Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Decodable but Not Clusterable: Coarse Supervision, Within-Class Geometry, and a Label-Free Diagnostic for Recoverable Subclass Order

Volkan Oban

volkanobn@gmail.com

Research Article

Keywords: subclass discovery, coarse supervision, representation geometry, k-means, whitening, label-free model diagnostics, neural collapse

Posted Date: November 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11708078/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Decodable but not clusterable: a spectral account of subclass loss under coarse supervision, with label-free detection and repair

Volkan Oban

Independent Researcher, Türkiye · ORCID 0000-0003-1046-9155 · volkanobn@gmail.com

Abstract

Networks trained with coarse labels often keep fine-label information that a linear probe can read but that clustering cannot recover. We give this dissociation a mechanism, a test and a remedy. Within a coarse class, the two-means optimum is captured by the leading nuisance eigendirection once the ratio of that eigenvalue to the semantic variance, the *spectral dose*, passes a crossover between 1 and $\pi/2$; beyond π the optimum is provably blind to the fine label. Within-class whitening removes this phase transition and recovers the Bayes semantic split. A label-free monitor, the k-means gain relative to a Gaussian with the same covariance, is bounded by $1 + \pi/(2\cdot\text{dose})$ and yields a repair rule that needs no fine labels. Across seven settings and 647 class snapshots, cross-entropy fine-tuning from in-domain self-supervised bases raised the dose and cut clusterability by 59–63% while the probe stayed flat; ImageNet and DINOv2 bases lowered the dose and gained. The monitor tracked clusterability changes (Spearman 0.77), and whitening recovered 48–59% of the loss on CIFAR-100 and all of it on Fashion-MNIST. The rule still helped when the number of clusters was also estimated, whereas covariance-adaptive clustering stayed trapped.

Keywords subclass discovery · coarse supervision · representation geometry · k-means · whitening · label-free model diagnostics · neural collapse

1 Introduction

Most labelled data are labelled coarsely. A radiology archive records *pneumothorax* but not its subtype; a product catalogue records *footwear* but not *sandal* versus *ankle boot*; a benchmark records *vehicle* but not its make. After a network has been trained on such labels, a practical question remains: what structure *inside* each label is still available to an analyst who has no finer annotation? Hidden stratification of this kind causes real failures (Oakden-Rayner et al 2020; Sohoni et al 2020), and a growing literature tries to recover fine categories from coarse supervision (Bukchin et al 2021; Touvron et al 2021; Xu et al 2021; Grcić et al 2024).

A recurring and puzzling observation motivates this paper. After cross-entropy fine-tuning, a linear probe trained *with* fine labels can still separate the subclasses about as well as before, yet unsupervised clustering of the same features recovers them much worse. In our shared-base controls on CIFAR-100, the within-class probe moves from 0.763 to 0.760 while oracle-order ARI falls from 0.224 to 0.093 (Section 7.1). The information has not gone; it has become inaccessible to the

geometry that clustering uses. The phenomenon is easy to state and hard to act on, because the two quantities that everybody measures, probe accuracy and cluster quality, point in different directions and nothing in between tells the analyst why, or what to do.

This paper supplies the missing middle. We make three claims, each with theory and an experiment designed to falsify it.

A mechanism. Inside a coarse class, two-means compares the variance it can explain along each direction. When the leading eigenvalue of the nuisance covariance, λ, is large relative to the variance v along the semantic direction, the optimum splits the nuisance axis and ignores the fine label. The decisive quantity is the *spectral dose* λ/v. The exact crossover between the semantic and nuisance splits lies in the band $(1, \pi/2]$ (Corollary 1), and beyond $\lambda = \pi v$ the global optimum is provably blind to the fine label (Theorem 1). On real encoders, coarse cross-entropy moves the dose from about 0.67 to about 1.0 on CIFAR-100 and to 1.27 on Fashion-MNIST, into the crossover band, and the two-means split turns away from the span of the fine-class means (Section 7.2).

A remedy that the mechanism predicts. If the failure is spectral, equalising the spectrum should repair it. Theorem 2 shows that after within-class whitening the problem no longer depends on the nuisance spectrum at all, and that the whitened optimum is the Bayes semantic split. This is a falsifiable prediction about trained networks, and it comes with a sharp control: whitening should help only where the dose has risen and should hurt where the geometry is already healthy. Both halves hold (Section 7.5): whitening recovers 48–59% of the ARI that cross-entropy removes on CIFAR-100 and all of it on Fashion-MNIST, and it lowers ARI on every self-supervised, PSL, ViT and ResNet-50 representation we tested.

A label-free monitor. Practitioners do not have fine labels, so neither the dose nor ARI can be computed. Proposition 3 shows that the ratio ρ of the k-means gain to the gain of a Gaussian with the same covariance obeys $1 \le \rho \le 1 + \pi/(2\,\mathrm{dose})$, so $\rho \to 1$ signals capture. Across seven settings the change in ρ during fine-tuning tracks the change in ARI with Spearman correlation 0.77 (0.82 for the mixture-model ARI), and the rule "whiten when ρ has dropped since the base model" (Algorithm 2) raises the pooled ARI from 0.278 to 0.303 without a single fine label (Section 7.5).

The empirical part is organised around a pre-specified dose–response experiment, an independent re-run of the original control, and four new training settings that vary the pretraining source (in-domain self-supervision, supervised ImageNet, DINOv2), the architecture (ResNet-18, ResNet-50, ViT-S/14), the hierarchy (CIFAR-100, CIFAR-10, Fashion-MNIST) and the length of fine-tuning (5 to 30 epochs). The results draw a *regime map*: cross-entropy damages clusterability when it starts from an in-domain self-supervised base and raises the dose, and improves it when it starts from a broadly pretrained base and lowers the dose. The sign of the dose change agrees with the sign of the ARI change in 75% of 580 class snapshots.

We also report what did not go our way. The dose is not the best scalar predictor of clusterability: the fine-level collapse statistic NC1 is better (|Spearman| 0.82 against 0.58), and given NC1 the dose adds little. NC1, however, rises from 1.2 to 2.8 during fine-tuning while the probe does not move, so it measures clusterability, not decodability; the dissociation itself survives. The spectral account explains *why* the two diverge and *which* intervention helps; NC1 does not. Section 8 discusses how the two views fit together.

Contributions. (i) Corollary 1 locates the crossover of the semantic and nuisance two-means splits in $(1, \pi/2]$, which reconciles Theorem 1 with the doses observed in trained encoders. (ii) Theorem 2 proves that within-class whitening removes the phase transition and recovers the Bayes semantic split. (iii) Proposition 3 bounds a label-free monitor by the dose, and Algorithm 2 turns it into a repair rule. (iv) A pre-specified dose–response experiment and an independent re-run reproduce the dissociation to the third decimal and show the dose rising and the two-means direction rotating epoch by epoch. (v) A seven-setting regime map with an interventional test separates spectral damage from other changes. Fig. 1 summarises the resulting workflow. (vi) The order-selection theory of the earlier sections (recoverable order, stability, identification and end-to-end consistency of a cross-fitted diagnostic) is retained and evaluated on identical frozen features. (vii) A fully label-free evaluation, with the order also estimated, and a comparison with spectral, Mahalanobis and tied-covariance clustering on the same features show that the repair keeps a significant advantage and that only covariance-adaptive Euclidean methods stay trapped (Section 7.7). Table 1 lists every formal result with its status.

Table 1. Formal results and their status. *Known* results are restated for completeness; *adapted* results specialise a known argument; *new* results are original to this work.

Result	Statement in brief	Status
Observation 1	unrestricted subclass order is not identifiable	known in spirit (Teicher 1963)
Proposition 1	equal information, different two-means recoverability	new construction
Proposition 2	axis-candidate comparison of semantic and nuisance splits	new
Corollary 1	the crossover dose lies in $(1, \pi/2]$	new
Theorem 1	phase transition of the global two-means optimum	new
Theorem 2	whitening removes the phase transition	new
Proposition 3	label-free monitor bounded by $1 + \pi/(2\,\mathrm{dose})$	new
Theorem 3	stability of recoverable order under a margin	adapted
Theorem 4	identification under separated components	new
Proposition 4	scale of the radial-null floor	new
Lemma 1	PCA retains the semantic direction under an eigengap	adapted (Davis and Kahan 1970; Yu et al 2015)
Theorem 5	end-to-end consistency of the implemented diagnostic	new
Proposition A.1	what the coarse-conditioned contrastive term preserves	adapted (von Kügelgen et al 2021)

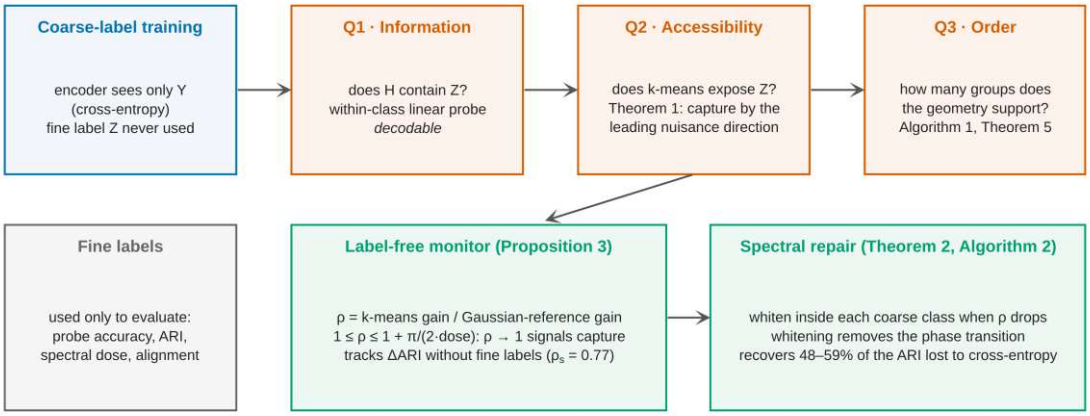


Fig. 1. Three questions and two tools. A representation trained on coarse labels is examined in three steps that fine labels never enter: whether it carries the fine distinction (Q1, a within-class probe), whether k-means can see it (Q2, the spectral mechanism of Theorem 1), and at what order the geometry supports groups (Q3, Algorithm 1). The label-free monitor of Proposition 3 detects the Q2 failure, and the within-class whitening of Theorem 2 repairs it. Fine labels are used only for evaluation.

2 Related work

Fine categories from coarse labels. Several methods learn fine-grained structure from coarse supervision: angular contrastive learning with coarse labels (ANCOR; Bukchin et al 2021), Grafit (Touvron et al 2021), weakly supervised coarse-label learning (Xu et al 2021), subclass distillation (Müller et al 2020), GEORGE for hidden stratification (Sohoni et al 2020) and FALCON, which jointly infers fine classes and their relation to the coarse ones (Grcić et al 2024). Generalised category discovery pursues a related goal with partial fine labels (Vaze et al 2022; Wang et al 2025; Dai et al 2025; Liu et al 2025). These works propose objectives; our aim is complementary. We explain when an ordinary objective hides fine structure from clustering although it keeps it decodable, and we give a label-free way to detect and undo the effect on a frozen encoder.

Collapse and within-class geometry. Neural collapse describes the shrinkage of within-class variability late in training (Papyan et al 2020), and subsequent work has qualified its reach (Hui et al 2022) and found fine-grained structure that survives it (Yang et al 2023; Wu and Mondelli 2025). Scalar collapse statistics such as NC1 summarise the *amount* of within-class spread; our analysis concerns its *orientation* relative to the semantic directions, which is what a clustering objective responds to. Section 7.4 compares the two empirically.

Two-means theory, principal points and whitening. Optimal two-cell partitions of a distribution are Voronoi half-spaces whose centres are cell means (Pollard 1981; Graf and Luschgy 2000); the two principal points of an elliptical law lie on its first principal axis (Flury 1990; Tarpey et al 1995). Spectral and Lloyd-type algorithms recover separated mixtures (Kumar and Kannan 2010; Awasthi and Sheffet 2012; Lu and Zhou 2016). Whitening is a standard preprocessing step in independent component analysis and projection pursuit, where non-Gaussian directions are sought after second-order structure has been removed (Friedman and Tukey 1974; Hyvärinen and Oja 2000; Kessy et al

2018). Theorem 2 connects these threads: it shows that whitening converts a variance competition, which Gaussian nuisance can win, into a shape competition, which it cannot.

Model order. Mixture order is identifiable only within a model class (Teicher 1963; Yakowitz and Spragins 1968; Keribin 2000). Practical selectors include BIC (Schwarz 1978), prediction strength (Tibshirani and Walther 2005) and dip-based splitting (Kalogeratos and Likas 2012). Clusters found in learned features invite selective inference (Gao et al 2024; Chen and Witten 2023), which we avoid by reporting stability margins rather than p-values.

3 Setting and two notions of access

Let $Y \in \{1, \dots, C\}$ be the observed coarse label and $Z \in \{1, \dots, K_Y\}$ an unobserved fine label, with $X \mid (Y, Z) \sim P_{Y,Z}$. An encoder h is trained on pairs (X, Y); Z never enters training and is used only to evaluate. We write $H = h(X)$ and work inside one coarse class at a time.

Observation 1 (Unrestricted order is not identifiable). *Every non-atomic distribution P on $\mathbb{R}^d$ can be written, for every $K \geq 2$, as a mixture of K components with positive weights.*

Partition the support into K sets of positive mass and condition on each. A subclass count is therefore meaningful only at a stated model class or resolution, which Section 5 makes explicit.

