# AMBIENT DIMENSION GOVERNS WHETHER FROZEN SINGLE-CELL EMBEDDINGS CAN BE EDITED AND INTERPRETED

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Frozen single-cell foundation models are increasingly used as off-the-shelf embeddings, and linear-probe benchmarks report that perturbation phenotypes are well "contained" in them. Yet the same embeddings rarely help unsupervised analysis, where perturbed and control cells fail to separate. We resolve this tension by distinguishing two axes the field conflates: recoverability, whether a supervised probe can read a label, and surfacing, whether the label occupies embedding variance without supervision. Across seven representations and six datasets, recoverability is high everywhere and, perturbation by perturbation, scales with the true transcriptional effect, whereas surfacing is low in every one of them, and whether the two axes come apart is governed by one geometric quantity, the ambient dimension of the representation. Ambient dimension thus leaves recoverability unchanged but gates whether a frozen embedding can be edited and interpreted, and this single quantity predicts three otherwise-separate phenomena. In high-ambient embeddings surfacing rather than recoverability predicts unsupervised usability (Spearman 0.77 vs 0.27), while at low ambient dimension the two axes coincide. Surfaced nuisances are removable at zero parameters everywhere, whereas aggressive edits and selective steering are safe only when ambient room exceeds the number of entangled directions: a compressed latent fails this criterion, a high-ambient one satisfies it, and an intermediate one sits on the boundary. And sparse-autoencoder features become perturbation-specific and recover the perturbed regulatory program only as ambient dimension grows (win rate 34% to 92%; regulon recovery 5% to 22%). One axis thus tells a practitioner in advance which operations a frozen embedding will support.

## 1 INTRODUCTION

A frozen single-cell foundation model is meant to be useful without labels (Cui et al., 2024; Theodoris et al., 2023; Karaletsos et al., 2026; Zhang et al., 2025): to let cell states cluster, datasets integrate, and structure emerge that no one annotated in advance. On exactly these unsupervised tasks, however, the embeddings underperform simple baselines (Kedzierska et al., 2025): plotted together, perturbed and control cells overlap, and clustering does not separate the perturbation. This is puzzling, because the same embeddings pass the field's usual test: a linear probe recovers a perturbation's identity or a cell's state at high accuracy (Wu et al., 2025; Baek et al., 2025), so the signal is reported as contained in the representation even as unsupervised use fails to find it. A probe and an unsupervised user therefore read different signals, and the standard evaluation cannot adjudicate between them. It asks only whether a label can be recovered, so it scores an embedding that presents a signal the same as one that merely retains it. The property that decides whether a frozen embedding is useful without supervision has gone unmeasured. Single-cell perturbation data are a controlled testbed for measuring it: the perturbation identity is known, and its transcriptional effect is measurable in gene space independently of any embedding, so what a representation recovers or surfaces can be checked against a ground truth the representation did not produce.

We argue the disagreement is not a contradiction but a conflation of two distinct properties. Recoverability asks whether a supervised reader can extract a label from the embedding; surfacing asks whether the label occupies embedding variance on its own, without supervision. The two need not



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

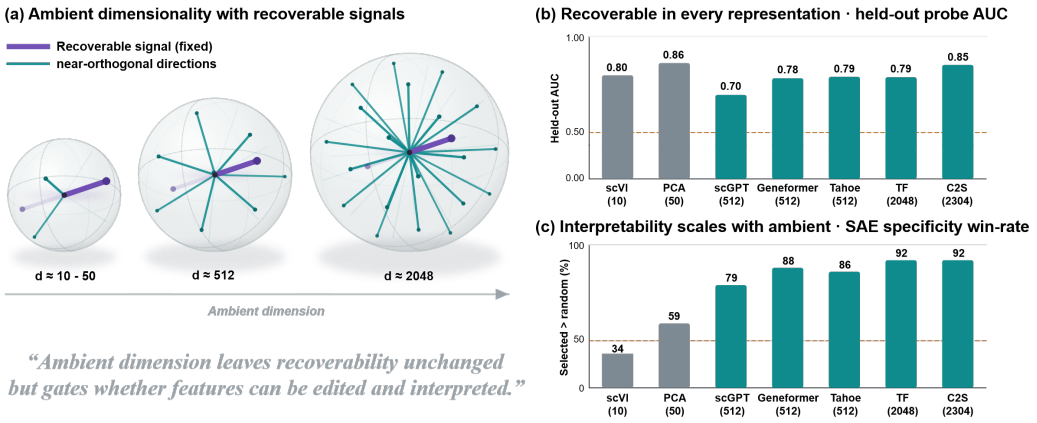


Figure 1: **Ambient dimension leaves recoverability unchanged but gates whether features can be edited and interpreted.** **(a)** Geometric intuition. A fixed, recoverable signal direction (violet) lies among near-orthogonal nuisance directions (teal) inside representations of increasing ambient dimension $d$. A linear probe recovers the signal at any $d$; the signal occupies a small fraction of the variance at every $d$, and as $d$ grows it becomes near-orthogonal to most other directions, which controls whether it can be edited or interpreted. **(b)** Recoverability. Held-out linear-probe AUC for the perturbation label is high in every representation and does not track ambient dimension. **(c)** Interpretability. SAE feature specificity rises with ambient dimension. Representations are ordered by ambient dimension and colored by group: low-ambient (scVI, PCA; gray) versus high-ambient ($\geq$ 512; teal). Bars in (b) and (c) are pooled over three perturbation datasets (norman2019, replogle_rpe1, replogle_k562_essential).

agree: a label can be perfectly recoverable by a probe while lying in a low-variance subspace that no unsupervised method will find. Superposition describes this same situation in other neural representations (Elhage et al., 2022). Once the axes are separated they behave very differently: recoverability holds in every representation, and surfacing stays low in every representation; what changes from one representation to the next is whether the two axes come apart or coincide.

That separation is governed by the ambient dimension of the representation, roughly ten for a compressed latent and hundreds to thousands for a foundation model. A high-ambient embedding has enough near-orthogonal axes to hold a perturbation signal in a low-variance subspace, separate from the dominant cell-identity structure, and it still has directions left over for editing or reading that signal without disturbing the rest. This is the recoverable-but-buried regime. A compressed latent has no such spare axes, so recoverability and surfacing collapse onto each other, editing is coarse, and features are not separable. Ambient dimension thus governs whether surfacing diverges from recoverability, together with editability and interpretability, but not recoverability itself (Figure 1b,c).

We make four contributions, each tying a different capability of a frozen embedding to this one axis. First, we separate recoverability from surfacing. Across seven representations and three perturbation datasets, recoverability is effect-calibrated and holds across representation classes while surfacing is low in all of them, and surfacing serves as a usability diagnostic exactly where ambient dimension pulls the two axes apart, which is where the field benchmarks with probing (Section 4.1, Figure 3). Second, we bring retraining-free editing operators (Turner et al., 2023; Zou et al., 2023; Ravfogel et al., 2020; Belrose et al., 2023; Ilharco et al., 2023) to these embeddings, together with a geometric criterion for when they are safe: ambient room versus the number of superposed directions (Section 4.2, Figure 5). Third, we show that the interpretability of sparse-autoencoder features (Bricken et al., 2023; Cunningham et al., 2024; Simon & Zou, 2024) on these embeddings is governed by the same axis: feature specificity rises with ambient dimension across three datasets, corroborated by regulon recovery on the one dataset with a clean ground truth (Section 4.3, Figure 6). Fourth, we show that ambient dimension is causal rather than a correlate of foundation-model pretraining: with learned structure removed, raising ambient dimension alone reproduces the editing and interpretability trends, which also resolves the scVI confound below (Section 4.4).

## 2 Related works

A growing body of work builds frozen embeddings of cells by self-supervised pretraining on large transcriptomic corpora (Cui et al., 2024; Theodoris et al., 2023; Karaletsos et al., 2026; Zhang et al., 2025; Levine et al., 2024; Rizvi et al., 2025), alongside the established variational scVI latent (Lopez et al., 2018) and cross-species embedding models (Rosen et al., 2023). These are typically assessed by how well a trained reader recovers cell type or perturbation identity, and the signal is reported to be well contained (Liu et al., 2026; Wu et al., 2025). A parallel line of work questions that reading: used zero-shot, the same embeddings often fail to separate cell types or correct batch effects and are matched or beaten by highly variable gene selection, Harmony, or scVI (Kedzierska et al., 2025). We take this discrepancy as our object of study and explain it by separating recoverability from surfacing, tied to a geometric property of the representation. That property is the geometry of the embedding space. Work on superposition shows that neural representations encode far more features than they have dimensions, placing low-importance features along low-variance, often non-orthogonal directions (Elhage et al., 2022); the variance a direction carries and the information it carries are therefore distinct quantities. We make this distinction operational for single-cell embeddings as the recoverability-versus-surfacing split, and we measure the room a representation has relative to its ambient dimension rather than by a variance-spread summary such as the participation ratio. Two families of retraining-free tools build on this geometry, both developed for language models: editing, by steering activations along concept directions (Turner et al., 2023; Zou et al., 2023), closed-form linear concept erasure (Ravfogel et al., 2020; Belrose et al., 2023), and task arithmetic on weights (Ilharco et al., 2023); and sparse autoencoders, which decompose activations into interpretable, near-monosemantic features and scale to frontier models (Bricken et al., 2023; Cunningham et al., 2024; Templeton et al., 2024; Gao et al., 2025), with early biological application to protein language models (Simon & Zou, 2024). We bring both to frozen single-cell embeddings and show that one quantity, a representation's ambient room, governs when an edit is safe and whether its features are interpretable.

