INTRINSIC DIMENSION READS WITHIN-CLASS STRUCTURE BELOW THE CLASS SIZE AND CLASS-MEAN GEOMETRY ABOVE IT

Anonymous authors
Paper under double-blind review

ABSTRACT

Intrinsic dimension is a popular one-number summary of a neural network's representation, used to argue that networks compress their representations during training and that lower dimension means better generalisation. We show that on labelled data this number measures one of two different things, and a rarely reported choice decides which: how many neighbours the estimator looks at, compared with how many images each class contributes. With few neighbours the number describes the shape of each class; with many, how the classes are laid out. The two geometries relate to accuracy in opposite ways. On 178 ImageNet classifiers, changing only the neighbour count turns the well-known correlation between dimension and error from $+0.80$ to -0.66 on the same networks and images, and in training runs the same checkpoints read as compression at one count and expansion at another. We prove two short theorems: one derives the switch from how a query's neighbours fill with classmates, the other shows that class means, class covariances and even the network's own logits do not determine what the estimator reports. The switch holds across architecture families and network scale, on four out-of-distribution test sets, in self-supervised encoders and in 26 language models. We close with a two-count check that any study can run: read at half and at four times the class size, and say which geometry the conclusion rests on.

1 INTRODUCTION

Two questions about learned representations are routinely answered with one number. Does a network compress its representation as it learns? Do lower-dimensional representations generalise better? The number is a nearest-neighbour estimate of intrinsic dimension, reported to end training lower than it starts, to rise and fall with depth, and to predict error across networks (Ansuini et al., 2019; Valeriani et al., 2023); lower dimension of the data themselves has been tied to easier learning (Pope et al., 2021). The best-known version of the last claim is that ImageNet classifiers whose last-layer representations have lower estimated dimension make fewer errors (Ansuini et al., 2019). This number has a hidden knob.

Nearest-neighbour estimators of intrinsic dimension work like this. Take a point, measure the distance to its nth nearest neighbour and to its $2n$th, and ask how much the radius had to grow to double the count: in a space with few directions it grows a lot, in one with many directions only a little. On labelled data the estimate is computed on a *probe*, a sample of m images from each of K classes, so a point has $m-1$ classmates. When the neighbour count n is small compared with the class size m, both neighbours are classmates and the estimator sees only the shape of that class. When n is large compared with m, both neighbours belong to other classes and the estimator sees how the classes are laid out. In between it sees a mixture. Figure 1 shows the three cases, and how the correlation with error changes sign between them.

Formally, write each point as the mean of its class plus what is left over,

$$x_i = \mu_{y_i} + r_i, \tag{1}$$

so a labelled cloud has two scales: how far points sit from their own class mean, and how far class means sit from each other.

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.

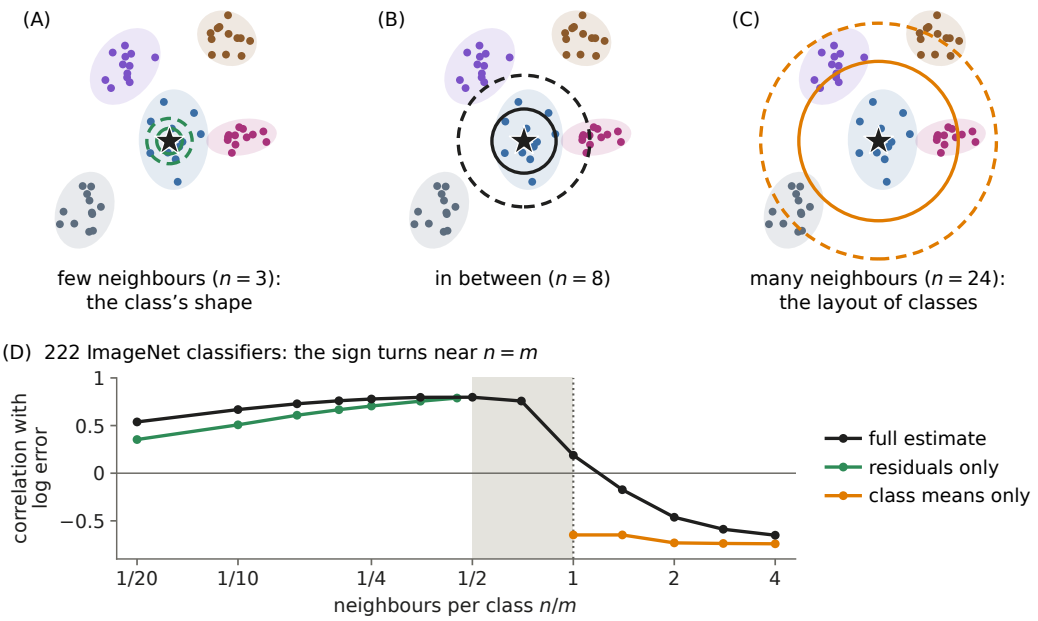

Figure 1: **Few neighbours read the class; many read the layout of classes; and the two relate to accuracy with opposite signs.** (A–C) A query ($\star$), its nth (solid circle) and $2n$th (dashed) nearest neighbours as the neighbour count grows past the class size ($m = 12$); circle colours match the curves in (D). (D) On 222 ImageNet classifiers (the 178 of Section 2 and 44 ResNet recipe checkpoints; $K = 100$, $m = 20$), correlation between estimated dimension and log top-5 error against neighbours per class, for the full estimate (black) and the estimator run on the within-class residuals alone (green) or on the class means alone (orange). Shaded: where the outer shell leaves the class.

Writing $D_{i,(k)}$ for the distance from x_i to its kth nearest neighbour among the $N = Km$ probe points, the estimator compares two shells around each point,

$$L_n(X) = \frac{1}{N}\sum_i \log \frac{D_{i,(2n)}}{D_{i,(n)}}, \qquad \widehat{d}_n(X) = \frac{\kappa_n}{L_n(X)}, \qquad \kappa_n = \psi(2n) - \psi(n), \tag{2}$$

so a small radial increase reads as a high dimension. This is a log-moment estimator on the ratio that GRIDE fits by maximum likelihood with the outer shell at twice the inner (Denti et al., 2022); at $n = 1$ the two coincide and give the maximum-likelihood form of TWONN (Facco et al., 2017). Under complete separation the switch is sharp: the estimate reads the class alone while both shells are among classmates ($n < m/2$) and the class means alone once both have left ($n \geq m$; Corollary 1). Real classes overlap, so on networks the class-means-only estimate takes over gradually and stands in for the full one only from about $2m$. Table 1 therefore names $n < m/2$ the *class regime*, $n \geq 2m$ the *layout regime*, and $m/2 \leq n < 2m$ the *mixture*, whose first half, $m/2 \leq n < m$, where the outer shell leaves the class, is the *crossover band* shaded in the figures.

The neighbour count therefore decides what the estimate measures, and the two measurements order

Table 1: **Three regimes of one estimator**, by neighbours per class on a stratified probe (m images from each of K classes); r is the deleted estimate's correlation with the full one.

neighbours per class	where the two shells sit	what the estimate sees	estimate that reproduces it	along training	sign against error (ImageNet collections)
class regime, $n < m/2$	both among classmates	the shape of each class	residuals only ($r \geq 0.92$)	little change, then falls	positive
mixture, $m/2 \leq n < 2m$ (first half: crossover band)	the outer, then both, leave the class	class spread against class spacing	neither	falls, then turns	positive, then crossing
layout regime, $n \geq 2m$	both among other classes	the layout of the class means	class means only ($r \geq 0.74$)	rises near $4m$	negative

networks by error in opposite directions. The published correlation is mostly a reading of the first kind and replicates on its own networks; read with many neighbours per class, the same estimator orders modern networks the other way. Both readings describe real geometry, so neither is wrong, but a sentence such as "lower dimension goes with lower error" has no sign until the neighbours per class are named. We make three contributions.

- **A published correlation flips with the neighbour count on modern networks.** On 178 ImageNet classifiers read on fixed images, the neighbour count alone reverses the sign of the dimension–error correlation, within architecture families, within one architecture trained with different recipes, and given network scale. The crossing moves with the class size, and in training runs the same checkpoints read as compression or expansion depending on the count. The original study's fourteen networks stay positive; that value goes with first-generation training recipes (Section 2).
- **What each regime measures, shown by deletion, intervention and two short theorems.** Keeping only the residuals reproduces the ordering below half the class size, and keeping only the class means reproduces it from twice the class size. Synthetic data that keeps only class means and each point's distance to its mean reproduces the correlation curve in seven populations, and equalising those distances removes the negative branch. One theorem says what moves the estimate as classes tighten; the other, that class means, covariances, the collapse statistics and even the logits do not determine it (Section 3).
- **Published readings change meaning, and a two-count check makes any of them interpretable.** Late-training "compression" is a range of neighbour counts set by the class size; choosing layers above the class size costs accuracy; the switch transfers to 26 language models and the sign inverts across the layers of one. Two readings, at $m/2$ and $4m$, tell the regimes apart (Section 4).

2 Changing the neighbour count flips the published correlation on modern networks

Same networks, same images. Only the neighbour count changes, and the sign of the correlation with error changes with it. Ansuini et al. (2019) read the estimator at one neighbour on 2,000 ImageNet (Deng et al., 2009) training images, about two per class, and found that last-layer dimension predicts top-5 error. We keep that design on validation images, fix the images, and change only n. On 178 networks from three model collections, torchvision and two drawn from timm (Wightman, 2019), the correlation between estimated dimension and log top-5 error runs from $+0.80$ at one neighbour to -0.66 at eighty, and each collection reverses on its own (Figure 2A). The ten networks with the lowest estimate at one neighbour share no member with the ten lowest at eighty, and the second group makes twice the errors of the first. Both signs are far outside a permutation null (Table 9, Appendix B), and on the original training images the reversal is unchanged (Table 2). On this design only about three images in five have two or more classmates in the probe (Appendix B), so one neighbour reads a mixture of regimes, in which the class regime is the positive one.

The reversal is not a difference between architecture families. With family means removed, networks in families of three or more still run from $+0.80$ to -0.63, and two ResNet recipe groups reverse on their own (Figure 2B). Architecture can be held almost fixed. Among 14 checkpoints of ResNet-101 and ResNet-152 trained with the same seven recipes, the correlation runs from $+0.85$ to -0.91 ($+0.81$ to -0.87 with the error convention fixed in advance), and the ResNet-18/34 recipes do the

Table 2: **The published design on its original images.** Correlation of the log estimate with log top-5 error, five 2,000-image draws per network; TwoNN fit is the original trimmed fit at one neighbour (published: $+0.94$, on raw rather than log values), the other columns Equation 2 at n neighbours.

images	networks	N	TwoNN fit	n=1	n=10	n=40	n=80
training	the original fourteen	14	+0.94	+0.68	+0.80	+0.63	+0.46
training	torchvision and four VGG-bn	55	+0.86	+0.84	−0.50	−0.67	−0.70
validation	the original fourteen	14	+0.93	+0.94	+0.80	+0.54	+0.39
validation	torchvision and four VGG-bn	55	+0.86	+0.86	−0.51	−0.67	−0.69

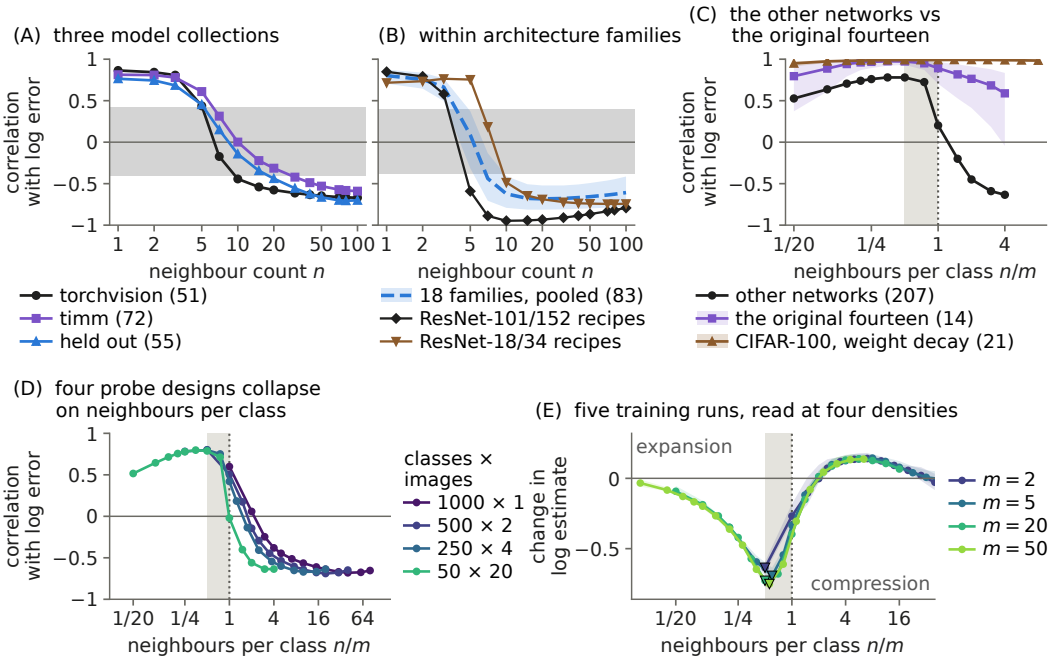

Figure 2: **Changing only the neighbour count flips the sign.** Correlation between estimated dimension and log top-5 error, against n in (A, B), as the published design fixes no class size. (A) Three model collections on the published design, read on validation images; grey, the 99% permutation band. (B) Within architecture families (blue, family-clustered 95% interval) and two ResNet recipe groups. (C) The other 207 networks, the original fourteen and a CIFAR-100 weight-decay sweep on one stratified probe (bands, 95% bootstrap intervals). (D) Four probe designs of 1,000 images. (E) Change in log estimate from epoch 48 to epoch 100 in five CIFAR-100 ResNet-18 runs, at four probe densities (median and range); the $m = 2$ cell lies outside the compression rule fixed in advance. In (C–E) the vertical grey band is the crossover band $m/2 \leq n < m$ and the dotted line $n = m$.

same (Appendix C). Nor is it a matter of scale. Within families and given parameter count, both branches keep their sign (Appendix C.3).

The fourteen networks of the original study stay positive through four times the class size on every design we tested (Figure 2C). The 15 torchvision networks with modern-recipe weights, three of them the fourteen's own ResNets retrained, flip (-0.86 at $4m$). The positive value goes with first-generation training recipes (Appendix B). With fourteen networks the check of Section 4.3 cannot resolve a sign at $4m$, but none of 100,000 random fourteen-network subsets of the other networks reaches the fourteen's value. That value is a contrast between four architecture types rather than an ordering within them. Their trimmed TwoNN fit reproduces the published value, so on them the published conclusion holds.

Two CIFAR-100 regulariser sweeps, in which the class-means term and the remainder move in opposite directions and nearly cancel, keep a positive sign at all neighbour counts (Table 10). The switch also holds for out-of-distribution error. On four shifted test sets and the relabelled validation set the correlation is positive at few neighbours and negative at many, in every case (Appendix C.4). The reversal is a property of populations that span both regimes; the switch is shown under control in Section 4.1. Tables 6 and 10 list the cohorts and boundary settings we tested and Table 31 the rules fixed before their outcomes, including the ones that failed.

Where the switch sits is set mainly by the class size. Holding the probe at 1,000 images and spreading them over fewer, larger classes, from one image in each of 1,000 classes to twenty in each of 50, gives profiles that differ against n and nearly coincide against n/m (Figure 2D). They are positive below about half a class and negative beyond three classes' worth. Over 20 designs the crossing doubles with m at a fixed number of classes and rises slowly with the number of classes (Appendix D).