Since $S = (Y, Z)$, the chain rule gives $I(H; S) = I(H; Y) + I(H; Z \mid Y)$. Cross-entropy rewards the first term and contains nothing that rewards the second. That does not force the fine information to disappear, and in our experiments it mostly does not; it means only that nothing protects the *form* in which it is kept. We separate two forms of access.

Decodability is measured by a within-class linear probe: a multinomial logistic regression trained with fine labels on frozen features of one coarse class and evaluated on held-out features. **Clusterability** is measured by the adjusted Rand index (Hubert and Arabie 1985) between the fine labels and a clustering of the same features at the taxonomy order K^* (oracle-order ARI). Following our earlier protocol, features are standardised per class and projected to ten principal components before a full-covariance Gaussian mixture is fitted; we also report k-means on twenty components, which is the object of the theory. Oracle-order ARI is a diagnostic of geometry, not a discovery result, and Section 5 treats order separately.

For a fixed coarse class, let Σ be the covariance of the standardised features, let μ_z be the fine-class means and let U be an orthonormal basis of the span of $\{\mu_z - \bar{\mu}\}$, the *semantic subspace*. We define

$$v = \lambda_{\max}(U^\top \Sigma U), \qquad \lambda = \lambda_{\max}\big((I - UU^\top)\,\Sigma\,(I - UU^\top)\big), \qquad \text{dose} = \lambda / v,$$

and the *alignment* $\| U^\top \theta \|^2 \in [0,1]$ of the normal θ of the empirical two-means split with the semantic subspace. Dose and alignment require fine labels and are used only for evaluation. With two subclasses and a Gaussian nuisance these reduce exactly to the quantities of Theorem 1.

4 Theory

4.1 Information does not fix geometry

Proposition 1 (Equal information, different two-means recoverability). *There are representations H_1, H_2 with $I(H_1;Z) = I(H_2;Z) = H(Z)$ for which optimal two-means recovers Z exactly from H_1 and has error $1/2$ on H_2.*

Proof. Let Z, U be independent and uniform on $\{-1,+1\}$, take $a > 1$, and set $H_1 = (aZ, U)$ and $H_2 = (aZU, U)$. Both determine Z, through $\mathrm{sign}(H_{1,1})$ and $\mathrm{sign}(H_{2,1}H_{2,2})$. Both are uniform on the same four points $(\pm a, \pm 1)$, so they have the same optimal two-means partition, the vertical split with distortion 1 (horizontal: a^2; diagonals: a^2+1; any 1–3 split: $2/3\,(a^2+1) > 1$). That split equals Z for H_1 and is independent of Z for H_2. ■

Mutual information is invariant under invertible recoding; a Voronoi partition is not. The rest of the section turns this logical point into a quantitative one.

4.2 The spectral dose and its crossover

Fix a coarse class with two subclasses and write

$$H = S\,u + \eta, \qquad S = mZ + \varepsilon_s,$$

where u is a unit vector, Z is uniform on $\{-1,+1\}$, $\varepsilon_s \sim \mathcal{N}(0, \sigma_s^2)$ and $\eta \sim \mathcal{N}(0, \Sigma_n)$ takes values in $u^{\perp}$, all independent. Let $v = m^2 + \sigma_s^2 = \mathrm{Var}(S)$, $\lambda = \lambda_{\max}(\Sigma_n)$ and $\kappa = m^2/v \in (0,1)$. The Bayes rule $\mathrm{sign}(u^{\top}H)$ has error $\Phi(-m/\sigma_s)$, which does not depend on Σ_n. Let $q_s = \mathbb{E}|S|$. A central role is played by the folded mean of the unit-variance two-component mixture

$$M_c = 1/2\,\mathcal{N}\big(-\sqrt{c},\, 1-c\big) + 1/2\,\mathcal{N}\big(\sqrt{c},\, 1-c\big), \qquad f(c) = \mathbb{E}|M_c|, \qquad c \in [0,1).$$

M_0 is standard normal, so $f(0)^2 = 2/\pi$, and $q_s = \sqrt{v}\,f(\kappa)$. Lemma A.1 shows that f is strictly increasing.

Proposition 2 (Axis candidates). *The semantic sign partition $\mathrm{sign}(u^{\top}H)$ has two-means gain q_s^2, and the sign partition along a leading eigenvector of Σ_n has gain $2\lambda/\pi$. Hence the nuisance split has strictly lower distortion whenever $\lambda > \pi q_s^2/2$, while the Bayes error is unchanged.*

Corollary 1 (Crossover band). *The two axis candidates exchange order at the dose*

$$\mathrm{dose}^{*} = \frac{\pi}{2}\frac{q_s^2}{v} = \frac{\pi}{2}\,f(\kappa)^2 \in \Big(1, \frac{\pi}{2}\Big],$$

which tends to 1 as $m/\sigma_s \to 0$ and to $\pi/2$ as $m/\sigma_s \to \infty$.

The bounds follow from $2/\pi = f(0)^2 < f(\kappa)^2 < 1$, the upper one being Jensen's inequality $q_s^2 < \mathbb{E}S^2 = v$. Candidate comparisons are not yet statements about the global optimum, which may use oblique directions; the next theorem closes that gap at a larger dose.

Theorem 1 (Phase transition of the global two-means optimum). *Every optimal population two-means partition of H is a pair of half-spaces whose unit normal θ satisfies*

$$\text{(i)}\ \ (u^\top\theta)^2 \le \frac{\pi v}{2\lambda}, \qquad \text{(ii)}\ \ e \ge \frac{1}{2} - \sqrt{\frac{\pi}{6}}\,\frac{m\sqrt{v}}{\lambda} \quad \text{whenever } \lambda \ge \pi v,$$

where e is the fine-label error after the best relabelling. Consequently (iii) the gap between the error of global two-means and the Bayes error is at least $1/2 - \Phi(-m/\sigma_s) - \sqrt{\pi/6}\,m\sqrt{v}/\lambda$ when $\lambda \ge \pi v$: the optimal normal becomes orthogonal to u at rate $O(\lambda^{-1/2})$ and the error tends to $1/2$ at rate $O(\lambda^{-1})$, although the Bayes error does not depend on λ.

Proposition 2 and Corollary 1 say where the competition is decided; Theorem 1 says that beyond $\lambda = \pi v$ no oblique configuration can rescue the semantic split. Numerically the global optimum switches exactly at the crossover dose of Corollary 1 (Fig. 2c): for $m/\sigma_s = 1$ and 2 the predicted crossovers are 1.07 and 1.17, and the fine-label error of the computed global optimum jumps from the Bayes level to 0.50 between doses 0.97 and 1.32. Theorem 1 is therefore a guarantee, and Corollary 1 is the operative threshold. This matters for real encoders, whose standardised doses after cross-entropy lie between about 1.0 and 1.3 (Section 7).

4.3 Whitening removes the phase transition

The mechanism is a variance competition between the semantic direction and the leading nuisance direction. Equalising variances should therefore change the outcome. Let $\Sigma = vuu^\top + \Sigma_n$ be the covariance of H, assume Σ_n is non-singular on $u^\perp$, and let $\widetilde{H} = \Sigma^{-1/2}H$.

Theorem 2 (Within-class whitening). *Under the model of Section 4.2:*

(i) $\widetilde{H} = (S/\sqrt{v})\,u + \xi$ with $\xi \sim \mathcal{N}(0, I - uu^\top)$ independent of S. The law of $\widetilde{H}$, and hence every population k-means partition of $\widetilde{H}$ and its fine-label error, depends on (m, σ_s) only through κ and not at all on Σ_n.

(ii) For every unit vector θ, $\theta^\top\widetilde{H}$ has the law M_c with $c = \kappa\,(u^\top\theta)^2$.

(iii) The semantic sign partition $\operatorname{sign}(u^\top\widetilde{H})$ has two-means gain $f(\kappa)^2$; every two-cell partition whose mean difference is orthogonal to u has gain at most $2/\pi < f(\kappa)^2$; and the symmetric one-dimensional gain $\left(\mathbb{E}\left|\theta^\top\widetilde{H}\right|\right)^2 = f(c)^2$ is maximised uniquely at $\theta = \pm u$.

(iv) If $m \le \sigma_s$, the semantic sign partition is the unique optimal population two-means partition of $\widetilde{H}$. Its fine-label error equals the Bayes error $\Phi(-m/\sigma_s)$ for every nuisance covariance, so whitened two-means has no procedure gap.

The proof (Appendix A.4) reduces the two-means problem to one dimension (Lemma A.2), uses (ii) to show that every projection is a member of the single family M_c, and uses the monotonicity of f

(Lemma A.1). For $m \leq \sigma_s$ every M_c is log-concave, so its optimal two-level quantiser is unique and symmetric (Kieffer 1983), which gives global optimality. For $m > \sigma_s$ we prove the statement for partitions by hyperplanes through the mean and verify the global claim numerically: across doses from 0.2 to 12 and $m/\sigma_s \in \{1,2\}$, the error of the computed global optimum of the whitened cloud stays within 0.006 of the Bayes error (Fig. 2c).

Two limits of the theorem shape the experiments. First, whitening equalises *all* directions, so it helps only when the competing directions are Gaussian-like; a nuisance direction that is itself bimodal can win after whitening. Healthy representations, whose nuisance is not Gaussian and whose semantic directions are already leading, should therefore lose from whitening, and they do (Section 7.5). Second, whitening needs the semantic direction inside the analysed subspace; Lemma 1 below says when principal-component truncation keeps it.

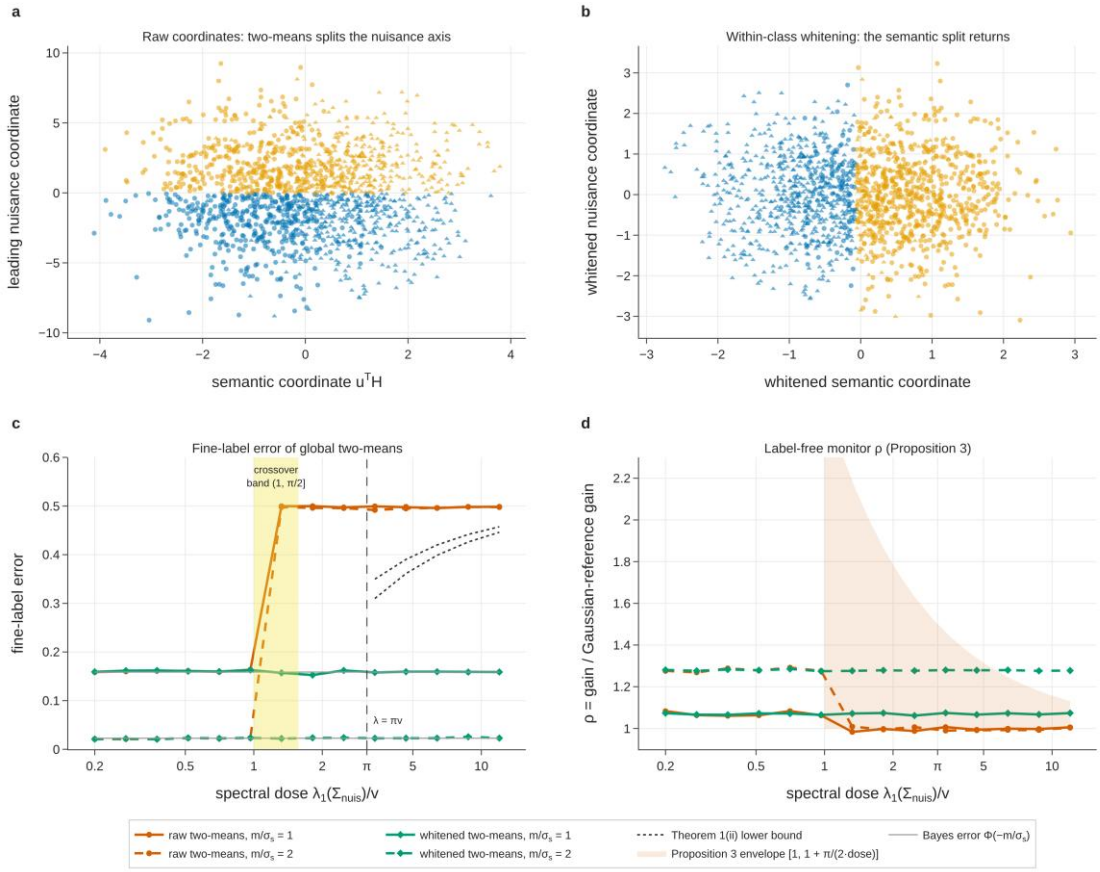


Fig. 2. The mechanism and its repair in the model of Section 4.2. (a) A cloud at dose 4 with $m = \sigma_s = 1$: colours are the two clusters of two-means, shapes the fine label. The split follows the nuisance axis, and its fine-label error is 0.50. (b) The same cloud after within-class whitening: the split follows the semantic axis and its error, 0.16, equals the Bayes error. (c) Fine-label error of the computed global two-means optimum (30,000 points, 20 restarts) against the dose, for $m/\sigma_s = 1$ (solid) and 2 (dashed). Raw coordinates switch inside the crossover band $(1, \pi/2]$ of Corollary 1; the dotted curves are the guaranteed lower bounds of Theorem 1(ii), valid beyond $\lambda = \pi v$. Whitened two-means stays at the Bayes error at every dose (Theorem 2). (d) The label-free monitor ρ in raw coordinates falls to 1 once the nuisance captures the split and stays inside the envelope of Proposition 3; in whitened coordinates it stays at its theoretical value $(\pi/2)f(\kappa)^2$.

