score L computed on the lift profile $q_v \propto p_v/p^*$. L fails both nulls ($\rho = 0.999$ on UMLS; 9–11 real-target wins of 20). D detects the training signal before and after occupancy control (17–19 wins of 20 for the continuous scores) and fails the pre-registered alignment test (partial correlations -0.35 to $+0.29$ against a threshold of $+0.30$). Post hoc, on the healthy dictionaries alone (live > 100) the four continuous forms of D do align ($+0.21$ to $+0.54$, bootstrap intervals excluding zero), which we report as exploratory; we do not recommend D as a selector.

MIG. M1 holds: MIG's partial correlation with the decodable fraction given the null is $+0.64$ ($+0.65$ against the probe AUROC and $+0.81$ against the sparse probe), falling to $+0.22$ on the healthy subset (live > 100), so part of the alignment comes from the contrast between collapsed and healthy dictionaries. M2 and M3 hold (Appendix J). M4 fails: MIG computed from a nine-bin quantile discretization of the activation agrees with MIG computed from binary firing at $\rho = 0.67$ over all 39 dictionaries and at 0.82 once dictionaries with fewer than 100 live features are excluded.

Table 4: **Partial correlation of each score with a probing anchor given the score's label-free null** (39 medical dictionaries; condition M1 of §3.3). The first two columns give the raw rank correlations of the score and of its null with the decodable fraction; the remaining columns give the partial correlation against each anchor.

	Raw ρ with decodable		Partial ρ given the label-free null		
Score	score	null	decodable	sparse probe (k=5)	probe AUROC
mass coverage	-0.429	-0.426	-0.071	-0.12	-0.050
MIG (binary estimator)	$+0.702$	-0.426	$\mathbf{+0.638}$	$\mathbf{+0.81}$	$+0.653$
MIG (quantile estimator)	–	–	$+0.398$	–	$+0.419$
modularity	$+0.584$	-0.426	$+0.442$	–	–

The two use different variables, magnitude and firing, so the disagreement mixes definition and estimation, and it is largest on collapsed and low-occupancy dictionaries; the alignment survives the change (partial $\rho = +0.40$). By the pre-registered rule MIG is not a selection scalar, and we use it as a diagnostic that concept structure is present (Figure A2, Appendix J).

5 Discussion

The result concerns rankings: a score can respond to the concept labels, as DCI-completeness and matched IoU do, while the order in which it places dictionaries is set by a property that the labels do not affect, and model selection depends on that order. In practice, a concept-conditioned score qualifies for comparing dictionaries once its ranking moves when the supports are randomized, separates real from permuted targets, and survives occupancy control; TWIN complements the model-side sanity checks of Korznikov et al. (2026) and Klotz et al. (2026) and the reliability audit of Chanin (2026), which test different properties. Capturing a concept with few features remains a reasonable requirement that sparse probing measures functionally (Karvonen et al., 2025; Chanin, 2026), on our pools as well (§4.2); its static scalarization fails the tests above.

The claims rest on evidence of different breadth: for the compactness scores, label-free tracking and ranking invariance under label randomization hold on all 65 dictionaries and three label systems; on the public pool, where mean pooling leaves mass coverage a range of 1.4–$1.6\times$, the invariance also holds under max and firing-rate pooling, 16 to 128 kept entries, and a co-occurrence-preserving permutation of the labels ($\rho \geq 0.84$; Appendices C, F, and G). By contrast, the model-selection consequences are established on the sweep we trained, because the training-time null requires concept supervision and the pre-registered anchor saturates on the public SAEs (Appendix I). Proposition 1 concerns mass coverage and constrains the score only where δ_v is small (median 0.095 on the medical pool); dictionaries that bind each concept to a supervised feature lie outside this regime and were not tested. The other compactness scores, DCI-completeness, and SAP fail empirically, outside the bound.

Limitations. The 39 medical dictionaries come from 12 configurations of unequal size, which leaves an effective sample size of about 6.5 (Appendix D), and the collapse-selection result requires collapsed dictionaries in the pool. DCI, SAP, and modularity are used outside their design regime, and three firing-rate substitutes are constant or nearly so on the short GoEmotions documents. MIG is computed from binary firing; a magnitude-based discretization changes its ranking, and its alignment with the anchors weakens on healthy dictionaries alone. Public SAEs are evaluated on pooled document codes, since the label systems are document-level (Appendix F), and five pre-registered predictions failed, each narrowing the claim it concerns (Appendix O).

In conclusion, a score computed from concept labels does not by that fact rank dictionaries by their concepts. Across 65 dictionaries and three label systems, randomizing the concepts leaves most rankings in place, and compactness goes on to select collapsed dictionaries and to miss the difference between real and permuted training targets. Under TWIN the probing anchors pass the tests that the compactness, importance, and matching scores fail.

AI USE STATEMENT

Generative AI tools were used solely to assist with language editing and to improve the clarity and readability of the manuscript. All scientific content, analyses, results, and conclusions were developed and verified by the authors.

REPRODUCIBILITY STATEMENT

The scores, TWIN, and the anchors are specified in §3 and Table 1, and the proof of Proposition 1 is in Appendix A. Dictionaries, data, splits, probes, and compute are described in Appendix C, the bootstrap and effect sizes in Appendix D, the label audit in Appendix E, the null-corrected scores in Appendix N, the held-out selection in Appendix L, and every pre-registered decision rule with its outcome in Appendix O. Every reported number is traceable to a single results file, and a verification script checks the manuscript against those files; every public-SAE quantity is recomputed end to end from the stored document codes by the released pipeline, which also regenerates those codes from the public corpora. Code and configurations will be released. The EuroVoc and GoEmotions supports are public and will be released with them. The UMLS concept supports are derived from MIMIC-CXR reports, which are distributed under the PhysioNet Credentialed Health Data License, so in their place we release the extraction and negation pipeline that regenerates them from a credentialed copy of MIMIC-CXR.

REFERENCES

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, 2018.

Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. Language models can explain neurons in language models. *OpenAI Blog*, 2023.

Olivier Bodenreider. The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl_1):D267–D270, 2004.

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

Marc-André Carbonneau, Julian Zaidi, Jonathan Boilard, and Ghyslain Gagnon. Measuring disentanglement: A review of metrics. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

Enrico Cassano, Riccardo Renzulli, Marco Nurisso, Mirko Zaffaroni, Alan Perotti, and Marco Grangetto. SAEmnesia: Erasing concepts in diffusion models with supervised sparse autoencoders. In *International Conference on Machine Learning*, 2026.

Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. Large-scale multi-label text classification on EU legislation. In *Annual Meeting of the Association for Computational Linguistics*, 2019.

David Chanin. Are sparse autoencoder benchmarks reliable? *arXiv preprint arXiv:2605.18229*, 2026.

David Chanin, James Wilken-Smith, Tomáš Dulka, Hardik Bhatnagar, Satvik Golechha, and Joseph Bloom. A is for absorption: Studying feature splitting and absorption in sparse autoencoders. In *Advances in Neural Information Processing Systems*, 2025.