## 3 Experimental setup

### 3.1 Recoverability, surfacing, and effect size.

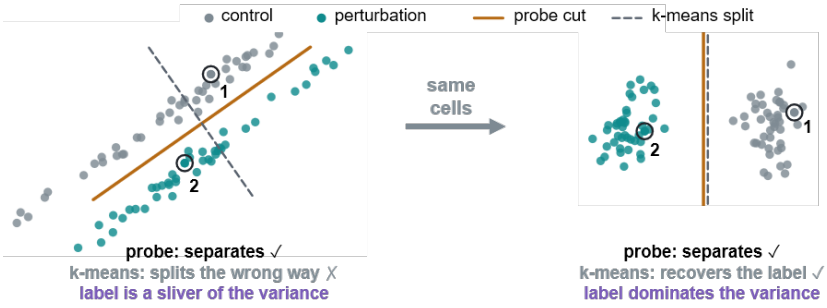


Figure 2: **Recoverability and surfacing are properties of the representation, not the data.** Illustrative schematic, not from data. The same cells carry the same control/perturbation label in both panels; only the representation differs, and two ringed cells mark identical points relocated. *Left:* the label occupies a sliver of the variance, so a supervised probe finds it (amber) but unsupervised $k$-means splits along the dominant nuisance axis (grey, dashed): *recoverable, not surfaced*. *Right:* an idealized representation where the label dominates the variance, so $k$-means recovers it too: *surfaced*. No frozen model reaches the right-hand regime, since surfaced variance is low everywhere (Fig. 3); what ambient dimension governs is whether recoverability and surfacing come apart (Section 4.1).

Recoverability asks whether a supervised reader can recover a label from the frozen embedding. We use a cross-validated linear probe: features are standardized, a multinomial logistic regression is fit under stratified $k$-fold cross-validation, and we report the macro one-vs-rest AUC; for a perturbation this is the binary probe of that perturbation against a fixed control subsample. Surfacing asks the

orthogonal question of whether the label occupies embedding variance without supervision. We define it as the standardized between-group variance fraction $\mathrm{ve} = \mathrm{tr}(S_b)/\,\mathrm{tr}(S_t)$, where $\mathrm{tr}(S_b) = \frac{1}{n}\sum_c n_c\|\mu_c - \mu\|_2^2$ is the group-mean separation and $\mathrm{tr}(S_t) = \sum_j \mathrm{Var}(z_{\cdot j})$ the total standardized variance. A whitened (Fisher) alternative (Appendix C) changes no conclusion. The two axes are formally independent: $\mathrm{AUC} \to 1$ is compatible with $\mathrm{ve} \to 0$ (Figure 2). Computing ve needs the labels, as does any measure of unsupervised usability, so its place is in a benchmark, reported next to the probe AUC, rather than in an unlabeled analysis. Both are properties of the embedding, so we need a third quantity that is not. Effect size measures how large a perturbation truly is in gene space: on lognormalized counts (Appendix A) we take per-group means and define $\mathrm{eff}_P = \|\bar{g}_P - \bar{g}_C\|_2$, the L2 norm of the pseudobulk log-fold-change between perturbation $P$ and control $C$. Short codes are used for these per-perturbation comparisons throughout: the low surfacing of even the recoverable perturbations (C1), the calibration of recoverability to the true effect size (C2), and the fraction crossing the recoverability threshold, whose over-separation signature we track separately (S1).

**Notation.** An embedding is a frozen map from a cell to $\mathbb{R}^d$, with $d$ the ambient dimension. For a labelling with classes $c = 1, \ldots, C$ we write group sizes $n_c$, total $n$, class means $\mu_c$ and grand mean $\mu$, all on the per-dimension standardized embedding $z$ unless stated otherwise.

**Representations and datasets.** We study seven frozen representations, grouped by ambient dimension rather than by model family, because ambient dimension is the axis that organizes every result below. The high-ambient group ($d \geq 512$) is scGPT, Geneformer, and Tahoe (all $d = 512$), TranscriptFormer ($d = 2048$), and C2S ($d = 2304$). The low-ambient group ($d \leq 50$) is PCA ($d = 50$ principal components of the lognormalized counts) and a trained scVI latent ($d \approx 10$). UCE is excluded because it soft-collapses on mixed-protocol data (Section 5). Two dataset families are used. For editing of nuisance factors we use three scIB integration benchmarks (Luecken et al., 2022) (pancreas, lung, immune), and for the perturbation analyses three scPerturb datasets (Norman et al., 2019; Replogle et al., 2022; Peidli et al., 2024): norman2019 (CRISPRa in K562), replogle_rpe1 (CRISPRi in RPE1), and replogle_k562_essential (CRISPRi of essential genes in K562). In each, every cell carries one targeted genetic perturbation or none (control), and the identity of that perturbation is the label used throughout. The perturbation set used by each analysis is given in Table 2 (batch and label keys in Appendix A). All representations embed the same fixed sample of $50{,}000$ cells per perturbation dataset, drawn from a shared base with a fixed seed.

**Effective dimension and editing operators.** Effective dimension is the participation ratio of the embedding covariance eigenvalues (Appendix D); it measures how spread the variance is, not the ambient room a representation has, and we report it only to falsify it as a predictor of editing safety. By ambient room we mean two ratios that are available before any experiment: the rank of the nuisance subspace to be removed relative to $d$, and the number of directions to be steered relative to $d$. Editing uses three operators, all applied at inference with no retraining. Mean-centering (E1) removes a surfaced low-rank nuisance in closed form: subtract each nuisance-group centroid and add back the global mean. Projection (E2) projects embeddings onto the orthogonal complement of the nuisance subspace; collateral is the drop in the cell-identity probe AUC. Additive steering (E3) induces a target perturbation $X$ by moving a control embedding along an edit direction $\mathrm{dir}(X)$, either the linear control-to-$X$ mean shift or a sparse-autoencoder feature direction. For steering we report off-target leakage, the fraction of that direction's norm lying inside the span of all other perturbation directions (Appendix D), which is 0 for an edit specific to $X$ and 1 for one indistinguishable from its entangled neighbours.

**Interpretation metrics.** We train a sparse autoencoder on each representation and select, for each perturbation $X$, the features specific to $X$. The specificity metric (I1) asks whether those features predict $X$'s own pseudobulk log-fold-change more specifically than a size-matched random feature set does; the win rate is the fraction of perturbations where they do. The regulon metric (I2) asks whether the perturbed transcription factor's own CollecTRI regulon is enriched in its selected feature, against a popularity-matched permutation null; it requires specific regulatory biology and is run on norman2019 only. Both constructions are in Appendix E.

**Usability and controls.** Unsupervised usability is the adjusted Rand index between $k$-means clustering ($k = C$) and the true labels, the quantity the surfacing diagnostic (C4) is meant to predict; the power and sequencing-depth controls are in Appendix B.

# 4 RESULTS

## 4.1 RECOVERABLE BUT NOT SURFACED, AND THE VALUE OF SURFACING

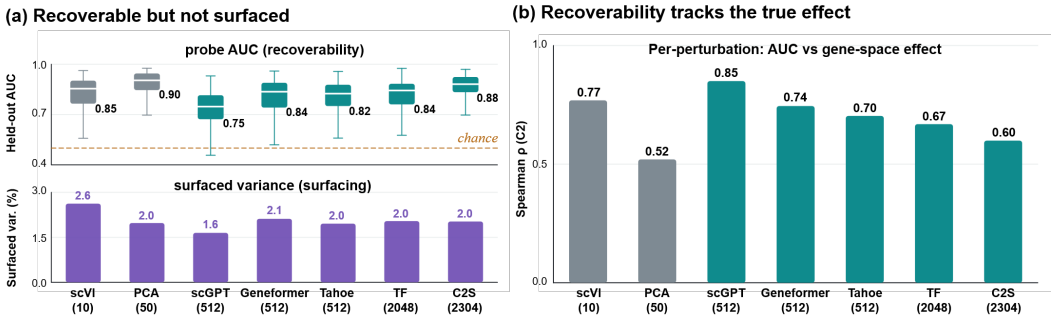


Figure 3: **Recoverability is high and effect-calibrated, but the phenotype is not surfaced (C1, C2).** Seven frozen representations, ordered by ambient dimension $d$. All quantities are per perturbation on norman2019. **(a)** *Top:* the median held-out linear-probe AUC is well above chance in every representation (boxes colored by ambient group; labels are medians). *Bottom:* the variance those same labels surface without supervision (ve) is only 1.6 to 2.6%; all representations sit far from surfaced (cf. the idealized regime in Figure 2). Recoverability and surfacing are therefore distinct axes. **(b)** Recoverability tracks the perturbation's true transcriptional effect: the Spearman correlation between probe AUC and the gene-space effect size is positive in every representation. The spread is not ordered by ambient dimension; PCA is lowest, the over-separation signature of S1.