4.4 A label-free monitor

The dose requires fine labels. What can be seen without them? For a distribution with covariance Σ_X, the two-means gain of the Gaussian $\mathcal{N}(0, \Sigma_X)$ is $2\lambda_1(\Sigma_X)/\pi$ (Proposition 4(ii)). Define the monitor

$$\rho(X) = \frac{G_2(X)}{2\lambda_1(\Sigma_X)/\pi},$$

the two-means gain G_2 (the reduction in distortion from one centre to two) relative to that Gaussian reference. A Gaussian class has $\rho = 1$; a well-separated mixture has ρ well above 1.

Proposition 3 (The monitor bounds the dose). *In the model of Section 4.2 with $\lambda \geq v$,*

$$1 \;\leq\; \rho(H) \;\leq\; 1 + \frac{\pi}{2\,\mathrm{dose}}, \qquad \mathrm{dose} = \lambda/v,$$

whereas after whitening $\rho(\tilde{H}) \geq (\pi/2)\, f(\kappa)^2 > 1$ independently of λ. In particular $\rho(H) \to 1$ when the dose grows, and an observed value $\rho(H) = 1 + \epsilon$ with $\epsilon > 0$ certifies dose $\leq \pi/(2\epsilon)$.

The upper bound uses only Cauchy–Schwarz for the semantic part and the optimality of the median split for the Gaussian part (Appendix A.5). Proposition 3 concerns two-means. On the 135 re-analysed pools, the change of the two-means monitor ρ_2, computed exactly with the Gaussian reference $2\lambda_1/\pi$, tracks the change of ARI with Spearman 0.42, and it is lower on cross-entropy features than on their bases (1.13 against 1.17). In practice we compute the monitor at the working order K rather than at two, which tracks the same changes more closely (0.71 on the same pools): $\rho_K(X) = G_K(X)/\overline{G_K(X^0)}$, where X^0 are $B = 3$ Gaussian surrogates with the empirical covariance of X. The proposition explains the observation that drives Algorithm 2: when fine-tuning pushes a class towards capture, its raw monitor falls towards 1, and the fall is visible without fine labels.

Algorithm 2 (Label-free spectral repair). Input: features of the base encoder and of the fine-tuned encoder, coarse labels, a working order K per coarse class (the taxonomy count or the output of Algorithm 1), principal-component dimension $r = 20$ and threshold $\delta = 0.10$. For each coarse class: (1) standardise and project the fine-tuned features to r principal components (X), and form the whitened coordinates $W = X\Lambda^{-1/2}$; (2) compute the monitor $\rho_K(X)$, and in the same way $\rho_K(X^{(0)})$ for the base encoder; (3) if $\rho_K(X^{(0)}) - \rho_K(X) > \delta$, cluster W, otherwise cluster X; (4) report the partition together with both monitor values. No step uses fine labels. The rule needs the base encoder, which is available whenever a model is fine-tuned, and its cost is that of $(B + 1)$ k-means fits per class.

5 Recoverable order and the order diagnostic

Accessibility at the taxonomy order is a diagnostic; a practitioner also needs to know how many groups the geometry supports. Observation 1 says this has an answer only at a stated resolution. For a coarse class with feature law P_y, let $D_y(k)$ be the optimal population k-point distortion and $G_y(k) =$

$D_y(k-1) - D_y(k)$ its gains. At resolution $\lambda_0 > 0$, the *recoverable order* is the largest k with $G_y(k) > \lambda_0$ (one if none), and the conservative *prefix order* stops at the first gain below the floor.

Theorem 3 (Stability). *If the empirical distortions are uniformly within ϵ of $D_y(k)$ for $k \le K_{\max}$, and the order K has margin $\gamma = \min\{G_y(K) - \lambda_0,\ \lambda_0 - \sup_{k>K} G_y(k)\} > 2\epsilon$, the empirical rule returns K.*

Theorem 4 (Identification under separated components). *Let $P_y = \sum_{z=1}^{K} w_z P_z$ with $w_z \ge w_{\min}$, component means μ_z at pairwise distance at least Δ, and largest within-component mean squared radius s^2. If $\Delta > 2s$ and $w_{\min}(\Delta/2 - s)^2 > 2s^2$, every resolution $s^2 < \lambda_0 < w_{\min}(\Delta/2 - s)^2 - s^2$ yields recoverable order K.*

Theorem 4 needs only second moments; by Wasserstein continuity of optimal distortion (Biau et al 2008) and the strong law, the empirical order is almost surely consistent under a positive margin, with a dimension-free $n^{-1/2}$ rate for bounded features (Appendix A.6).

Algorithm 1 (Cross-fitted distortion-gain diagnostic). For each coarse class: repeat $S = 6$ discovery/validation splits; fit standardisation and PCA to $r = 10$ dimensions on the discovery half only; fit k-means for $k = 1, \dots, K_{\max}$ on the discovery half and evaluate distortion on the validation half; average the held-out gains; calibrate a k-specific floor as the $q = 0.90$ quantile of the gains of $B_0 = 20$ *radial-null* clouds, built by recolouring each centred observation as $\mu + \Sigma^{1/2}(r_i u_i)$ with its Mahalanobis radius r_i and a uniform direction u_i, and passed through the identical pipeline; report the prefix order and the full margin curve.

Proposition 4 (Scale of the null floor). *For $X^0 = \mu + \Sigma^{1/2} RU$ with U uniform on the sphere, $R \perp U$ and $\mathbb{E}R^2 = d$: (i) $\lambda_1(\mathbb{E}R)^2(\mathbb{E}|U_1|)^2 \le G^0(2) \le \lambda_1$; (ii) if X^0 is Gaussian, $G^0(2) = 2\lambda_1/\pi$; (iii) a two-component mixture with weights $w, 1-w$, mean difference Δ and average within-component covariance Σ_w exceeds the Gaussian reference by at least $(1 - 2/\pi)w(1-w)\Delta^2 - (2/\pi)\lambda_1(\Sigma_w)$.*

Lemma 1 (PCA keeps the semantic direction under an eigengap). *If $\Sigma = vuu^\top + \Sigma_n$ with nuisance eigenvalues $\lambda_1 \ge \lambda_2 \ge \cdots$ on $u^\perp$, every top-r principal subspace contains u iff $v > \lambda_r$, and an estimate $\hat{\Sigma}$ moves u out of the estimated subspace by at most $2\sqrt{r}\ \|\hat{\Sigma} - \Sigma\|_{\mathrm{op}}/g_r$, where $g_r = \min\{v, \lambda_{r-1}\} - \lambda_r$.*

Lemma 1 and Theorem 1 rest on different comparisons, and the difference is the crux of the paper: truncation discards the semantic direction only when r nuisance eigenvalues exceed v, whereas two-means is captured as soon as one does. Keeping $r = 10$ or 20 components therefore protects the semantic direction from truncation but not from the clustering objective, which is exactly the regime where whitening inside the retained subspace helps.

Theorem 5 (End-to-end consistency of Algorithm 1). *Assume bounded features with per-coordinate variances bounded away from zero (A1), the eigengap of Lemma 1 with a vanishing ridge in the null construction (A2), a margin* $\gamma > 0$ *of the population gains over the population null floor (A3), and clustering within* $\varepsilon_{\mathrm{opt}} < \gamma/8$ *of the empirical optimum (A4). Then, with probability at least* $1 - \delta$*, all estimated gains and floors are within*

$$C\left(\sqrt{d/n} + \sqrt{\log(1/\delta)/n} + \sqrt{\log(1/\delta)/B_0}\right) + 2\varepsilon_{\mathrm{opt}}$$

of their population targets, and Algorithm 1 returns the target prefix order whenever this quantity is below $\gamma/2$.

The proof (Appendix A.7) handles in turn the estimated analysis map, held-out evaluation with near-optimal centres, the estimated null floor and the threshold crossing. Assumption (A4) is an oracle condition, certifiable from the achieved distortion and provable under separation (Arthur and Vassilvitskii 2007; Kumar and Kannan 2010; Lu and Zhou 2016). Because the floor rewards gains above a no-preferred-direction reference, a class captured by Gaussian nuisance (dose large) has gains close to the floor, which is the same fact that makes the monitor of Proposition 3 informative. We report the order as an *interval* between Algorithm 1 and a variant (Algorithm 1-S) that additionally requires a density valley between neighbouring clusters (Kalogeratos and Likas 2012): the lower end counts valley-separated groups, the upper end reproducible distortion structure.

6 Experimental design

Settings. Table 2 lists the seven settings. Settings S1–S3 and S5 come from our shared-base controls, whose frozen features we re-analysed; S2, S4, S6 and S7 were trained for this paper in one pre-specified run. In every setting a single base encoder is fine-tuned with coarse cross-entropy only, with several seeds, and the base is evaluated with the same protocol. On CIFAR-100 (Krizhevsky 2009) the coarse classes are the 20 superclasses of five fine classes each; on Fashion-MNIST (Xiao et al 2017) four groups (tops $K = 4$, footwear $K = 3$, trouser and dress $K = 2$, bag $K = 1$); on CIFAR-10 two groups, vehicles ($K = 4$) and animals ($K = 6$). The ResNet-18 bases were pretrained without labels with SimCLR (Chen et al 2020) on the same images; the DINOv2 ViT-S/14 (Oquab et al 2024) and ResNet-50 (He et al 2016) bases are public ImageNet-scale models.

Table 2. Settings. "SSL" denotes in-domain SimCLR pretraining on the training images; S2 retrained this base from scratch with a different seed and run. Snapshots are the CE epochs at which features were frozen.

	Base encoder	Data and hierarchy	CE fine-tuning	Seeds	Snapshots
S1	ResNet-18, SSL 120 ep.	CIFAR-100, 20 × 5	30 ep., SGD	3	0, 30
S2	ResNet-18, SSL 120 ep. (independent rerun)	CIFAR-100, 20 × 5	30 ep., SGD	3	0, 2, 5, 10, 20, 30
S3	ResNet-18, SSL 80 ep.	Fashion-MNIST, 4 groups	20 ep., SGD	3	0, 20
S4	ResNet-18, SSL 120 ep. (from S2)	CIFAR-10, vehicles 4 / animals 6	30 ep., SGD	3	0, 5, 10, 30

	Base encoder	Data and hierarchy	CE fine-tuning	Seeds	Snapshots
S5	DINOv2 ViT-S/14, 112 px	CIFAR-100, 20 × 5	5 ep., AdamW	3 (1 re-analysed)	0, 5
S6	DINOv2 ViT-S/14, 112 px	CIFAR-100, 20 × 5	30 ep., AdamW	2	0, 5, 10, 20, 30
S7	ResNet-50, supervised ImageNet, 96 px	CIFAR-100, 20 × 5	15 ep., SGD	2	0, 5, 15

Protocol. The ResNet-18 fine-tuning recipe is that of the controls: SGD with Nesterov momentum 0.9, learning rate 0.1 with cosine decay, weight decay 5×10^{-4}, batch 512, random crop and flip, mixed precision. ViT used AdamW (learning rate 2×10^{-5}, weight decay 0.05, batch 128) and ResNet-50 SGD (0.01, 10^{-4}, batch 128). At each snapshot all training and test images were encoded, and for every coarse class we computed on test features: the probe (trained on training features), GMM ARI (PCA-10), k-means ARI (PCA-10 and PCA-20), whitened k-means ARI (PCA-20 whitened), dose, alignment, NC1 $= \mathrm{tr}(\Sigma_W \Sigma_B^{+})/K$ and the trace ratio $\mathrm{tr}\Sigma_W/\mathrm{tr}\Sigma_B$ at the fine level, and the monitors ρ_K in raw and whitened coordinates. Fine labels enter only the probe, ARI, dose, alignment, NC1 and trace ratio. The entire new run took 4.98 hours on one T4 GPU; the analysis plan, including the dose–response snapshots, was fixed before it started. Code, results and features are archived at Zenodo (Oban 2026). Unless stated otherwise, class-level statistics are Spearman correlations over class snapshots, and change scores compare a fine-tuned snapshot with the base of the same setting.

7 Results

7.1 The dissociation on shared bases

Table 3 and Fig. 3 reproduce the shared-base controls. In both ResNet settings cross-entropy leaves the probe where it was and cuts oracle-order ARI by 59% (CIFAR-100) and 63% (Fashion-MNIST). Every PSL and PSL-C seed exceeds every CE seed in ARI on CIFAR-100 (Welch $t = 39.5$ and 105.6). The DINOv2 setting behaves differently: all objectives raise both probe and ARI, and no pairwise ARI difference is significant ($p \geq 0.19$).

The independent rerun S2 retrained the SSL base from scratch. Its base has a lower probe (0.722) but the same ARI (0.225 against 0.224), and after 30 epochs of cross-entropy its ARI is 0.093 ± 0.008, equal to the original 0.093 ± 0.002 to the third decimal, while its probe moves from 0.722 to 0.709.

Table 3. Shared-base controls (mean ± s.d. over fine-tuning seeds). Probe: within-class multinomial logistic regression; ARI: full-covariance GMM at the taxonomy order after per-class standardisation and PCA-10. PSL and PSL-C are the contrastive objectives of Appendix A.8.