Ricky T. Q. Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, 2018.

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

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. GoEmotions: A dataset of fine-grained emotions. In *Annual Meeting of the Association for Computational Linguistics*, 2020.

Cian Eastwood and Christopher K. I. Williams. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations*, 2018.

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 *International Conference on Learning Representations*, 2025.

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

Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

Thomas Heap, Tim Lawson, Lucy Farnik, and Laurence Aitchison. Automated interpretability metrics do not distinguish trained and random transformers. In *International Conference on Learning Representations*, 2026.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Conference on Empirical Methods in Natural Language Processing*, 2019.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *AAAI Conference on Artificial Intelligence*, 2019.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(317), 2019.

Adam Karvonen, Can Rager, Johnny Lin, Curt Tigges, Joseph Bloom, David Chanin, Yeu-Tong Lau, Eoin Farrell, Callum McDougall, Kola Ayonrinde, et al. SAEBench: A comprehensive benchmark for sparse autoencoders in language model interpretability. In *International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 29223–29264, 2025.

Jonas Klotz, Cassio F. Dantas, Pallavi Jain, Diego Marcos, and Begüm Demir. Evaluating the interpretability of sparse autoencoders with concept annotations. In *European Conference on Computer Vision*, 2026.

Anton Korznikov, Andrey Galichin, Alexey Dontsov, Oleg Rogov, Ivan Oseledets, and Elena Tutubalina. Sanity checks for sparse autoencoders: Do SAEs beat random baselines? *arXiv preprint arXiv:2602.14111*, 2026.

Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. In *International Conference on Learning Representations*, 2018.

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

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

Hyesu Lim, Jinho Choi, Jaegul Choo, and Steffen Schneider. Sparse autoencoders reveal selective remapping of visual concepts during adaptation. In *International Conference on Learning Representations*, 2025.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, 2019.

George Ma, Zhongyuan Liang, Irene Y. Chen, and Somayeh Sojoudi. Falsifying sparse autoencoder reasoning features in language models. *arXiv preprint arXiv:2601.05679*, 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*, 2025.

Gonçalo Paulo, Alex Mallen, Caden Juang, and Nora Belrose. Automatically interpreting millions of features in large language models. In *International Conference on Machine Learning*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

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.

Karl Ridgeway and Michael C. Mozer. Learning deep disentangled embeddings with the F-statistic loss. In *Advances in Neural Information Processing Systems*, 2018.

Luca Soldaini and Nazli Goharian. QuickUMLS: A fast, unsupervised approach for medical concept extraction. In *MedIR Workshop, SIGIR*, 2016.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, et al. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*, 2024.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11:2837–2854, 2010.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. AxBench: Steering LLMs? Even simple baselines outperform sparse autoencoders. In *International Conference on Machine Learning*, 2025.

Minglai Yang, Xinyu Guo, Zhengliang Shi, Jinhe Bi, Steven Bethard, Mihai Surdeanu, and Liangming Pan. AlignSAE: Concept-aligned sparse autoencoders. *arXiv preprint arXiv:2512.02004*, 2025.

A PROOF OF PROPOSITION 1

Lipschitz step. Let p, q be probability vectors and fix r. For every set A of features, $|\sum_{j\in A} p_j - \sum_{j\in A} q_j| \le \frac{1}{2}\|p-q\|_1$, the total-variation distance. Since $C_p(r) = \max_{|A|=r} \sum_{j\in A} p_j$, let A_p attain the maximum. Then $C_p(r) = \sum_{j\in A_p} p_j \le \sum_{j\in A_p} q_j + \frac{1}{2}\|p-q\|_1 \le C_q(r) + \frac{1}{2}\|p-q\|_1$, and by symmetry $|C_p(r) - C_q(r)| \le \frac{1}{2}\|p-q\|_1$ for every r. Applied to p_v and p^* this gives $\delta_v \le \frac{1}{2}\varepsilon_v$.

Sandwich step. By definition of δ_v, $C_{p_v}(r) \ge C_{p^*}(r) - \delta_v$ for all r. Hence $C_{p_v}(r) \ge \alpha$ whenever $C_{p^*}(r) \ge \alpha + \delta_v$, so the smallest r with $C_{p_v}(r) \ge \alpha$ is at most the smallest r with $C_{p^*}(r) \ge \alpha + \delta_v$: $N_\alpha(p_v) \le N_{\alpha+\delta_v}(p^*)$. The lower bound follows from $C_{p_v}(r) \le C_{p^*}(r) + \delta_v$ in the same way. Given δ_v, neither bound depends on which documents S_v contains, and no rate for δ_v or ε_v as a function of $|S_v|$ is assumed. □

Measured quantities. For each dictionary we compute ε_v, δ_v, the interval $[N_{\alpha-\delta_v}(p^*), N_{\alpha+\delta_v}(p^*)]$, its width relative to the label-free null $N_\alpha(p^*)$, and whether equation 2 holds, for every concept with support ≥ 20. Concepts with $\delta_v \ge 0.2$ are not excluded: for them $\alpha + \delta_v \ge 1$, the upper endpoint equals the live count, and the interval is uninformative on that side (the lower endpoint reaches 1 on 17 pairs). There are none on the public pool and 4,999 of the 29,207 concept–dictionary pairs on the medical pool (0–439 per dictionary, median 62; none on 13 dictionaries). Across the 39 medical dictionaries: violations $0/39$, as the proposition requires; median $\varepsilon_v = 0.341$, median $\delta_v = 0.095$, median $\delta_v/\varepsilon_v = 0.18$; median interval width 42.2% of the label-free null; $N_\alpha(p^*)$ ranges over 1–778; the interval midpoint predicts the measured score at $\rho = 0.995$. The interval is widest on the collapsed dictionaries. Table A1 gives the same quantities on the public pool. The bound holds on all 91 dictionary–label-system pairs, which checks the computation and is not independent evidence; the informative quantities are the width and the midpoint. On the public pool δ_v is smaller than on the medical pool, but so is the spread of $N_\alpha(p^*)$ across dictionaries ($1.6\times$ and $1.4\times$ against $778\times$), so there the interval is narrower than the spread without being negligible against it; the ranking result on the public pool rests on the measured correlations of Table 3, which the interval midpoint reproduces ($\rho = 0.977, 0.972$).

Table A1: **The quantities of Proposition 1 on the three pools** (medians over concepts, then over dictionaries). $\varepsilon_v = \|p_v - p^*\|_1$; $\delta_v = \max_r |C_{p_v}(r) - C_{p^*}(r)|$; width: width of the interval of equation 2 relative to the label-free null $N_\alpha(p^*)$, with the maximum over dictionaries in parentheses; violations: dictionaries with any concept outside its interval; midpoint: rank correlation of the interval midpoint with the measured score; dense share: mass of p_v on the top $N_\alpha(p^*)$ features of p^*, under real and random supports.