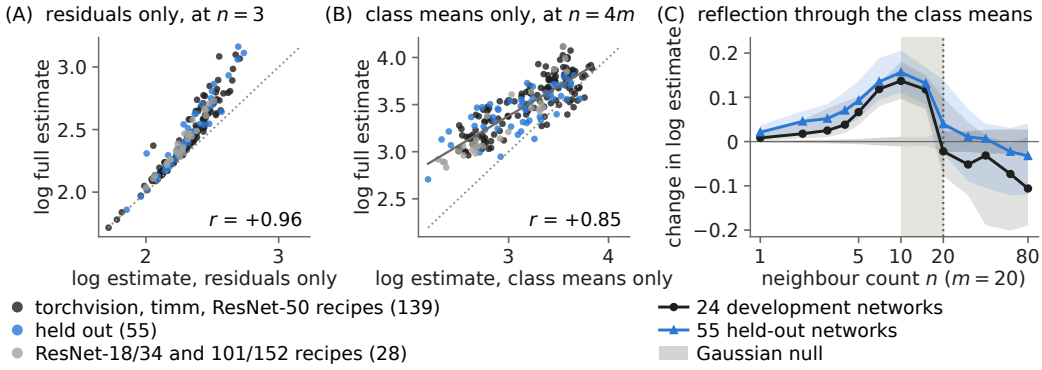

Figure 3: **Deleting residuals or class means reproduces the estimate on either side of the class size; class summaries do not pin it down.** (A) Residuals only against the full estimate at $n = 3$ and (B) class means only against the full estimate at $n = 4m$ (222 networks, $K = 100$, $m = 20$; dotted, identity). (C) Change in $\log \widehat{d}_n$ when each point is reflected through its class mean, on 24 development networks (used to develop the analysis) and 55 held-out ones (read after its rules were fixed); medians and interquartile bands; pale band, Gaussian classes with the same covariances; grey, the crossover band.

3 Why the estimate switches, and what each regime measures

We show this three ways: by deleting one geometry or the other, with synthetic data that keeps only each class's mean and each point's distance to it, and with two short theorems.

The two deleted estimates run the same estimator on parts of the same probe:

$$\underbrace{\widehat{d}_n^{\,\mathrm{res}} = \widehat{d}_n\big(\{r_i\}_{y_i=c}\big) \text{ pooled over classes,}}_{\text{residuals only}} \qquad \underbrace{\widehat{d}_n^{\,0} = \widehat{d}_n\big(\{\mu_{y_i}\}_i\big)}_{\text{class means only}}. \tag{3}$$

The first runs inside each class's residual cloud, where no neighbour from another class exists. The second runs with all residuals deleted, so each class contributes m coincident copies of its mean and the estimate is defined only for $n \geq m$.

3.1 Deleting residuals or class means

Across 222 networks the residuals-only estimate orders networks almost exactly as the full estimate does below half the class size (r above 0.9; Figure 3A), and reads the shape of each class's residual cloud, not its size. In the crossover band the estimate tracks a statistic with no estimator in it, the share of a query's $m/2$ nearest neighbours from other classes ($r = +0.94$; Appendix D). The share tracks the branch but does not cause it: with one image per class, where no neighbour is a classmate, the branch is still there (Section 5). From about twice the class size the class-means-only estimate does the same ($r = +0.85$ at $4m$; Figure 3B), and tracks error more strongly than the full estimate does (Table 17).

3.2 Synthetic data, and an intervention on real features

Replace each feature by

$$\tilde{x}_i = \mu_{y_i} + \|r_i\| \frac{Lg_i}{\|Lg_i\|}, \qquad g_i \sim \mathcal{N}(0, I), \qquad LL^\top = \Sigma_W, \tag{4}$$

synthetic data that keeps each class's mean and each point's distance to it, with residual directions drawn at random and shaped by the pooled within-class covariance Σ_W. In seven populations on ImageNet and CIFAR-100, including self-supervised encoders, the synthetic data's correlation curve tracks the real one (r from 0.82 to 0.99; Figure 4A–D), even in the one population that never reverses.

To test whether those distances are needed, we intervene on real features, keeping class means and residual directions and changing only each point's distance to its class centre (Figure 4E). Equalising

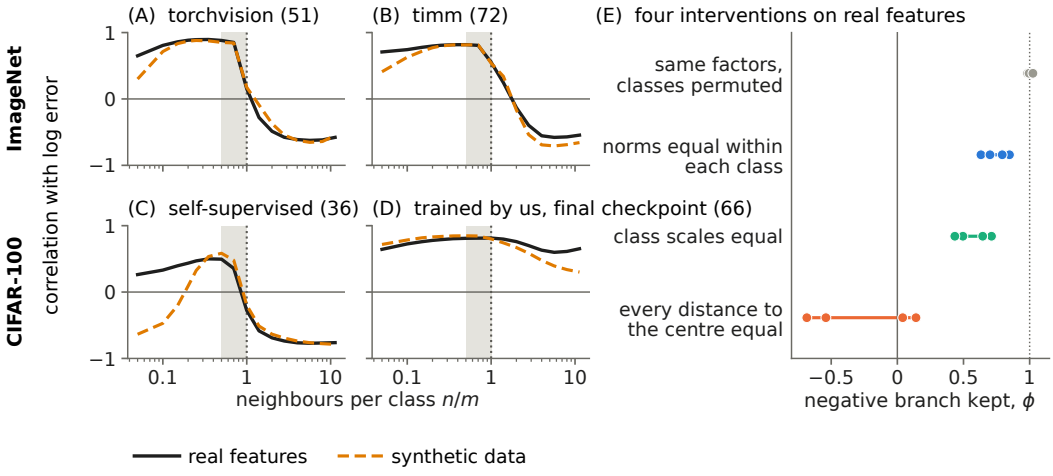

Figure 4: **Synthetic data keeping only class means and each point's distance to them reproduces the curve; equalising those distances removes the negative branch.** (A–D) Correlation with log error, real (black) and synthetic (orange, dashed; Equation 4), on ImageNet (A, B) and CIFAR-100 (C, D). (E) Share of the negative branch kept, $\phi = r_{\text{after}}/r_{\text{before}}$ at the most negative neighbour count; each row's dots are torchvision and timm at $m = 2$ and 20.

norms within each class, or the scale of each class, keeps much of the negative branch; equalising all distances to the centre removes it in all four designs and on timm turns the correlation positive, while the same scale factors applied to randomly permuted classes leave it intact. Equalising also shifts each class mean slightly. With the probe points re-centred so that each class mean stays where it was, the stricter test we fixed in advance, the removal survives at twenty images per class but not at two (Appendix E).

3.3 What moves the estimate, and what does not pin it down

Two questions decide what the estimate can be trusted to mean. What moves it, and what fails to pin it down?

Theorem 1 (what moves the estimate). *Shrink every residual by a factor t, so $x_i(t) = \mu_{y_i} + t\, r_i$, and let $\widehat{P}_j(t)$ be the share of points whose jth nearest neighbour is a classmate. For every labelled cloud and almost every t, the log gap of Equation 2 satisfies*

$$\frac{dL_n}{d\log(1/t)} = \widehat{P}_n(t) - \widehat{P}_{2n}(t) + R(t),$$

where R collects pairs in different classes and vanishes as classes separate (Appendix A.2 states and proves the same identity for every zero-sum statistic of the neighbour distances). The estimate responds to within-class scale through the classmate share of its two shells.

When classes are fully separated, with both shells inside the class the two shares are equal and the estimate does not move; once the outer shell has left, it falls; once both have left, it converges to the class-means-only value (Corollary 1). On real networks the classmate term is the larger of the two near the class size, and the cross-class term overtakes it only at four times the class size, where the response has faded (Appendix A.2). Under complete separation the log gap in the crossover band splits into within-class scale, the distance to the nearest other class mean and residual shape, so moving class means towards a simplex should raise it. With residuals contracted it does, in all 24 networks tested (Appendix A.3).

For the second question, reflecting each point through its class mean leaves the usual class summaries unchanged and moves the estimate anyway.

Theorem 2 (what does not pin it down). *Reflecting every point through its class mean preserves every class mean, every within-class distance and every class covariance, hence NC1 and NC2; a*

Table 3: **Published readings, re-read**: the probe each used, the regime it sits in, and what changing the neighbours per class shows (Appendices C and F).

published reading	probe as published	regime	what changing n/m shows
last-layer dimension predicts ImageNet error (Ansuini et al., 2019)	$n = 1$; $m \approx 2$	mixed	$+0.80 \to -0.66$ from 1 to 80 neighbours (178 networks); the original fourteen stay positive
dimension rises then falls with depth (Ansuini et al., 2019)	$n = 1$; one class at a time	class	within one class the profile falls by 0.46 of its peak, on a pooled probe by 0.82
dimension falls late in training, read as compression	one n; m set by the sample	depends on n/m	the same checkpoints compress near $m/2$ and expand near $4m$ (20 runs)
relative dimension minima mark the most semantic layers (Valeriani et al., 2023)	$n = 1$, $n/m \ll 1$	class	read at $n/m = 2$ it costs a median 36 points of probe accuracy (32 encoders)
a dimension peak over language-model layers (Cheng et al., 2025)	one n per model and corpus, from a scale scan; unlabelled text	none defined	the peak layer moves with n/m (6 of 7 models); across layers the sign inverts

second reflection preserves every logit. Any predictor of the estimate built from those summaries errs by at least half the reflection's shift on one of the two clouds (Appendix A.5).

The reflection moves the estimate far more than it moves Gaussian classes with the same covariances, whose expected shift is zero (Figure 3C; Proposition 2). Across 222 networks the reflection shifts the estimate at four times the class size by 0.18 in log units (root mean square). Half of that shift is the floor Theorem 2 puts on any prediction from the summaries across the paired clouds. It is 30% of the estimate's spread across the original networks there, against 10% at one neighbour, so the limit matters most where the estimate reads the layout of classes. On the 55 held-out networks the median absolute shift is 0.08, below the gate fixed in advance but well above the Gaussian null. The logit-preserving reflection moves the estimate in the same direction in all 44 networks tested (Appendix G). The class-mean reflection leaves NC1, NC2, CDNV, the spectral statistics and both deleted estimates unchanged, and moves every other neighbour statistic, 1-NN and nearest-class-mean accuracy included (Table 30).

Once the residuals are gone, the estimate is set by the relative gaps between each class mean's nearest neighbours among the other means, and grows as those means become more evenly spaced (Proposition 1, Appendix A.4). So the negative branch means that better networks space their class means more evenly. Within architecture families the class-means-only estimate tracks error at -0.81, more strongly than the three collapse statistics we tested (Table 18). The remainder is not negatively related to error in any ImageNet group (Table 19), so evenness of class spacing is a correlate of accuracy that the rest of the estimate does not carry.

4 What the switch changes

Three published readings change meaning once the neighbours per class are known, and one rule makes any of them interpretable (Table 3).

4.1 Compression during training

Read one training run's penultimate features at two checkpoints and write $\Delta_n = \log \widehat{d}_n(e_{100}) - \log \widehat{d}_n(e_{48})$. In 20 pre-registered CIFAR runs, drawing m images per class puts the strongest compression near $n = m/2$, in each seed of each included cell, while the same checkpoints expand near $n = 4m$ in each cell whose grid reaches it (Figure 2E). The neighbour count of strongest compression grows in proportion to m (fitted exponent 1.04; Table 15), and the pattern recurs at three further class sizes and in two further settings (Appendix C.7). Across the checkpoints of 51 CIFAR-100 runs, the expansion at $4m$ accompanies the class means approaching equal spacing in every run, as Proposition 1 predicts, but the fall at $m/2$ does not track within-class shrinkage (Appendix C.8). A fall in the estimate at one neighbour count is therefore not evidence that a representation compressed.

4.2 Depth profiles, layer choice and language models

The published depth profile was computed within single categories, so its analogue here is the residuals-only estimate; read on a pooled probe the profile falls nearly twice as far, and the extra fall

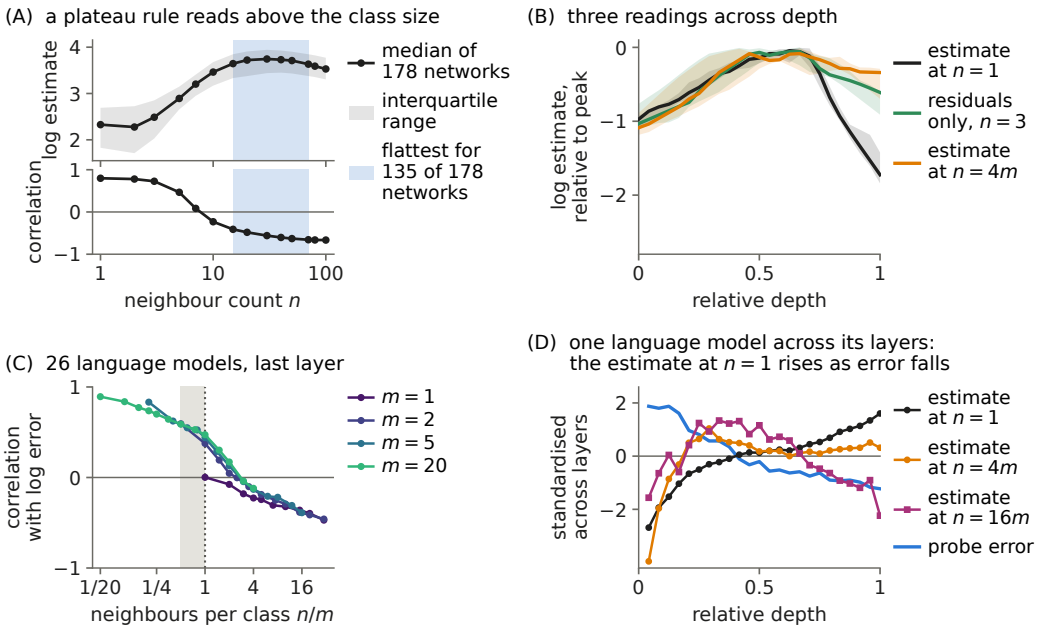

Figure 5: **Three consequences: a plateau rule reads the layout of classes, a layer profile depends on the reading, and language models switch across models and invert across layers.** (A) Published design, 178 networks. (B) Thirteen torchvision classifiers, relative to each one's peak. (C) 26 language models, last layer, banking77: correlation across models between the estimate and probe error at four probe sizes m. (D) The clinc model (Pythia-1.4B, $m = 5$) with the median across-layer correlation at $n = 4m$; z-scores.

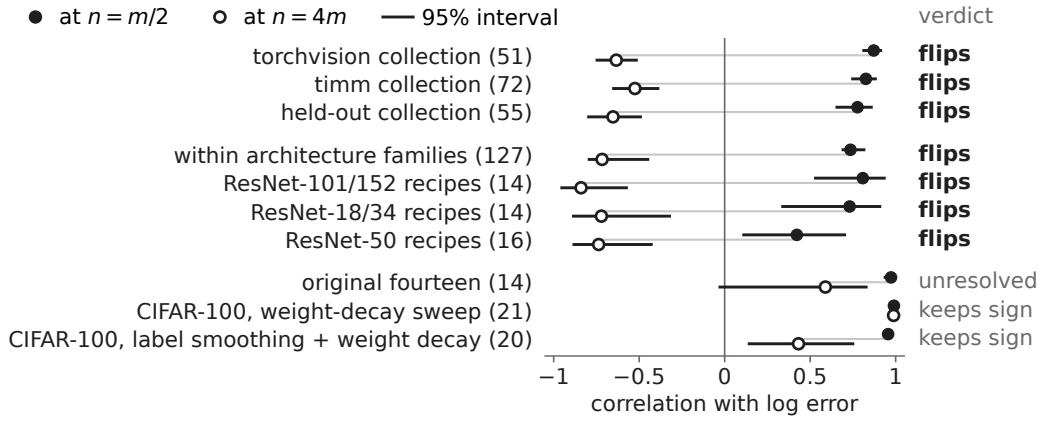

Figure 6: **Reading at two neighbour counts tells the regimes apart.** Correlation with log error at $n = m/2$ (filled) and $n = 4m$ (open), with 95% intervals, for ten groups; at right, the verdict.

arrives late, as classes form (Figure 5B; Table 24).