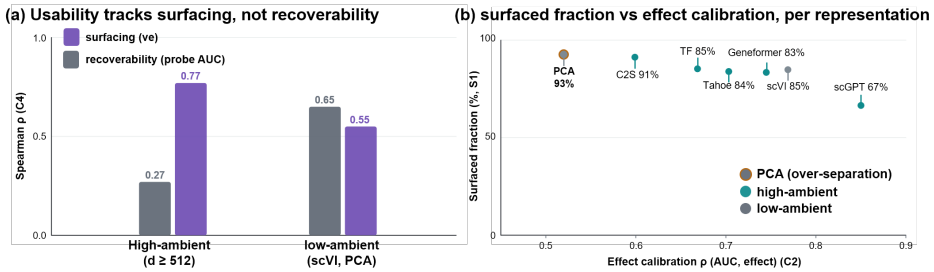


Figure 4: **Separating the two axes is diagnostic: surfacing predicts unsupervised usability, and PCA over-separates (C4, S1).** Per-perturbation quantities; (a) pooled over the three datasets, (b) on norman2019. **(a)** Among the recoverable perturbations (probe AUC $> 0.8$), unsupervised usability (2-means cluster ARI) is predicted by surfacing (ve) and not by recoverability in the high-ambient representations, while the relationship reverses in the low-ambient ones, where the two axes largely coincide. Spearman correlations pooled over the relevant representations (Table 5). **(b)** Each point is one representation: surfaced fraction (AUC $> 0.7$, S1) against effect calibration (C2). PCA (amber ring) combines the highest surfaced fraction with the lowest calibration: it separates perturbations more than their true effects warrant. The signature is directional, not a calibration ranking; surfaced fraction is not ordered by ambient dimension.

In every representation the perturbation phenotype is recoverable by a supervised linear probe yet occupies at most a few percent of embedding variance. Even among the most recoverable perturbations (probe AUC $> 0.7$) the median surfacing fraction is only ve $\approx 0.015$ pooled across the three datasets and $\approx 0.020$ on norman. Recoverability and surfacing are therefore distinct axes: the phenotype is decodable but not geometrically dominant, and the gap is depth-independent, since

per-perturbation AUC has no positive partial correlation with sequencing depth once effect size is controlled (Table 4). Probe benchmarks report the signal as contained because recoverability is high, and the same embeddings do not visibly help because surfacing is low (Figure 3a).

Recoverability is not arbitrary. It is calibrated to a perturbation's true transcriptional effect, measured independently in gene space as the L2 norm of the pseudobulk log-fold-change. The Spearman correlation between probe AUC and gene-space effect is positive in every representation and dataset: on norman it ranges from $0.52$ to $0.85$ across the seven representations, and from $0.52$ to $0.89$ across all three perturbation datasets, with low-effect perturbations sitting near chance (Table 3, Figure 3b). The correlation is robust to statistical power (fixing each perturbation to a common cell count leaves it essentially unchanged) and to sequencing depth (Table 4). The gap and its effect-calibration hold identically across the pretrained foundation models, the trained variational latent and PCA, so they are a property of single-cell representation geometry rather than of pretraining.

PCA departs from this pattern, combining the highest fraction of perturbations crossing the recoverability threshold with the lowest effect-calibration: it separates perturbations regardless of their true effect magnitude (Table 3, Figure 4b). The signature is directional rather than a calibration ranking, and it is dataset-dependent (Appendix B), so we claim only the direction.

The payoff of separating the two axes is a diagnostic. We measure unsupervised usability as the 2-means clustering agreement (adjusted Rand index) between the $\{$perturbation, control$\}$ split and the true labels, and ask whether usability is better predicted, among the recoverable perturbations (probe $\mathrm{AUC} > 0.8$), by surfacing or by recoverability. In the high-ambient representations it is predicted by surfacing and not by recoverability: pooled across the three datasets, $\rho(\mathrm{ARI}, \mathrm{ve}) = 0.77$ versus $\rho(\mathrm{ARI}, \mathrm{probe}) = 0.27$ (per dataset, ve / probe: norman $0.73$ / $0.28$, k562 $0.92$ / $0.59$, rpe1 $0.61$ / $0.14$; Table 5, Figure 4a). The relationship reverses in the low-ambient representations: for PCA and scVI, pooled $\rho(\mathrm{ARI}, \mathrm{ve}) = 0.55$ versus $\rho(\mathrm{ARI}, \mathrm{probe}) = 0.65$, because in low ambient dimension recoverability and surfacing largely coincide, so probing is not misleading there. Per representation, ve predicts usability in every representation, low- and high-ambient alike (Table 6); what distinguishes the low-ambient group is that probe predicts it as well. Because the comparison is restricted to perturbations with probe AUC above $0.8$, the range of the probe axis is compressed while that of ve is not, so we read the gap between the two correlations as a direction rather than as a magnitude. The diagnostic value of surfacing is thus specific to exactly the high-ambient representations that the field benchmarks with probing, and raw ve is at least as predictive there as a null-corrected whitened alternative (Table 6, Appendix C). That the two axes come apart only in high-ambient representations is the first appearance of the paper's axis; the next two sections show that the same axis governs whether the buried signal can be edited and read.

### 4.2 Retraining-free editing is gated by ambient room

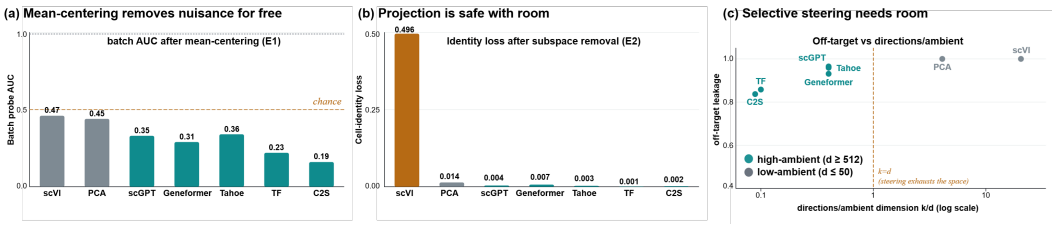


Figure 5: **Whether a buried perturbation can be edited is gated by ambient room, across three editing operations (E1–E3).** scIB batch task, max over pancreas/lung/immune, for (a,b); norman2019 for (c). **(a)** Mean-centering removes the surfaced nuisance (batch) at zero parameters in every representation: the batch probe drops to chance while cell identity is essentially untouched. **(b)** Subspace projection preserves cell identity only where there is ambient room, collapsing only in the compressed scVI latent; predicting this needs the deleted-subspace rank relative to ambient dimension, not the effective dimension or the entanglement overlap (Appendix Figure 8). **(c)** Selective steering saturates once the number of steered directions approaches the ambient dimension ($k/d = 1$, vertical guide). Values in Tables 7 and 9.

The surfaced part of a representation is the part that is easy to edit. A surfaced low-rank nuisance such as batch can be removed in closed form, with no training, by subtracting each batch centroid and adding back the global mean. Across all seven representations and the three scIB datasets this drops the batch probe AUC from about $0.99$ to chance or below in every representation, at most $0.46$ in any representation, while leaving cell identity essentially intact, the identity probe AUC falling by at most $0.022$ (worst case scVI on lung; Table 7, Figure 5a). Removing a surfaced nuisance therefore incurs no parameter cost.

Editing what is not surfaced is where representations diverge, and the distinguishing factor is ambient room rather than variance concentration. Aggressive subspace-projection removal of the nuisance is safe in every representation with sufficient ambient dimensions and damages the preserved signal only in the compressed scVI latent ($d \approx 10$), where cell identity drops by $0.051$, $0.190$, and $0.496$ (to chance) on pancreas, immune, and lung, while every representation of dimension at least $50$ loses at most $0.014$ anywhere (Table 7, Figure 5b). Within scVI the severity tracks the target-preserve entanglement across the three datasets ($0.726/0.865/1.000$ maps to $0.051/0.190/0.496$).

Two tempting predictors of safety are explicitly false (Table 8, Appendix D). Neither the variance participation ratio nor the entanglement overlap separates the fragile representation from the safe ones: the high-ambient models and PCA take participation-ratio and overlap values that overlap scVI's, yet only scVI is damaged. What actually differs is the size of the deleted subspace relative to the whole. The batch subspace has a near-fixed rank set by the number of batches, which is a large fraction of scVI's ten dimensions but a small one at fifty or more; the identity probe is then re-fit on the surviving directions, and only scVI has too few of them left to recover identity. scVI is also the only trained VAE here, so projection alone cannot separate low ambient dimension from a VAE that entangles batch and biology; the ambient-room reading is carried by PCA remaining safe at $50$ dimensions, by the next result, where $d$ is fixed and only the number of directions changes, and by the causal controls of Section 4.4.

Inducing one perturbation's phenotype by additive steering while avoiding its entangled neighbours is achievable only when the representation has room beyond the number of perturbation directions. In the high-ambient models a sparse autoencoder finds perturbation-specific features and reduces off-target leakage across all three datasets (median reductions: norman $0.03$ to $0.14$, rpe1 $0.12$ to $0.18$, k562 $0.09$ to $0.19$; helping $94$ to $100\%$ of perturbations pooled; Table 9) while remaining at least as faithful as the linear edit. In scVI no perturbation-specific features exist and off-target is exactly $1.000$ everywhere. PCA ($d = 50$) sits on the boundary: it disentangles on replogle_k562 (26 perturbations, off-target $0.82 \rightarrow 0.70$) and saturates on norman (205 perturbations, off-target $1.000$).

The boundary is real but not a clean numeric "perturbations fewer than dimensions" threshold: replogle_rpe1 has $43$ perturbations, also below $50$, yet PCA does not disentangle there either, so the clean flip is k562 versus norman (Appendix D). Plotting off-target against the ratio of the number of directions to the ambient dimension collapses all seven representations onto a single monotone trend, selective well below a ratio of one and saturated as it approaches and exceeds one (Figure 5c). The trend also sits well above the leakage a direction drawn at random would show, about $k/d$ for $k$ directions in $d$ dimensions (Appendix D), which is between $0.01$ and $0.4$ for the high-ambient models across the three datasets against an observed linear off-target of $0.97$ or above (Table 9), so the perturbation directions of a learned representation are more entangled than random ones. This reproduces the projection result's ambient-room gate with a different operation, while ambient dimension varies and the model family does not matter.