Pool	ε_v	δ_v	δ_v/ε_v	width (max)	violations	midpoint ρ	range of $N_\alpha(p^*)$	dense share real / random
UMLS, 39 medical	0.341	0.095	0.18	42.2% (217%)	0/39	0.995	1–778	0.797 / 0.801
EuroVoc, 26 Gemma Scope	0.144	0.017	0.11	14.9% (36.4%)	0/26	0.977	17–27	0.796 / 0.806
GoEmotions, 26 Gemma Scope	0.057	0.006	0.12	4.7% (14.3%)	0/26	0.972	19–27	0.806 / 0.807

B SYNTHETIC PLANTED-CONCEPT CHECK

We build dictionaries in which the answer is known: 20,000 documents, $m = 2{,}048$ features, 200 concepts with supports of 100–1,000, and $n_{\text{bg}} \in \{32, 128, 512, 1848\}$ background features, with three seeds per cell (48 dictionaries). Each concept v has a dedicated planted feature that fires on exactly $|S_v|$ documents at every signal strength $s \in \{0, 0.25, 0.5, 1\}$: a fraction s of them is drawn from S_v and the rest uniformly from the corpus, so that $s = 0$ is a size-matched random support and $s = 1$ the concept itself. The live count is therefore $n_{\text{bg}} + 200$ at every s, and the signal axis s is separated from the occupancy axis n_{bg}. A first version of this check let the planted feature fire only when $s > 0$, which moved the live count along the signal axis, and drew the remaining documents outside S_v, which excludes the concept at $s = 0$; both were corrected before the run reported here, and the first run is archived with the results. TWIN (§3.2) and the anchors of §3.3 are applied unchanged, and the decision rules were fixed before the results were computed. Table A2 reports means over n_{bg} and seeds at each s.

Table A2: Synthetic dictionaries with the live count fixed at $n_{\text{bg}} + 200$: means over $n_{\text{bg}} \in \{32, 128, 512, 1{,}848\}$ and three seeds at each signal strength s. Compactness changes little along the signal axis; MIG and the decodable fraction follow it.

Signal strength s	live@k32	mass coverage (real labels)	MIG	decodable
0 (random support)	$n_{\text{bg}} + 200$	452	0.0003	0.007
0.25	$n_{\text{bg}} + 200$	450	0.102	0.73
0.5	$n_{\text{bg}} + 200$	449	0.297	1.00
1 (the concept)	$n_{\text{bg}} + 200$	444	0.994	1.00

Across the 48 dictionaries, mass coverage tracks its label-free null at $\rho = 0.962$ and spans $38\times$ along the occupancy axis, whereas along the signal axis it moves by a median of 1.6% (at most 3.2%) and random labels return it to 0.998 of its value. MIG and the decodable fraction track the signal ($\rho = 1.00$ and 0.95 with s); MIG under random labels is 0.001 of its real-label value at $s > 0$; and at $s = 0$, where the planted feature fires on a size-matched random support by construction, both anchors are at their random-label values (MIG ratio 1.015, decodable difference $+0.005$), so they do not manufacture a signal where none was planted. One pre-registered criterion fails as written: we required $|\rho(\text{mass coverage}, s)| < 0.5$ within each n_{bg}–seed cell, but with four values of s per cell the rank correlation is -1 for any monotone decrease, and the 1.6% decrease is one (11 of 12 cells). The criterion should have been an effect size (1.6% against $38\times$); we report the failure as written.

C Dictionaries, data, and compute

Features and splits. Table A3 lists the 12 configurations. Image features are CLIP ViT-B/16 patch embeddings averaged within a 2×2 grid of regions (3,072 dimensions). The 242,727 images are split at the study level into 169,891 / 48,579 / 24,257 (train / validation / held-out; seed 0). The UMLS label system has 767 concepts with support ≥ 20 on the held-out set and 1,045 on the validation set. The 26 Gemma Scope SAEs are layers 5, 12, and 20 at width 16k and layer 12 at widths 32k and 65k, each at five or six target ℓ_0 values between 12 and 445; on GoEmotions their realized ℓ_0 is 1.7–40$\times$ the target. Ridge probes use a study-level 70/30 split of the held-out set (16,986 / 7,271 images). One training run of the report-aligned dictionary takes 58–95 s (mean 70 s) on a single NVIDIA H200. Probes are solved in the dual form when the features outnumber the training documents; restricting a probe to live features leaves its solution unchanged. New score implementations were cross-checked against the existing pipeline with exact agreement on all 39 medical and 26 public dictionaries.

Number of kept entries. Changing the number of kept entries from 32 to 16, 64, or 128 leaves every compactness verdict in place: $\rho(\text{real}, \text{random})$ is at least 0.84 (EuroVoc mass coverage at $k = 16$) and at least 0.93 in every other cell, and mass coverage tracks its label-free null at $\rho = 0.86$–0.999. Verdicts of other scores move on the public pool, where the median live count changes 13-fold across k on EuroVoc (432 to 5,633) and 4-fold on GoEmotions (2,399 to 9,629), against 302 to 893 on UMLS: DCI-completeness passes under EuroVoc at $k = 16$ (0.40), best-feature AUROC and max-activation purity pass there at $k \geq 64$ (0.31–0.38), modularity fails there at $k = 16$ (0.61), and MIG fails under GoEmotions at $k = 64$ (0.62).

Sparse probe. The k-sparse anchor follows SAEBench (Karvonen et al., 2025): for each concept, the $k \in \{1, 2, 5\}$ latents with the largest difference in mean activation between documents in and out of S_v on the training part are selected, and an ℓ_2-regularized logistic probe is refit on them with standardized inputs. It was adopted under a rule fixed before it was run: real labels beat size-matched random supports on at least 37 of 39 dictionaries with a random-support median in $[0.47, 0.53]$, and real-target training beats permuted targets on a majority of seeds at $p < 0.05$. All three k satisfy it; at $k = 5$ the median AUROC is 0.571 under real supports (0.50–0.62 across dictionaries) and 0.499 under random ones, real labels win on 38 of 39, and `poincare` beats `shuffled` on 20 of 20 seeds (0.626 vs. 0.585, $d = 5.0$; all four conditions in Table 2 and Appendix K), with the raw and matched-budget values equal to four decimals because the selected latents lie within the top 115 by activation mass. A masked ridge probe, the full probe with all but its k largest-magnitude weights set to zero and no refitting, has a median of 0.499 on this pool and carries signal only on the

trained conditions (0.56–0.60); it is not an anchor and appears only in the geometry comparison of Appendix K.