Setting	Representation	Coarse acc.	Probe	Oracle-order ARI
CIFAR-100, ResNet-18 (S1)	SSL base	0.690	0.763	0.224
	CE, 30 ep.	0.840 ± 0.003	0.760 ± 0.003	**0.093 ± 0.002**
	PSL, 30 ep.	0.815 ± 0.002	0.790 ± 0.002	0.238 ± 0.006
	PSL-C, 30 ep.	0.810 ± 0.001	0.786 ± 0.002	0.236 ± 0.001
CIFAR-100, ResNet-18 (S2, rerun)	SSL base	–	0.722	0.225

Setting	Representation	Coarse acc.	Probe	Oracle-order ARI
	CE, 30 ep.	0.840 ± 0.001	0.709 ± 0.007	**0.093 ± 0.008**
Fashion-MNIST, ResNet-18 (S3)	SSL base	0.978	0.927	0.484
	CE, 20 ep.	0.990 ± 0.001	0.919 ± 0.003	**0.178 ± 0.009**
	PSL, 20 ep.	0.987 ± 0.002	0.928 ± 0.004	0.481 ± 0.019
	PSL-C, 20 ep.	0.985 ± 0.000	0.922 ± 0.001	0.532 ± 0.040
CIFAR-100, DINOv2 ViT-S/14 (S5)	base	0.846	0.869	0.448
	CE, 5 ep.	0.938 ± 0.001	0.896 ± 0.002	0.546 ± 0.028
	PSL, 5 ep.	0.937 ± 0.001	0.902 ± 0.001	0.515 ± 0.006
	PSL-C, 5 ep.	0.938 ± 0.000	0.900 ± 0.003	0.524 ± 0.018

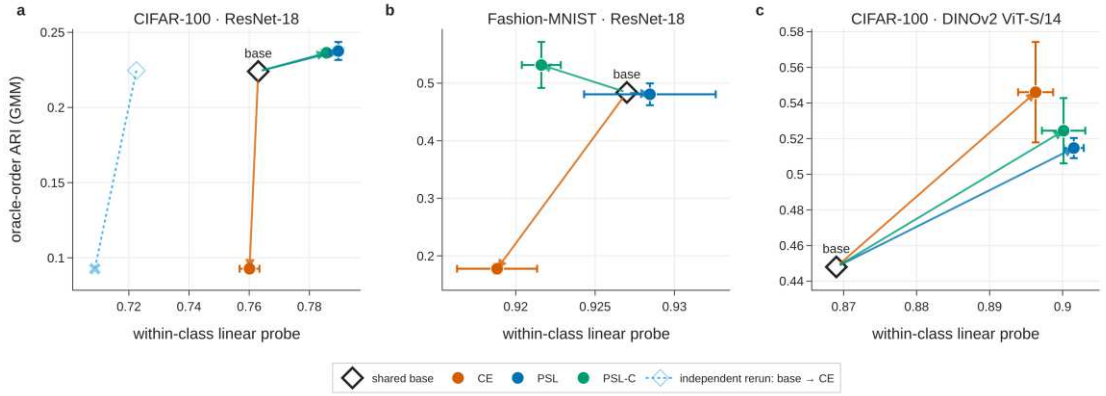


Fig. 3. Decodability against clusterability on shared bases. Open diamonds mark the base encoder, arrows lead to each objective's mean (three seeds; bars are one standard deviation). (a) CIFAR-100 with a ResNet-18 SSL base: cross-entropy keeps the probe and loses 59% of ARI; the dotted line is the independent rerun S2, which lands on the same ARI. (b) Fashion-MNIST: the same pattern, with a 63% loss. (c) DINOv2 ViT-S/14: every objective improves both quantities.

7.2 Dose–response: the mechanism epoch by epoch

The dose–response protocol fixed in advance asked whether the damage tracks the spectral mechanism during training. Table 4 and Fig. 4 give the answer for S2. The probe stays between 0.709 and 0.732 at every snapshot. GMM ARI drops in two phases, from 0.225 to 0.165 within two epochs and from 0.164 to 0.093 between epochs 10 and 30. Over the same epochs the median dose rises from 0.675 to 0.998, into the crossover band of Corollary 1, and the alignment of the two-means split with the semantic subspace falls from 0.85 to 0.53. The label-free monitor falls from 1.30 to 1.12. At epoch 30 the ARI has fallen in 19, 18 and 20 of the 20 superclasses in the three seeds, and the dose has risen in 17, 16 and 17.

Two details are informative. The early drop at epoch 2 coincides with a jump of NC1 from 1.2 to 2.4 but not with a dose increase (0.649), so the first damage is a change in the amount of within-class spread, not its orientation. The later drop coincides with the dose increase and the rotation of the two-means split, which is the spectral mechanism. Whitening behaves accordingly: it is worse than raw k-means at the base (0.112 against 0.198), breaks even at epoch 10, and from epoch 20 on recovers about half of the loss (0.149 against 0.093 at epoch 30; 48%, 50% and 59% of the lost ARI in the three seeds).

Table 4. Dose–response in setting S2 (CIFAR-100, ResNet-18 SSL base, three CE seeds; mean ± s.d. over seeds of the per-class means; dose and NC1 are per-class medians). Coarse accuracy at epoch 0 is that of an untrained head and is omitted.

CE epoch	0	2	5	10	20	30
Coarse accuracy	–	0.610 ± 0.004	0.675 ± 0.023	0.748 ± 0.009	0.819 ± 0.004	0.840 ± 0.001
Within-class probe	0.722	0.724 ± 0.003	0.725 ± 0.005	0.732 ± 0.002	0.718 ± 0.006	0.709 ± 0.007
ARI, GMM (PCA-10)	0.225	0.165 ± 0.010	0.170 ± 0.013	0.164 ± 0.007	0.117 ± 0.009	0.093 ± 0.008
ARI, k-means (PCA-20)	0.198	0.155 ± 0.004	0.167 ± 0.019	0.169 ± 0.001	0.114 ± 0.010	0.093 ± 0.019
ARI, whitened k-means	0.112	0.136 ± 0.005	0.145 ± 0.001	0.172 ± 0.008	0.143 ± 0.011	0.149 ± 0.003
Spectral dose λ/v	0.675	0.649 ± 0.014	0.731 ± 0.076	0.750 ± 0.073	0.918 ± 0.094	0.998 ± 0.032
Alignment $\| U^\top \theta \|^2$	0.852	0.791 ± 0.026	0.715 ± 0.066	0.717 ± 0.075	0.560 ± 0.029	0.527 ± 0.020
NC1	1.208	2.370 ± 0.075	2.286 ± 0.185	2.227 ± 0.149	2.581 ± 0.114	2.828 ± 0.069
Monitor ρ_K (raw)	1.296	1.220 ± 0.019	1.219 ± 0.008	1.201 ± 0.012	1.140 ± 0.009	1.122 ± 0.005

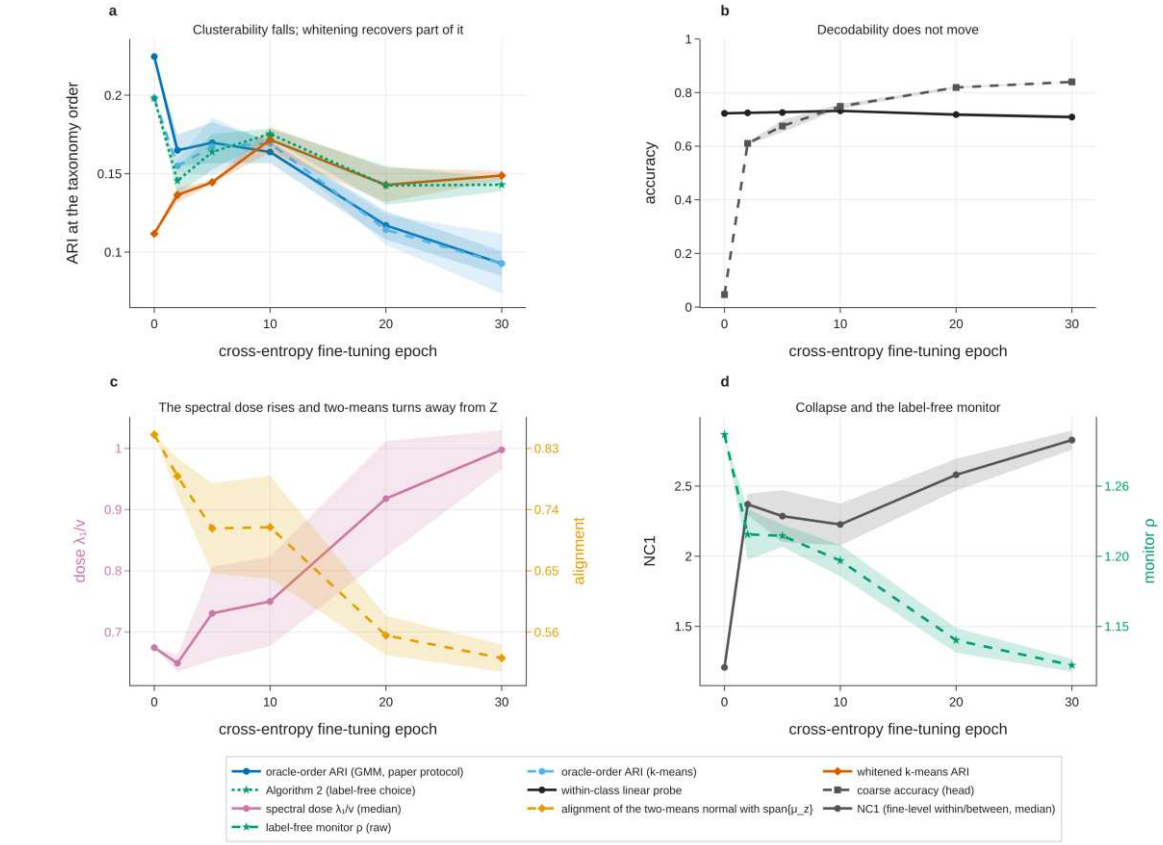


Fig. 4. Dose–response in setting S2 (three seeds; bands are one standard deviation over seeds). (a) Clusterability: GMM and k-means ARI fall in two phases; whitened k-means overtakes raw k-means after epoch 10; Algorithm 2 follows the better of the two without fine labels. (b) Decodability does not move while coarse accuracy rises from chance to 0.84. (c) The spectral dose rises into the crossover band and the two-means split rotates out of the semantic subspace. (d) NC1 jumps in the first two epochs and keeps rising; the label-free monitor falls steadily.

7.3 A regime map across seven settings

Table 5 and Fig. 5 place all settings on one map. The three in-domain SSL settings with ResNet-18 on CIFAR-100 and Fashion-MNIST lie in the quadrant the theory predicts for capture: the dose rises (mean change in log dose 0.33 to 0.49), alignment falls by 0.29 to 0.52, and ARI falls in 78–100% of the classes. The three settings with broadly pretrained bases lie in the opposite quadrant: cross-entropy lowers the dose (−0.18 to −0.35 in log), raises alignment, and raises ARI by 0.12 to 0.22, with the

probe rising by about 0.05. The ViT result of the controls therefore does not depend on short fine-tuning: after 30 epochs the dose is still lower than at the base (0.41 against 0.58) and ARI is higher (0.588 against 0.365 for k-means).

Setting S4 is the exception. The dose rises strongly (from 0.70 to 1.38) and alignment collapses (0.89 to 0.34); GMM ARI falls (0.220 to 0.170), as predicted, but k-means ARI rises (0.164 to 0.240). With only two coarse classes and six class snapshots this setting carries little weight, and Section 8 discusses a plausible reason. Pooling every class snapshot of every setting, the change in log dose and the change in ARI have Spearman correlation −0.58, and their signs agree in 75% of the 580 snapshots in which the dose moved by more than 5% (95% cluster-bootstrap interval 68–82%, resampling the 105 setting–class units, so that repeated snapshots of one class are not counted as independent).

Table 5. Regime map. Class-level changes from the base to the last snapshot (mean over classes and seeds). n: class snapshots; Δ probe is unavailable for re-analysed settings. "Whitening gain" is whitened minus raw k-means ARI at the last snapshot.

Setting	n	Δ log dose	Δ alignment	Δ ARI (k-means)	Δ ARI (GMM)	Δ probe	Whitening gain
S1 ResNet-18 SSL, CIFAR-100	60	+0.37	−0.29	−0.107	−0.128	–	**+0.055**
S2 same, independent rerun	60	+0.33	−0.33	−0.105	−0.132	−0.014	**+0.056**
S3 ResNet-18 SSL, Fashion-MNIST	9	+0.49	−0.52	−0.433	–	–	**+0.440**
S4 ResNet-18 SSL, CIFAR-10 hierarchy	6	+0.72	−0.55	+0.076	−0.049	+0.008	+0.006
S5 DINOv2 ViT-S/14, 5 ep.	20	−0.35	+0.07	+0.116	–	–	−0.082
S6 DINOv2 ViT-S/14, 30 ep.	40	−0.30	+0.08	+0.224	+0.161	+0.050	−0.032
S7 ResNet-50 ImageNet, 15 ep.	40	−0.18	+0.18	+0.180	+0.143	+0.057	−0.090

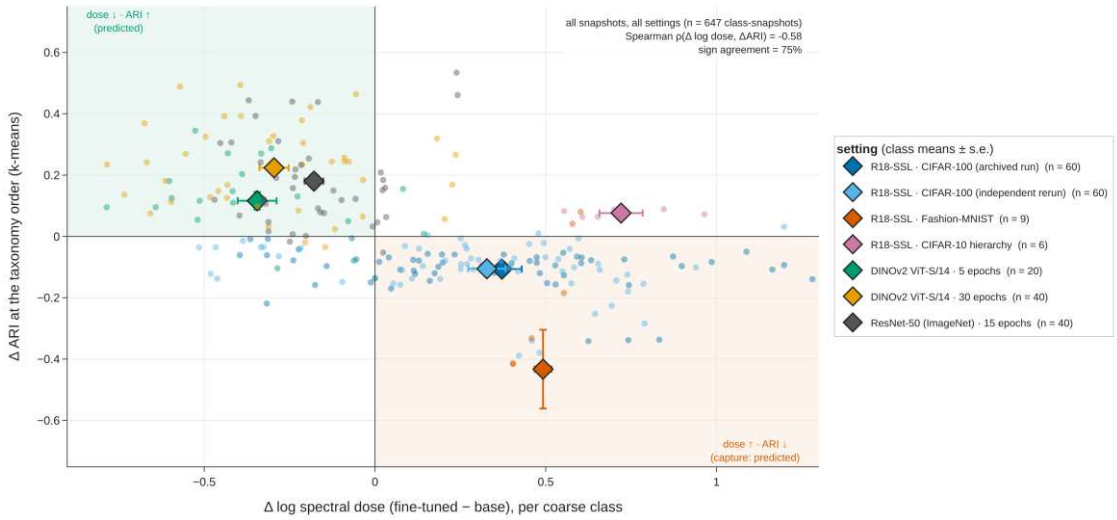


Fig. 5. Regime map. Each small point is one coarse class at the last snapshot of one seed; diamonds are setting means with standard errors. Cross-entropy from in-domain SSL bases moves classes into the lower-right quadrant (dose up, clusterability down), as Theorem 1 predicts; from broadly pretrained bases it moves them into the upper-left quadrant. The annotation pools all snapshots of all settings.