No representation offers both faithful and selective editing at once: the compressed latent is faithfully but coarsely editable (linear residual about $0.61$, off-target $1.000$), the high-ambient representations are selectively editable, and linear steering is a modest fidelity primitive everywhere (residuals $0.6$ to $0.88$; Table 9). The sparse autoencoder buys selectivity without degrading fidelity, not perfect induction.

### 4.3 Interpretability is gated by the same axis

The third operation is reading the embedding rather than editing it. We train a sparse autoencoder on each frozen representation and ask two questions of its features. The primary one, tested across all three perturbation datasets, is whether the features are specific to individual perturbations; the

secondary one, which requires a clean regulon ground truth and so runs on norman alone, is whether they recover the perturbation's named regulatory program. Both rise with ambient dimension and both fail in the two low-ambient representations (Table 1), the first as the load-bearing evidence and the second as corroboration. For specificity we score, per perturbation, whether the selected feature is more aligned to that perturbation than to its neighbours, against a size-matched random feature. The win rate rises with ambient dimension, from $34\%$ for scVI and $59\%$ for PCA to 79–$92\%$ for the high-ambient models (Table 11). The decomposition shows where the gain comes from and what it is not: in the high-ambient representations the selected feature preserves on-target alignment ($0.20$ to $0.28$, against $0.18$ to $0.25$ for random) while suppressing off-target alignment (near zero or negative, against a positive $0.09$ for random), so the specificity margin roughly doubles relative to the random baseline ($0.26$ versus $0.13$ pooled; Table 13). The two sides of that comparison move separately: the selected feature's own margin rises with ambient dimension, from $0.18$ in scVI to $0.28$ in C2S, and the random baseline's margin falls over the same range, from $0.19$ to $0.11$, so the absolute margin rises with $d$ on its own and the win rate, which compares the two, is the more generous of the two readings. The features become specific by avoiding neighbours, not by attenuating everything, and the low-ambient representations show no such margin. For program recovery we test, on norman ($n = 73$ eligible perturbations with a known regulator; perturbation sets per analysis in Table 2), whether the perturbed gene's regulon is enriched in the selected feature against a popularity-matched null. The real hit rate rises with ambient dimension, low in the two low-ambient representations and higher in the high-ambient ones (Table 12). The boundary is precise: most high-ambient models clear the null, while Geneformer is only marginal ($p = 0.069$) despite a high specificity win rate, so high ambient dimension makes interpretable features available without guaranteeing that every high-ambient model surfaces the exact program. This is the one interpretation result that rests on a single dataset, since the essential-gene screens lack a clean regulon ground truth, and we hold it as suggestive rather than settled.

Table 1: What each representation supports, read off its ambient dimension. The rule, stated as the observed boundaries rather than as thresholds: effect-calibrated recoverability (C2) and mean-centering (E1) hold at any $d$; subspace projection (E2) is safe once the deleted-subspace rank is a small fraction of $d$ (below $0.18$ for every representation of at least $50$ dimensions, against $0.5$ to $0.7$ for scVI); selective steering (E3) is selective while the number of steered directions stays well below $d$, which the high-ambient models meet on every dataset and PCA on one; and the surfacing diagnostic (C4) and SAE interpretability (I1, I2) separate the high-ambient group from the two low-ambient representations, with the location of that boundary between $50$ and $512$ not resolved by these seven. Values are $\rho_{\text{eff}}$ on norman (the spread is the PCA over-separation of S1, not ambient dimension), the I1 win rate pooled over the three perturbation datasets, and the I2 real-hit rate on norman.

| | | not gated | | ambient-gated | | | | |
|---|---|---|---|---|---|---|---|---|
| Representation | $d$ | $\rho_{\text{eff}}$ (C2) | mean-center (E1) | project (E2) | selective steer (E3) | surfacing diag. (C4) | win% (I1) | real% (I2) |
| scVI | 10 | 0.77 | ✓ | × | × | × | 34 | 5 |
| PCA | 50 | 0.52 | ✓ | ✓ | ∼ | × | 59 | 8 |
| scGPT | 512 | 0.85 | ✓ | ✓ | ✓ | ✓ | 79 | 15 |
| Geneformer | 512 | 0.74 | ✓ | ✓ | ✓ | ✓ | 88 | 14 |
| Tahoe | 512 | 0.70 | ✓ | ✓ | ✓ | ✓ | 86 | 18 |
| TranscriptFormer | 2048 | 0.67 | ✓ | ✓ | ✓ | ✓ | 92 | 21 |
| C2S | 2304 | 0.60 | ✓ | ✓ | ✓ | ✓ | 92 | 22 |

### 4.4 Ambient dimension is causal, not foundation-model structure

The results so far show that surfacing, editing safety, and interpretability all track ambient dimension across representations, but the high-ambient representations are also the pretrained transformers, and scVI is at once the only fragile editing target and the only trained variational model. Ambient dimension is therefore confounded with learned structure. We separate the two with embeddings that have a chosen ambient dimension and no learned structure (Table 10, Appendix Figure 7). A random Johnson–Lindenstrauss projection of the standardized gene space has a chosen $d$ and no learned

structure at all; random orthogonal padding instead keeps a real low-dimensional embedding's axes intact (scVI, PCA) and appends random directions to raise $d$, so the encoded signal is fixed while only $d$ changes. The random projection carries no learned structure, yet it reproduces the scVI pattern: subspace-projection identity loss and selective-steering off-target both fall as $d$ grows, so a structureless $d = 10$ embedding is as fragile as scVI and the fragility is gone by $d = 2048$. Editing safety is gated by ambient room, not by the fact that scVI is a VAE. SAE feature specificity rises with $d$ under every construction that carries signal at all, tracking the real representations, so interpretability is ambient-gated as well, with learned structure adding only a marginal bump above the structureless controls; since the win rate is scored against a random feature from the same dictionary, and padding appends directions whose features carry no signal, we read the control win rates as the direction of the dimensional effect and rely on the absolute margin of Table 13 for its size. Recoverability is flat across the real representations regardless of $d$ (probe AUC $0.71$ to $0.85$), whereas the controls move it in opposite directions, the probe rising from $0.53$ to $0.78$ under projection and falling from $0.83$ to $0.57$ under padding, while interpretability rises with $d$ under both; the eightfold fall in projection identity loss shares its range with the rising probe and so retains an information-amount component that we report rather than remove. Recoverability is thus the learned contribution and editing room and interpretation the dimensional one, which resolves the scVI confound, since a structureless representation of the same ambient dimension is fragile in the same way.

## 5 Discussion and conclusion

One geometric quantity organizes three operations that the literature treats separately. The same ambient dimension that creates the recoverable-but-buried gap also governs whether a buried signal can be surfaced as a diagnostic, removed or induced by a retraining-free edit, and read off as an interpretable feature, so that three quantities available before any experiment, $d$, the rank of the nuisance subspace relative to $d$, and the number of directions to be steered relative to $d$, indicate which of these operations a representation will support (Table 1). The result also reframes what a benchmark score means. Recoverability is the axis the field optimizes and reports, and it is the one axis ambient dimension does not gate. A high linear-probe score is consistent with a signal that occupies a few percent of the variance and that no unsupervised method will find, so a strong probe is not evidence that an embedding will help where supervision is absent. The operations that do fail in low ambient dimension are exactly those that rely on geometric structure rather than a trained reader. The confound that scVI is both the only projection-fragile representation and the only trained variational latent is resolved by Section 4.4, where structureless embeddings reproduce the fragility at low $d$ and lose it as $d$ grows.

The remaining qualifications are narrower. The steering criterion is representation-level: the ratio of directions to ambient dimension collapses the off-target curves onto one trend, but a per-perturbation entanglement predictor is not robust across datasets, so we do not claim which single edit will be clean. Program recovery rests on norman alone, the only dataset with a clean regulator-to-regulon mapping, where one high-ambient model reaches high specificity yet only marginal program recovery; more regulator-annotated datasets would settle it. The interpretability claim is relative specificity against a size-matched random feature, not absolute monosemanticity, gained by suppressing off-target alignment rather than improving on-target completeness, and it rests on the absolute margin of the selected feature, which rises with ambient dimension on its own, rather than on the win rate alone (Table 13). We provide no per-model ranking, since the leading representation varies across datasets and axes; the claim is the ambient-dimension structure, not a best model. Whether the gap extends beyond perturbation and batch, for instance to a graded cell-state axis, is an open check we expect to hold but have not run at scale, and UCE is excluded because it soft-collapses on mixed-protocol data, a limitation of that representation rather than a result about ambient dimension. One caveat carries over from the causal controls: they do not hold recoverability exactly fixed, so the trends across $d$ retain a small information-amount component, which we report rather than remove. What the present results establish is narrower than a new method: for a frozen single-cell embedding, its ambient dimension and two ratios computed from it indicate, before any edit or probe is attempted, which retraining-free operations it will support.

## AI USE STATEMENT

We used a large language model as a writing aid: to tighten prose, shorten captions, and check internal consistency between the reported numbers and the tables. All research ideas, experimental design, implementation, analysis, and claims are the authors' own, and the authors verified every number and statement in the paper against the source data.