Codes and live features. The 32 entries of largest magnitude are kept and negative entries among them are set to zero. On the medical pool a median of 51% of the kept entries (47–100% across dictionaries) are negative, so a code has a median of 16 nonzero entries; in the five collapsed dictionaries, linear encoders (Table A3), 96–99.99% of the kept entries are negative, which leaves 9–16 live features. The public SAEs use JumpReLU encoders with nonnegative codes, on which the truncation has no such effect. **Matched IoU.** The Hungarian assignment runs over all features, including those with zero IoU, so every concept receives a feature; the mean is over all concepts, and a concept assigned no live feature contributes zero, so under $L = 115$ at most 115 of the 767 concepts contribute.

Table A3: The 12 configurations of the report-aligned dictionary that make up the 39 medical dictionaries. Granularity: alignment per sample or per feature. Sparsification: train-time or post-hoc top-k. The five collapsed dictionaries (live@k32 ≤ 16) are all linear-ReLU encoders with $\lambda_{\text{align}} = 0$. The seed grid uses the sample / Poincaré / MLP / post-hoc and sample / Euclidean-tangent / MLP / post-hoc configurations.

Granularity	Geometry	Encoder	Sparsification	ReLU	n	live@k32	top-k at training	λ_{align}
sample	Poincaré	MLP	train	–	13	137–1,951	32, 64	0, 10^{-4}–10^{-1}
sample	Poincaré	MLP	post-hoc	–	10	50–747	32, 64	0, 0.01
sample	Poincaré	linear	train	yes	4	9–16	32, 64	0
sample	Euclidean tangent	MLP	train	–	4	712–1,560	64	10^{-4}–10^{-1}
feature	Poincaré	MLP	train	–	1	1,244	64	0.01
feature	Poincaré	MLP	post-hoc	–	1	400	32	0.01
feature	Euclidean tangent	MLP	train	–	1	1,244	64	0.01
feature	Euclidean tangent	MLP	post-hoc	–	1	400	32	0.01
sample	Poincaré	linear	post-hoc	yes	1	16	32	0
sample	Euclidean tangent	MLP	post-hoc	–	1	305	32	0.01
sample	Euclidean ambient	MLP	post-hoc	–	1	409	32	0.01
sample	Poincaré	linear	train	–	1	677	32	0

D STATISTICS

Cluster bootstrap. A bootstrap sample draws clusters with replacement and recomputes the correlation; CIs are the 2.5th and 97.5th percentiles of $B = 4{,}000$ resamples (seed 0). Clusters: the 12 configurations for the 39 medical dictionaries (mean size 3.25), the 20 seeds for the 80-run grid (size 4), and the five layer–width groups for the 26 public SAEs (mean size 5.2). **Effective sample size.** With intraclass correlation ICC and cluster sizes m_c summing to N, the design effect for unequal clusters is $1 + (\sum_c m_c^2/N - 1)\,\mathrm{ICC}$. Medical: $\mathrm{ICC} = 0.722$ and cluster sizes 13, 10, 4, 4, and eight singletons give a design effect of 6.00 and $n_{\text{eff}} = 6.5$. Public: $\mathrm{ICC} = 0.056$ and five groups of five or six give a design effect of 1.24 and $n_{\text{eff}} = 21.0$. All intervals in the paper are cluster bootstraps and do not rely on n_{eff}. **AUROC.** Ties receive average ranks, so a constant probe scores exactly 0.5. **Aggregation.** Per-concept scores are aggregated to a dictionary-level score by the median over concepts; with the mean, the label-free correlation on the public SAEs is 0.9814 against 0.9748 with the median, so the choice does not affect the result. **Effect sizes.** For the paired comparisons of Appendix K we report $d = |\Delta\text{mean}|/\text{pooled sd}$ as a descriptive effect size (mass coverage on `poincare` vs. `shuffled`: 0.296; decodable fraction: 5.50; MIG: 2.35). It is not a power analysis for the paired Wilcoxon test, whose smallest two-sided p-value with n pairs is 2^{1-n}, and we make no claim about the number of seeds a test would need. p-values are nominal and unadjusted for the number of scores; the conclusions rest on the win counts and on the comparisons with $p < 10^{-3}$.

E LABEL AUDIT AND REBUILD

The UMLS extractor was audited against the CheXpert labeler (Irvin et al., 2019) on the 13 findings the two share: consolidation, pneumothorax, atelectasis, pneumonia, edema, fracture, pleural effusion, lung opacity, lung lesion, support devices, enlarged cardiomediastinum, pleural other, and cardiomegaly. Before the rebuild the false-positive rate on negated mentions was 0.998 for consolidation and 0.945 for pneumothorax, 0.549 averaged over ten findings; after phrase-level negation

the three values are 0.030, 0.009, and 0.108. The rebuild removed concepts with support ≥ 20 from 1,912 to 1,045 on the validation set and from 1,346 to 767 on the held-out set. The correlation between mass coverage and its label-free null on the held-out set moved from 0.9982 to 0.9983; the dictionary rankings before and after agree at $\rho = 0.9997$ on the held-out set and 1.0000 on the validation set.

F Encoding order and pooling on public SAEs

Document codes for public SAEs are computed by encoding each token and averaging codes, $\boldsymbol{z}_i = \frac{1}{T}\sum_t \mathrm{SAE}(\boldsymbol{h}_t)$. Averaging hidden states first and encoding once collapses the code (Table A4); the label-free correlation is unchanged.

Table A4: Encoding-order ablation on the 26 Gemma Scope SAEs under EuroVoc.

Order	realized/target ℓ_0	live@k32	ρ(mass coverage, p^*)
encode tokens, then average (used)	1.176	1,128	0.9748
average hidden states, then encode	0.092	96	0.9874

Pooling. Mean pooling leaves mass coverage an integer in a narrow range on the public pool (Table A1), so a ranking that does not move could reflect ties. We therefore recomputed the evaluation-time null under max pooling and under firing-rate pooling, in which a document's code for a feature is the fraction of its tokens (padding and the BOS token excluded, as for mean pooling) on which the feature's JumpReLU output is positive; every pooling keeps the 32 largest entries of each document code, with ties in firing rate broken by mean activation, with the pre-registered criterion that $\rho(\text{real}, \text{random})$ exceed 0.5 under all three poolings (Table A5). The criterion is met, with a minimum of 0.948. Ties do not produce the result: Kendall τ_b is 0.91–1.00; the four compactness scores that take 26 distinct values across the 26 SAEs (exp-entropy, inverse Simpson, Gini impurity, top-10 mass) give $\rho(\text{real}, \text{random}) = 0.968$–$1.000$ over the three poolings and both corpora; and under firing-rate pooling on GoEmotions mass coverage ranges over 73–826 with 25 distinct values while $\rho = 0.995$. The re-encoding used the settings of the stored document codes; token counts match, and the recomputed mean-pooled codes differ from the stored ones by a maximum relative error of 4.9×10^{-4}.

Table A5: **Pooling ablation for mass coverage on the 26 Gemma Scope SAEs.** Range of the score under real labels with the number of distinct values in parentheses, and the agreement between the rankings under real and size-matched random supports.