7.4 What tracks clusterability?

Table 6 and Fig. 6 compare the candidate explanations over all 647 class snapshots. NC1 is the strongest single predictor of the ARI change (Spearman −0.82), followed by the label-free monitor (+0.77), the alignment (+0.65) and the dose (−0.58). Given NC1 the dose retains a partial correlation of only −0.15, whereas NC1 given the dose retains −0.76. We therefore do not claim that the dose is the best scalar summary.

Two facts keep the scalar result from absorbing the paper's thesis. First, NC1 measures fine-level within-to-between spread, so it is a measure of how separable the fine classes are, and it is no surprise that it tracks a separability score such as ARI. What it cannot do is explain the dissociation: in the dose–response run NC1 more than doubles (1.21 to 2.83) while the probe does not move (Table 4). A scalar that rises when ARI falls and stays silent when the probe does not change describes the clustering failure but not why the information remains linearly readable. Second, NC1 does not say what to do. The spectral account predicts an intervention, and the intervention discriminates: whitening gains 0.039 ARI on average in snapshots whose dose rose by more than 10% and loses 0.041 in snapshots whose dose fell by more than 10%.

Table 6. Spearman correlation of class-level changes with the change in k-means and GMM ARI over all class snapshots ($n = 647$; GMM correlations exclude snapshots without GMM ARI). Brackets: 95% cluster-bootstrap intervals that resample the 105 setting–class units (2000 replicates). Only the monitor is computed without fine labels.

Predictor	Uses fine labels	ρ with Δ ARI (k-means)	ρ with Δ ARI (GMM)
Δ log spectral dose	yes	−0.58 [−0.68, −0.45]	−0.59
Δ alignment of the two-means split	yes	+0.65 [0.56, 0.73]	+0.67
Δ log NC1	yes	−0.82 [−0.88, −0.74]	−0.82
Δ monitor ρ_K	**no**	**+0.77 [0.68, 0.83]**	**+0.82**
Δ log dose, partial on Δ log NC1	yes	−0.15	–
Δ log NC1, partial on Δ log dose	yes	−0.76	–

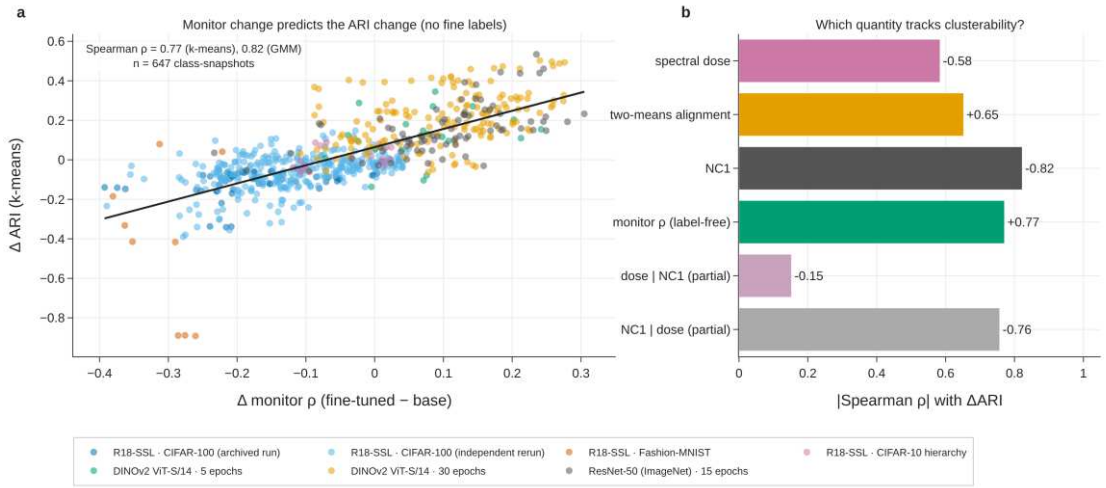


Fig. 6. The label-free monitor. (a) Change of the raw monitor ρ_K against the change of k-means ARI for every class snapshot of every setting; the line is a least-squares fit. (b) Absolute Spearman correlation of each candidate quantity with the ARI change (signed values printed); the last two bars are partial correlations.

7.5 Spectral repair without fine labels

Fig. 7 and Table 7 report the intervention for every representation. Whitening helps exactly where the dose rose: on CIFAR-100 CE features it raises k-means ARI from 0.090 to 0.145 (S1) and from 0.093 to 0.149 (S2), and on Fashion-MNIST CE features from 0.093 to 0.533, which is the level of the SSL base (0.526). On every healthy representation it lowers ARI, by 0.02 to 0.11: the SSL bases, PSL, PSL-C, the DINOv2 and ResNet-50 encoders before and after fine-tuning. This double dissociation is the interventional evidence for the spectral mechanism that correlations alone cannot provide.

Algorithm 2 converts the dissociation into a rule. Over 386 coarse-class pools, raw k-means averages 0.278 and always whitening 0.266; Algorithm 2 averages 0.303, close to the 0.316 of an oracle that picks the better option per pool with fine labels. On the damaged CE representations of S1–S3 it reaches 0.135, 0.143 and 0.533. Its advantage is $+0.025$ ARI over raw k-means (95% bootstrap interval 0.016–0.035 over pools) and $+0.037$ over always whitening (0.027–0.046). The rule is insensitive to its threshold: $\delta \in \{0.05, 0.10, 0.15\}$ gives 0.303, 0.303 and 0.302. Because δ could have been tuned on the same pools, we also chose it by leave-one-setting-out: for each held-out setting, δ was picked from a grid of 0.02 to 0.30 on the other six. Every fold chose 0.08, and the pooled held-out ARI is 0.304. The repair is specific to Euclidean objectives, as Theorem 2 states. A full-covariance Gaussian mixture is invariant to affine maps in population, so whitening cannot change its optimum, and fitting it in whitened coordinates (same model, different initialisation) leaves the damaged ARI essentially where it was: 0.099 against 0.092 on CIFAR-100 CE features and 0.185 against 0.174 on Fashion-MNIST CE features, where whitened k-means reaches 0.533 on the same coordinates. A more flexible clustering model is therefore not a substitute for the repair; its likelihood can be raised by components shaped along non-Gaussian nuisance, which is consistent with the NC1 increase it tracks. Its one costly decision is Fashion-MNIST PSL, where the monitor of one of the three pools dropped by more than δ and whitening lowered the mean ARI from 0.521 to 0.472.

Table 7. Within-class whitening and Algorithm 2 (mean k-means ARI at the taxonomy order, PCA-20). "Monitor drop" is $\rho_K(\text{base}) - \rho_K(\text{representation})$; Algorithm 2 whitens when it exceeds $\delta = 0.10$. The oracle chooses per pool with fine labels.

Setting	Representation	Pools	Raw	Whitened	Algorithm 2	Oracle	Monitor drop
S1 CIFAR-100 ResNet-18 SSL	base	20	0.196	0.123	0.196	0.196	0
	CE	60	0.090	0.145	**0.135**	0.148	+0.16
	PSL	20	0.239	0.212	0.239	0.246	+0.02
	PSL-C	20	0.247	0.222	0.250	0.261	+0.04
S2 same, rerun	base	20	0.198	0.112	0.198	0.199	0
	CE	60	0.093	0.149	**0.143**	0.151	+0.17
S3 Fashion-MNIST ResNet-18 SSL	base	3	0.526	0.415	0.526	0.526	0
	CE	9	0.093	0.533	**0.533**	0.536	+0.30
	PSL	3	0.521	0.426	0.472	0.521	+0.09
	PSL-C	3	0.517	0.423	0.517	0.593	+0.05
S4 CIFAR-10 ResNet-18 SSL	base	2	0.164	0.147	0.164	0.164	0

Setting	Representation	Pools	Raw	Whitened	Algorithm 2	Oracle	Monitor drop
	CE	6	0.240	0.245	0.240	0.255	+0.03
S5 DINOv2, 5 ep.	base	20	0.449	0.385	0.449	0.468	0
	CE	20	0.565	0.484	0.565	0.576	−0.07
S6 DINOv2, 30 ep.	base	20	0.365	0.296	0.365	0.371	0
	CE	40	0.588	0.556	0.588	0.622	−0.11
S7 ResNet-50 ImageNet	base	20	0.222	0.175	0.222	0.233	0
	CE	40	0.402	0.312	0.401	0.415	−0.12
All pools		386	0.278	0.266	**0.303**	0.316	

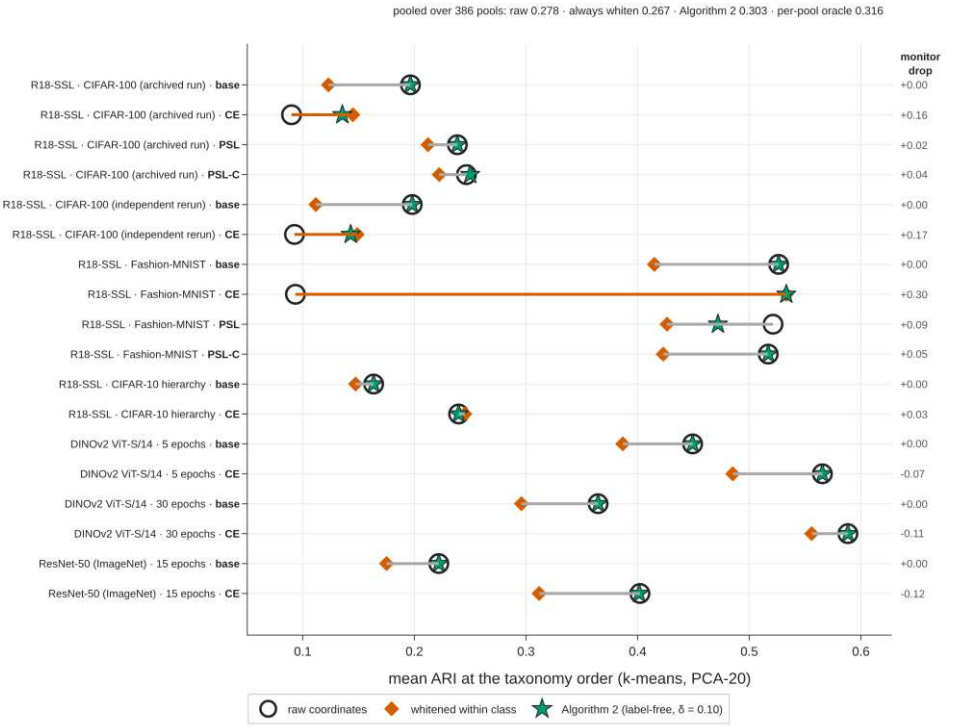


Fig. 7. The interventional test. For every setting and representation, black rings mark raw k-means ARI, diamonds whitened k-means ARI and stars the choice of Algorithm 2; orange connectors mark representations where whitening helps. Whitening repairs exactly the cross-entropy representations of the in-domain SSL settings and hurts every healthy representation; the monitor drop printed on the right is what Algorithm 2 sees without fine labels.

7.6 Order selection on identical frozen features

Because frozen features were kept, order selectors can be compared on identical inputs (Fig. 8). On a theory-controlled test with planted orders, Algorithm 1 returned order one for every cloud whose fine label was a linear code of a single Gaussian (decodable, not clusterable; probe 0.953) and recovered the planted order in 90% of 40 geometry-aligned clouds. On CIFAR-100 ResNet features, Algorithm 1 and BIC both register the drop caused by cross-entropy (ARI 0.05 against 0.14 for the contrastive objectives); prediction strength is too conservative to separate the objectives. On the ViT features Algorithm 1 returns the best-aligned partitions of all selectors (ARI 0.46–0.49). On Fashion-MNIST the contrastive features carry many small stable modes: Algorithm 1 and BIC saturate near $K_{\max} = 8$, while the valley-separated Algorithm 1-S selects 3.2 and 3.7 groups with ARI 0.43 and 0.45 and returns one group on CE features. The order interval of Algorithm 1-S and Algorithm 1 brackets the taxonomy count in seven of nine setting–objective cells; the two exceptions are the CE cells whose

geometry Sections 7.1–7.5 show to be captured. On Fashion-MNIST, UMAP (McInnes et al 2018) followed by silhouette selection reaches ARI 0.56 even on CE features, where a Euclidean mixture reaches 0.18: the fine information that Euclidean geometry hides is still reachable by a neighbourhood graph, an independent confirmation that the damage concerns access, not information.