Choosing a layer where the estimate is lowest turns on the same knob. In 32 vision encoders, choosing it at twice the class size rather than half picks a worse layer in 24 and a better one in none, at a median cost of 36 points of probe accuracy against the best block (Appendix F). An automatically chosen neighbour count can land in the layout regime unannounced. A simple plateau rule lands at or above the class size for most of the 178 networks on the published design, where the correlation with error is negative (Figure 5A; Appendix F.3; Table 27 places eleven published designs on this axis).

Language models behave in the same way across models and the opposite way across layers. Across 26 models at their last layer, on the intent sets banking77 and clinc (77 and 150 intents; error of a

linear probe), the correlation with probe error on banking77 falls from about $+0.9$ in the class regime to near zero at four times the class size and -0.39 at sixteen times (Figure 5C). Across the layers of one model the signs invert. On clinc each sign inverts in most of the 26 models; on banking77 only the class-regime sign does (Appendix F). Deeper clinc layers have residual clouds of higher dimension and lower probe error, so there the class regime reads negative (Figure 5D).

4.3 A TWO-COUNT CHECK

A single reading cannot tell the two geometries apart, so we read twice: at $n = m/2$, the edge of the class regime, because the residuals-only estimate still reproduces the ordering there (r above 0.9 just below it), and at $n = 4m$ because both shells have left the class. Applied to ten groups (Figure 6), the check finds that seven flip sign, the original fourteen get an unresolved verdict and the two CIFAR-100 sweeps keep their sign. Among the statistics we tested on the same probe, nearest-class-mean and 1-NN accuracy track error most closely ($r \approx 0.95$ and 0.93; Table 18), and at half the class size the estimate is the closest of those that need no labels (Table 21). The class regime needs no labels either: with k-means clusters in place of classes the residuals-only estimate still reproduces it (Appendix D).

To read a nearest-neighbour dimension estimate on labelled data.
1. Report the estimator, the neighbour counts read, and the probe's K, m and class balance. **2.** Read twice, at $n = m/2$ and at $n = 4m$, keeping $2n \leq N/4$ (with fewer than about 32 classes the upper reading is unavailable; say so). If the two signs agree, the conclusion does not depend on the neighbour count; if not, report which geometry was read, not a dimension result. **3.** Report the residuals-only and class-means-only estimates beside the full one; the one the full estimate follows names the geometry. **4.** Below half the class size the estimate is a label-free measure of class shape that tracks error; from twice the class size, a class-layout statistic that adds little beyond nearest-class-mean accuracy, so for a claim about generalisation also report nearest-class-mean and 1-NN accuracy. **5.** If the neighbour count was chosen automatically, report its n/m.

5 RELATED WORK AND LIMITATIONS

That neighbour-based estimates depend on scale is well known (Levina & Bickel, 2004; Facco et al., 2017; Amsaleg et al., 2019; Denti et al., 2022; Di Noia et al., 2026), and we do not claim it is new. Concurrently, Osman et al. (2026) index a within-class/global switch by the neighbour count relative to class occupancy in language models and derive its location; on labelled text our two deleted estimates reproduce those regimes (Appendix B). That work does not reverse a published conclusion on fixed vision networks and images, establish the two regimes by intervention, or give the two theorems of Section 3.3. Schulte & Rügamer (2026) show that common estimates need not track a true dimension; we name the two quantities they track on labelled data. Recent work relates intrinsic dimension to accuracy across language-model layers (Joshi et al., 2025), to compositionality over training (Lee et al., 2025), to self-supervised probe accuracy (Mordacq et al., 2026) and, at an explicit neighbourhood size, to generalisation (Yu et al., 2026); on labelled data such readings should report the neighbours per class. Neural collapse (Papyan et al., 2020; Kothapalli, 2023) describes the class means the layout regime reads. Elmoznino & Bonner (2024) report the reverse association, higher dimension in better models of visual cortex, with a spectral measure that has no neighbour count.

Five limitations. All cross-network results are correlations. The crossing scales with the class size but is fitted, not derived. The reversal is absent on the original fourteen, 64-pixel ImageNet training, one ViT family and three CIFAR-100 cohorts (Table 10). With one image per class the positive branch remains, and its source is open. Parameter count explains part of one branch's size but neither sign.

6 CONCLUSION

On labelled data the estimate joins two measurements near the class size, the shape of each class below it and the layout of classes above. On modern ImageNet collections they order networks by error in opposite directions. The published correlation is real and reads mostly the first. A neighbour count chosen by a plateau, or a training change read at one count, can read the second without saying so. Read at $m/2$ and at $4m$, report n, m and K, compute the two deleted estimates, and say which geometry the conclusion rests on.

AI USE STATEMENT

The question, the main claims and the decisions in the protocol files are ours. We used generative AI tools throughout the work: to help develop the framework and refine hypotheses, to help formulate the mathematical claims and to work out and write up their proofs, to give feedback on the experimental design and the pre-registered criteria, to implement methods and write the analysis code (including the scripts that draw the synthetic data and every figure), to clean and reformat the feature data, to help interpret results, to find and summarise related work and format references, and to edit the text for readability. Translation and qualitative analysis do not apply. We reviewed all of it. Every measured number in the paper comes from recorded per-cell measurements: most are regenerated by the scripts in the supplementary material and the rest are copied from verification records. The ledger (Appendix H) lists, for every main-text result, the rules fixed in advance with their outcomes, failures included, and names the protocol file that recorded each rule where one is included in the supplementary material. The tests in the supplementary material check Corollary 1 and Theorem 1 numerically, and we checked every reference against its source. We take responsibility for the final content.

REPRODUCIBILITY STATEMENT

Every measured number in the paper comes from recorded per-cell measurements; most are set as macros by the manifest script in the supplementary material and the rest are copied from verification records into one marked block that it reads (Appendix B.6). The supplementary material also lists model identifiers, weight sources, probe draws and seeds, and, in the ledger, the protocol file that fixed each pre-registered rule where one is included (Appendices B.6 and H). Recomputing a measurement from features requires the listed checkpoints and the extraction hook of Appendix B.2. The reading rule is implemented as `tool/read_id.py`. Code, manifests and protocol files are provided as supplementary material.

REFERENCES

Laurent Amsaleg, Oussama Chelly, Michael E Houle, Ken-ichi Kawarabayashi, Miloš Radovanović, and Weeris Treeratanajaru. Intrinsic dimensionality estimation within tight localities. In *SIAM International Conference on Data Mining (SDM)*, pp. 181–189, 2019.

Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. In *NeurIPS*, 2019. arXiv:1905.12784.

Emily Cheng, Diego Doimo, Corentin Kervadec, Iuri Macocco, Jade Yu, Alessandro Laio, and Marco Baroni. Emergence of a high-dimensional abstraction phase in language transformers. In *International Conference on Learning Representations (ICLR)*, 2025.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *International Conference on Learning Representations (ICLR)*, 2018.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

Francesco Denti, Diego Doimo, Alessandro Laio, and Antonietta Mira. The generalized ratios intrinsic dimension estimator. *Scientific Reports*, 12:20005, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019.

Antonio Di Noia, Iuri Macocco, Aldo Glielmo, Alessandro Laio, and Antonietta Mira. Scale adaptive and robust intrinsic dimension estimation via optimal neighbourhood identification. *Scientific Reports*, 16(1):18559, 2026. doi: 10.1038/s41598-026-48005-4.

Diego Doimo, Aldo Glielmo, Alessio Ansuini, and Alessandro Laio. Hierarchical nucleation in deep neural networks. In *Advances in Neural Information Processing Systems*, 2020.

Diego Doimo, Alessandro Serra, Alessio Ansuini, and Alberto Cazzaniga. The representation landscape of few-shot learning and fine-tuning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024. arXiv:2409.03662.

Eric Elmoznino and Michael F. Bonner. High-performing neural network models of visual cortex benefit from high latent dimensionality. *PLOS Computational Biology*, 20(1):e1011792, 2024. doi: 10.1371/journal.pcbi.1011792.

Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, 7:12140, 2017.

Tomer Galanti, András György, and Marcus Hutter. On the role of neural collapse in transfer learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=SwIp410B6aQ.

Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann LeCun. RankMe: Assessing the downstream performance of pretrained self-supervised representations by their rank. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10929–10974. PMLR, 2023. URL https://proceedings.mlr.press/v202/garrido23a.html.

Abhinav Joshi, Divyanshu Bhatt, and Ashutosh Modi. Geometry of decision making in language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2511.20315.

Chon-Fai Kam, Miloud Bessafi, and Frédéric Cadet. Symmetry without a manifold: intrinsic dimension on orbits. *arXiv preprint arXiv:2609.17926*, 2026.

Nicholas Konz and Maciej A. Mazurowski. The effect of intrinsic dataset properties on generalization: Unraveling learning differences between natural and medical images. In *International Conference on Learning Representations*, 2024.

Simon Kornblith, Ting Chen, Honglak Lee, and Mohammad Norouzi. Why do better loss functions lead to less transferable features? In *Advances in Neural Information Processing Systems*, 2021.

Vignesh Kothapalli. Neural collapse: A review on modelling principles and generalization. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=QTXocpAP9p.

Jin Hwa Lee, Thomas Jiralerspong, Lei Yu, Yoshua Bengio, and Emily Cheng. Geometric signatures of compositionality across a language model's lifetime. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 5292–5320, 2025.

Elizaveta Levina and Peter J. Bickel. Maximum likelihood estimation of intrinsic dimension. In *NeurIPS*, 2004.

Xingjun Ma, Bo Li, Yisen Wang, Sarah M. Erfani, Sudanthi Wijewickrema, Grant Schoenebeck, Dawn Song, Michael E. Houle, and James Bailey. Characterizing adversarial subspaces using local intrinsic dimensionality. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://arxiv.org/abs/1801.02613.

Julie Mordacq, Vicky Kalogeiton, and Steve Oudot. IdEst: Assessing self-supervised learning representations via intrinsic dimension. *arXiv preprint arXiv:2606.03338*, 2026.

Arwa Osman, Marco Baroni, and Iuri Macocco. Local and global regimes of geometric complexity in language model representations. *arXiv:2608.14361*, 2026.

Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *PNAS*, 117(40):24652–24663, 2020.

Phillip Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://arxiv.org/abs/2104.08894.

Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. Hubs in space: popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11:2487–2531, 2010.

Stefano Recanatesi, Matthew Farrell, Madhu Advani, Timothy Moore, Guillaume Lajoie, and Eric Shea-Brown. Dimensionality compression and expansion in deep neural networks, 2019. URL https://arxiv.org/abs/1906.00443.

Benjamin Matthias Ruppik, Julius von Rohrscheidt, Carel van Niekerk, Michael Heck, Renato Vukovic, Shutong Feng, Hsien-chin Lin, Nurul Lubis, Bastian Rieck, Marcus Zibrowius, and Milica Gašić. Less is more: Local intrinsic dimensions of contextual language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025. arXiv:2506.01034.

Rickmer Schulte and David Rügamer. Rethinking intrinsic dimension estimation in neural representations. In *AISTATS*, 2026. arXiv:2604.20276.

Yifan Tang, Qiquan Wang, Inés García-Redondo, and Anthea Monod. Topological signatures of grokking. *arXiv preprint arXiv:2605.06352*, 2026.

Lucrezia Valeriani, Diego Doimo, Francesca Cuturello, Alessandro Laio, Alessio Ansuini, and Alberto Cazzaniga. The geometry of hidden representations of large transformer models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023. arXiv:2302.00294.

Ross Wightman. Pytorch image models. GitHub repository, https://github.com/huggingface/pytorch-image-models, 2019. doi:10.5281/zenodo.4414861.

Junjie Yu, Wenxiao Ma, Chen Wei, Jianyu Zhang, Haotian Deng, Zihan Deng, and Quanying Liu. Local intrinsic dimension of representations predicts alignment and generalization in AI models and human brain. arXiv preprint arXiv:2601.22722, 2026.

A Formal statements and proofs

Supports Section 3, mainly Section 3.3.

Guide to the appendix. **A** The results of Section 3, each stated with its proof, and the contraction and the two reflections they use. **B** Data, networks and model collections. **C** The reversal and the training results in full, failures included. **D** The two deleted estimates, the class-mean gap, other estimators, probe designs and the horse race. **E** The surrogate and the interventions. **F** Layer profiles, language models, the plateau rule, published designs and layer choice. **G** Reflections on real networks. **H** The ledger, which records every rule fixed in advance and its outcome, failures included (Table 31). Analyses in every appendix are post hoc unless marked as fixed in advance.

A.1 Notation

Class c holds m_c probe points ($m_c = m$ unless stated) and $N = \sum_c m_c$. In $x_i = \mu_{y_i} + r_i$, μ_c is the empirical class mean, so the residuals of each class sum to zero. Class means are distinct, no two points coincide, and ties between equal distances are broken by a fixed index order. Table 4 lists the symbols and terms used throughout.

A.2 What moves the estimate: Theorem 1 in general form

Write a zero-sum statistic of the neighbour distances as $F_w = \frac{1}{N}\sum_i \sum_{j\le q} w_j \log D_{i,(j)}$ with $\sum_j w_j = 0$. The log gap L_n of Equation 2 is the case $w_{2n} = +1$, $w_n = -1$. Shrink every residual, $x_i(t) = \mu_{y_i} + t\, r_i$. Let $Z_{i(j)} = 1$ when the jth neighbour of x_i is a classmate, and let the shell occupancy $\widehat{P}_j$ be the fraction of points whose jth neighbour is a classmate. For a neighbour k of x_i in another class, $\pi_{ik} = \langle x_i - x_k, \mu_{y_i} - \mu_{y_k}\rangle / \|x_i - x_k\|^2$ is the coefficient of the projection of the pair's class-mean offset on the pair's offset. It equals one when residuals contribute nothing to the pair's distance. It is not confined to $[0, 1]$.

Table 4: Notation and terms.

$x_i = \mu_{y_i} + r_i$	point, class mean, residual	$\widehat{P}_n$	shell occupancy: share of points whose nth neighbour is a classmate
K, m, N	classes, class size, Km	$\widehat{d}_n = \kappa_n / L_n$	two-shell estimate (Eq. 2)
n, n/m	neighbour count, neighbours per class	$\widehat{d}_n^{\mathrm{res}}$, $\widehat{d}_n^{0}$	residuals-only and class-means-only estimates (Eq. 3)
$D_{i,(k)}$	distance to the kth neighbour	q_1, q_2	which nearest class means the two shells read
$a_i(\cdot)$, $b_i(\cdot)$	sorted distances to classmates and to points of other classes	G_n, $\rho_{\pm}$	class-mean gap and its bound (Prop. 1)
$I_i(k)$	points of other classes among the k nearest	A_{ij}, B_{ij}	$D_{ij}^2 = A_{ij} + B_{ij}$; B the alignment term
κ_n, L_n	$\psi(2n) - \psi(n)$, mean log gap	π_{ik}, R	projection coefficient, remainder (Thm. 1)
probe	m images from each of K classes	regimes	class regime $n < m/2$; mixture $m/2 \le n < 2m$, whose first half $m/2 \le n < m$ is the crossover band; layout regime $n \ge 2m$

Theorem 1 (what moves the estimate, general form)**.** *For every labelled cloud and almost every* $t \in (0, 1]$,

$$\frac{dF_w}{d\log(1/t)} = -\sum_{j \le q} w_j \widehat{P}_j(t) + R(t), \qquad R(t) = -\frac{1}{N}\sum_i \sum_{j \le q} w_j \big(1 - Z_{i(j)}\big)\big(1 - \pi_{i(j)}\big). \tag{5}$$

For the two-shell weights the occupancy term is $\widehat{P}_n - \widehat{P}_{2n}$, which is the form stated in Section 3.3. No separation is assumed.