## REFERENCES

Seungbyn Baek, Kyungwoo Song, and Insuk Lee. Single-cell foundation models: bringing artificial intelligence into cell biology. *Experimental & Molecular Medicine*, 57(10):2169–2181, 2025. doi: 10.1038/s12276-025-01547-5.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, et al. Towards monosemanticity: decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21:1470–1480, 2024. doi: 10.1038/s41592-024-02201-7.

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 (ICLR)*, 2024.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *Transformer Circuits Thread*, 2022.

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 (ICLR)*, 2025.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations (ICLR)*, 2023.

Theofanis Karaletsos, Stephen R. Quake, et al. TranscriptFormer: a generative cell atlas across 1.5 billion years of evolution. *Science*, 2026. doi: 10.1126/science.aec8514. Preprint: bioRxiv 2025.04.25.650731.

Kasia Z. Kedzierska, Lorin Crawford, Ava P. Amini, and Alex X. Lu. Zero-shot evaluation reveals limitations of single-cell foundation models. *Genome Biology*, 26, 2025. doi: 10.1186/s13059-025-03574-x. Preprint: bioRxiv 2023.10.16.561085.

Daniel Levine, Syed Asad Rizvi, Sacha Lévy, Nazreen Pallikkavaliyaveetil, David Zhang, Xingyu Chen, Sina Ghadermarzi, Ruiming Wu, Zihe Zheng, Ivan Vrkic, Anna Zhong, Daphne Raskin, Insu Han, Antonio Henrique de Oliveira Fonseca, Josue Ortega Caro, Amin Karbasi, Rahul M. Dhodapkar, and David van Dijk. Cell2Sentence: teaching large language models the language of biology. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *PMLR*, pp. 27299–27325, 2024.

Tianyu Liu et al. Evaluating the utilities of foundation models in single-cell data analysis. *Advanced Science*, 2026. doi: 10.1002/advs.202514490. scEval; preprint: bioRxiv 2023.09.08.555192.

Romain Lopez, Jeffrey Regier, Michael B. Cole, Michael I. Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15:1053–1058, 2018. doi: 10.1038/s41592-018-0229-2.

Malte D. Luecken, Maren Büttner, Kridsadakorn Chaichoompu, others, and Fabian J. Theis. Benchmarking atlas-level data integration in single-cell genomics. *Nature Methods*, 19:41–50, 2022. doi: 10.1038/s41592-021-01336-8.

Thomas M. Norman, Max A. Horlbeck, Joseph M. Replogle, Alex Y. Ge, Albert Xu, Marco Jost, Luke A. Gilbert, and Jonathan S. Weissman. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science*, 365:786–793, 2019. doi: 10.1126/science.aax4438.

Stefan Peidli, Tessa D. Green, Ciyue Shen, Torsten Gross, Joseph Min, Samuele Garda, Bo Yuan, Linus J. Schumacher, Jake P. Taylor-King, Debora S. Marks, Augustin Luna, Nils Blüthgen, and Chris Sander. scPerturb: harmonized single-cell perturbation data. *Nature Methods*, 21:531–540, 2024. doi: 10.1038/s41592-023-02144-y.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

Joseph M. Replogle, Reuben A. Saunders, Angela N. Pogson, Jeffrey A. Hussmann, Alexander Lenail, Alina Guna, Lauren Mascibroda, Eric J. Wagner, Karen Adelman, Thomas M. Norman, and Jonathan S. Weissman. Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. *Cell*, 185(14):2559–2575.e28, 2022. doi: 10.1016/j.cell.2022.05.013.

Syed Asad Rizvi, Daniel Levine, Aakash Patel, Shiyang Zhang, Eric Wang, et al. Scaling large language models for next-generation single-cell analysis (Cell2Sentence-Scale). *bioRxiv*, 2025.

Yanay Rosen, Yusuf Roohani, Ayush Agarwal, Leon Samotorčan, Tabula Sapiens Consortium, Stephen R Quake, and Jure Leskovec. Universal cell embeddings: A foundation model for cell biology. *BioRxiv*, pp. 2023–11, 2023.

Elana Simon and James Zou. InterPLM: discovering interpretable features in protein language models via sparse autoencoders. *bioRxiv*, 2024.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, et al. Scaling monosemanticity: extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*, 2024.

Christina V. Theodoris, Ling Xiao, Anant Chopra, Mark D. Chaffin, Zeina R. Al Sayed, Matthew C. Hill, Helene Mantineo, Elizabeth M. Brydon, Zexian Zeng, X. Shirley Liu, and Patrick T. Ellinor. Transfer learning enables predictions in network biology. *Nature*, 618:616–624, 2023. doi: 10.1038/s41586-023-06139-9.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Activation addition: steering language models without optimization. *arXiv preprint arXiv:2308.10248*, 2023.

Jialu Wu, Qing Ye, Yilin Wang, Renling Hu, Yiheng Zhu, Mingze Yin, Tianyue Wang, Jike Wang, Chang-Yu Hsieh, and Tingjun Hou. Biology-driven insights into the power of single-cell foundation models. *Genome Biology*, 26(1):334, 2025. doi: 10.1186/s13059-025-03781-6.

Jesse Zhang, Airol A. Ubas, Richard de Borja, Valentine Svensson, Nicole Thomas, Neha Thakar, Ian Lai, Aidan Winters, Umair Khan, Matthew G. Jones, et al. Tahoe-100M: a giga-scale single-cell perturbation atlas for context-dependent gene function and cellular modeling. *bioRxiv*, 2025. doi: 10.1101/2025.02.20.639398.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, et al. Representation engineering: a top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

Table 2: Perturbation set used by each analysis, per dataset. The sets differ by design (selection threshold and cell-count requirements differ across pipelines), so the counts are not expected to match; the qualitative claims are stable across these sets. Effect calibration and the surfacing fraction (C1, C2) use the spectrum set; the usability diagnostic (C4) uses the surfacing set; SAE interpretability (I1) uses the interpretation set; regulon recovery (I2) uses the eligible-regulator subset of norman.

| Analysis | norman | rpe1 | k562 |
|---|---|---|---|
| Spectrum / followup (C1, C2) | 218 | 117 | 145 |
| Surfacing (C4) | 60 | 117 | 120 |
| SAE interpretation (I1) | 205 | 43 | 26 |
| Regulon recovery (I2) | 73 | – | – |

# A Experimental details

This appendix collects the per-analysis tables, controls, and ablations that support the results, grouped by the section they extend. Table 2 records the perturbation set each analysis draws on; the sets differ by design because selection thresholds and cell-count requirements differ across pipelines, and the qualitative claims are stable across them. The per-perturbation surfacing spectrum admits every perturbation with at least $50$ cells ($218/117/145$ for norman / rpe1 / k562), whereas the editing and interpretation analyses use the stricter editing-threshold set ($205/43/26$).

**Dataset keys.** The three scIB integration benchmarks are used with the batch and label keys they ship with: pancreas (batch key `tech`, label `celltype`), lung (`batch`, `cell_type`), and immune (`batch`, `final_annotation`). In the three scPerturb perturbation datasets, controls are labelled `control`.

**Count normalization.** Gene-space quantities, including the effect size of Section 3.1, are computed on lognormalized counts: raw counts are total-count normalized to $10^4$ per cell and then transformed by $\log(1+x)$. Because the inputs are normalized this way, recoverability is not expected to track sequencing depth, which is the control reported in Appendix B.

# B Recoverability and surfacing

Table 3: C2/S1 per-perturbation spectrum. Effect-calibration Spearman between probe AUC and gene-space effect, and the surfaced fraction (AUC$>$0.7), per representation and dataset.

| | | Spearman $\rho_{\text{eff}}$ | | | surfaced % | | |
|---|---|---|---|---|---|---|---|
| Representation | $d$ | norman | rpe1 | k562 | norman | rpe1 | k562 |
| scVI | 10 | 0.77 | 0.69 | 0.87 | 85 | 29 | 67 |
| PCA | 50 | 0.52 | 0.55 | 0.82 | 93 | 53 | 66 |
| scGPT | 512 | 0.85 | 0.71 | 0.89 | 67 | 19 | 47 |
| Geneformer | 512 | 0.74 | 0.57 | 0.88 | 83 | 44 | 59 |
| Tahoe | 512 | 0.70 | 0.63 | 0.84 | 84 | 40 | 57 |
| TranscriptFormer | 2048 | 0.67 | 0.66 | 0.86 | 85 | 37 | 59 |
| C2S | 2304 | 0.60 | 0.54 | 0.84 | 91 | 54 | 66 |

These extend Section 4.1. Table 3 gives the per-representation effect-calibration and surfaced fraction behind C2 and the S1 over-separation signature, Table 4 is the depth-independence control for recoverability, and Table 5 is the surfacing diagnostic (C4) pooled and per dataset.

**Statistical power.** Recoverability could in principle track how many cells a perturbation has rather than how large its effect is. We control this by subsampling each perturbation to a fixed cell count and recomputing the effect-calibration correlation, which is essentially unchanged (for example scGPT stays at about $0.85$ before and after).