Corpus	Pooling	$N_{0.8}$ range (distinct)	$\rho(\text{real}, \text{random})$	Kendall τ_b
EuroVoc	mean (used)	17–28 (10)	0.975	0.933
	max	24–46 (15)	0.979	0.927
	firing-rate	23–31 (7)	0.948	0.913
GoEmotions	mean (used)	19–27 (8)	0.987	0.971
	max	18–24 (7)	1.000	1.000
	firing-rate	73–826 (25)	0.995	0.968

G Full battery values

Draw variability. Every evaluation-time cell was recomputed under ten independent draws of the random supports. Thirty-seven of the thirty-nine defined cells keep their verdict on all ten draws: the compactness and importance scores range over at most 0.06 in ρ (mass coverage 0.9991–0.9993 on UMLS, 0.969–0.975 on EuroVoc, 0.987–0.994 on GoEmotions), and MIG passes on all ten draws under UMLS (0.09–0.34) and EuroVoc (-0.42–0.07). Two cells are draw-dependent and are marked borderline in Figure 2: MIG under GoEmotions (0.17–0.61, above 0.5 on two draws) and max IoU under GoEmotions (-0.02–0.63, above 0.5 on one). Tables and figures report the pre-registered draw. Cluster-bootstrap intervals for that draw are in the released results; with five clusters they are wide on the public pool (MIG under GoEmotions: $[-0.12, 0.72]$).

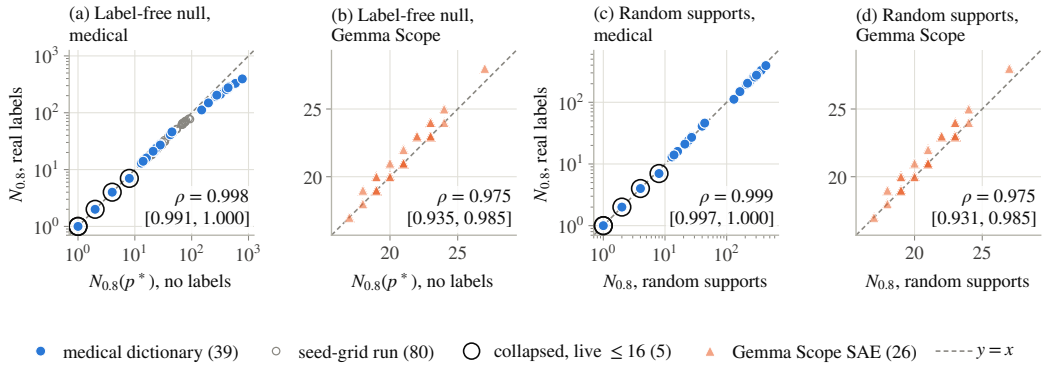

Figure A1: **Mass coverage against its label-free null and against random supports.** One point per dictionary. (a, b) The score computed with the concept labels against the same statistic computed on the marginal profile, on the 39 medical dictionaries and the 26 Gemma Scope SAEs under EuroVoc; the 80 seed-grid runs are hollow and the collapsed dictionaries are circled. (c, d) The score under real labels against size-matched random supports. Scores on the public SAEs are integers, so markers overlap. Rank correlations with cluster-bootstrap intervals are printed in each panel; those of panels (a) and (b) are listed in Table 3.

Table A6: Evaluation-time null for all fourteen scores under the three label systems: $\rho(\text{real}, \text{random})$ with, in parentheses, real-label wins and the median relative effect, signed so that positive favors the real labels. n.d.: the score takes at most two distinct values across the 26 SAEs on this corpus, so a rank correlation is uninformative; such cells are not counted as passes. Brackets give the range of ρ over ten draws of the random supports for the two draw-dependent cells.

Family	Score	UMLS (39)	EuroVoc (26)	GoEmotions (26)
Compactness	mass coverage $N_{0.8}$	0.999 (21/39, +4.5%)	0.975 (0/26, 0.0%)	0.987 (1/26, 0.0%)
	coverage $N_{0.7}$	1.000 (22/39, +6.3%)	0.936 (1/26, 0.0%)	0.993 (2/26, 0.0%)
	exp-entropy	0.998 (28/39, +4.5%)	0.993 (5/26, −1.3%)	0.996 (22/26, +0.4%)
	inverse Simpson	0.999 (28/39, +3.5%)	0.990 (3/26, −2.8%)	0.996 (10/26, −0.4%)
	Gini impurity	0.999 (28/39, +0.0%)	0.990 (3/26, −0.2%)	0.996 (10/26, −0.0%)
	top-10 mass	0.998 (25/39, +2.4%)	0.975 (5/26, −0.9%)	0.999 (24/26, +0.3%)
Importance	DCI-completeness	0.985 (35/39, +2.7%)	0.905 (9/26, −1.9%)	0.971 (8/26, −0.9%)
Mutual information	MIG	0.257 (38/39, +132%)	−0.197 (26/26, +9,751%)	0.483 [0.17, 0.61] (26/26, +8,730%)
	SAP	0.972 (10/39, −7.0%)	0.884 (0/26, −80.1%)	n.d.
	modularity	0.432 (37/39, +0.9%)	0.329 (26/26, +1.3%)	0.984 (26/26, +1.4%)
Max-over-features	max-activation purity	0.985 (32/39, +10.0%)	0.855 (20/26, +1.7%)	n.d.
	best-feature AUROC	0.987 (32/39, +0.1%)	0.855 (20/26, +0.0%)	n.d.
	max IoU	0.973 (39/39, +16.1%)	0.751 (26/26, +442%)	0.048 [−0.02, 0.63] (26/26, +121%)
Matching	matched IoU	0.980 (39/39, +16.7%)	0.869 (26/26, +174%)	0.768 (26/26, +96.5%)

Healthy subset. Restricting the medical pool to dictionaries with live@k32 > 100 ($n = 32$) leaves the diagnosis unchanged: the six compactness scores keep their ranking of the subset under random labels ($\rho(\text{real}, \text{random}) = 0.996$–$0.999$; real wins 20–24/32), DCI-completeness at 0.978 (31/32), SAP at 0.953 (7/32), max IoU and matched IoU at 0.951 and 0.965 (32/32), while MIG keeps a label-dependent ranking (0.295; 32/32) and modularity sits at the threshold (0.503; 32/32). On the same subset the partial correlation of MIG with the decodable fraction given its null is $+0.222$ (mass coverage: -0.181), below the pre-registered M1 threshold of 0.30 (Appendix J).

Row-permutation null. As a second evaluation-time null, the rows of the label matrix are permuted: every document receives the full label set of another document, so support sizes, label co-occurrence, and the number of labels per document are preserved and only the correspondence between documents and labels is broken. Table A8 lists $\rho(\text{real}, \text{row-permuted})$ for the first of three permutations; all three give the same verdict as the size-matched null in every defined cell, the compactness scores staying within ± 0.007 of their size-matched values, and MIG passing under all three label systems (UMLS 0.18–0.28, EuroVoc -0.31–0.14, GoEmotions 0.32–0.43). The one change is that max-activation purity under GoEmotions is undefined on all three permutations.