Proof. Idea. Along the contraction every pairwise distance changes at a rate fixed by whether the pair shares a class, and summing these rates with the estimator's weights gives the identity.

Write $s = \log(1/t)$ and, for a pair (i, k), $v_{ik}(t) = x_i(t) - x_k(t) = \Delta\mu_{ik} + t\,\Delta r_{ik}$, with $\Delta\mu_{ik} = \mu_{y_i} - \mu_{y_k}$ and $\Delta r_{ik} = r_i - r_k$.

Step 1: one pair. Wherever $v_{ik} \neq 0$, the log distance of the pair moves at the rate

$$\frac{d\log\|v_{ik}\|}{ds} = -\frac{t\langle v_{ik}, \Delta r_{ik}\rangle}{\|v_{ik}\|^2} = -\frac{\langle v_{ik}, v_{ik} - \Delta\mu_{ik}\rangle}{\|v_{ik}\|^2} = -(1 - \pi_{ik}), \tag{6}$$

which is -1 for classmates, since then $\Delta\mu_{ik} = 0$ and both points carry the factor t. For a pair in different classes it is positive when $\pi_{ik} > 1$, that is, when the residual offset opposes the class-mean offset strongly enough.

Step 2: order statistics. Each $\|v_{ik}(t)\|^2$ is a quadratic in t, so two of them either agree identically or cross at most twice. Each vanishes at most once, because no two points coincide. Finitely many t therefore split $(0, 1]$ into intervals, on each of which $D_{i,(j)}(t) = \|v_{ik}(t)\|$ for one fixed neighbour $k = i(j)$ (ties are broken by index). On each interval, summing Step 1 with the weights gives

$$\frac{dF_w}{ds} = -\frac{1}{N}\sum_i \sum_j w_j \Big[Z_{i(j)} + \big(1 - Z_{i(j)}\big)\big(1 - \pi_{i(j)}\big)\Big]. \tag{7}$$

Step 3: occupancy. Since $\frac{1}{N}\sum_i Z_{i(j)} = \widehat{P}_j$, the bracket splits into the occupancy term and the remainder,

$$\frac{dF_w}{ds} = -\sum_j w_j \underbrace{\frac{1}{N}\sum_i Z_{i(j)}}_{\widehat{P}_j} \underbrace{- \frac{1}{N}\sum_i \sum_j w_j \big(1 - Z_{i(j)}\big)\big(1 - \pi_{i(j)}\big)}_{R}, \tag{8}$$

which is Equation 5. □

Because $\sum_j w_j = 0$, the same derivative can be written over neighbours in other classes alone, $dF_w/ds = \frac{1}{N}\sum_i \sum_j w_j(1 - Z_{i(j)})\,\pi_{i(j)}$. For every such pair $|1 - \pi_{ik}| \le t\|\Delta r_{ik}\|/\|v_{ik}\|$ and $\|v_{ik}\| \to \|\Delta\mu_{ik}\| > 0$, so R vanishes as $t \to 0$. If labels are exchangeable given the neighbour ordering, $\mathbb{E}[\widehat{P}_j] = \sum_c \frac{m_c}{N}\frac{m_c - 1}{N-1}$ for every j. The occupancy term then has expectation zero, and what it carries is the association between labels and geometry.

The two terms on real networks. Every residual is shrunk by 10% on each of 194 ImageNet networks. The measured change in $\log \widehat{d}_n$ is compared with two predictions from Equation 5, the occupancy term alone and the occupancy term plus the remainder R (Table 5).

Table 5: **Near the class size the occupancy term is the larger of the two terms, and adding the remainder brings the prediction close to the measured change.** A 10% contraction of every residual, 194 ImageNet networks, stratified probe ($K = 100$, $m = 20$). Changes are in $\log \widehat{d}_n$ and errors are per network. Every entry is a median over networks.

	$n = m$	$n = 2m$	$n = 4m$
measured change	-0.164	-0.077	-0.008
predicted by the occupancy term alone	-0.233	-0.074	-0.025
predicted by the identity (occupancy term and remainder)	-0.161	-0.069	-0.003
error of the occupancy term alone	0.066	0.024	0.028
error of the identity	0.008	0.009	0.009
$\lvert R\rvert$ relative to the occupancy term	0.31	0.32	1.16

A.3 Three regimes under separation: Lemma 1 and Corollary 1

Lemma 1 (intruder identity)**.** *For every labelled cloud, every query i and every neighbour count k, $D_{i,(k)} = \max\big(a_i(k - I_i(k)),\ b_i(I_i(k))\big)$, with $a_i(0) = b_i(0) = -\infty$. Every estimator built from a query's neighbour distances is therefore a function of its classmate list, its list of points in other classes and its intruder counts.*

Proof. Fix i and k. The k nearest neighbours of x_i contain $I_i(k)$ points of other classes and $k - I_i(k)$ classmates. They are the $I_i(k)$ nearest points of other classes and the $k - I_i(k)$ nearest classmates, because a closer point of either kind would itself be among the k nearest. The kth neighbour is the farthest of the k, hence the farther of $a_i(k - I_i(k))$ and $b_i(I_i(k))$. The convention $a_i(0) = b_i(0) = -\infty$ covers a neighbourhood with no classmate or no point of another class. Nothing about the cloud is used. □

Corollary 1 (three regimes under complete separation)**.** *If every query's classmates are nearer than every point of another class, the occupancy term is $\mathbf{1}\{n \le m - 1 < 2n\}$. As residuals shrink by a factor $\varepsilon \to 0$, $\widehat{d}_n$ (i) is unchanged when $2n \le m - 1$, (ii) falls to zero as $1/\log(1/\varepsilon)$ when $n \le m - 1 < 2n$, and (iii) tends to the class-means-only estimate $\widehat{d}^{\,0}_n$ for $n \ge m$ (generically).*

Separation, query by query. Complete separation means that for every query the largest classmate distance is below the smallest distance to another class, $a_i(m_{y_i} - 1) < b_i(1)$. By Lemma 1 this is $I_i(k) = \max(0, k - m_{y_i} + 1)$ for every k. The first $m_{y_i} - 1$ neighbours are then classmates and every later one is in another class, so

$$D_{i,(k)} = a_i(k) \quad (k \le m_{y_i} - 1), \qquad D_{i,(k)} = b_i(k - m_{y_i} + 1) \quad (k \ge m_{y_i}).$$

Under the contraction $x_i(\varepsilon) = \mu_{y_i} + \varepsilon r_i$, $\varepsilon \in (0, 1]$, classmate distances are multiplied by ε. A distance between classes, $\|\mu_{y_i} - \mu_{y_j} + \varepsilon(r_i - r_j)\|$, moves by at most $2(1-\varepsilon)\bar{r}$, with $\bar{r} = \max_i \|r_i\|$. Separation holds at ε whenever $\varepsilon\, a_i(m_{y_i} - 1) + 2(1 - \varepsilon)\bar{r} < b_i(1)$ for every i. This condition is affine in ε and holds at $\varepsilon = 1$, so it holds near $\varepsilon = 1$, and on all of $(0, 1]$ if also $b_i(1) \ge 2\bar{r}$ for every i. Whatever the cloud at $\varepsilon = 1$, separation also holds for every small enough ε. As $\varepsilon \to 0$ classmate distances vanish, while distances between classes tend to $\|\mu_{y_i} - \mu_{y_j}\|$, which is positive because the class means are distinct.

Proof of Corollary 1. Idea. Under separation the lemma fixes which list each shell reads. The contraction multiplies classmate distances by ε and moves distances to other classes towards distances between class means.

Take equal class sizes m; unequal sizes are treated below. Write $a_i^r(k) > 0$ for the classmate distances at $\varepsilon = 1$, so that $a_i(k) = \varepsilon\, a_i^r(k)$. Let $\beta_{i,p} > 0$ be the pth smallest entry of the list of distances from μ_{y_i} to the other class means, each repeated m times. Each distance to another class is within $2\varepsilon\bar{r}$ of the corresponding distance between class means, and order statistics are 1-Lipschitz in the maximum norm, so $b_i(p) = \beta_{i,p} + O(\varepsilon)$.

Case (i): $2n \le m - 1$, both shells among classmates. The two read distances are the nth and $2n$th neighbour distances of r_i inside its class's residual cloud, scaled by ε, which cancels. At every ε at which separation holds,

$$L_n = \frac{1}{N}\sum_i \log\frac{a_i(2n)}{a_i(n)} = \frac{1}{N}\sum_i \log\frac{a_i^r(2n)}{a_i^r(n)}, \qquad \text{so} \qquad \widehat{d}_n = \widehat{d}_n^{\,\mathrm{res}}.$$

Case (ii): $n \le m - 1 < 2n$, the outer shell has left the class. The inner shell is the classmate distance $\varepsilon\, a_i^r(n)$, and the outer shell is $b_i(2n - m + 1) = \beta_{i,2n-m+1} + O(\varepsilon)$, so

$$L_n = \log\frac{1}{\varepsilon} + \frac{1}{N}\sum_i \log\frac{\beta_{i,2n-m+1}}{a_i^r(n)} + O(\varepsilon), \qquad \text{so} \qquad \widehat{d}_n = \frac{\kappa_n}{L_n} \to 0 \text{ as } \frac{1}{\log(1/\varepsilon)}.$$

Case (iii): $n \ge m$ and $2n \le N - 1$, both shells outside the class. Both read distances are to other classes and converge to entries of the same list. In the class-means-only cloud a query has its $m - 1$ own copies at distance zero followed by this list, so its nth and $2n$th neighbour distances are the same two entries. If the two entries differ, the log gap L_n^0 of that cloud is positive, and

$$D_{i,(n)} \to \beta_{i,n-m+1}, \quad D_{i,(2n)} \to \beta_{i,2n-m+1}, \qquad \text{so} \qquad L_n \to L_n^0, \quad \widehat{d}_n \to \frac{\kappa_n}{L_n^0} = \widehat{d}_n^{\,0}.$$

Position p of the list lies in the block of the $\lceil p/m \rceil$th nearest other class mean, so the two shells read the q_1th and q_2th nearest class means of Proposition 1. □

Band additivity. Case (ii), the crossover band, can be written out. The outer shell reads position $2n - m + 1 \le m - 1$ of the list, which lies in the block of the class mean nearest to μ_{y_i}, so $\beta_{i,2n-m+1} = \Delta_{y_i}$ with $\Delta_c = \min_{b \ne c} \|\mu_c - \mu_b\|$. With m points in every class the log gap is then a sum of three terms,

$$L_n(\varepsilon) = \log\frac{1}{\varepsilon} + \frac{1}{K}\sum_c \log\Delta_c - \frac{1}{N}\sum_i \log a_i^r(n) + O(\varepsilon), \tag{9}$$

with weights $(+1, +1, -1)$ on within-class scale, the log distance to the nearest other class mean and residual shape at rank n. The sum is additive in L_n, not in $\log\widehat{d}_n$, and exact only in the limit. It predicts that moving the class means towards a regular simplex at fixed energy, which raises the smallest distances between class means, raises L_n in the band and so lowers the estimate there. We tested this on stored runs of the 24 development networks ($K = 100$, $m = 20$), comparing L_n at the start and 99% of the way along a path that moves the class means to a regular simplex at fixed energy. With residuals contracted to a tenth, L_n rises in the band in 24 of 24 networks at $n = 10$ and 24 of 24 at $n = 15$, which passes the criterion fixed in advance (a median rise at both counts, at least 18 of 24 rising). Outside the band the median change is zero at $n = 5$ ($+0.00$) and negative at $n = m$ (-0.09), where both shells have left the class. Results at the two larger residual scales are in Table 31.

Scope. Case (i) uses only that both read distances are classmate distances, so it holds for any scale-invariant statistic of a query's first $k \le m - 1$ neighbour distances, such as the Levina–Bickel estimator, GRIDE at any pair of neighbour counts below m, or a correlation-dimension estimate at neighbour scale. Cases (ii) and (iii) are about the two-shell statistic and do not transfer to estimators that average over all neighbour counts up to some $k \ge m$. The Levina–Bickel estimator is treated under "Other estimators" in Appendix D.

Unequal class sizes. With class sizes m_c each query falls in the case set by its own class size m_{y_i}. Since L_n averages over queries, a probe with unequal classes mixes the cases at one neighbour count in the proportions m_c/N instead of switching at the average class size. In the class-means-only cloud each other class mean b is repeated m_b times, and the two shells read the weighted list $\{D_{y_i b}$ with multiplicity $m_b\}_{b \ne y_i}$ where its cumulative multiplicity first reaches $n - m_{y_i} + 1$ and $2n - m_{y_i} + 1$.

Where the class-means-only estimate does not exist. In the class-means-only cloud every query has $m-1$ copies of itself at distance zero, so for $n \le m-1$ the log ratio is $0/0$ ($2n \le m-1$) or infinite ($n \le m-1 < 2n$). Such cells are excluded, not jittered, and below m the residuals-only estimate is used instead.

Figure 7 shows the classmate shares of the two shells and the level of the estimate on the probe of Figure 1.

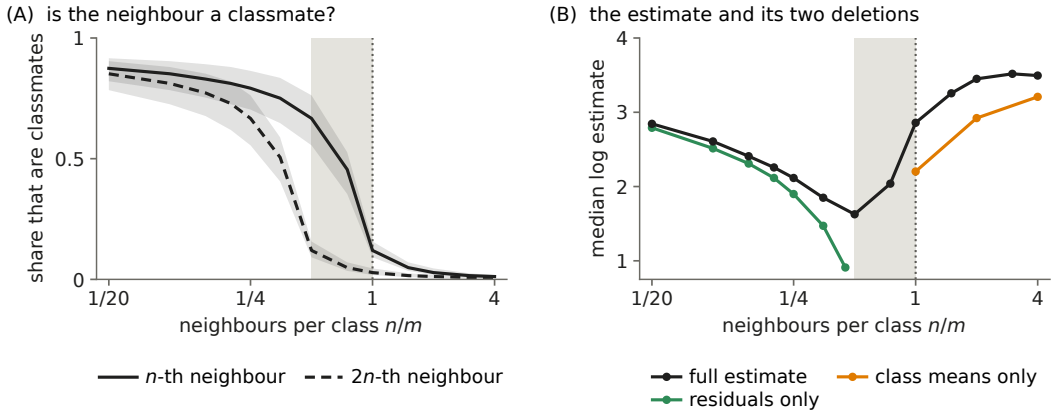

Figure 7: Classmate shares of the two shells and the level of the estimate on one probe (222 ImageNet classifiers, $K = 100$, $m = 20$). (A) Share of points whose nth and $2n$th neighbours are classmates, $\widehat{P}_n$ and $\widehat{P}_{2n}$ (medians, interquartile bands). $\widehat{P}_{2n}$ stays above zero from $n = m/2$ on because classes overlap: a point of another class can come before the last classmate. (B) The full estimate and the two deleted estimates where defined. Grey marks the crossover band, where the outer shell has left the class and the inner has not, so neither deleted estimate stands in for the full one.