Table 4: Depth-independence control. Partial Spearman correlation between probe AUC and per-cell sequencing-depth difference, controlling for gene-space effect size, per representation and dataset ($^* p < 0.05$). Every value is near zero and none is positive and significant; the only significant ones are negative, the opposite of a depth-inflation confound (the largest positive value across all cells is $+0.07$). Recoverability tracks true effect, not depth.

| Representation | $d$ | norman | rpe1 | k562 |
|---|---|---|---|---|
| scVI | 10 | $-0.109$ | $+0.038$ | $+0.044$ |
| PCA | 50 | $-0.186^*$ | $+0.017$ | $+0.005$ |
| scGPT | 512 | $+0.024$ | $-0.008$ | $+0.071$ |
| Geneformer | 512 | $-0.184^*$ | $-0.054$ | $+0.031$ |
| Tahoe | 512 | $-0.074$ | $-0.006$ | $+0.009$ |
| TranscriptFormer | 2048 | $-0.123$ | $+0.001$ | $-0.019$ |
| C2S | 2304 | $-0.150^*$ | $-0.011$ | $-0.056$ |

Table 5: C4 surfacing diagnostic. Across perturbations, unsupervised usability (2-means ARI) is predicted by surfacing (ve) and not by recoverability (probe) in the high-ambient representations; the relationship reverses in the low-ambient ones, where the two axes coincide. Spearman correlations pooled over the relevant representations, with per-dataset high-ambient values shown.

| Group | dataset | $\rho(\mathrm{ARI}, \mathrm{ve})$ | $\rho(\mathrm{ARI}, \mathrm{probe})$ |
|---|---|---|---|
| high-ambient | norman | 0.73 | 0.28 |
| high-ambient | k562 | 0.92 | 0.59 |
| high-ambient | rpe1 | 0.61 | 0.14 |
| high-ambient | pooled | **0.77** | **0.27** |
| low-ambient | pooled | 0.55 | 0.65 |

**Sequencing depth.** We compute a partial Spearman correlation of recoverability against the per-cell sequencing-depth difference while controlling for gene-space effect size. It is near zero throughout, which is expected because the inputs are total-count normalized; Table 4 reports it per representation and dataset, where the only significant values are negative, the opposite of a depth-inflation confound.

## C WHITENED SURFACING ABLATION

Table 6: C4 surfacing diagnostic and the whitened-surfacing ablation (Appendix C). $\rho(\mathrm{ARI}, \mathrm{ve})$ uses raw ve; $\rho(\mathrm{ARI}, \mathrm{wht})$ uses the null-corrected whitened Fisher alternative. In the high-ambient representations raw ve is the stronger predictor and the whitened metric degrades with ambient dimension; in the low-ambient representations the two are comparable, and the whitened alternative never improves on raw ve where the diagnostic matters. The k562 run omits C2S.

| | | $\rho(\mathrm{ARI}, \mathrm{ve})$ | | | $\rho(\mathrm{ARI}, \mathrm{wht})$ | | |
|---|---|---|---|---|---|---|---|
| Representation | $d$ | norman | rpe1 | k562 | norman | rpe1 | k562 |
| scVI | 10 | 0.80 | 0.55 | 0.93 | 0.74 | 0.55 | 0.94 |
| PCA | 50 | 0.83 | 0.74 | 0.85 | 0.94 | 0.70 | 0.92 |
| scGPT | 512 | 0.82 | 0.71 | 0.85 | 0.66 | 0.45 | 0.51 |
| Geneformer | 512 | 0.88 | 0.66 | 0.81 | 0.39 | 0.26 | 0.40 |
| Tahoe | 512 | 0.47 | 0.63 | 0.86 | 0.24 | 0.46 | 0.46 |
| TranscriptFormer | 2048 | 0.85 | 0.54 | 0.86 | 0.27 | 0.27 | 0.35 |
| C2S | 2304 | 0.92 | 0.64 | – | 0.12 | 0.26 | – |

Section 3.1 uses the standardized between-group variance fraction ve as its only surfacing measure. Here we check that the surfacing diagnostic of Section 4.1 does not depend on that choice. The

whitened (Fisher) alternative replaces the isotropic normalization of ve with one that accounts for within-group scatter. For a {perturbation, control} unit we form the between-group scatter $S_b$ and the within-group scatter $S_w$, estimate $S_w$ with Ledoit–Wolf shrinkage so it is well conditioned when the cell count is below the ambient dimension, and take the generalized eigenvalue $\lambda$ of $(S_b, S_w)$, which is the between-versus-within signal strength after whitening by $S_w$. The whitened surfacing scalar is $\lambda/(1+\lambda) \in [0, 1]$, in contrast to ve, which normalizes the same between-group separation by the total variance and is therefore inflated by a dominant nuisance axis. The "null-corrected" value subtracts, per unit, the mean whitened surfacing over label permutations, removing the $d/n$ baseline that both surfacing measures carry when the ambient dimension is large relative to the cell count. Table 6 compares the two definitions. In the high-ambient representations raw ve is the stronger predictor of unsupervised usability and the whitened metric degrades with ambient dimension; in the low-ambient representations the two are comparable. The whitened alternative therefore never improves on raw ve where the diagnostic matters, so the simpler raw-ve measure of the main text is the conservative choice, and no conclusion in Section 4.1 changes under the alternative.

# D EDITING

Table 7: Editing safety on the scIB batch task (max over pancreas, lung, immune). Batch is recoverable at AUC $\approx 0.99$ in every representation before editing. Mean-centering (E1) drives the batch probe to near chance at zero parameters everywhere while the cell-identity loss stays negligible. Subspace projection (E2) is safe in every representation with ambient room and damages identity only in the compressed scVI latent.

| Representation | $d$ | batch AUC (mean-center) | identity loss (mean-center) | identity loss (projection) |
|---|---|---|---|---|
| scVI | 10 | 0.46 | 0.022 | 0.496 |
| PCA | 50 | 0.44 | 0.006 | 0.014 |
| scGPT | 512 | 0.33 | 0.003 | 0.004 |
| Geneformer | 512 | 0.29 | 0.004 | 0.007 |
| Tahoe | 512 | 0.34 | 0.003 | 0.003 |
| TranscriptFormer | 2048 | 0.22 | 0.001 | 0.001 |
| C2S | 2304 | 0.16 | 0.002 | 0.002 |

Table 8: E2 falsification: what predicts projection safety. Ranges are over the three scIB datasets; identity loss is the max. The variance participation ratio (eff. dim) does not predict safety (scVI is middling while Geneformer and TranscriptFormer are lower yet safe, and PCA is higher yet safe), and the entanglement overlap is comparably high in scVI and in the safe high-ambient representations yet only scVI is damaged. The deleted-subspace rank relative to ambient dimension (rank/$d$) is the quantity that tracks the damage: it is $0.5$–$0.7$ for scVI and below $0.18$ for everything else.

| Representation | $d$ | eff. dim | overlap | rank/$d$ | identity loss | safe |
|---|---|---|---|---|---|---|
| scVI | 10 | 8.9–9.3 | 0.73–1.00 | 0.51–0.75 | 0.496 | $\times$ |
| PCA | 50 | 8.8–15.2 | 0.34–0.68 | 0.10–0.17 | 0.014 | ✓ |
| scGPT | 512 | 5.5–14.6 | 0.73–0.79 | 0.01–0.01 | 0.004 | ✓ |
| Geneformer | 512 | 4.8–12.2 | 0.59–0.86 | 0.01–0.01 | 0.007 | ✓ |
| Tahoe | 512 | 6.9–9.3 | 0.63–0.83 | 0.01–0.01 | 0.003 | ✓ |
| TranscriptFormer | 2048 | 5.2–6.3 | 0.62–0.87 | 0.00–0.00 | 0.001 | ✓ |
| C2S | 2304 | 9.5–13.2 | 0.63–0.81 | 0.00–0.00 | 0.002 | ✓ |

These extend Section 4.2. Table 7 reports the mean-centering (E1) and projection (E2) safety on the scIB batch task, Table 8 and Figure 8 falsify the two rejected safety predictors against the deleted-subspace rank, and Table 9 gives the selective-steering (E3) off-target and fidelity residuals.

Table 9: E3 selective steering, medians pooled over the three perturbation datasets at the perturbation level, so norman (205 perturbations) carries most of the weight. Linear additive steering cannot avoid off-target leakage anywhere (off-target $\approx 1$). A sparse autoencoder restricted to perturbation-specific features reduces off-target in the high-ambient representations (helping nearly every perturbation) while keeping fidelity (lower residual is better) at least as good as linear; the compressed scVI latent has no specific features (off-target stays 1.0, helps few) and PCA is on the boundary. Fidelity residuals show the underlying tradeoff: scVI is faithfully but coarsely editable, the high-ambient representations are selectively editable but linear fidelity is only moderate.

| | | off-target | | | fidelity residual | |
|---|---|---|---|---|---|---|
| Representation | $d$ | linear | SAE | helps% | linear | SAE |
| scVI | 10 | 1.000 | 1.000 | 29 | 0.606 | 0.611 |
| PCA | 50 | 1.000 | 1.000 | 67 | 0.709 | 0.820 |
| scGPT | 512 | 0.994 | 0.958 | 99 | 0.796 | 0.681 |
| Geneformer | 512 | 0.982 | 0.912 | 94 | 0.834 | 0.783 |
| Tahoe | 512 | 0.996 | 0.945 | 100 | 0.851 | 0.794 |
| TranscriptFormer | 2048 | 0.984 | 0.837 | 99 | 0.837 | 0.771 |
| C2S | 2304 | 0.968 | 0.808 | 97 | 0.869 | 0.833 |

**Effective dimension.** The participation ratio of the embedding covariance eigenvalues $\lambda_i$ is

$$\text{eff_dim} = \Big(\sum_i \lambda_i\Big)^2 \Big/ \sum_i \lambda_i^2.$$

It summarizes how spread the variance is and is reported only as a falsified predictor of projection safety (Table 8).