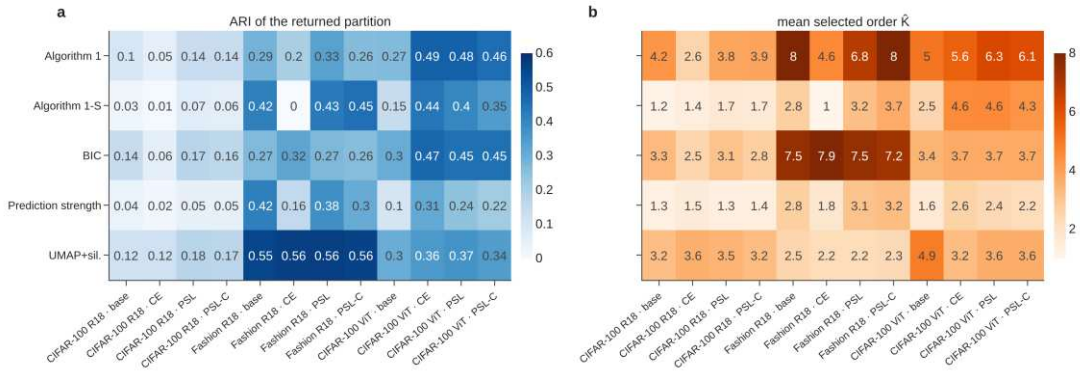


Fig. 8. Order selection on identical frozen features (null quantile $q = 0.90$; three seeds). (a) ARI of the partition returned by each selector at its selected order. (b) Mean selected order. Algorithm 1-S counts valley-separated groups and Algorithm 1 reproducible distortion structure; together they form the reported order interval.

7.7 Fully label-free evaluation and other clustering families

Sections 7.5 and 7.6 evaluated repair and order separately. Here they are combined, and Algorithm 2 is compared with clustering families that could plausibly avoid the damage on their own. All 190 coarse-class pools of the archived features (S1, S3, S5; 18 encoders) were reprocessed with a single script that is part of the reproducibility package (`code/endtoend.py`). At the taxonomy order it reproduces Table 7 to the third decimal (CIFAR-100 CE: raw 0.090, Algorithm 2 0.135; Fashion-MNIST CE: 0.093 and 0.533). The comparison methods are spectral clustering on a 10-nearest-neighbour graph of the PCA-20 coordinates (Ng et al 2002), with QR-based assignment (Damle et al 2019); adaptive Mahalanobis k-means, which re-whitens by the pooled within-cluster covariance of the current partition until the partition stops changing (Melnykov and Melnykov 2014); and a Gaussian mixture with a tied covariance. Each method was run once with its default settings and none was tuned. Table 8 gives the results at the taxonomy order, and Table 9 the fully label-free results.

Table 8. Clustering families at the taxonomy order on the archived features (mean ARI over pools). Mahal.: adaptive Mahalanobis k-means; Tied GMM: Gaussian mixture with a tied covariance. "Healthy" pools are the base, PSL and PSL-C representations. Bold marks the best entry of each row.

Setting	Pools	Raw k-means	Alg. 2	Spectral	Mahal.	Tied GMM
S1 CIFAR-100 CE	60	0.090	0.135	**0.137**	0.108	0.072
S1 CIFAR-100 healthy	60	0.227	**0.228**	0.225	**0.228**	0.176
S3 Fashion-MNIST CE	9	0.093	0.533	**0.581**	0.102	0.099
S3 Fashion-MNIST healthy	21	0.534	0.524	0.511	**0.537**	0.529
S5 DINOv2 CE	20	**0.565**	**0.565**	0.555	0.563	0.518
S5 DINOv2 base	20	**0.449**	**0.449**	0.418	0.446	0.361
All pools	190	0.270	**0.305**	0.301	0.276	0.234

Other clustering families. The two families that model covariance do not repair the damage. On CIFAR-100 CE features adaptive Mahalanobis k-means reaches 0.108 and the tied mixture 0.072; on Fashion-MNIST CE features they reach 0.102 and 0.099, where Algorithm 2 reaches 0.533. Theorem 1 explains why. Both methods estimate the metric from the *within-cluster* covariance of the current partition. When that partition is the captured nuisance split, the nuisance direction is between-cluster variance, so the estimated metric does not shrink it and the captured partition is a fixed point. Theorem 2 whitens by the covariance of the whole coarse class, which needs no partition, and this is why it escapes. Spectral clustering does escape. At the taxonomy order it matches Algorithm 2 over all pools (difference −0.004, 95% bootstrap interval −0.014 to 0.006) and exceeds it on Fashion-MNIST CE features (0.581). It is slightly worse on the healthy pools (−0.011, −0.024 to 0.003), most clearly on the DINOv2 base (0.418 against 0.449). A neighbourhood graph depends on local distances rather than on the global ranking of variance directions that drives Theorem 1. This agrees with the UMAP result of Section 7.6: the damage blocks variance-based access to the fine labels, not neighbourhood-based access. The two repairs are interchangeable once the monitor has fired. Using spectral clustering instead of whitening in step (3) of Algorithm 2 gives the same pooled ARI (0.305). The label-free part of the procedure is therefore the monitor, which fires on 80% of the damaged pools and 4% of the healthy ones. The repair step can use either method.

Fully label-free. Table 9 combines the two steps. Algorithm 1 returns the taxonomy order in only 6% of pools, so every method loses ARI when the order is estimated (raw k-means falls from 0.270 to 0.190). Algorithm 2 still beats raw k-means at the same estimated order, by +0.010 (0.001 to 0.022), and always-whitening by +0.018 (0.010 to 0.025). The gain is smaller than at the taxonomy order (+0.035, 0.019 to 0.053), and the reason is mechanistic: capture also hides the order. On CIFAR-100 CE features Algorithm 1 selects 2.3 groups against 4.0 on the healthy representations, and one group in 35% of pools. At such small orders the monitor drop is small, and the rule fires on only 26% of the damaged pools. Algorithm 2 needs the base encoder in any case, so the order can be read from the base encoder instead. Then the rule fires on 48% of the damaged pools, its whitening decision agrees with the taxonomy-order decision in 82% of all pools, and the damaged-pool ARI rises from 0.103 to 0.132. On Fashion-MNIST CE features the label-free gain is large at either estimated order (0.228 to 0.402, and 0.372 to 0.490). At the representation's own order spectral clustering is level with Algorithm 2 (0.195 against 0.200), but it is more sensitive to an over-specified order: with the base-encoder order it trails Algorithm 2 by 0.016 (0.007 to 0.027) over all pools and reaches 0.280 against 0.490 on Fashion-MNIST CE features. End to end, then, the repair keeps a significant advantage without any fine label. Its size is bounded by order selection, which remains the weakest link of the pipeline.

Table 9. Fully label-free evaluation (mean ARI over pools). "Order" is the mean order chosen by Algorithm 1 on the representation itself, and the next three columns are at that order. "Order (base)" is chosen on the base encoder, and

the last three columns are at that order. At every estimated order, Algorithm 2 decides between raw and whitened k-means without fine labels. Bold marks the best entry of each block.

Setting	Order	Raw	Alg. 2	Spectral	Order (base)	Raw	Alg. 2	Spectral
S1 CIFAR-100 CE	2.3	0.038	0.046	**0.062**	4.1	0.063	**0.079**	0.075
S1 CIFAR-100 healthy	4.0	**0.152**	**0.152**	0.140	4.1	**0.119**	**0.119**	0.109
S3 Fashion-MNIST CE	5.2	0.228	**0.402**	0.313	8.0	0.372	**0.490**	0.280
S3 Fashion-MNIST healthy	7.8	**0.284**	0.275	0.280	8.0	**0.273**	0.259	0.272
S5 DINOv2 CE	5.8	**0.484**	**0.484**	0.480	5.8	**0.437**	**0.437**	0.420
S5 DINOv2 base	5.8	**0.349**	**0.349**	0.333	5.8	**0.349**	**0.349**	0.333
All pools	4.3	0.190	**0.200**	0.195	5.0	0.188	**0.197**	0.181

8 Discussion

Information, orientation and amount. The experiments separate three things that are often merged. Decodability, measured by the probe, barely moves under cross-entropy in any setting. The *amount* of within-class spread, measured by NC1, changes early and is the best scalar correlate of clusterability. The *orientation* of that spread relative to the semantic subspace, measured by dose and alignment, changes later in the in-domain settings and is what an intervention can act on. The dose–response run shows the two phases directly: an early NC1 jump with no dose change, and a later dose rise with a rotation of the two-means split, which is the phase that whitening repairs. A complete account needs both, and only the second comes with a mechanism and a remedy.

Why the base matters. Cross-entropy from an in-domain SSL base damages clusterability, and from ImageNet or DINOv2 bases it improves it, at every fine-tuning length we tried. A plausible reading is that an in-domain SSL encoder has already organised each class along view-invariant directions that include the subclasses, and cross-entropy then stretches directions that serve the coarse decision, which become the leading within-class nuisance. A broadly pretrained encoder starts with many directions unrelated to the current classes, and cross-entropy prunes those, which lowers the dose. The regime map records this behaviour; it does not establish the reading, which a future study could test by varying the pretraining distribution continuously.

The CIFAR-10 exception. In S4 the dose rose and alignment collapsed, as in the damaged settings, and GMM ARI fell, but k-means ARI rose. Each of the two coarse classes contains four or six fine classes with very unequal visual similarity (for example, cats and dogs against frogs and horses), so a single dominant nuisance direction may still separate some fine classes from others. The two-subclass model of Section 4.2 does not cover this case, and we report it as the boundary of the present theory.

For practitioners. The results suggest a short procedure for a coarse-label model whose subclasses matter. Keep the base encoder. After fine-tuning, compute the monitor ρ_K of each coarse class in both encoders. Where the monitor has dropped, cluster in whitened coordinates (Algorithm 2) or on a nearest-neighbour graph, but not with a covariance-adaptive Euclidean method, which keeps the captured partition. Read the working order from the base encoder rather than the damaged one (Section 7.7), report the order as an interval (Algorithm 1 and 1-S), and treat clusters as hypotheses to

be inspected rather than validated subtypes. Where clustering matters at training time, a contrastive term with in-class negatives (PSL-C, Appendix A.8) prevents the damage at a cost of 2.5–3.0 points of coarse accuracy on CIFAR-100 and under half a point elsewhere.

Scope. Our evidence comes from image data, three architectures and three pretraining sources; whether text and tabular encoders show the same regimes remains open. The theory assumes Gaussian nuisance and, in its global statements, two subclasses; the monitor's guarantee is stated for two-means while the practical version uses the working order. Algorithm 2 needs the base encoder. With the order estimated, its gain is about a third of the gain at the taxonomy order. The order diagnostic does not recover benchmark taxonomies reliably, and no inferential p-value is attached to any discovered cluster.

9 Conclusion

A coarse-label network can keep a fine distinction in its features and still hide it from clustering. We traced the hiding to a spectral competition inside each coarse class, located its crossover in $(1, \pi/2]$, proved that within-class whitening removes it, and bounded a label-free monitor that detects it. A pre-specified dose–response experiment, an independent rerun and four new settings showed the mechanism at work where the theory says it should be, showed its absence where the dose falls, and confirmed it by an intervention that helps only the damaged representations. The practical message is that the question "has my model lost its subclasses?" has three parts, information, access and order, and that the second part can now be checked and repaired without fine labels.

Statements and Declarations

Funding. No funding was received for this work.

Competing interests. The author declares no competing interests.

Data and code availability. All datasets are public (CIFAR-10, CIFAR-100, Fashion-MNIST); the DINOv2 and ResNet-50 weights are those released by their authors. The training notebook, the analysis code, the raw and derived results, the frozen test-split features of the archived encoders and Figures 1–8 are archived at Zenodo, https://doi.org/10.5281/zenodo.24146068 (Oban 2026); running its scripts reproduces the reported numbers. The training-split features of the archived run (2.3 GB) are available from the author on request.

Ethics approval and consent to participate. Not applicable: the study uses public benchmark image datasets and involves no human participants or animals.

Consent for publication. Not applicable.

Use of AI tools. Large-language-model tools were used for language editing, figure preparation and coding assistance. The author verified all mathematical and empirical claims and takes full responsibility for the content.

Author contributions. Sole author.