A.4 The class-mean gap: Proposition 1

The class-means-only estimate. For $n \ge m$ a query's two shells read its q_1th and q_2th nearest class means, $q_1 = \lceil (n-m+1)/m \rceil$ and $q_2 = \lceil (2n-m+1)/m \rceil$. Write class mean a's distances to the others as $D_{ab} = \bar{D}_a(1+\delta_{ab})$, with $\bar{D}_a$ their mean, and let $\delta_{a,(q)}$ be the qth smallest relative deviation.

Proposition 1 (the class-mean gap, exact bound)**.** *Assume equal class sizes, distinct class means and neighbour counts $m \le n$ with $2n \le Km-1$. Let $G_n = K^{-1}\sum_a \left(\delta_{a,(q_2)} - \delta_{a,(q_1)}\right) \ge 0$, $\rho_- = \max_a \max(0, -\delta_{a,(q_1)}) < 1$ and $\rho_+ = \max_a \max(0, \delta_{a,(q_2)})$. Whenever $G_n > 0$,*

$$(1-\rho_-)\,\frac{\kappa_n}{G_n} \;\le\; \widehat{d}_n^{\,0} \;\le\; (1+\rho_+)\,\frac{\kappa_n}{G_n}. \tag{10}$$

The estimate diverges exactly when $\delta_{a,(q_1)} = \delta_{a,(q_2)}$ for every class mean a. Equidistant class means suffice but are not necessary.

The class-means-only estimate is an ordered statistic of relative distance gaps between class means, not a dimension of the data. On real networks the bound is loose (Table 18).

Proof. Idea. In the class-means-only cloud each query reads two distances from its own class mean. Written relative to that mean's average distance, their log ratio is an integral of $1/(1+u)$ over an interval of length $\delta_{a,(q_2)} - \delta_{a,(q_1)}$, and bounding the integrand bounds the gap.

Take $n \ge m$ and $2n \le Km-1$, which gives $q_2 \le K-1$.

Step 1: the log gap. In the class-means-only cloud all m queries of class a read, by Case (iii) of the proof of Corollary 1, the distances $D_{a,(q_1)}$ and $D_{a,(q_2)}$ to their q_1th and q_2th nearest other class means, so

$$L_n^0 = \frac{1}{K}\sum_a \log \frac{D_{a,(q_2)}}{D_{a,(q_1)}}.$$

Step 2: relative distances. Write $D_{ab} = \bar{D}_a(1+\delta_{ab})$, with $\bar{D}_a > 0$ the mean of a's distances. The map $\delta \mapsto \bar{D}_a(1+\delta)$ is increasing, so $D_{a,(q)} = \bar{D}_a(1+\delta_{a,(q)})$, the scale $\bar{D}_a$ cancels, and

$$L_n^0 = \frac{1}{K}\sum_a \big[\log(1+\delta_{a,(q_2)}) - \log(1+\delta_{a,(q_1)})\big] = \frac{1}{K}\sum_a \int_{\delta_{a,(q_1)}}^{\delta_{a,(q_2)}} \frac{du}{1+u}.$$

Step 3: the bound. Every selected distance is positive, so $\delta_{a,(q_1)} > -1$ for every a and $\rho_- < 1$. Since $q_1 < q_2$, each interval of integration has non-negative length, and it lies in $[-\rho_-, \rho_+]$, where $1/(1+\rho_+) \le 1/(1+u) \le 1/(1-\rho_-)$. Multiplying by the interval lengths and averaging over the K class means gives

$$\frac{G_n}{1+\rho_+} \;\le\; L_n^0 \;\le\; \frac{G_n}{1-\rho_-}.$$

Step 4: inversion. When $G_n > 0$ all three quantities are positive, $\widehat{d}_n^0 = \kappa_n / L_n^0$, and inverting gives Equation 10. Finally, L_n^0 is a mean of non-negative terms, so it vanishes, and $\widehat{d}_n^0$ diverges, exactly when every term vanishes, that is when $\delta_{a,(q_1)} = \delta_{a,(q_2)}$ for every a. Equidistant class means, with $\delta_{ab} = 0$ throughout, are one such case. □

For $K = 4$ class means at the corners of a unit square and $n = m$, both shells read distance 1 ($q_1 = 1$, $q_2 = 2$), so $\widehat{d}_n^0$ diverges although the means are not equidistant (cf. Kam et al., 2026).

A.5 What does not pin it down: Theorem 2 and Gaussian classes

Reflections. For a pair in different classes the squared distance splits as $D_{ij}^2 = A_{ij} + B_{ij}$, with $A_{ij} = \|\mu_{y_i} - \mu_{y_j}\|^2 + \|r_i - r_j\|^2$ and the *alignment term* $B_{ij} = 2\langle \mu_{y_i} - \mu_{y_j}, r_i - r_j\rangle$. Within a class $B_{ij} = 0$. Reflecting every point through its class mean, $x_i \mapsto 2\mu_{y_i} - x_i$, flips every B_{ij} and fixes every A_{ij}. It also fixes every class mean and every class covariance, hence the NC1 and NC2 statistics of neural collapse. The head-null reflection $r_i \mapsto (I - 2P_{\ker W})r_i$ fixes every logit, every class mean and every within-class distance.

Theorem 2 (what does not pin it down)**.** *Let X^- be a reflection of X that preserves a summary S, $z = \log \widehat{d}_n$, and $o = \frac{1}{2}[z(X) - z(X^-)]$. Every predictor $f(S)$ errs by at least $|o|$ on X or on X^-. For the reflection through the class means, S may be any function of the class means and class covariances. For the head-null reflection, S may be any function of the logits, the class means and the within-class distances.*

The generic identity. Let T be an involution on labelled clouds, z a statistic finite at X and TX, and S a summary with $S(TX) = S(X)$. Put $e = \frac{1}{2}[z(X) + z(TX)]$ and $o = \frac{1}{2}[z(X) - z(TX)]$, the even and odd parts of z on the pair, so $z(X) = e + o$ and $z(TX) = e - o$. A predictor $f(S)$ takes one value on the pair, and expanding $(f - e \mp o)^2$ gives

$$\tfrac{1}{2}\big([f - z(X)]^2 + [f - z(TX)]^2\big) = (f-e)^2 + o^2 \;\ge\; o^2. \tag{11}$$

Proof of Theorem 2. Step 1: the floor. The larger of two squares is at least their mean, so Equation 11 gives

$$\max\{|f - z(X)|, |f - z(TX)|\} \;\ge\; |o|,$$

and over a set of pairs the paired root-mean-square error is at least $\sqrt{\operatorname{mean} o^2}$. It remains to check that the two reflections are involutions and what they preserve.

Step 2: reflection through the class means, $Tx_i = 2\mu_{y_i} - x_i = \mu_{y_i} - r_i$. Applying it twice returns x_i. The residuals of each class sum to zero, so the class means of TX are the μ_c. The residuals $-r_i$ have the same class covariances and the same Gram matrix as r_i, so every within-class distance, Σ_W, Σ_B, the total covariance, $\mathrm{NC1} = \operatorname{tr}(\Sigma_W \Sigma_B^+)/K$ ($^+$ the pseudo-inverse) and every function of the class means, NC2 included, are unchanged. S may therefore be any function of the class means and the class covariances. With A_{ij} and B_{ij} as above,

$$D_{ij}^2(X) = A_{ij} + B_{ij}, \qquad D_{ij}^2(TX) = A_{ij} - B_{ij}, \qquad B_{ij} = 0 \text{ for classmates.}$$

Step 3: reflection in the head's null space. Let $W \in \mathbb{R}^{K\times D}$ be the classifier head, P the orthogonal projector onto $\ker W$, and $Tx_i = \mu_{y_i} + (I - 2P)r_i$. The matrix $I - 2P$ is symmetric and orthogonal

with $(I-2P)^2 = I$, so T is an involution and every within-class distance is kept. Residual means are zero, so class means are kept. Because $WP = 0$,

$$WTx_i = W\mu_{y_i} + Wr_i - 2WPr_i = Wx_i,$$

so every logit is kept, bias included, and S may be any function of the logits, the class means and the within-class distances. Class covariances become $(I-2P)\Sigma_c(I-2P)$. Their traces and spectra are kept, but their fixed-coordinate entries and NC1 generally are not. If $D \le K$ and W has full rank, $\ker W = \{0\}$ and T is the identity, so the construction needs $D > K$. □

The floor bounds paired clouds only. It is not a bound on the error of predicting z across ordinary trained networks, which do not come in reflected pairs, and TX need not be a representation any network produces.

Proposition 2 (Gaussian classes do not move on average)**.** *Let the classes be independent, class c holding m_c i.i.d. points from $\mathcal{N}(\mu_c, \Sigma_c)$, and let g replace every residual about its empirical class mean, r_i, by $Q_{y_i} r_i$, with each Q_c a fixed orthogonal matrix that commutes with Σ_c. Then gX has the same law as X, so $\mathbb{E}\log \widehat{d}_n(gX) = \mathbb{E}\log \widehat{d}_n(X)$. In particular the reflection through the class means ($Q_c = -I$) has zero expected effect for Gaussian classes of any covariance.*

Proof. Idea. For Gaussian classes the residuals are independent of the empirical mean, and a rotation that commutes with the covariance leaves their law unchanged.

Write $x_i = \mu_c + \Sigma_c^{1/2} z_i$ with z_i i.i.d. $\mathcal{N}(0, I)$, so $\bar{x}_c = \mu_c + \Sigma_c^{1/2}\bar{z}_c$ and $r_i = \Sigma_c^{1/2}(z_i - \bar{z}_c)$. Because Q_c commutes with Σ_c, it commutes with $\Sigma_c^{1/2}$, and

$$Q_c r_i = \Sigma_c^{1/2} Q_c (z_i - \bar{z}_c).$$

The deviations of an i.i.d. Gaussian sample from its mean are independent of that mean. The $(Q_c z_i)_i$ are again i.i.d. $\mathcal{N}(0, I)$, with mean $Q_c\bar{z}_c$, so $(Q_c(z_i - \bar{z}_c))_i$ has the law of $(z_i - \bar{z}_c)_i$ and is likewise independent of $\bar{z}_c$. Hence $(\bar{x}_c, (Q_c r_i)_i)$ has the law of $(\bar{x}_c, (r_i)_i)$ in each class, independently across classes, so gX has the law of X and every integrable statistic has the same expectation on both. □

The Gaussian clouds of Section 3 and Appendix G are a null whose expected shift is zero. Their median $|o|$ of 0.013 is sampling variation. A systematic shift on real networks requires non-Gaussian residuals.

B Data and networks

This appendix supports Section 2 and describes the data behind every section.

B.1 Networks and model collections

The ImageNet networks come from three model collections and two sets of ResNet training recipes. The collections are torchvision, timm with one checkpoint per architecture, and further timm classifiers held out until their test was fixed. The recipes are the ResNet-50 recipes and the recipe groups, which take seven recipes at each of ResNet-18, 34, 101 and 152 (the same seven at 101 and 152; nine distinct across 18 and 34). Their networks are analysed as a ResNet-101/152 group and a ResNet-18/34 group (Appendix C). Model identifiers and weight sources are listed in the supplementary material.

On the published design the reversal is tested in the 178 networks of the three collections and, separately, in the recipe groups. The stratified probes use all 222 head-checked networks, and the reflection through class means (Appendix G) uses 24 development and 55 held-out networks for its gate and all 222 for its floor. The layer profiles (Appendix F) use 13 torchvision classifiers: AlexNet, VGG-11/13/16/19, ResNet-18/34/50/101/152 and DenseNet-121/161/169. These are not the original fourteen, since they include three DenseNets and none of the fourteen's four VGG-bn networks. Every result for the original fourteen uses the fourteen themselves. Table 6 gives every ImageNet cohort on both designs, and Table 10 in Appendix C the settings where the reversal does not appear.

Table 6: **Every ImageNet collection among the 222 head-checked networks reverses on both designs, and the original fourteen stay positive.** Each ImageNet cohort appears once, followed by two checks of the torchvision collection and by the original fourteen. r is the correlation of $\log \widehat{d}_n$ with log top-5 error, with 224-pixel matched errors unless noted. The published design reads five uniform draws of 2,000 validation images from $n = 1$ to 80, and the stratified design takes $m = 20$ validation images from each of $K = 100$ classes, read from $n = m/2$ to $4m$. Under "fixed in advance?", "yes" means the rule was recorded before the outcome was computed; "before features" means before any feature of those networks was read. Boundary settings are in Table 10, the training runs in Table 15 and Appendix C.7, the probe arms in Table 20 and the language models in Table 25.

cohort	N	composition	fixed in advance?	published, $n = 1 \to 80$	stratified, $n = \frac{m}{2} \to 4m$
ImageNet, the 222 head-checked networks					
torchvision collection	51	torchvision classifiers (ten of the original fourteen among them)	no	$+0.86 \to -0.67$; sign change at $n = 5$–7	$+0.87 \to -0.63$
timm collection	72	one timm checkpoint per architecture	no	$+0.81 \to -0.58$; sign change at $n = 10$–15	$+0.83 \to -0.53$
ResNet-50 recipes	16	timm ResNet-50 training recipes	no	$+0.53 \to -0.74$	$+0.42 \to -0.74$
main collection	139	torchvision, timm and the ResNet-50 recipes	no	$+0.82 \to -0.64$	$+0.83 \to -0.59$
held-out collection	55	further timm classifiers, held out until their test	yes; held out	$+0.76 \to -0.71$; sign change at $n = 7$–10	$+0.78 \to -0.65$
ResNet-101/152 recipes	14	mixed-depth recipe group (Table 14)	before features	$+0.85 \to -0.91^a$	$+0.81 \to -0.84$
ResNet-18/34 recipes	14	mixed-depth recipe group (Table 14)	before features	$+0.72 \to -0.74^a$	$+0.73 \to -0.72$
pooled 178	178	torchvision, timm and held-out collections	no	$+0.80 \to -0.66$	$+0.81 \to -0.62$
within families	83	pooled 178, families with ≥ 3 members (18), family means removed	yes	$+0.80 \to -0.63$	$+0.78 \to -0.62$
within families, all	127	all 222, families with ≥ 3 members (18), family means removed	no	$+0.76 \to -0.72$	$+0.74 \to -0.72$
all 222	222	every head-checked network	no	$+0.79 \to -0.68$	$+0.80 \to -0.65$
one draw[b]	222	the same networks	no	$+0.788 \to -0.686$	
error inside the interquartile range[b]	110	networks inside the pooled interquartile error range, 14 of them from the recipe groups	no	$+0.457 \to -0.436$	
error under 10%[b]	205	networks with error under 10%	no	$+0.758 \to -0.703$	
other 207	207	the 222 without the ten original networks and five timm copies of their weights	no	$+0.77 \to -0.67$	$+0.78 \to -0.63$
Checks of the torchvision collection					
on training images	55	torchvision and the four VGG-bn, on a fixed pool of training images (Table 2)	no	$+0.84 \to -0.70$	
re-extracted	51	torchvision, features extracted again independently (Table 7)	no	$+0.868 \to -0.660$	
original fourteen	14	AlexNet; VGG-11/13/16/19 with and without batch norm; ResNet-18/34/50/101/152	question only	$+0.95 \to +0.41^c$	$+0.97 \to +0.59$

[a] Most negative count $n \geq 30$, as fixed in advance (at $n = 30$, 80; -0.82, -0.74 at $n = 80$), with 224-pixel errors (post hoc). Group pass rules and the error-range test used the first row of timm's results table, at 256–320 pixels for 27 of 88 timm and 12 of 28 recipe-group networks (Table 14). [b] One draw, errors as in the error-range test. [c] Table 2: validation draws $+0.94 \to +0.39$, training images $+0.68 \to +0.46$.