**Off-target leakage.** For additive steering of perturbation $X$ along an edit direction $\mathrm{dir}(X)$,

$$\text{off-target}(X) \;=\; \frac{\|\Pi_{\mathrm{span}(D_{\neg X})}\,\mathrm{dir}(X)\|_2^2}{\|\mathrm{dir}(X)\|_2^2} \;\in [0,1],$$

the fraction of the edit direction's energy lying inside the span $D_{\neg X}$ of all other perturbation directions, computed with an orthonormal basis of that span. For a direction drawn uniformly at random in $\mathbb{R}^d$ and a subspace of dimension $k-1$, the expected value of this quantity is $(k-1)/d$, which is the random-direction reference used in Section 4.2.

**Editing protocol.** E1 and E2 operate in the native embedding coordinates; the probe standardizes its input internally. For E2 the nuisance subspace is spanned by the size-weighted centroid offsets of the nuisance groups, $\sqrt{n_g/n}\,(\mu_g - \mu)$, and the deleted-subspace rank of Table 8 is the participation ratio of their singular values; the entanglement overlap is the fraction of the identity centroid-offset energy that lies in that subspace, and the effective dimension is computed on a subsample of $5{,}000$ cells. For E3 every perturbation with at least 80 cells is subsampled to 80 and split in half, as is a subsample of $2{,}000$ control cells; $\mathrm{dir}(X)$ is the mean shift from the control half to the perturbation half in native coordinates, the edited cells are the held-out control half plus $\mathrm{dir}(X)$, and $D_{\neg X}$ is spanned by the other perturbations' mean shifts. The fidelity residual is the three-fold two-sample probe AUC between the edited cells and the held-out perturbed cells, so 0.5 is a perfect edit and the unedited control-versus-perturbed AUC is the gap an edit must close. For the sparse-autoencoder edit, the $X$-specific features are the 32 features whose mean activation shift under $X$, relative to control, most exceeds the largest such shift of any other perturbation, restricted to features whose shift is positive; the edit decodes those 32 shifts through the decoder, rescales them to native coordinates, and is scored for off-target against the same $D_{\neg X}$.

**Falsified safety predictors.** The two predictors ruled out in Section 4.2 fail on the numbers in Table 8. The variance participation ratio does not track projection safety: scVI's is middling (8.9 to 9.3), the safe high-ambient models reach lower values (Geneformer 4.8, TranscriptFormer 5.2 to 6.3), and PCA is comparable or higher (8.8 to 15.2), yet all but scVI are safe. The entanglement

overlap, the fraction of identity-offset energy lying inside the batch subspace, is comparably high in scVI (0.73 to 1.00) and in the safe high-ambient models (0.6 to 0.87), so the two groups share it while only scVI is damaged. The predictor that does track the damage is the deleted-subspace rank relative to ambient dimension: the batch subspace rank of about 4 to 8 is 51 to 75% of scVI's ten dimensions, 10 to 17% at PCA's fifty, and under 2% at any dimension of at least 512 (Figure 8).

## E INTERPRETABILITY

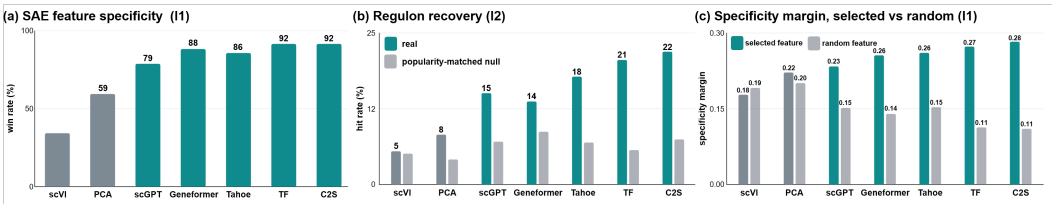


Figure 6: **Sparse-autoencoder interpretability rises with ambient dimension and fails in the compressed representations.** Seven frozen representations, ordered by ambient dimension $d$; low-ambient (scVI, PCA) in gray, high-ambient in teal. **(a)** I1 feature specificity, the fraction of perturbations whose selected SAE feature beats a size-matched random feature, pooled over the three perturbation datasets. **(b)** I2 regulon recovery on norman2019 ($n = 73$): the real hit rate (teal) rises above a popularity-matched null (gray) only in the high-ambient models, with Geneformer marginal ($p = 0.069$). **(c)** The specificity gain is selectivity, not attenuation: the selected feature suppresses off-target alignment while preserving on-target alignment (Table 13).

These extend Section 4.3. Table 11 is the I1 feature-specificity win rate per dataset and pooled, Table 13 decomposes it into on- and off-target alignment, Table 12 is the I2 regulon recovery on norman2019, and Figure 6 plots all three.

**Sparse autoencoder.** One TopK sparse autoencoder (Gao et al., 2025) is trained per dataset and representation on the per-dimension standardized embedding of every cell in the dataset, the same 50,000 cells for each of the three perturbation datasets and for every representation. The encoder keeps the $k$ largest pre-activations and rectifies them, so at most $k = 32$ features are active per cell and there is no sparsity coefficient to tune; the decoder is linear, initialized as the transpose of the encoder and trained separately. The dictionary has $4d$ features, so its size follows the ambient dimension (40 for scVI, 200 for PCA, 2048 at $d = 512$, 8192 for TranscriptFormer, 9216 for C2S), and the active fraction $k/4d$ falls from 0.8 in scVI, where the top-$k$ constraint is nearly inactive, to 0.16 in PCA and below 0.02 in every high-ambient representation. Decoder columns are held at unit norm, the decoder bias is initialized at the standardized mean, and training minimizes the mean squared reconstruction error with Adam (learning rate $10^{-3}$, batch size 4096, 120 epochs, one fixed seed). The same rule defines the dictionary for the structureless constructions of Appendix F. For each perturbation $X$ the $X$-specific features are the 32 features whose mean activation shift under $X$ most exceeds the largest such shift of any other perturbation (Appendix D); the number of them whose specificity score is positive is reported as $n_{\text{specific}}$, and it tracks ambient dimension directly, with a median of 0 in scVI and PCA, 8 to 10 in the 512-dimensional models, and 32, the cap, in TranscriptFormer and C2S. The absence of positive-specificity features in the two low-ambient representations is what leaves their steering off-target at 1.000 (Table 9). The dictionary is therefore defined relative to $d$ rather than held fixed across representations: a compressed latent cannot host a sparse overcomplete dictionary at all, which is part of what ambient room means for feature interpretability, and every win rate in Table 11 compares the selected feature against a random feature drawn from the same dictionary.

**Feature specificity (I1).** Every sparse-autoencoder feature is given a gene signature by Spearman-correlating its activation with each gene over cells. For a perturbation $X$ the $X$-specific features are aggregated, effect-weighted, into a predicted gene direction, which is compared to $X$'s true pseudobulk log-fold-change. This yields own-gene alignment $a_{\text{own}}$, the Spearman correlation with $X$'s own genes, and off-target alignment $a_{\text{off}}$, the mean Spearman correlation with other perturbations' genes.

The win rate is the fraction of perturbations satisfying $\left(a_{\text{own}} - a_{\text{off}}\right)_{\text{selected}} > \left(a_{\text{own}} - a_{\text{off}}\right)_{\text{random}}$ against a size-matched random feature set. The $X$-specific features are selected by the rule of Appendix D, the features whose activation shift under $X$ most exceeds the largest such shift of any other perturbation. Because the comparison is relative, Table 13 also reports the two margins separately; the selected margin is the absolute quantity, and the random margin is the baseline it is compared against.

**Regulon recovery (I2).** For each perturbation that activates a transcription factor with a defined CollecTRI regulon, we ask whether that TF's own regulon is enriched among the top terms of the $X$-specific feature. The feature is the single most $X$-specific one under the selection rule above; its gene signature is the correlation of its activation with each of the $2{,}000$ highly variable genes over all cells, and its top $50$ genes by absolute signature are tested by hypergeometric over-representation against the CollecTRI regulons restricted to that panel, keeping regulons with at least five targets in the panel and taking as universe the panel genes that belong to any testable regulon. The ten regulons with the smallest $p$-values are kept, a hit is the perturbed TF's own regulon among them, and the median rank of Table 12 is its rank within those ten when hit. A perturbation is eligible when its own TF has a testable regulon, which gives $n = 73$ on norman. The hit rate is scored against a popularity-matched permutation null that shuffles the perturbation-to-TF assignment among the eligible perturbations ($2{,}000$ permutations), so a regulon that is simply large or frequently annotated is hit under the null as often as under the real assignment; $p$ is the fraction of permutations that reach the real hit rate. Because the analysis requires a clean mapping from the perturbed regulator to a known regulon, it is run on norman2019 only.