Table A7: **Training-time null on the medical seed grid** (`poincare` vs. `shuffled`, 20 paired seeds), raw and after occupancy control ($L = 115$). Wins count paired seeds on which the real-target condition scores better in the score's own direction; p is a paired Wilcoxon test, nominal and unadjusted.

Family	Score	raw wins	raw p	controlled wins	controlled p	Verdict
Compactness	mass coverage $N_{0.8}$	6/20	0.232	6/20	0.274	does not favor real targets
	coverage $N_{0.7}$	5/20	0.312	6/20	0.533	does not favor real targets
	exp-entropy	5/20	0.044	6/20	0.154	favors permuted targets (raw)
	inverse Simpson	5/20	0.114	6/20	0.189	does not favor real targets
	Gini impurity	5/20	0.097	6/20	0.133	does not favor real targets
	top-10 mass	7/20	0.105	7/20	0.143	does not favor real targets
Importance	DCI-completeness	12/20	0.231	4/20	0.011	favors permuted targets after control
Mutual information	MIG	19/20	6×10^{-6}	20/20	$< 10^{-5}$	passes
	SAP	2/20	1.7×10^{-4}	7/20	0.027	favors permuted targets
	modularity	20/20	$< 10^{-5}$	20/20	$< 10^{-5}$	passes
Max-over-features	max-activation purity	17/20	1.3×10^{-4}	5/20	0.030	reverses after control
	best-feature AUROC	17/20	3.9×10^{-4}	5/20	0.022	reverses after control
	max IoU	18/20	1.6×10^{-4}	12/20	0.927	lost after control
Matching	matched IoU	19/20	9.5×10^{-6}	20/20	$< 10^{-5}$	passes

Table A8: Evaluation-time null with row-permuted labels: ρ between the rankings under real labels and under a common permutation of the rows of the label matrix (first of three permutations). n.d.: the score takes at most two distinct values across the 26 SAEs.

Family	Score	UMLS (39)	EuroVoc (26)	GoEmotions (26)
Compactness	mass coverage $N_{0.8}$	0.999	0.975	0.994
	coverage $N_{0.7}$	1.000	0.936	0.993
	exp-entropy	0.998	0.995	0.996
	inverse Simpson	0.998	0.990	0.996
	Gini impurity	0.999	0.990	0.996
	top-10 mass	0.998	0.975	0.999
Importance	DCI-completeness	0.986	0.922	0.962
Mutual information	MIG	0.277	-0.075	0.321
	SAP	0.865	0.625	n.d.
	modularity	0.397	0.312	0.988
Max-over-features	max-activation purity	0.986	0.882	n.d.
	best-feature AUROC	0.986	0.882	n.d.
	max IoU	0.967	0.733	-0.066
Matching	matched IoU	0.977	0.891	0.884

H GoEmotions stress case

GoEmotions documents are short comments (40,000 documents, 28 labels, 705,584 tokens); the Gemma Scope SAEs run at $8.1\times$ their target ℓ_0 on them (range 1.7–40). The firing-rate substitutes for SAP, best-feature AUROC, and max-activation purity take at most two distinct values across the 26 SAEs on this corpus (SAP 0, best-feature AUROC 0.970, max-activation purity two values near 32.49) and are reported as undefined (n.d.); MIG is well defined (real 0.0011–0.0245 against random 0–0.0002). Mass coverage beats random supports on 1 of 26 SAEs with median effect 0.000% ($\rho(\text{real}, \text{random}) = 0.987$); MIG keeps a label-dependent ranking on eight of ten draws of the random supports (0.483 on the reported draw, range 0.17–0.61; 26/26 wins); modularity does not (0.984); max IoU passes on this corpus alone, on nine of ten draws (0.048 on the reported draw, range -0.02–0.63; Appendix G). Full values are in Table A6.

I Public-SAE anchors

On the 26 Gemma Scope SAEs under EuroVoc, live@k32 ranges over 199–3,168 against 100 labels. The decodable fraction is 0.990–1.000, the k-sparse masked probe (k=5) is 0.500–0.527, and the median probe AUROC is 0.807–0.885. The pre-registered anchor is therefore saturated and the model-selection tests of §4.3 were declared not adjudicable before any post-hoc analysis. Post hoc, the partial correlation of mass coverage with each anchor given its null is $+0.41$ (probe AUROC),

$+0.37$ (masked probe), and -0.34 (decodable fraction); the signs disagree across anchors, and we treat the result as inconclusive.

J MIG VERIFICATION

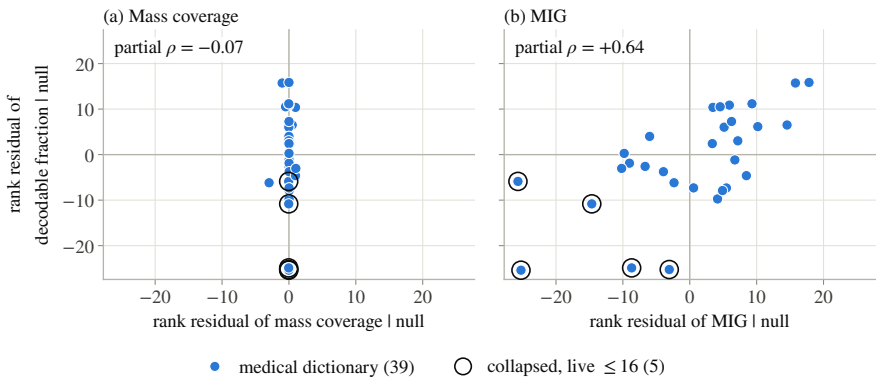

Figure A2: **Rank residuals given the label-free null.** Rank residuals of the score and of the decodable fraction after regressing each on the score's label-free null, one point per medical dictionary, with the five collapsed dictionaries marked. (a) Mass coverage shows little residual rank association with recoverability (partial $\rho = -0.07$). (b) MIG shows one ($+0.64$); its sensitivity to discretization is largest on the collapsed dictionaries.