B.2 Features and their checks

A network is admitted only if its own classifier head, applied to its cached last-layer features under that checkpoint's evaluation transform, reproduces its published accuracy. Features are always taken at the input of the linear head. All 222 networks and the training-image caches pass this check. Only the four VGG-bn networks of the original fourteen, which are not among the 222, fall short. An independent re-extraction of 51 torchvision networks reproduces the sweep, and Table 7 lists every feature cache and what it was used for.

B.3 Designs and image pools

The *published* design adapts that of Ansuini et al. (2019), who drew ImageNet *training* images five times and fitted TwoNN. Here it uses five fixed uniform draws of 2,000 of the 50,000 validation images, read with the two-shell estimator at every neighbour count. A *stratified* design draws m distinct source images from each of K classes. No arm mixes two pools. The validation pool of 50 images per class serves the stratified probe ($K = 100$, $m = 20$) and the fixed-N designs. For

Table 7: Feature caches and their checks.

cache	networks	check	use
main cache, validation and training images	all 222	every head reproduces its published accuracy	every analysis
re-extraction (`weights.transforms()`, hook on the input of the final linear layer)	51 torchvision	top-1 within 0.6 points of published; agrees with the main cache at $r = +0.997$ ($n = 1$) and $+0.988$ ($n = 80$); sweep $+0.868 \to -0.660$	check only
VGG-bn features, extracted from already-cropped images with a second resize and crop	the four VGG-bn of the original fourteen	heads 1.7–2.3 points below published top-1 on validation images	the original fourteen; dropping them makes those results stronger
an earlier cache, images preprocessed differently	every network	heads at 0.20–0.27% top-1, chance	none
one ViT family	22	class-structure check only	boundary tables, Table 19 and the horse race
a second ResNet-50 recipe cache	30	class-structure check only	none as evidence

the training-image bridge (Table 2) the 55 torchvision networks were re-extracted on a fixed pool of 10,000 training images (seed 20260913), and both estimators read the same five draws. The training-image arms use a disjoint fixed pair (seed 20260914): 300 images from each of 100 classes (298 after removing exact duplicates, so m can exceed 50), and five uniform 2,000-image draws.

A uniform draw leaves class sizes unequal. Over our five draws, 2,000 images cover 864 to 879 classes, with 2.29 images per occupied class on average and 2.95 in a query's own class. Of the queries, 59% have two or more classmates, 27% one and 14% none. At one neighbour both shells of the first group can stay in the class, the outer shell of the second must leave it, and both shells of the third read other classes. So the published reading at one neighbour mixes the class regime, the crossover band and the layout regime. How it mixes them also depends on the network. Measured on 12 networks, the nearest neighbour is a classmate for 11% of queries in AlexNet and for up to 69% in the others.

B.4 The original fourteen

The original fourteen are the networks of the accuracy experiment of Ansuini et al. (2019), as listed in the legend of their accuracy figure (Table 6). Their trimmed TwoNN fit on these fourteen and on training images gives $+0.938$ here on logs, against their $+0.94$ on raw values. They stay positive through $4m$ on every design we tested (Tables 8 and 6). In the fixed-N designs their correlation drifts to about zero only far past m. Sample size does not explain this, since none of 100,000 random fourteen-network subsets of the other networks is as positive at $4m$. The positive value goes with first-generation training recipes, and we do not know why. A forecast fixed in advance from two class statistics, fitted on the other 207 networks, predicted a reversal for the fourteen. It failed.

B.5 Conventions, and the primary test of the reversal

At a fixed neighbour count $z_n = \log \widehat{d}_n = \log \kappa_n - \log L_n$, so a signed change in log dimension is minus the change in the mean log gap. Every signed change, correlation sign and direction count in this paper, including those of the two deleted estimates, is stated in z_n. The estimate $\widehat{d}_n$ is always κ_n over the query-mean log gap, never a mean of per-query inverse gaps. Correlations are Pearson, across networks, of $\log \widehat{d}_n$ with log top-5 error. The estimate is averaged over image draws first; per-draw correlations agree to 0.004. Every p for a branch of a neighbour-count sweep uses the null extreme over all scanned counts. Intervals for the deleted estimates are percentile bootstraps over networks (3,000 resamples), and for training the unit is the run. Tests other than those below are reported one by one, uncorrected.

The primary test of the reversal permutes error across networks 10,000 times in each collection and compares the correlations at both ends of the sweep with the null maximum and minimum over the 14 scanned counts. Holm's correction runs over the three collections and both branches, and the test is repeated within the architecture families of at least three members with family fixed effects (Table 9; family map in Appendix C). We fixed this test before running it, but after the raw correlations across networks had been seen. Its part across networks adds a declared null and multiplicity control. The evidence is in its within-family part.

Table 8: The original fourteen against random subsets of the other networks and against related cohorts. Stratified probe; correlations of $\log \widehat{d}_n$ with log top-5 error. In the last row, the class-means-only estimate of the fourteen lacks the error ordering that produces the negative branch in the other networks. [a]All 222 networks less the ten original-study networks and eight timm copies of torchvision weights (the other 207 less the copies of the v2 ResNet-50/101/152), read at $n = m/2$ and $4m$ with bootstrap intervals. [b]The question and its pass rule were recorded before these profiles were examined; the published-design result was already known.

comparison	result	status
the fourteen at $n = 1$, m, $4m$ (95% interval at $4m$)	$+0.80$, $+0.89$, $+0.59$ $[-0.04, +0.83]$	question fixed in advance[b]
$4m$: 100,000 random fourteen-network subsets of the other 207	none as positive as the fourteen; median -0.65	post hoc
$4m$: 2,000 random groups of fourteen from 204 networks[a]	fourteen $+0.59$; groups median -0.66, maximum $+0.20$; 82% flip, 18% unresolved, none keeps its sign	post hoc
$4m$: within the four architecture types (AlexNet, VGG, VGG-bn, ResNet); without AlexNet; without VGG-bn	-0.10; $+0.29$; $+0.76$	post hoc
$4m$: the 10 other torchvision CNNs with first-generation weights	$+0.31$	exploratory
$m/2$ and $4m$, 224-pixel errors: those 10 networks; the 15 torchvision networks with modern-recipe weights (three of them the fourteen's own ResNets retrained)	$+0.88 \to +0.31$ $[-0.30, +0.76]$; $+0.86 \to -0.86$ $[-0.95, -0.67]$	post hoc
forecast of a reversal from two class statistics (a within-class tightness and the dispersion of distances between class means) fitted on the other 207	correct in 8 of 16 sign cells, extrapolating far outside its fitted range: **failed**	fixed in advance
class-means-only estimate against error, $n = m$ and $4m$	fourteen $+0.67$, -0.03; other 207 -0.68, -0.73	post hoc

Table 9: **Both branches of the reversal pass the primary test in each collection and within architecture families.** Published design; correlation with log top-5 error from $n = 1$ to $n = 80$. Eighteen clusters are few for a cluster bootstrap. Fixed effects remove family means only, not size or recipe within a family.

part	result	reference
across networks, each collection: both branches	Holm-adjusted $p \leq 6 \times 10^{-4}$, the resolution floor	
across networks, each collection: $r_1 - r_{80}$	1.39–1.53	null 99% quantile 0.49–0.61
within architecture families (83 networks, 18 families): r, family-clustered 95% intervals (5,000 resamples)	$+0.80$ $[+0.70, +0.87] \to -0.63$ $[-0.78, -0.45]$	within-family permutation $p \leq 10^{-4}$, both branches
within families, negative branch: jackknife over families; family × collection fixed effects	$[-0.81, -0.33]$; -0.77	
pooled 178, published design: the ten lowest estimates at $n = 1$ and at $n = 80$	no network in common; median top-5 error 3.6% against 7.1%	post hoc
every test whose rule was fixed before its outcome and that bears on the reversal or the training change	9 of 14 reject at family-wise 0.05 after Holm	grouping chosen after the outcomes were known

B.6 Reproduction

Most numbers in the text are macros in `numbers/`, produced from the per-cell CSVs by the manifest scripts in the supplementary material or, for a small set of values, copied from verification records into one marked block. The rest, in a few tables and sentences, are copied from tables and records that those scripts generate, or quoted from the cited papers. An independent re-run of the main manifest reproduced every macro the paper uses. Recomputing a CSV needs features extracted from the listed checkpoints. The reading rule is runnable. For given features, labels and neighbour counts, `tool/read_id.py` returns $\widehat{d}_n$, the residuals-only and class-means-only estimates and the shell occupancies $\widehat{P}_n$, $\widehat{P}_{2n}$. It also evaluates the two sides of Theorem 1. Its tests check the separated limits of Corollary 1 and the identity against finite differences on synthetic clouds.

In language models, too, the two deleted estimates reproduce the regimes that Osman et al. (2026) report. On labelled text (clinc), the residuals-only estimate against the one-neighbour estimate gives $r = +0.977$ over 588 model–layer units. The class-means-only estimate against the estimate at $n = 3m$ gives $+0.853$ over 26 last layers.

C THE REVERSAL AND THE TRAINING RESULTS IN FULL

This appendix supports Sections 2 and 4.1. The fixed-N designs are in Appendix D.

C.1 WHERE THE REVERSAL HOLDS AND WHERE IT FAILS

Table 6 (Appendix B) gives, for each cohort, the correlation across networks between $\log \widehat{d}_n$ and log top-5 error, from $n = 1$ to $n = 80$ on the published design and from $n = m/2$ to $4m$ on a stratified design. The reversal holds in each of the three collections. It also holds on training images, after an independent re-extraction, within architecture families and in both ResNet recipe groups. Inside the interquartile error range it is not established collection by collection. The settings where it fails, or appears only in part, are mapped in Table 10. Three further CIFAR results, all post hoc, show the switch only in part: the contrast of five ResNet-18 against five VGG-11-BN runs is negative at $n = 1$ to 3, positive near $m/2$ and negative above m on CIFAR-100 (Table 19); on CIFAR-10 it reverses above m, but through the residual term, with a class-means term indistinguishable from zero; and the CIFAR-10 training change is a slight expansion, not compression, at $n/m \leq \frac{1}{4}$.

Table 10: **The reversal fails in one ViT family, on ImageNet at 64 pixels, in the CIFAR-100 collection and with local PCA, and the original fourteen never turn negative.** Settings where the reversal is absent or partial are listed with their status; three post hoc CIFAR results are described in the text. The last column says whether the measured log estimate at $4m$ was split into its class-means term and a remainder (Table 19; for the original fourteen, Table 8). [a]Note a of Table 6 explains the first-row and 224-pixel errors.

setting	what happens	status	splits into terms?
original fourteen	positive through $4m$ in every tested design, never significantly negative ($+0.59$ $[-0.04, +0.83]$ at $4m$); the value is a contrast between four architecture types (within types -0.10)	question fixed in advance; reason post hoc	yes: class-means term absent (-0.03)
one ViT family, 22 checkpoints	fails its test for a positive correlation at one neighbour on the published design ($+0.30$, $p = 0.31$); its features passed only a class-structure check, not the head check (Appendix B)	fixed before the features were read; **failed**	yes: class-means term not significant
ImageNet at 64 px, 42 networks	negative at every neighbour count on the stratified probe (-0.69 to -0.77); a classmate is the nearest neighbour for 37% of queries	fixed before the features were read; **failed**	no
CIFAR-100 collection, 66 networks	positive at every neighbour count	post hoc	no
CIFAR-100 weight-decay sweep, 21 runs (ResNet-18; seven weight decays, three seeds)	positive at every neighbour count ($+0.99$ at $m/2$, $+0.99$ at $4m$)	post hoc	yes: large residual term
CIFAR-100 label-smoothing and weight-decay sweep, 20 runs (ResNet-18; four settings, five seeds)	weakens from $+0.96$ at $m/2$ to $+0.43$ $[+0.13, +0.76]$ at $4m$, never negative	post hoc	yes: large residual term
each collection in its interquartile error range[a]	the negative branch is marginal on torchvision ($p = 0.049$); timm fails with first-row errors ($p = 0.23$), passes unadjusted with 224-pixel errors ($p = 0.026$) but not after Holm adjustment ($p = 0.07$); so the error-range confound is not excluded per collection	pre-registered; 224-pixel values post hoc	no
local PCA in place of the estimator	never reverses	fixed in advance; predicted to reverse; **failed**	not applicable

C.2 ARCHITECTURE FAMILIES

Networks of one architecture family are not independent replicates, so we use two family maps. The fine map merges version suffixes of the architecture name mechanically; its families with at least three members carry the primary within-family test (Appendix B). The broad map takes families from checkpoint metadata and was recorded with its statistic before the analysis. The reversal is as strong between families as within them, and every leave-one-out keeps both branches (Table 11). These analyses address composition and leave independence open, since the model collections are not a random sample of architectures and training recipe within a family is not controlled.

Table 11: Family analyses on the published design, correlating $\log \widehat{d}_n$ with log top-5 error. The leave-one-out rows and the timm sibling families were not fixed in advance.

analysis	networks (families)	$n = 1 \to 80$	crossing
fine map, family means (between families)	178 (70)	$+0.79 \to -0.65$	
fine map, leaving out one family at a time: within-family negative branch	83 (18)	-0.71 to -0.55 at $n = 80$	
timm sibling families with at least two members (post hoc)	120 (52)	$+0.78 \to -0.77$	
broad map, every network weighted equally	178 (24)	$+0.80 \to -0.66$	$n \approx 7.7$
broad map, every family weighted equally	178 (24)	$+0.73 \to -0.66$	$n \approx 6.4$
broad map, leaving out one family at a time		positive branch $\geq +0.73$; negative branch ≤ -0.61	$n = 6.5$ to 8.6
broad map, without the 37 ViT networks (the largest single change)		$+0.82 \to -0.61$	

C.3 Network scale

Parameter counts for all 222 networks come from torchvision weight metadata and, for timm, from instantiation. Within the 18 architecture families with at least three members, pooled with family fixed effects, we compute the partial correlation of $\log \widehat{d}_n$ with log top-5 error given log parameter count. This is done on the published design ($n = 1$ and 80, 83 networks) and on the stratified probe ($n = m/2$ and $4m$, 127 networks), with family-clustered bootstrap intervals (2000 resamples). Larger networks make fewer errors (within families $r = -0.69$), and they have a lower estimate at one neighbour and a higher one at eighty. The criterion fixed in advance asked that both signs be kept, with intervals excluding zero, on both designs. It passed (Table 12). In one family, EfficientNet, parameter count and error are almost collinear, and there the partial correlation reverses.

Table 12: **Network scale does not produce either sign.** Partial correlation with log error given log parameter count, within architecture families (fixed effects), and the same without the control, with family-clustered 95% intervals from 2000 resamples shared by both columns (Table 9 gives the values without the control from 5,000).

design	neighbour count	given log parameters	without the control
published	$n = 1$	$+0.71$ $[+0.48, +0.81]$	$+0.80$ $[+0.71, +0.87]$
published	$n = 80$	-0.63 $[-0.80, -0.32]$	-0.63 $[-0.78, -0.44]$
stratified	$n = m/2$	$+0.56$ $[+0.39, +0.75]$	$+0.74$ $[+0.68, +0.83]$
stratified	$n = 4m$	-0.70 $[-0.79, -0.33]$	-0.72 $[-0.80, -0.43]$

C.4 Out-of-distribution error

Published top-5 errors on four shifted test sets (ImageNet-V2 matched frequency, ImageNet-Sketch, ImageNet-R and ImageNet-A) and on the relabelled validation set ImageNet-ReaL come from timm's results tables, at each model's evaluation resolution. They are joined by exact model name to the networks of the two designs (150 on the published design, 194 on the stratified probe); torchvision networks enter only where timm lists the same weights. We correlate the log estimate with log out-of-distribution error. A second statistic is the partial correlation of the class-means-only estimate at $4m$ with that error given log in-distribution error, with family-clustered intervals. The report and a new-finding bar were fixed in advance, the bar asking for a partial correlation with $|r| \geq 0.3$ and an interval excluding zero on at least two sets, in the same direction. The switch holds on every test set and both designs, and all 20 family-clustered intervals exclude zero (Table 13 shows the correlations). Given in-distribution error, the class-means-only estimate adds nothing detectable about out-of-distribution error except on ImageNet-A. The bar failed, and the layout reading is not a separate robustness signal.