References

Arthur D, Vassilvitskii S (2007) k-means++: the advantages of careful seeding. In: Proceedings of the 18th annual ACM-SIAM symposium on discrete algorithms. SIAM, Philadelphia, pp 1027–1035

Awasthi P, Sheffet O (2012) Improved spectral-norm bounds for clustering. In: Approximation, randomization, and combinatorial optimization. Lecture Notes in Computer Science 7408. Springer, Berlin, pp 37–49

Biau G, Devroye L, Lugosi G (2008) On the performance of clustering in Hilbert spaces. IEEE Transactions on Information Theory 54(2): 781–790

Bukchin G, Schwartz E, Saenko K et al (2021) Fine-grained angular contrastive learning with coarse labels. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. IEEE, Piscataway, pp 8730–8740

Chen T, Kornblith S, Norouzi M et al (2020) A simple framework for contrastive learning of visual representations. In: Proceedings of the 37th international conference on machine learning. PMLR 119, pp 1597–1607

Chen YT, Witten DM (2023) Selective inference for k-means clustering. Journal of Machine Learning Research 24(152): 1–41

Dai Q, Huang H, Wu Y et al (2025) Adaptive part learning for fine-grained generalized category discovery: a plug-and-play enhancement. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. IEEE, Piscataway, pp 25444–25453

Damle A, Minden V, Ying L (2019) Simple, direct and efficient multi-way spectral clustering. Information and Inference 8(1). https://doi.org/10.1093/imaiai/iay008

Davis C, Kahan WM (1970) The rotation of eigenvectors by a perturbation. III. SIAM Journal on Numerical Analysis 7(1): 1–46

Flury BA (1990) Principal points. Biometrika 77(1): 33–41

Friedman JH, Tukey JW (1974) A projection pursuit algorithm for exploratory data analysis. IEEE Transactions on Computers C-23(9): 881–890

Gao LL, Bien J, Witten D (2024) Selective inference for hierarchical clustering. Journal of the American Statistical Association 119(545): 332–342

Graf S, Luschgy H (2000) Foundations of quantization for probability distributions. Lecture Notes in Mathematics 1730. Springer, Berlin

Grcić M, Gadetsky A, Brbić M (2024) Fine-grained classes and how to find them. In: Proceedings of the 41st international conference on machine learning. PMLR 235

He K, Zhang X, Ren S et al (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. IEEE, Piscataway, pp 770–778

Hubert L, Arabie P (1985) Comparing partitions. Journal of Classification 2: 193–218

Hui L, Belkin M, Nakkiran P (2022) Limitations of neural collapse for understanding generalization in deep learning. arXiv:2202.08384

Hyvärinen A, Oja E (2000) Independent component analysis: algorithms and applications. Neural Networks 13(4–5): 411–430

Kalogeratos A, Likas A (2012) Dip-means: an incremental clustering method for estimating the number of clusters. In: Advances in neural information processing systems 25. Curran Associates, Red Hook, pp 2393–2401

Keribin C (2000) Consistent estimation of the order of mixture models. Sankhyā Series A 62(1): 49–66

Kessy A, Lewin A, Strimmer K (2018) Optimal whitening and decorrelation. The American Statistician 72(4): 309–314

Kieffer JC (1983) Uniqueness of locally optimal quantizer for log-concave density and convex error weighting function. IEEE Transactions on Information Theory 29(1): 42–47

Krizhevsky A (2009) Learning multiple layers of features from tiny images. Technical report, University of Toronto

Kumar A, Kannan R (2010) Clustering with spectral norm and the k-means algorithm. In: Proceedings of the 51st IEEE symposium on foundations of computer science. IEEE, Piscataway, pp 299–308

Liu D, Tan Z, Zhao L et al (2025) Generalized category discovery via reciprocal learning and class-wise distribution regularization. In: Proceedings of the 42nd international conference on machine learning. PMLR 267, pp 39651–39668

Lu Y, Zhou HH (2016) Statistical and computational guarantees of Lloyd's algorithm and its variants. arXiv:1612.02099

McInnes L, Healy J, Melville J (2018) UMAP: uniform manifold approximation and projection for dimension reduction. arXiv:1802.03426

Melnykov I, Melnykov V (2014) On K-means algorithm with the use of Mahalanobis distances. Statistics & Probability Letters 84: 88–95. https://doi.org/10.1016/j.spl.2013.09.026

Müller R, Kornblith S, Hinton G (2020) Subclass distillation. arXiv:2002.03936

Ng AY, Jordan MI, Weiss Y (2002) On spectral clustering: analysis and an algorithm. In: Advances in neural information processing systems 14. MIT Press, Cambridge, pp 849–856

Oban V (2026) Decodable but not clusterable: reproducibility package (code, results and frozen features), version 2. Zenodo. https://doi.org/10.5281/zenodo.24146068

Oakden-Rayner L, Dunnmon J, Carneiro G et al (2020) Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In: Proceedings of the ACM conference on health, inference, and learning. ACM, New York, pp 151–159

Oquab M, Darcet T, Moutakanni T et al (2024) DINOv2: learning robust visual features without supervision. Transactions on Machine Learning Research

Papyan V, Han XY, Donoho DL (2020) Prevalence of neural collapse during the terminal phase of deep learning training. Proceedings of the National Academy of Sciences 117(40): 24652–24663

Pollard D (1981) Strong consistency of k-means clustering. Annals of Statistics 9(1): 135–140

Poole B, Ozair S, van den Oord A et al (2019) On variational bounds of mutual information. In: Proceedings of the 36th international conference on machine learning. PMLR 97

Schwarz G (1978) Estimating the dimension of a model. Annals of Statistics 6(2): 461–464

Sohoni NS, Dunnmon JA, Angus G et al (2020) No subclass left behind: fine-grained robustness in coarse-grained classification problems. In: Advances in neural information processing systems 33. Curran Associates, Red Hook

Tarpey T, Li L, Flury BD (1995) Principal points and self-consistent points of elliptical distributions. Annals of Statistics 23(1): 103–112

Teicher H (1963) Identifiability of finite mixtures. Annals of Mathematical Statistics 34(4): 1265–1269

Tibshirani R, Walther G (2005) Cluster validation by prediction strength. Journal of Computational and Graphical Statistics 14(3): 511–528

Touvron H, Sablayrolles A, Douze M et al (2021) Grafit: learning fine-grained image representations with coarse labels. In: Proceedings of the IEEE/CVF international conference on computer vision. IEEE, Piscataway, pp 874–884

Tsai YHH, Li T, Ma MQ et al (2022) Conditional contrastive learning with kernel. In: International conference on learning representations

van den Oord A, Li Y, Vinyals O (2018) Representation learning with contrastive predictive coding. arXiv:1807.03748

Vaze S, Han K, Vedaldi A et al (2022) Generalized category discovery. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. IEEE, Piscataway, pp 7492–7501

Vershynin R (2018) High-dimensional probability: an introduction with applications in data science. Cambridge University Press, Cambridge

Villani C (2009) Optimal transport: old and new. Springer, Berlin

von Kügelgen J, Sharma Y, Gresele L et al (2021) Self-supervised learning with data augmentations provably isolates content from style. In: Advances in neural information processing systems 34. Curran Associates, Red Hook

Wang M, Zhong Z, Gong X (2025) Prior-constrained association learning for fine-grained generalized category discovery. In: Proceedings of the AAAI conference on artificial intelligence 39(20). AAAI Press, Washington, pp 21162–21170

Wu D, Mondelli M (2025) Neural collapse beyond the unconstrained features model: landscape, dynamics, and generalization in the mean-field regime. In: Proceedings of the 42nd international conference on machine learning. PMLR 267, pp 67499–67536

Xiao H, Rasul K, Vollgraf R (2017) Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. arXiv:1708.07747

Xu Y, Qian Q, Li H et al (2021) Weakly supervised representation learning with coarse labels. In: Proceedings of the IEEE/CVF international conference on computer vision. IEEE, Piscataway

Yakowitz SJ, Spragins JD (1968) On the identifiability of finite mixtures. Annals of Mathematical Statistics 39(1): 209–214

Yang Y, Steinhardt J, Hu W (2023) Are neurons actually collapsed? On the fine-grained structure in neural representations. In: Proceedings of the 40th international conference on machine learning. PMLR 202

Yu Y, Wang T, Samworth RJ (2015) A useful variant of the Davis–Kahan theorem for statisticians. Biometrika 102(2): 315–323

Appendix A. Proofs

Throughout, the gain of a two-cell partition $\{B, B^c\}$ of a centred random vector X with $p = P(B) \in (0,1)$ is the reduction of distortion from one centre to the two cell means,

$$G(B) = p(1-p)\ \| \Delta \|^2 = \frac{\| \mathbb{E}[X(\mathbf{1}_B - p)] \|^2}{p(1-p)}, \qquad \Delta = \mathbb{E}[X \mid B] - \mathbb{E}[X \mid B^c],$$

and $G_2(X) = \sup_B G(B)$. For a real random variable W let $g(W)$ be the supremum of $G(B)$ over two-cell partitions of $\mathbb{R}$.

A.1 Two lemmas

Lemma A.1. $f(c) = \mathbb{E}|M_c|$ *is strictly increasing on* $[0,1)$*, with* $f(0) = \sqrt{2/\pi}$ *and* $f(c) < 1$.

Proof. By symmetry $f(c) = \mathbb{E}|\alpha + \beta N|$ with $N \sim \mathcal{N}(0,1)$, $\alpha = \sqrt{c}$, $\beta = \sqrt{1-c}$. Differentiating under the expectation, $\partial_\alpha \mathbb{E}|\alpha + \beta N| = 1 - 2\Phi(-t)$ and $\partial_\beta \mathbb{E}|\alpha + \beta N| = \mathbb{E}[N\text{sign}(\alpha + \beta N)] = 2\varphi(t)$ with $t = \alpha/\beta$. Since $d\alpha/dc = 1/(2\alpha)$ and $d\beta/dc = -1/(2\beta)$,

$$2\alpha\, f'(c) = 2\Phi(t) - 1 - 2t\varphi(t) =: h(t).$$

Now $h(0) = 0$ and $h'(t) = 2\varphi(t) - 2\varphi(t) - 2t\varphi'(t) = 2t^2\varphi(t) > 0$ for $t > 0$, so $f'(c) > 0$ on $(0,1)$. The value at 0 is the folded-normal mean, and $f(c)^2 < \mathbb{E}M_c^2 = 1$ by Jensen's inequality. ■

Lemma A.2 (Reduction to one dimension). *For a centred random vector* X *with finite second moment,* $G_2(X) = \sup_{\|\theta\|=1} g(\theta^\top X)$.

Proof. For any partition with $\Delta \neq 0$ put $\theta^* = \Delta/\| \Delta \|$. Then $G(B) = p(1-p)(\theta^{*\top}\Delta)^2$, which is the between-cell variance of $\theta^{*\top}X$ under the same partition and hence at most $g(\theta^{*\top}X)$. Conversely, a threshold partition of $\theta^\top X$ is a half-space partition of X, and its gain in $\mathbb{R}^d$ is at least its between-cell variance along θ. ■

A.2 Proposition 2 and Corollary 1

For the semantic sign partition, $p = 1/2$ and, because η is centred and independent of S, the cell means are $\pm q_s u$; the gain is q_s^2. For the sign partition along a unit leading eigenvector w of Σ_n, the cell means are $\pm\mathbb{E}|w^\top\eta|\, w = \pm\sqrt{2\lambda/\pi}\, w$ and the gain is $2\lambda/\pi$. A larger gain is a smaller distortion, which proves Proposition 2. Writing $S/\sqrt{v} = \sqrt{\kappa}Z + \sqrt{1-\kappa}\, N$ shows $S/\sqrt{v} \sim M_\kappa$, so $q_s = \sqrt{v}\, f(\kappa)$ and the crossover $\lambda^* = \pi q_s^2/2$ corresponds to the dose $(\pi/2)f(\kappa)^2$. Lemma A.1 gives $2/\pi < f(\kappa)^2 < 1$ for $\kappa \in (0,1)$, and the limits follow from continuity of f and $f(c) \to 1$ as $c \to 1$. ■

A.3 Theorem 1

An optimal pair of centres exists, both cells have positive mass, the cells are Voronoi half-spaces and each centre is its cell mean (Pollard 1981; Graf and Luschgy 2000). Hence the normal θ is parallel to Δ and the gain is $G = p(1-p)\| \Delta \|^2$. For any unit a, the law of total variance gives $p(1-p)(a^\top\Delta)^2 \leq \text{Var}(a^\top H)$; with $a = u$, $p(1-p)(u^\top\Delta)^2 \leq v$. The optimum gains at least the $2\lambda/\pi$ of the nuisance sign split, so $(u^\top\theta)^2 = (u^\top\Delta)^2/\| \Delta \|^2 \leq \pi v/(2\lambda)$, which is (i).

For (ii) put $c = u^\top\theta$. Given $Z = z$, $\theta^\top H$ is Gaussian with mean cmz and variance $\tau^2 = c^2\sigma_s^2 + \theta^\top\Sigma_n\theta$. The partition is a threshold rule on $\theta^\top H$, and no threshold separates $\mathcal{N}(\pm cm, \tau^2)$ with error below $\Phi(-|c|m/\tau)$. Taking $a = \theta$ in the variance inequality, $2\lambda/\pi \leq G \leq \tau^2 + c^2m^2$. If $\lambda \geq \pi v$, (i) gives $c^2m^2 \leq \lambda/(2\pi)$, so $\tau^2 \geq 3\lambda/(2\pi)$ and $|c|m/\tau \leq \pi m\sqrt{v}/(\sqrt{3}\lambda)$; since $\Phi(-x) \geq 1/2 - x/\sqrt{2\pi}$, (ii) follows. Subtracting the Bayes error $\Phi(-m/\sigma_s)$ gives (iii). ■

A.4 Theorem 2

(i) Since $\Sigma_n u = 0$ and Σ_n maps $u^\perp$ to itself, $\Sigma^{-1/2} = v^{-1/2}uu^\top + \Sigma_n^{-1/2}$, with $\Sigma_n^{-1/2}$ the inverse square root on $u^\perp$. Hence $\widetilde{H} = \big(S/\sqrt{v}\big)u + \xi$ with $\xi = \Sigma_n^{-1/2}\eta \sim \mathcal{N}(0, I - uu^\top)$, independent of S. As $S/\sqrt{v} \sim M_\kappa$ conditionally coupled to Z through $\sqrt{\kappa}Z$, the joint law of $\big(\widetilde{H}, Z\big)$ depends only on κ.

(ii) With $a = u^\top\theta$, $\theta^\top\widetilde{H} = a\,S/\sqrt{v} + \theta^\top\xi$, where $\theta^\top\xi \sim \mathcal{N}(0, 1-a^2)$ is independent of S. Given $Z = z$ this is Gaussian with mean $a\sqrt{\kappa}z$ and variance $a^2(1-\kappa) + 1 - a^2 = 1 - a^2\kappa$, i.e. $\theta^\top\widetilde{H} \sim M_{a^2\kappa}$.