M1. Table 4 and its sparse-probe and probe-AUROC columns ($+0.38$ against the masked probe of Appendix C), with the residuals plotted in Figure A2; on the healthy subset (live > 100, $n = 32$) the partial correlation with the decodable fraction is $+0.222$. **M2.** Condition means over 20 seeds rank the four conditions `euclid_matched` > `poincare` > `shuffled` > `noalign` by MIG and `poincare` > `euclid_matched` > `shuffled` > `noalign` by the decodable fraction (the two orderings differ by one adjacent swap; with $n = 4$ we do not treat this as a test); MIG exceeds `shuffled` on 19/20 paired seeds. The sparse probe of §3.3 orders the four conditions as MIG does (Appendix K); against the masked probe, which places `poincare` above `euclid_matched`, the condition-order correlation is $+0.40$. **M3.** `poincare` vs. `shuffled`: MIG 0.00307 vs. 0.00235, $d = 2.35$, $p = 6 \times 10^{-6}$; within-condition coefficient of variation 9.1–13.9%. **M4.** Binary vs. nine-bin quantile MIG over the 39 dictionaries: $\rho = 0.670$; over dictionaries with live > 100 ($n = 32$): 0.819; live > 200 ($n = 31$): 0.804. The largest rank moves are collapsed or small dictionaries: 37 positions for a 16-live-feature dictionary, 20 for a 201-feature one, 17 and 10 for 16- and 9-feature ones; a 12-feature dictionary has quantile MIG exactly 0, because eight quantile bins have no support. An adaptive-bin estimator (bins $= \mathrm{clip}(\mathrm{nnz}/10, 2, 8)$) agrees with the quantile estimator at $\rho = 0.997$ and with the binary one at 0.680 (0.833 for live > 100), and gives partial correlation $+0.400$ with the decodable fraction; the disagreement is largest on collapsed and low-occupancy dictionaries and is not repaired by bin design.

K FOUR-CONDITION PAIRED TESTS

Table A9: **The four conditions of the report-aligned dictionary** (20 seeds each). λ·align is the alignment term's contribution to the final loss; matching it across geometries is what makes the geometry comparison below interpretable.

Condition	Distance	Targets	λ_{align}	final align	λ·align
`poincare`	Poincaré	real	0.0100	12.498	0.1250
`shuffled`	Poincaré	permuted	0.0100	14.586	0.1459
`euclid_matched`	Euclidean (tangent)	real	0.2352	0.5212	0.1226
`noalign`	–	–	0	0	0

Targets. Each study's target is the arithmetic mean of the embeddings of the concepts in its report; the embeddings are node2vec vectors (Grover & Leskovec, 2016) of the UMLS subgraph spanned by the corpus, treated as tangent vectors at the origin and mapped by $\exp_0$ into the Poincaré ball, so the targets carry no hyperbolic structure beyond that map.

Table A10 lists the paired tests. The λ-matching check: the ratio of the alignment term's loss contribution in `poincare` to that in `euclid_matched` is 1.004–1.032 on 20/20 seeds. Feature budget after control: 115.0 and 114.45 live features for the two conditions. The remaining pairs are in the released results file.

Table A10: Paired Wilcoxon tests over 20 seeds, raw and after occupancy control ($L = 115$); d is the paired effect size.

Pair	Anchor	Mode	wins	d	p
`poincare` vs. `shuffled`	mass coverage	raw	6/20	0.296	0.232
	decodable	raw / controlled	20/20 / 20/20	5.50 / 5.47	2×10^{-6}
	sparse probe (k=5)	raw / controlled	20/20 / 20/20	5.03 / 5.02	2×10^{-6}
	masked probe (k=5)	raw / controlled	20/20 / 20/20	3.88 / 4.15	2×10^{-6}
	MIG	raw	19/20	2.35	6×10^{-6}
`poincare` vs. `euclid_matched`	mass coverage	raw	8/20	0.166	0.793
	decodable	raw / controlled	12/20 / 12/20	0.355 / 0.358	0.203 / 0.231
	sparse probe (k=5)	raw / controlled	3/20 / 3/20	1.07 / 1.07	0.001 / 0.001
	masked probe (k=5)	raw / controlled	20/20 / 20/20	5.05 / 4.95	2×10^{-6}
	probe AUROC	raw / controlled	11/20 / 11/20	0.138 / 0.166	0.546 / 0.475
`euclid_matched` vs. `shuffled`	decodable	raw / controlled	20/20 / 20/20	5.65 / 5.64	2×10^{-6}
	sparse probe (k=5)	raw / controlled	20/20 / 20/20	6.0 / 6.0	2×10^{-6}
	masked probe (k=5)	raw / controlled	2/20 / 2/20	1.75 / 1.10	6×10^{-4} / 0.003
	probe AUROC	raw / controlled	20/20 / 20/20	5.46 / 5.56	2×10^{-6}
`poincare` vs. `noalign`	decodable	raw / controlled	20/20 / 20/20	10.80 / 8.90	2×10^{-6}

Geometry comparison. The λ-matched condition also bears on the geometry of report-supervised dictionaries. At the same λ_{align} the alignment term differs $23.5\times$ between the Poincaré and Euclidean conditions of §3.4 (Table A9); once matched, the whole-dictionary anchors do not separate the geometries (decodable fraction 12/20, $p = 0.20$; probe AUROC 11/20, $p = 0.55$), and both geometries beat permuted targets on 20/20 seeds. The two k-sparse probes order them in opposite directions. Under the masked probe the Poincaré condition is higher (20/20, $d = 5.0$) and the matched Euclidean condition scores below permuted-target training (2/20), whereas under the sparse probe of §3.3 the Euclidean condition is higher (17/20, $d = 1.07$, $p = 0.001$) and beats permuted targets on 20/20 seeds ($d = 6.0$), with raw and matched-budget values identical. The two probes read different things from the same codes, the k largest weights of a full probe and the k most class-separating latents, so we draw no conclusion about geometry from either; since the targets are $\exp_0$-mapped node2vec vectors and carry no hyperbolic structure beyond that map, the question is open.

L Selection regret on held-out data

Each score is computed on the validation split (48,579 images) for the 39 medical dictionaries, the dictionary it ranks best is selected, and the selection is scored by the held-out decodable fraction; the regret of a selector is the gap between the best held-out dictionary (0.682) and its pick, and the expected regret of a uniformly random choice is 0.319 (Table A11). Five compactness scores, modularity, and the label-free null $N_\alpha(p^*)$, which uses no labels, pick the same collapsed dictionary (16 live features), so the selection is closed on held-out data: it does not depend on the labels, and its regret equals that of a random choice. Top-10 mass and SAP pick a collapsed dictionary with 9 live features, DCI-completeness picks the collapsed dictionary with 12 live features, the worst in the pool, and MIG picks the fifth-best dictionary. Averaged over each selector's three top picks, the regret is 0.48 for the compactness scores, 0.04 for MIG, and 0.02 for the validation anchor.

M Anchors under the evaluation-time null

Every support S_v is replaced by a uniformly random subset of documents of the same size, probes are retrained, and the anchors are recomputed (Table A12; three independent draws, draw 0 shown, the others agree). AUROC uses average ranks for ties. On the 39 medical dictionaries the one tie is

Table A11: **Selection regret.** Each score picks the dictionary it ranks best on the validation split; the pick is scored by the held-out decodable fraction (best 0.682, median 0.370). Regret is the gap to the best dictionary; the random-choice regret is 0.319.