C.5 The two ResNet recipe groups

Seven training recipes at each of four ResNet depths form two groups, the ResNet-101/152 recipes and the ResNet-18/34 recipes, fixed with their neighbour counts and pass rule before any of their features was read. Each group spans two depths and recipe tags recur across depths (all seven in

Table 13: **The switch holds on four shifted test sets and the relabelled validation set.** Correlation of $\log \widehat{d}_n$ with log top-5 error on each test set. The last column, with 95% family-clustered intervals, is the partial correlation of the class-means-only estimate at $4m$ with that error given in-distribution error.

	published design		stratified probe		
test set	$n = 1$	$n = 80$	$n = m/2$	$n = 4m$	partial, class means $4m$
ImageNet-V2	$+0.79$	-0.67	$+0.80$	-0.66	$-0.08\ [-0.23, +0.12]$
ImageNet-Sketch	$+0.74$	-0.61	$+0.73$	-0.55	$+0.07\ [-0.13, +0.21]$
ImageNet-R	$+0.67$	-0.56	$+0.63$	-0.50	$+0.17\ [-0.02, +0.30]$
ImageNet-A	$+0.81$	-0.70	$+0.80$	-0.67	$-0.35\ [-0.47, -0.18]$
ImageNet-ReaL	$+0.76$	-0.66	$+0.77$	-0.65	$-0.03\ [-0.19, +0.17]$

the ResNet-101/152 group, five of seven in the ResNet-18/34 group, which has nine distinct), so these are mixed-depth recipe groups rather than single-architecture controls. On the published design both groups pass both branches of the rule, a positive branch at $n = 1$ and a negative one at the most negative count $n \geq 30$, each against a scan-corrected permutation null (Table 14). On the stratified probe the switch is close to a step near $n = m$. With the two group means removed the value at $n = m$ stays negative, so part of the step lies between the groups and part within them.

Table 14: The two ResNet recipe groups. On the published design the endpoints are $n = 1$ and the most negative count $n \geq 30$, with recipe-clustered 95% intervals; the stratified probe ($K = 100$, $m = 20$) gives the correlation around $n = m$. The published-design endpoints use each timm network's first-row published error (256–320 pixels for 12 of the 28 networks). With 224-pixel errors they are $+0.85 \to -0.91$ and $+0.72 \to -0.74$, both still passing (Table 6, note a).

		published design (fixed in advance)		stratified probe (post hoc)		
group	N	$n = 1$	most negative	$n = 15$	$n = 20 = m$	$n = 30$
ResNet-101/152 recipes (7 distinct tags)	14	$+0.81$	$-0.87\ (n = 30)\ [-0.98, -0.51]$	$+0.659$	-0.914	-0.929
ResNet-18/34 recipes (9 distinct tags)	14	$+0.83$	$-0.86\ (n = 70)\ [-0.95, -0.65]$	$+0.867$	$+0.545$	-0.707
both, pooled	28			$+0.780$	-0.741	-0.870
both, group means removed	28				-0.47	

C.6 The training experiment

Twenty runs train ResNet-18 and VGG-11-BN on CIFAR-10 and CIFAR-100, five seeds each, with SGD (learning rate 0.1, momentum 0.9, weight decay 5×10^{-4}, batch 256), a cosine schedule over 100 epochs, and random crop and flip. Penultimate test-set features are read at epochs 48 and 100 on five class-stratified probe draws of S images, and Δ_n is each run's change in $\log \widehat{d}_n$ between the two, at the neighbour counts $1, 2, 3, 4, 5, 7, 10, 14, 20, 28, 40, 56, 80, 113, 160, 226, 320$ with $2n \leq S - 1$. The sixteen cells cross the two architectures with $m \in \{5, 20, 50, 200\}$ on CIFAR-10 and $\{2, 5, 20, 50\}$ on CIFAR-100, so m spans $\{2, 5, 20, 50, 200\}$, a 100-fold range; the CIFAR-100 $m = 2$ cells are outside the compression rule. The design, the cells and the four rules were recorded before any of these runs was trained. Every rule holds (Table 15). In each of the sixteen cells all five runs compress at the grid count nearest $m/2$ and expand at the one nearest $4m$. For CIFAR-10 at $m = 200$ that count is 320, not within 25% of $4m$, so neither of those two cells is counted as a $4m$ test. The most-compressing neighbour count moves in proportion to m on both exponent statistics.

C.7 Replications of the training change

The twenty runs were read again at three new class sizes, $m \in \{7, 13, 35\}$, with a query–reference form of the estimator (a close relative of Equation 2) in which a few test images per class are queries among m reference images per class from a disjoint pool; ten further runs, ResNet-34 on CIFAR-100 and WRN-16-4 on CIFAR-10 with five seeds each, were trained after the density protocol was fixed and read the same way. Counting a change of more than 0.05 in $\log \widehat{d}_n$, all 60 run–class-size designs compress at $n/m = \frac{1}{2}$ and expand at $n/m = 4$ in each of two overlapping query–reference allocations, and the further runs do the same in 30 of 30 designs, a count made afterwards. The switch criterion was fixed before the same-run count. As the twenty runs had been studied before, that count shows robustness to class size and to the images, not independent replication.

Table 15: The training experiment: the four rules fixed before training, their outcomes, and one post hoc fit. S is the number of images in a probe. The compression rule takes the count nearest $m/2$ on a log scale; the linearly nearest count (2 instead of 3 at $m = 5$) also gives five of five.

rule	fixed before training	outcome
compression	for $m \geq 5$, $\Delta_n < 0$ at the grid count nearest $m/2$ (on a log scale) in at least four of five runs, in at least 12 of the 14 such cells; the two CIFAR-100 cells at $m = 2$, whose target count is the bottom of the grid, are outside it	holds in 14 of 14 cells, each five of five; as a run-level sign test, all 20 runs compress in every included cell, $p = 9.5 \times 10^{-7}$
expansion	$\Delta_n > 0$ at the grid count nearest $4m$ in at least four of five runs, in at least 7 of the 8 CIFAR-100 cells; CIFAR-10 excluded in advance, because at $K = 10$ the count $4m$ reaches 80% of the probe	8 of 8 (all 10 runs, $p = 9.8 \times 10^{-4}$); CIFAR-10 is reported only descriptively
matching	the most-compressing count agrees within a factor two at matched m across datasets and differs by at least four at matched sample size	8 of 8 comparisons
exponent	regress $\log n_{\min}$ on $\log m$ across all sixteen cells, $n_{\min}$ the grid count that minimises the run-averaged Δ_n, searched over $2n \leq S/4$ because the ratio estimator falls sharply as $2n$ approaches S; bar: slope in $[0.6, 1.4]$	slope 1.04 $[1.02, 1.06]$ (bootstrap over runs): passes
per-run exponent (post hoc)	each run's own most-compressing count over all counts, in the 14 cells with $m \geq 5$ (70 run–density pairs)	$n^* \approx 0.57\, m^{1.02}$, interval $[0.99, 1.04]$; slopes 0.96 to 1.04 within each dataset–architecture pair

C.8 Collapse statistics along training

Under rules fixed in advance, we test whether, along training, the change in $\log \widehat{d}_n$ at $n = m/2$ tracks the change in log NC1 and the change at $n = 4m$ tracks that in NC2, within each run of a checkpoint cohort and across the 20 runs of the training experiment (Table 16). The estimate at $4m$ tracks NC2 in every run, and NC1 falls in every run too, but the estimate at $m/2$ falls with NC1 in the density runs and first rises in the intervention and weight-decay runs, so the NC1 half failed, as did the test of a positive branch at $n = 1$ late in training. With six checkpoints per run, a within-run Spearman correlation of ± 1 says only that both quantities are monotone in training time. The across-run correlations on two-checkpoint differences clear both bars but are reported, not counted, because the registered check of the stored estimates failed at $n = 1$. The expansion at $4m$ accompanies the class means' approach to equal spacing but is not identified with the two collapse statistics.

Table 16: **Along training the estimate at $4m$ tracks NC2, but the NC1 half of the collapse test failed, and so did the late-branch and within-run sign-pair tests.** All three tests were fixed in advance. NC1 is $\mathrm{tr}(\Sigma_W \Sigma_B^+)/K$ and NC2 the spread of the cosines between centred class means, on the estimate's probe draws ($m = 20$). Within-run ρ is a Spearman correlation over one run's checkpoints, in the 51 cohort runs with stored features (306 checkpoints; the other 10 kept only their checkpoints, from which the surrogate of Appendix E extracts its own features). The $n = 1$ rows correlate the estimate with log error across the cohort at one epoch. The last block asks log error to move with the estimate at $m/2$ and against it at $4m$ within each run. *Reported, not counted: the registered check of the stored estimates failed at $n = 1$ (duplicate CIFAR-100 test images).

	required	result
median within-run ρ, change at $m/2$ vs NC1	$\geq +0.70$	$+0.49$ **(failed)**
density runs only		$+0.71$
runs with the expected sign, NC1	$\geq \frac{3}{4}$	30 of 51 **(failed)**
median within-run ρ, change at $4m$ vs NC2	≤ -0.70	-1.00
runs with the expected sign, NC2	$\geq \frac{3}{4}$	51 of 51
across the 20 pre-registered runs, NC1; NC2	$\geq +0.60$; ≤ -0.60	$+0.96^*$; -0.86^*
r at $n = 1$ with log error, earliest epoch (16)	≤ 0	-0.92
the same, last epoch (80)	$\geq +0.40$	$+0.24$ $[-0.03, +0.51]$ **(failed)**
runs with sign pair $(+, -)$ vs log error	$\geq \frac{3}{4}$	28 of 51 **(failed)**
pre-registered runs with that pair	all	20 of 20

D The two deleted estimates, the class-mean gap and the probe designs

Supports Section 3 and the two-count check of Section 4.3. Unless stated otherwise, each analysis uses the stratified probe of Section 3 ($K = 100$ classes, $m = 20$ validation images each, three

draws averaged) on the 222 ImageNet classifiers. Correlations are Pearson r across networks, with percentile bootstrap intervals over networks.

D.1 THE TWO DELETED ESTIMATES

We computed the residuals-only estimate $\widehat{d}_n^{\mathrm{res}}$ below m and the class-means-only estimate $\widehat{d}_n^0$ from $n = m$ (Equation 3) on the draws of the full estimate, with classes and again with k-means clusters (Table 17). Below m the residuals-only estimate orders the networks as the full one does, with clusters as with classes, though at the published choice $n = 1$ it relates to error more weakly. As n grows, foreign points enter the outer shell and lift the full estimate above the residuals-only one; for separated classes Corollary 1(i) predicts equality, so the high correlations there partly reflect distance from separation. From $n = m$ to $4m$ the class-means-only estimate follows the full one more and more closely, within training runs too, and correlates with error at least as strongly; cluster means keep up only to about $2m$, since clusters are not classes and differ in size. An occupancy-weighted mixture of the two deleted estimates has the wrong sign at $n = m$ unless a crossover-band term is fitted. Replacing each query's intruder counts (Lemma 1) by the network's mean places the crossing within about one neighbour of the measured one in all four designs of Table 17. Rescaling every network's residuals to a common spread lowers the correlation with error at $n = 1$, barely changes it at $n = 10$ and moves the crossing towards $n = m$; reshaping their spectrum moves the estimate below m and hardly above (Table 31).

Table 17: **Deleting residuals or class means reproduces the estimate on either side of the class size.** The top block gives the residuals-only estimate below m with the median level gap, the middle block the class-means-only estimate from $n = m$, and the fourth column the full estimate's own correlation with error. The bottom block holds checks; the foreign share in it contains no estimator. Errors are first-row (Table 6, note a). The designs compared are $(K, m) = (250, 4)$, $(200, 5)$, $(50, 20)$ and $(100, 20)$. Both deleted estimates agree between them, the residuals-only one only at equal m, as a reading at the density set by m should. Clusters are k-means with $k = 100$ on the centred probe points of two separate draws, which also give the class values in parentheses (adjusted Rand index with the classes, quartiles 0.54–0.70). The share at $n = 10$ is not zero although each query has only 19 classmates: the 20th neighbour is a classmate whenever a point of another class comes earlier.

n	$r(\log \widehat{d}_n, \log \widehat{d}_n^{\mathrm{res}})$	$r(\log \widehat{d}_n^{\mathrm{res}}, \log \mathrm{err})$	$r(\log \widehat{d}_n, \log \mathrm{err})$	median $\log \widehat{d}_n - \log \widehat{d}_n^{\mathrm{res}}$
1	$+0.966$ $[+0.95, +0.98]$	$+0.351$ $[+0.21, +0.48]$	$+0.536$	0.03
3	$+0.960$ $[+0.95, +0.97]$	$+0.598$ $[+0.48, +0.70]$	$+0.723$	0.09
5	$+0.954$ $[+0.94, +0.97]$	$+0.691$ $[+0.60, +0.77]$	$+0.772$	0.17
9	$+0.925$ $[+0.90, +0.94]$	$+0.775$ $[+0.71, +0.83]$	$+0.796$	0.69
n	$r(\log \widehat{d}_n, \log \widehat{d}_n^0)$	$r(\log \widehat{d}_n^0, \log \mathrm{err})$	$r(\log \widehat{d}_n, \log \mathrm{err})$	
$20 = m$	$+0.237$ $[+0.10, +0.38]$	-0.676 $[-0.74, -0.61]$	$+0.169$	
40	$+0.744$ $[+0.67, +0.80]$	-0.732 $[-0.79, -0.67]$	-0.471	
80	$+0.850$ $[+0.82, +0.88]$	-0.738 $[-0.79, -0.68]$	-0.651	
160	$+0.740$ $[+0.68, +0.80]$	-0.682 $[-0.74, -0.62]$	-0.647	
share of queries whose nth and $2n$th neighbours are both classmates, $n = 10$				0.08
r(foreign share among the $m/2$ nearest neighbours, $\log \widehat{d}_{m/2}$)				$+0.938$
Spearman $\rho(\log \widehat{d}_n, \log \widehat{d}_n^0)$ within 20 CIFAR runs (14 checkpoints), $n = m$; $4m$				-0.347; $+0.956$
r between designs: class-means-only, $n = 80$				0.90–0.99
r between designs: residuals-only, $n = 1$; equal m, different m				0.98; 0.74–0.82
clusters: residuals-only vs full, $n \leq 7$ (classes $+0.946$ to $+0.972$)				$+0.977$ to $+0.983$
clusters: residuals-only vs log error, $n = 5$ (classes $+0.678$)				$+0.715$
clusters: cluster-means-only vs full, $n = 40$; 80				$+0.720$; $+0.283$

D.2 THE CLASS-MEAN GAP AND NEURAL COLLAPSE ON REAL NETWORKS

On 1,998 cells (222 networks, three draws, $n \in \{m, 2m, 4m\}$) we checked the bound of Equation 10, and across networks we correlated $\log \widehat{d}_n^0$ with the gap statistic G_n and with NC2, the standard deviation of the cosines between centred class means (Table 18). The bound holds in every cell and is loose at $n = m$; what it shows is which quantity the statistic computes. Both correlations are reproduced by 240 random class-mean configurations, an exploratory null whose families, seed and statistic were fixed in advance, so neither says anything about the trained networks. The gap correlation is forced, since $\log \widehat{d}_n^0$ and $\log(\kappa_n / G_n)$ differ only at second order. Within architecture

families the class-means-only estimate tracks error more strongly than each collapse statistic, and given the error of either of the probe's own classifiers its association with error largely disappears.