(iii) The semantic sign partition has cell means $\pm f(\kappa)u$, hence gain $f(\kappa)^2$. If $\Delta \perp u$, the argument of Lemma A.2 bounds the gain by $g\big(\theta^{*\top}\widetilde{H}\big)$ with $\theta^* \perp u$; then $\theta^{*\top}\widetilde{H} \sim \mathcal{N}(0,1)$, whose optimal two-cell partition is the median split with gain $2/\pi$ (Flury 1990). The symmetric gain along θ is $\big(\mathbb{E}\big|M_{c(\theta)}\big|\big)^2 = f\big(c(\theta)\big)^2$ with $c(\theta) = \kappa(u^\top\theta)^2 \le \kappa$, equality iff $\theta = \pm u$; Lemma A.1 gives uniqueness.

(iv) If $m \le \sigma_s$ then $\kappa \le 1/2$ and $c(\theta) \le 1/2$ for every θ. The density of M_c is proportional to $\exp\{-x^2/(2s^2)\}\cosh(ax/s^2)$ with $a = \sqrt{c}$ and $s^2 = 1 - c$, so $(\log)'' = -s^{-2} + a^2 s^{-4}\operatorname{sech}^2(ax/s^2) \le -s^{-2} + a^2 s^{-4} \le 0$ when $a^2 \le s^2$, i.e. $c \le 1/2$. For a log-concave density the locally optimal two-level quantiser is unique (Kieffer 1983); by symmetry it is the median split, so $g(M_c) = f(c)^2$. By Lemma A.2, $G_2\big(\widetilde{H}\big) = \max_\theta f\big(c(\theta)\big)^2 = f(\kappa)^2$, attained by the semantic sign partition. Any optimal partition is a half-space pair with normal $\theta^* \parallel \Delta$ and gain at most $f\big(c(\theta^*)\big)^2$, which is below $f(\kappa)^2$ unless $\theta^* = \pm u$; along u the optimal threshold is unique, so the optimum is the semantic sign partition. Because Z affects $\widetilde{H}$ only through $u^\top\widetilde{H}$, whose class-conditional laws are $\mathcal{N}\big(\pm\sqrt{\kappa}, 1-\kappa\big)$, that partition is the Bayes rule, with error $\Phi\big(-\sqrt{\kappa/(1-\kappa)}\big) = \Phi(-m/\sigma_s)$. ■

A.5 Proposition 3

Because $\lambda \ge v$, $\lambda_1(\Sigma_H) = \lambda$ and the Gaussian reference gain is $2\lambda/\pi$. The nuisance sign split attains $2\lambda/\pi$, so $\rho \ge 1$. For an arbitrary partition write $\mathbb{E}[H(\mathbf{1}_B - p)] = \mathbb{E}[S(\mathbf{1}_B - p)]\,u + \mathbb{E}[\eta\mathbf{1}_B]$, an orthogonal decomposition. By Cauchy–Schwarz, $\mathbb{E}[S(\mathbf{1}_B - p)]^2 \le \mathrm{Var}(S)\mathrm{Var}(\mathbf{1}_B) = v\,p(1-p)$. For the second term take the unit direction w of $\mathbb{E}[\eta\mathbf{1}_B]$ and $X = w^\top\eta \sim \mathcal{N}(0, \sigma_w^2)$ with $\sigma_w^2 \le \lambda$. With $\beta(X) = P(B \mid X) \in [0,1]$ and $\mathbb{E}\beta = p$, the bathtub principle gives $\mathbb{E}[X\beta(X)] \le \sigma_w\varphi\big(z_p\big)$, where $z_p = \Phi^{-1}(p)$. The threshold split of $\mathcal{N}(0,1)$ at its p-quantile has between-cell variance $\varphi\big(z_p\big)^2/\big(p(1-p)\big)$, which is at most $2/\pi$ by the optimality of the median split; hence $\| \mathbb{E}[\eta\mathbf{1}_B] \|^2 \le (2\lambda/\pi)\,p(1-p)$. Dividing by $p(1-p)$, every gain is at most $v + 2\lambda/\pi$, and $\rho \le 1 + \pi v/(2\lambda)$. After whitening the covariance is the identity, the reference gain is $2/\pi$, and the semantic sign partition attains $f(\kappa)^2$ (Theorem 2(iii)). Solving the upper bound for the dose gives the certificate. ■

A.6 Order results

Theorem 3. Every empirical gain is within 2ϵ of its population value, so the gain at K stays above λ_0 and all later gains stay below it. ■

Theorem 4. With the K component means as centres, $D_y(K) \le s^2$. Any set C of $K-1$ centres leaves some mean μ_z at distance at least $\Delta/2$ (two means share a nearest centre, which cannot be within $\Delta/2$ of both). For $H \sim P_z$, $\mathrm{dist}(H, C) \ge (\Delta/2 - \| H - \mu_z \|)_+$, and Jensen's inequality for the convex non-increasing map $t \mapsto (\Delta/2 - t)_+^2$ with $\mathbb{E} \| H - \mu_z \| \le s$ gives $\mathbb{E}\mathrm{dist}(H, C)^2 \ge (\Delta/2 - s)^2$. So $D_y(K-1) \ge w_{\min}(\Delta/2 - s)^2$ and $G_y(K) > \lambda_0$; for $k > K$, $G_y(k) \le D_y(K) \le s^2 < \lambda_0$. ■

Consistency. For finite second moments, $\left|D_k(P)^{1/2} - D_k(Q)^{1/2}\right| \le W_2(P, Q)$ (Minkowski over couplings; Biau et al 2008), and $W_2\left(\hat{P}_n, P\right) \to 0$ almost surely (Villani 2009), so Theorem 3 applies eventually. For $\| H \| \le R$, a Rademacher bound on the class $\{\min_{c \in C} \| x - c \|^2\}$ and McDiarmid's inequality give $\max_{k \le K_{\max}} \left|\widehat{D}(k) - D(k)\right| \le R^2\left[6K_{\max} + 2\sqrt{2\log(2K_{\max}/\delta)}\right]/\sqrt{n}$ with probability $1 - \delta$.

Proposition 4. Any two-cell gain equals a between-cell variance along $\Delta/\| \Delta \|$, hence is at most λ_1. Splitting by the sign of $e_1^\top X^0$ for a leading eigenvector e_1 gives centroids $\pm\sqrt{\lambda_1}\, \mathbb{E}R\, \mathbb{E}|U_1|\, e_1$ by the symmetry $U_j \mapsto -U_j$, which is the lower bound. In the Gaussian case every projection is Gaussian and the median split of $\mathcal{N}(0, \tau^2)$ has gain $2\tau^2/\pi$ (Flury 1990), maximised at $\tau^2 = \lambda_1$. For (iii), one centre per component gives $G_P(2) \ge w(1-w)\Delta^2$, and Weyl's inequality gives $\lambda_1(\Sigma_P) \le w(1-w)\Delta^2 + \lambda_1(\Sigma_w)$. ■

Lemma 1. The eigenvalues of Σ are v (eigenvector u) and $\lambda_1, \lambda_2, \ldots$ on $u^\perp$, which gives the retention condition. For the perturbation bound, $\| \left(I - \hat{P}_r\right)u \| \le \| \sin\Theta\left(\hat{V}, V\right) \|_{\mathrm{op}}$ for the top-r eigenbases, and the Davis–Kahan theorem in the form of Yu et al (2015) bounds this by $2\sqrt{r}\, \| \hat{\Sigma} - \Sigma \|_{\mathrm{op}}/g_r$. ■

A.7 Theorem 5

Fix a coarse class and let $T(x) = P_r D_{sc}(x - \mu)$ be the population analysis map. *Step 1 (estimated map).* Under (A1) the centring and scaling estimates converge at rate $n^{-1/2}$; the covariance estimate satisfies $\| \hat{\Sigma} - \Sigma \|_{\mathrm{op}} \le C\left(\sqrt{d/n} + \sqrt{\log(1/\delta)/n}\right)$ (Vershynin 2018), and Lemma 1 transfers this to the projector. For affine maps A_1, A_2 and Q supported in a ball of radius R, $W_2(A_{1\#}Q, A_{2\#}Q) \le R\, \| A_1 - A_2 \|$, so by the Wasserstein continuity above all distortions in estimated coordinates are within $C\epsilon_T$ of those in population coordinates. *Step 2 (held-out evaluation).* For discovery-fitted centres $\hat{C}$ within $\varepsilon_{\mathrm{opt}}$ of the empirical optimum, $\left|\hat{P}_{\mathrm{val}} f_{\hat{C}} - D_k\right| \le \left|\hat{P}_{\mathrm{val}} f_{\hat{C}} - P f_{\hat{C}}\right| + \varepsilon_{\mathrm{opt}} + 2\sup_C \left|\hat{P}_{\mathrm{disc}} f_C - P f_C\right|$, and both suprema are $O\left(\left(K_{\max} + \sqrt{\log(1/\delta)}\right)/\sqrt{n}\right)$ by the Rademacher bound. *Step 3 (null floor).* The recolouring $x \mapsto \hat{\mu} + \hat{\Sigma}_\gamma^{1/2}(ru)$ is Lipschitz in $\left(\mu, \Sigma^{1/2}\right)$ on bounded support, matrix square

roots are Lipschitz under the ridge of (A2), and the empirical radius law converges in W_2; hence each null gain concentrates around its population value, and the empirical q-quantile of B_0 draws is within $C\sqrt{\log(1/\delta)/B_0}$ of the population floor by the Dvoretzky–Kiefer–Wolfowitz inequality. *Step 4 (crossing).* When all gains and floors are within $\gamma/2$ of their targets, (A3) forces every crossing through K and rejects the next, so the prefix rule returns K. ■

A.8 Training-time insurance: PSL and PSL-C

PSL adds an NT-Xent term (Chen et al 2020) to the coarse cross-entropy; PSL-C replaces half of it by a term whose negatives are drawn only from the anchor's coarse class:

$$\alpha L_{\mathrm{CE}} + \beta\left[1/2\, L_{\mathrm{NCE}} + 1/2\, L^{Y}_{\mathrm{NCE}}\right], \qquad L^{Y}_{\mathrm{NCE}} = -\frac{1}{2B}\sum_{i=1}^{2B} \log \frac{\exp\left(s_{i,p(i)}/\tau\right)}{\sum_{j\neq i:\, Y_j = Y_i} \exp\left(s_{ij}/\tau\right)}.$$

L^{Y}_{NCE} lower-bounds the conditional view information $I(H; H' \mid Y)$ (van den Oord et al 2018; Poole et al 2019; Tsai et al 2022). **Assumption V.** *Given Y, the views are $X = g(Z, \eta)$ and $X' = g(Z, \eta')$ with $\eta' \perp (Z, \eta) \mid Y$ and $g(\cdot, \eta)$ injective.* **Proposition A.1.** *Under Assumption V, any representation that maximises $I(H; H' \mid Y)$ determines Z almost surely; if instead augmentation preserves a nuisance factor and resamples Z-relevant detail, the same argument protects the nuisance.* The proof is the conditional data-processing inequality $I(H; H' \mid Y) \leq I(X; X' \mid Y) = H(Z \mid Y)$, with equality iff H determines Z given Y, in the style of von Kügelgen et al (2021). The converse is testable: on Fashion-MNIST, without intensity-changing augmentations the first within-class principal component of PSL features has absolute correlation 0.94 with mean image intensity and oracle-order ARI 0.19, whereas contrast and gamma jitter with resized crops reduce the correlation to 0.59 and raise ARI to about 0.50.

Appendix B. Implementation details

Features and metrics. Features are the 512-dimensional (ResNet-18), 384-dimensional (ViT-S/14) or 2048-dimensional (ResNet-50) pooled outputs before the classification head. Within each coarse class: the probe is a multinomial logistic regression (L2, $C = 1$, 1000 iterations) on standardised training features, scored on test features; GMM ARI uses a full-covariance Gaussian mixture (three initialisations, covariance ridge 10^{-4}) on PCA-10 of standardised test features; k-means ARI uses ten initialisations on PCA-10 or PCA-20; whitened k-means divides the PCA-20 scores by the square roots of their eigenvalues. The dose, alignment, NC1 and trace ratio are computed on the full standardised test features of the class. The monitor ρ_K divides the k-means gain $D(1) - D(K)$ (four initialisations) by the mean gain of $B = 3$ Gaussian clouds with the same covariance and size.

Training. ResNet-18 encoders use the CIFAR stem (3×3 first convolution, no max-pooling) and SimCLR pretraining with batch 512, temperature 0.2 and a 128-dimensional projector; fine-tuning uses the recipe of Section 6, random crop with 4-pixel padding and horizontal flip performed on the

GPU, and mixed precision with the contrastive loss computed in single precision. DINOv2 ViT-S/14 is evaluated at 112 pixels and ResNet-50 (timm `resnet50.a1_in1k`) at 96 pixels, both with ImageNet normalisation. All runs used one NVIDIA T4 GPU; the new settings (S2, S4, S6, S7) took 4.98 hours in total, of which the SSL base of S2 took about one hour.

Author biography

Volkan Oban received his PhD from Yıldız Technical University, Istanbul, in 2018, with a thesis on inverse problems for the Burgers and Black–Scholes equations. He is an independent researcher in Türkiye working on the theory of representation learning, clustering and retrieval-augmented language models, and he serves as a reviewer for Information Processing & Management.