Selector (validation)	held-out decodable	rank / 39	regret
mass coverage, coverage $N_{0.7}$, exp-entropy, inverse Simpson, Gini impurity, modularity	0.386	18	0.296
label-free null $N_\alpha(p^*)$ (same pick)	0.386	18	0.296
top-10 mass, SAP	0.083	38	0.598
DCI-completeness	0.000	39	0.682
max-activation purity, best-feature AUROC	0.475	11	0.207
max IoU, matched IoU	0.370	19	0.312
MIG	0.558	5	0.124
probing anchor on validation	0.682	1	0.000
random choice (expected)	0.363	–	0.319

the dictionary with 12 live features, whose probes are nearly constant and score 0.5 under real and random supports alike.

Table A12: Anchors under real and size-matched random supports.

Pool	Anchor	real (median)	random (median [range])	real wins (ties)	ρ(real, random)
UMLS, 39 trained	decodable fraction	0.370	0.070 [0.000, 0.094]	38/39 (1)	-0.06
	median probe AUROC	0.569	0.500 [0.497, 0.505]	38/39 (1)	-0.30
	sparse probe (k=5)	0.571	0.499 [0.497, 0.502]	38/39 (0)	–
	masked probe (k=5)	0.499	0.499 [0.496, 0.503]	37/39 (2)	$+0.30$
EuroVoc, 26 Gemma Scope	median probe AUROC	0.843	0.500 [0.495, 0.504]	26/26	$+0.19$
GoEmotions, 26 Gemma Scope	median probe AUROC	0.686	0.499 [0.494, 0.512]	26/26	-0.04

Under random supports the ranking of the 39 dictionaries by any anchor is noise, so $\rho(\text{real}, \text{random})$ near zero is expected and is not the verdict; the verdict rests on the win counts and on the collapse of the random-label values to chance. The masked probe (Appendix C) is at chance under real labels on this pool (median 0.499) and is not an anchor; the sparse probe of §3.3 wins 38 of 39.

N NULL-CORRECTED COMPACTNESS

The decision rules below were recorded before the corrected scores were computed. Table A13 reports two corrections of the compactness scores: D, the score under real supports minus the score under size-matched random supports, and L, the score computed on the lift profile $q_v \propto p_v/p^*$. T1 is the evaluation-time null, T2 the training-time null, and T3 the alignment with recoverability required of MIG in condition M1. For D the null of T1 is a second random draw minus the first, which shares its baseline with D under real labels, so T1 is not an independent test for D; one cell exceeds the threshold (top-10 mass under EuroVoc, $+0.74$), and the evidence for D is T2. The largest T3 value, $+0.29$ for Gini impurity, falls just below the threshold and is reported as a failure under the rule (for mass coverage, D is zero on 13 of 39 dictionaries); its cluster-bootstrap interval is $[-0.08, 0.68]$, and the intervals of all D and L scores include zero on the full pool. *Post hoc (not pre-registered).* On the healthy subset (live@k32 > 100, $n = 32$, 10 clusters) T3 for D is $+0.54$ $[0.17, 0.87]$ for Gini impurity, $+0.39$ $[0.07, 0.78]$ for top-10 mass, $+0.30$ $[0.07, 0.72]$ for exp-entropy, $+0.21$ $[0.08, 0.55]$ for inverse Simpson, and $+0.06$ $[-0.07, 0.47]$ and $+0.09$ $[-0.04, 0.46]$ for the two integer-valued coverage scores; for L and for every uncorrected compactness score the interval includes zero on the full pool and on the subset. The failure of D on the full pool therefore depends on the seven dictionaries with at most 100 live features. Under the stricter threshold live@k32 > 200 ($n = 31$) the intervals for exp-entropy, inverse Simpson, and top-10 mass still exclude zero, and that for Gini impurity ($+0.59$) is $[-0.03, 0.87]$; both thresholds were examined, and neither was fixed in advance.

O DEVIATIONS FROM PRE-REGISTERED PREDICTIONS

Every confirmatory test was pre-registered with a decision rule before its results were seen; the pre-registration documents will be released with the code. Five predictions that bear on the claims failed as written, and the paper reports each in the place it affects. The binary and quantile MIG rankings were predicted to agree at $\rho \geq 0.80$ and agree at 0.670, which is why MIG is a diagnostic and

Table A13: **Null-corrected compactness under the pre-registered tests.** T1: ρ between the ranking under real supports and under the null, pass ≤ 0.5; n.d.: the score takes at most two distinct values across the pool and is not counted as a pass. T2: seeds (of 20) on which real-target training scores better than permuted-target training, raw / occupancy-controlled, pass: majority and $p < 0.05$. T3: partial correlation with the decodable fraction given $N_\alpha(p^*)$, pass $\geq +0.30$; T3 is signed so that more compact and more decodable is positive, the opposite of Table 4 for scores in which low is compact.

Score	T1 UMLS	T1 EuroVoc	T1 GoEmotions	T2 raw / controlled	T3	Verdict
mass coverage, uncorrected	0.999	0.975	0.987	6 / 6	$+0.07$	fails all
D, mass coverage	-0.11	n.d.	n.d.	12 / 12 ($p = 0.007$ / 0.006)	-0.35	fails T3
D, coverage $N_{0.7}$	-0.59	n.d.	n.d.	9 / 11 ($p = 0.013$ / 0.005; 10 / 8 ties)	-0.33	fails T2 (raw), T3
D, exp-entropy	-0.58	-0.09	$+0.16$	19 / 19	-0.17	fails T3
D, inverse Simpson	-0.35	-0.31	$+0.17$	18 / 18	-0.17	fails T3
D, Gini impurity	-0.10	-0.23	$+0.21$	18 / 18	$+0.29$	fails T3 (borderline)
D, top-10 mass	-0.45	$+0.74$	$+0.40$	17 / 17	$+0.08$	fails T1 (EuroVoc), T3
L, mass coverage	0.999	0.977	0.987	11 / 8	$+0.11$	fails all
L, exp-entropy	0.996	0.967	0.990	9 / 6	$+0.08$	fails all

not a selector (§4.5, Appendix J). The decodable fraction was predicted to span at least $2\times$ across the public SAEs and spans $1.01\times$, which is why the consequences of §4.3 are established on the medical pool only (Appendix I). The k-sparse masked probe was predicted to beat random supports on at least 37 of 39 dictionaries under the pre-registered tie-breaking and did so on 26 on one of three draws (on 37 of 39 on all three draws with average-rank AUROC, Appendix M), so it is not used as a positive control; the sparse probe of §3.3, adopted under a rule fixed before it was run (Appendix C), wins 38 of 39. At least one max-over-features score was predicted to survive occupancy control and none did (§4.2). The geometry comparison was predicted to give the same verdict on every anchor and does not: whole-dictionary anchors do not separate the geometries while the masked probe does, and the k-sparse ordering itself reverses under the refit sparse probe, so Appendix K draws no conclusion about geometry. Four further pre-registered predictions concerned intermediate checks (a monotone dependence of the label-free correlation on label structure, the recall target of the label rebuild, the trigger threshold of the label audit, and one criterion of the synthetic check) and are documented with the pre-registration files.