## F AMBIENT CAUSAL CONTROL

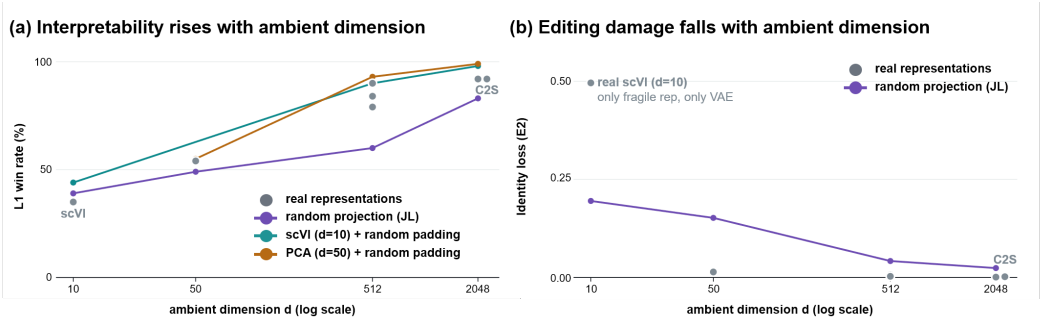


Figure 7: **Ambient dimension causally gates editing and interpretability, independent of learned structure.** To separate ambient dimension from architecture, we evaluate structureless embeddings: random projection (JL, purple), which carries no learned structure, and dimension padding (scVI $d$=10 or PCA $d$=50 augmented with random orthogonal directions, teal and orange), which preserves a real low-dimensional embedding's axes while raising $d$. **(a)** The I1 SAE feature-specificity win rate rises with $d$ under every construction that carries signal (JL $39 \to 83$, scVI+pad $44 \to 98$, PCA+pad $55 \to 99$, real $35 \to 92$), so interpretability is ambient-gated; learned structure adds only a marginal bump on top of the dimensional trend. **(b)** The E2 subspace-projection identity loss falls with $d$ for both real representations (gray) and JL (purple). Real scVI ($d$=10, the only fragile real representation and the only VAE) sits at $0.496$, yet a structureless $d$=10 embedding already reproduces the fragility ($0.195$), which decays to $0.024$ by $d$=2048; the fragility is thus dimensional, not architectural. Together the panels resolve the confound between ambient dimension and learned structure: raising $d$ in embeddings with no learned structure moves editing and interpretability in the same direction as the real representations.

Table 10 is the full structureless-control grid for Section 4.4, spanning random projection and dimension padding across ambient dimension, and Figure 7 plots the interpretability and editing trends it contains. The grid also shows what learned structure does at fixed $d$: at $d = 512$ the random projection's off-target is $0.50$ for the linear edit and $0.43$ for the sparse-autoencoder edit, while every learned representation of that dimension is at $0.98$ and $0.91$ or above respectively (Table 9),

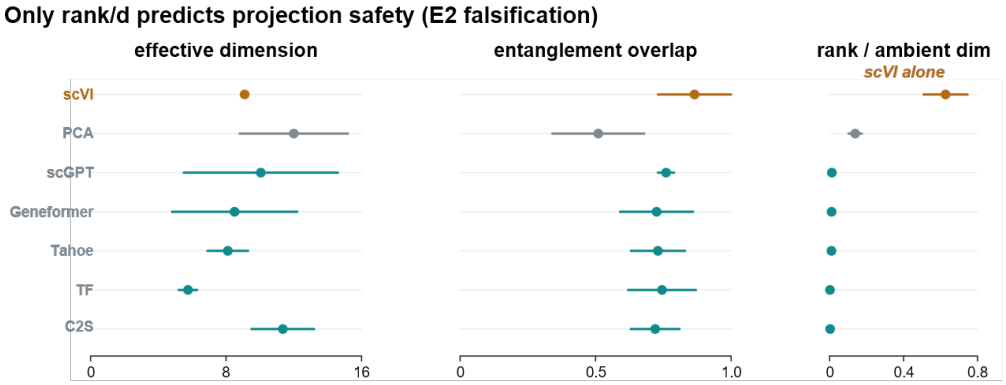


Figure 8: **The deleted-subspace rank relative to ambient dimension, not the effective dimension or the entanglement overlap, predicts projection safety (E2 falsification).** Three candidate predictors of subspace-projection safety, one per column, with representations on the vertical axis ordered by ambient dimension; each point is the mean and each whisker the min–max over the three scIB datasets (pancreas, lung, immune). scVI is the only representation damaged by projection (identity loss 0.496; Table 8). Reading across its row, scVI's effective dimension and entanglement overlap fall inside the cloud of the safe high-ambient representations, so neither separates the fragile case, whereas the deleted-subspace rank relative to ambient dimension places scVI apart (0.5–0.7 versus below 0.18 for every safe representation). Variance spread, overlap fraction, and available room are three different quantities, and only room governs safety.

so pretraining concentrates the perturbation directions into a shared subspace rather than spreading them over the available room, which is the geometry of the recoverable-but-buried regime. Ambient dimension sets the room; learned structure decides how much of it the perturbation directions share.

Table 10: Ambient causal control. Structureless embeddings at a chosen ambient dimension $d$ isolate ambient room from learned structure. Random projection (JL) is a Johnson–Lindenstrauss map of the standardized gene space; dimension padding appends random orthogonal directions to a real low-dimensional embedding (scVI $d$=10, PCA $d$=50), raising $d$ while leaving the original axes intact. Probe (recoverability), ve (surfacing), I1 win rate, and E3 off-target are on norman2019; E2 identity loss is on scIB (max over pancreas/lung/immune). The E3 column is the sparse-autoencoder steering off-target (median over perturbations); the linear off-target of the same random projections is 1.000, 1.000, 0.504 and 0.249 at $d = 10$, 50, 512 and 2048, so at every dimension with room the sparse-autoencoder edit is the more selective of the two. Editing operators are not run on the padding constructions (em-dash). Under every construction that carries signal, I1 rises and E2/E3 damage falls with $d$, matching the real representations, while probe is the axis the learned representations hold flat and the controls do not. The I1 win rate is relative to a random feature from the same dictionary; under padding the appended directions supply features without signal, so the random baseline weakens as $d$ grows and the win rate bounds the dimensional effect from above.

| Construction | $d$ | probe | ve | I1 win% | E2 id. loss | E3 off-target |
|---|---|---|---|---|---|---|
| random projection (JL) | 10 | 0.529 | 0.008 | 39 | 0.195 | 1.000 |
| random projection (JL) | 50 | 0.545 | 0.008 | 49 | 0.152 | 1.000 |
| random projection (JL) | 512 | 0.668 | 0.008 | 60 | 0.042 | 0.425 |
| random projection (JL) | 2048 | 0.780 | 0.008 | 83 | 0.024 | 0.123 |
| scVI + random padding | 10 | 0.828 | 0.050 | 44 | – | – |
| scVI + random padding | 512 | 0.603 | 0.007 | 90 | – | – |
| scVI + random padding | 2048 | 0.567 | 0.007 | 98 | – | – |
| PCA + random padding | 50 | 0.845 | 0.045 | 55 | – | – |
| PCA + random padding | 512 | 0.788 | 0.010 | 93 | – | – |
| PCA + random padding | 2048 | 0.708 | 0.007 | 99 | – | – |

Table 11: I1 feature-specificity win rate (%), the fraction of perturbations where the selected SAE feature beats a size-matched random feature in specificity. Per dataset, and pooled at the perturbation level over the three datasets.

| Representation | $d$ | norman | rpe1 | k562 | pooled |
|---|---|---|---|---|---|
| scVI | 10 | 35 | 49 | 8 | 34 |
| PCA | 50 | 54 | 77 | 77 | 59 |
| scGPT | 512 | 79 | 70 | 96 | 79 |
| Geneformer | 512 | 90 | 74 | 100 | 88 |
| Tahoe | 512 | 84 | 86 | 100 | 86 |
| TranscriptFormer | 2048 | 92 | 84 | 100 | 92 |
| C2S | 2304 | 92 | 86 | 100 | 92 |

Table 12: I2 regulon recovery on norman ($n = 73$ eligible perturbations). Real vs popularity-matched null hit rate, with the permutation $p$-value. High-ambient models clear the null; Geneformer is marginal and the two low-ambient representations do not.

| Representation | $d$ | real% | null% | $p$ | median rank |
|---|---|---|---|---|---|
| scVI | 10 | 5 | 5 | 0.549 | 6.5 |
| PCA | 50 | 8 | 4 | 0.062 | 2.0 |
| scGPT | 512 | 15 | 7 | 0.006 | 5.0 |
| Geneformer | 512 | 14 | 9 | 0.069 | 4.5 |
| Tahoe | 512 | 18 | 7 | <0.001 | 3.0 |
| TranscriptFormer | 2048 | 21 | 6 | <0.001 | 3.0 |
| C2S | 2304 | 22 | 7 | <0.001 | 3.0 |

Table 13: I1 mechanism decomposition, pooled over the three perturbation datasets. For the selected specific SAE feature (sel) versus a size-matched random feature (rnd): own-gene (on-target) alignment, off-target alignment, and the specificity margin (on-target minus off-target). The high-ambient specificity gain comes from suppressed off-target alignment (sel near zero or negative while rnd is positive) with own-gene alignment preserved, not from a general attenuation. The two low-ambient representations show no margin gain. The selected margin rises with $d$ on its own ($0.178$ to $0.283$) while the random margin falls ($0.191$ to $0.110$); the win rate of Table 11 compares the two, so the selected column is the absolute reading.

| | | own-gene | | off-target | | specificity | |
|---|---|---|---|---|---|---|---|
| Representation | $d$ | sel | rnd | sel | rnd | sel | rnd |
| scVI | 10 | 0.304 | 0.308 | +0.126 | +0.117 | 0.178 | 0.191 |
| PCA | 50 | 0.311 | 0.284 | +0.089 | +0.083 | 0.222 | 0.201 |
| scGPT | 512 | 0.203 | 0.243 | -0.030 | +0.092 | 0.234 | 0.152 |
| Geneformer | 512 | 0.240 | 0.234 | -0.016 | +0.095 | 0.256 | 0.140 |
| Tahoe | 512 | 0.250 | 0.246 | -0.011 | +0.093 | 0.261 | 0.153 |
| TranscriptFormer | 2048 | 0.276 | 0.186 | +0.003 | +0.073 | 0.273 | 0.113 |
| C2S | 2304 | 0.267 | 0.183 | -0.016 | +0.073 | 0.283 | 0.110 |