Table 18: **Random class-mean configurations reproduce the gap and NC2 correlations.** In the top block the bound holds with $\rho_+ = 0$ in every cell, and the gap and NC2 correlations use 194 networks; the null configurations are Gaussian means in 64–2048 and in 12–60 dimensions, uniform, and a perturbed simplex. The middle block works within families (127 networks in 18 families, family bootstrap, $n = 4m$), with tightness the median residual length over the median class-mean distance and CV the coefficient of variation. The bottom block gives partial r given the error of the probe's nearest-class-mean (NCM) or 1-NN classifier, whose own errors track log error at $+0.95$ and $+0.93$.

	$n = m$	$n = 4m$	all measured n	null
median ρ_-	0.60	0.21	0.32	
median $\lvert\widehat{d}_n^0/(\kappa_n/G_n) - 1\rvert$	0.27	0.11		
$r(\log \widehat{d}_n^0, \log G_n)$		-0.998		-0.9999 to -0.9982
$r(\log \widehat{d}_n^0, \text{NC2})$		-0.925		-0.98 to -0.99

statistic	r with log error	$\lvert r_{\text{cen}}\rvert - \lvert r\rvert$
class-means-only estimate	-0.81 $[-0.87, -0.78]$	—
tightness (NC1-type)	$+0.46$ $[+0.30, +0.79]$	$+0.35$ $[+0.02, +0.52]$
CV of class-mean distances (NC2-type)	$+0.63$ $[+0.34, +0.71]$	$+0.18$ $[+0.11, +0.48]$
CV of class-mean norms (NC2-type)	$+0.45$ $[+0.04, +0.56]$	$+0.37$ $[+0.26, +0.75]$
class-means-only estimate against log error	given NCM error	given 1-NN error
all 222 networks (raw -0.74)	$+0.15$ $[+0.01, +0.32]$	$+0.05$ $[-0.11, +0.24]$
within families (raw -0.81)	-0.02 $[-0.55, +0.16]$	-0.07 $[-0.52, +0.18]$

D.3 Other estimators

The Levina–Bickel MLE (Levina & Bickel, 2004) of order k averages over every neighbour up to the kth, including neighbours still inside the class, so under separation the occupancy term of Theorem 1 gives its inverse a response $(m-1)/(k-1)$ to within-class scale for every $k \geq m$, which predicts that it flips later than the two-shell estimate but not where. On the 222 networks it first changes sign at $k = 77$ for $m = 5$ ($K = 200$), against $n = 6.5$ for the two-shell estimate on the same three draws (6.87 on the five draws of Table 20), and not by $k = 80$ for $m = 20$. In a test fixed in advance on the torchvision and timm collections in two designs each, GRIDE and three ratio estimators reverse in all four cells, as predicted; local PCA reverses in none, where at least two were predicted, and the moment-form MLE, predicted not to reverse at $m = 20$, does so once the neighbour count passes 80 (Table 31).

D.4 The class-means term and the remainder, group by group

Write $\log \widehat{d}_n = X + Y$, with $X = \log \widehat{d}_n^0$ and Y the remainder, so that for a group of networks with log error E, $\mathrm{Cov}(\log \widehat{d}_n, E) = \mathrm{Cov}(X, E) + \mathrm{Cov}(Y, E)$, an identity in a chosen basis that says where the class-means reading applies, not which term causes a sign. Because Y contains the measurement and both correlations use the group's own errors, the split forecasts a new group's sign only if both terms are predicted independently. On the ImageNet groups the class-means term has the larger variance and is negatively related to error, and the full estimate follows the class-means-only one, so the association is negative (Table 19). In the two CIFAR-100 regulariser sweeps the two terms nearly cancel, so the full estimate varies far less than either (sd 0.10 against 0.21–0.41); it does not follow the class-means-only one, so it is positive while the class-means term is strongly negatively related to error. A test fixed in advance asked that in both sweeps the median classmate share $\widehat{P}_{4m}$ be at least twice the ImageNet value and correlate with $\log \widehat{d}_{4m}$ at $|r| \geq 0.7$; the share held (0.038 and 0.033 against 0.010) and so did the correlation in the weight-decay sweep ($r = +0.97$), but not in the label-smoothing sweep ($r = +0.25$), so the test failed and nothing is attributed to occupancy.

D.5 The probe designs and the crossing

Across the probe arms of Table 20, the crossing doubles with m at a fixed number of classes, lies near $n = m$ and rises slowly with K; the slope in K is fitted, not derived, and the one-image arm

Table 19: **The class-means term is negative in every group, and the full estimate is positive only in the two sweeps.** All at $n = 4m$ ($m = 20$). $X = \log \widehat{d}_n^0$, $Y = \log \widehat{d}_n - X$, $Z = \log \widehat{d}_n$, and E is the log error; r is a correlation and s a standard deviation. The sign of r_{ZE} is that of $s_X r_{XE} + s_Y r_{YE}$, computable from the columns shown, and r_{ZX} says how closely the full estimate follows the class-means-only one. The ImageNet rows use the stratified probe, with main and held out the two collections, recipes the ResNet recipe groups, and ViT the ViT family, whose class-means term is not significant and whose features were never head-checked. The ImageNet rows use the three probe draws on which the class-means term was computed (Table 6 averages five, so r_{ZE} can differ by 0.01), and the all and recipes rows use first-row errors (Table 6, note a). The CIFAR-100 rows are final checkpoints of the label-smoothing and weight-decay (LS+WD) and weight-decay (WD) sweeps, and five ResNet-18 against five VGG-11-BN seeds (ordinary, a two-group contrast), each read on one draw of the probe (Table 10 averages five). The last row, the median within-run Spearman correlation over checkpoints of the CIFAR runs (in parentheses, runs with a negative class-means term), is not an independent group. A class-density CIFAR-100 group repeats the ordinary-runs design and is omitted.

group	N	r_{ZE}	r_{XE}	r_{YE}	s_X^2	s_Y^2	r_{XY}	s_Z	r_{ZX}
ImageNet, all	222	-0.651	-0.738	$+0.478$	0.159	0.046	-0.69	0.30	$+0.850$
ImageNet, main	139	-0.60	-0.73	$+0.52$					
ImageNet, held out	55	-0.66	-0.73	$+0.47$					
ImageNet, recipes	28	-0.840	-0.874	$+0.379$	0.165	0.022	-0.50	0.36	—
ImageNet, ViT	22	-0.55	-0.38	$+0.01$					
CIFAR-100, LS+WD	20	$+0.388$	-0.917	$+0.972$	0.152	0.165	-0.97	0.10	-0.05
CIFAR-100, WD	21	$+0.986$	-0.965	$+0.982$	0.043	0.092	-0.995	0.10	-0.95
CIFAR-100, ordinary	10	-1.00	-0.98	$+0.81$					
CIFAR, within run	20	-0.96	-0.92 (20)						

lies off the line. Equalising the spread of residual norms or the tightness of classes, or replacing images by means of 49 images, does not remove the one-image branch (Table 31). A mechanism in which near-duplicate class means delay the crossing failed all three of its tests fixed in advance; only a sibling count chosen afterwards gives the direction of a sibling-heavy against sibling-light comparison, and not its size (Table 31).

Table 20: Every probe arm and its crossing. All arms read the same 222 networks on validation images, with five draws for the first four arms and three for the rest. The crossing n^* is the log-interpolated zero of the correlation profile with log top-5 error. The left half lists the first ten arms, which mostly hold $N = Km$ at 1000 and vary how it is spread over classes; the right half lists ten later arms, which vary K at fixed m, and m at fixed K, over wider ranges. The line $n^*/m = 0.98 + 0.00158\,(K - 1)$, fitted afterwards to the 19 arms with $m \geq 2$ ($R^2 = 0.98$), puts the one-per-class arm at $n^*/m = 2.56$.

arm	K	m	N	n^*	n^*/m	arm	K	m	N	n^*	n^*/m
one per class	1000	1	1000	2.11	2.11	K=50, m=5	50	5	250	4.75	0.95
two per class	500	2	1000	3.47	1.73	K=100, m=5	100	5	500	5.67	1.13
four per class	250	4	1000	5.98	1.50	K=500, m=5	500	5	2500	8.93	1.79
K=200, m=5	200	5	1000	6.87	1.37	K=1000, m=5	1000	5	5000	12.54	2.51
K=50, m=20	50	20	1000	19.75	0.99	K=25, m=10	25	10	250	9.27	0.93
K=100, m=20	100	20	2000	23.51	1.18	K=500, m=10	500	10	5000	18.58	1.86
K=50, m=10	50	10	500	9.56	0.96	K=1000, m=10	1000	10	10000	25.21	2.52
K=100, m=10	100	10	1000	12.17	1.22	K=200, m=20	200	20	4000	27.39	1.37
K=200, m=10	200	10	2000	13.35	1.34	K=50, m=40	50	40	2000	41.96	1.05
K=25, m=20	25	20	500	20.26	1.01	K=100, m=40	100	40	4000	46.78	1.17

D.6 THE HORSE RACE

The horse race of Table 21 sets the estimate among other statistics on the same probe draws. On the stratified probe the probe's own nearest-class-mean and nearest-neighbour accuracies, which use the labels, track error more closely ($|r| = 0.95$ and 0.93; Table 18) than any reading of the estimate (at most $+0.80$, for $\widehat{d}_{10}$). The partial correlations are descriptive and do not test mediation. Whether the estimate adds held-out information beyond the nearest-neighbour error is unresolved, since the gain's interval includes zero when whole architecture families are left out and excludes it over finer name-prefix groups. We chose the neighbour count of $\widehat{d}_{10}$ on all 222 outcomes, so the cross-validation is retrospective, although every other choice was made inside the training folds.

Table 21: **Every statistic that tracks error more closely than the estimate uses the labels.** Entries are Pearson r between the log predictor and log top-5 error (224-pixel matched, first-row for the 28 recipe-group networks; Table 6, note a), with accuracy-type predictors entered as $\log(1 - \text{acc})$; "labels" says whether the statistic uses the probe labels. The pooled column covers all 222 head-checked ImageNet networks, with a 95% bootstrap interval. Held-out R^2 comes from a log-linear fit on the main collection (139) applied to the other 83 networks (the 55 held-out and the 28 ResNet-18/34 and 101/152 recipe networks). ViT is the frozen family of 22, whose features were never head-checked, and the uniform design draws 2,000 validation images uniformly over the 1,000 classes. The neighbour count of $\widehat{d}_{10}$ was chosen after seeing these outcomes. The bottom rows give partial correlations and the held-out gain in R^2 from adding $\log \widehat{d}_{10}$ to the 1-NN error (nested cross-validation, 95% intervals).

predictor	pooled r	95% CI	held-out r	held-out R^2	ViT	labels
1-NN accuracy, probe (K=100, m=20)	$+0.93$	$[+0.91, +0.95]$	$+0.94$	0.83	$+0.93$	yes
1-NN accuracy, 10 per class $\times$ 1000	$+0.92$	$[+0.90, +0.94]$	$+0.94$	0.82	$+0.84$	yes
5-NN accuracy, probe	$+0.90$	$[+0.87, +0.92]$	$+0.92$	0.74	$+0.80$	yes
1-NN accuracy, uniform design	$+0.88$	$[+0.86, +0.91]$	$+0.91$	0.72	$+0.72$	yes
NC1 (Papyan et al., 2020)	$+0.85$	$[+0.82, +0.88]$	$+0.90$	0.77	$+0.56$	yes
foreign fraction among $m/2$ neighbours	$+0.85$	$[+0.81, +0.88]$	$+0.88$	0.63	$+0.70$	yes
$\widehat{d}_{10}$ (below m)	$+0.80$	$[+0.76, +0.84]$	$+0.81$	0.50	$+0.34$	no
$\widehat{d}_1$, uniform design	$+0.79$	$[+0.75, +0.84]$	$+0.80$	0.48	$+0.31$	no
residuals-only $\widehat{d}_9^{\,\mathrm{res}}$	$+0.79$	$[+0.73, +0.84]$	$+0.81$	0.52	$+0.27$	yes
class-means-only $\widehat{d}_{4m}^{\,0}$	-0.74	$[-0.79, -0.69]$	-0.80	0.50	-0.38	yes
gap statistic G_{4m}	$+0.74$	$[+0.68, +0.79]$	$+0.80$	0.50	$+0.35$	yes
within-class spread ratio τ	$+0.68$	$[+0.62, +0.74]$	$+0.66$	0.28	$+0.38$	yes
class separation R^2 (Kornblith et al., 2021)	$+0.68$	$[+0.62, +0.73]$	$+0.64$	0.26	$+0.43$	yes
NC2 equiangularity	$+0.66$	$[+0.59, +0.73]$	$+0.76$	0.41	$+0.21$	yes
$\widehat{d}_{4m}$	-0.65	$[-0.72, -0.58]$	-0.75	0.40	-0.55	no
spectral decay α	$+0.62$	$[+0.55, +0.69]$	$+0.71$	0.30	-0.25	no
class-mean-distance CV	$+0.61$	$[+0.52, +0.69]$	$+0.67$	0.22	$+0.57$	yes
CDNV (Galanti et al., 2022)	$+0.60$	$[+0.51, +0.67]$	$+0.43$	0.12	$+0.33$	yes
participation ratio	-0.34	$[-0.47, -0.19]$	-0.44	0.03	-0.23	no
RankMe (Garrido et al., 2023)	-0.17	$[-0.31, -0.02]$	-0.51	-0.38	-0.44	no

partial r with log error: $\log \widehat{d}_{10}$ given 1-NN error -0.44; 1-NN error given $\log \widehat{d}_{10}$ $+0.84$
gain in held-out R^2, leaving out one of 24 architecture families: $+0.021$ $[-0.013, +0.050]$; one of 78 name-prefix groups: $+0.022$ $[+0.003, +0.044]$

E THE SURROGATE AND THE INTERVENTIONS IN FULL

Section 3 states the surrogate as Equation 4 and summarises the interventions in Figure 4E; both are given in full here. `figE_4.py` in the supplementary material regenerates Table 22 and Figure 8 from the recorded curves.

E.1 THE SURROGATE

The surrogate of Equation 4 keeps each class mean and each point's distance to it, and draws residual directions as $Lg_i/\|Lg_i\|$ with nothing fitted. It discards individual directions and real neighbourhoods but keeps the orientation of Σ_W relative to the class means, though after normalisation its residual covariance is not exactly Σ_W. On seven populations, about four of them independent, it tracks the real curve (Figure 8) and beats all three baselines nearly everywhere (Table 22), so the curve is carried by the distribution of distances to the class centres and not by class means or scales alone. It gets the reversal status right in 6 of 7 populations, as does the guess that every population reverses. On the top 16 or 32 principal components, too few to hold the 99 directions spanned by the 100 ImageNet class means, the positive branch survives; the negative branch largely disappears with 16 and returns only weakly with 32, so it is carried by the many directions the class means span. A follow-up fixed in advance passed all four of its tests: on the top 16 and 32 components the surrogate beats the mean-curve baseline in 6 and 7 of 7 populations and matches the reversal status at 16 in 7 of 7, and hubness-reduced ranks (Radovanović et al., 2010; Conneau et al., 2018) leave the reversal status unchanged in 7 of 7.

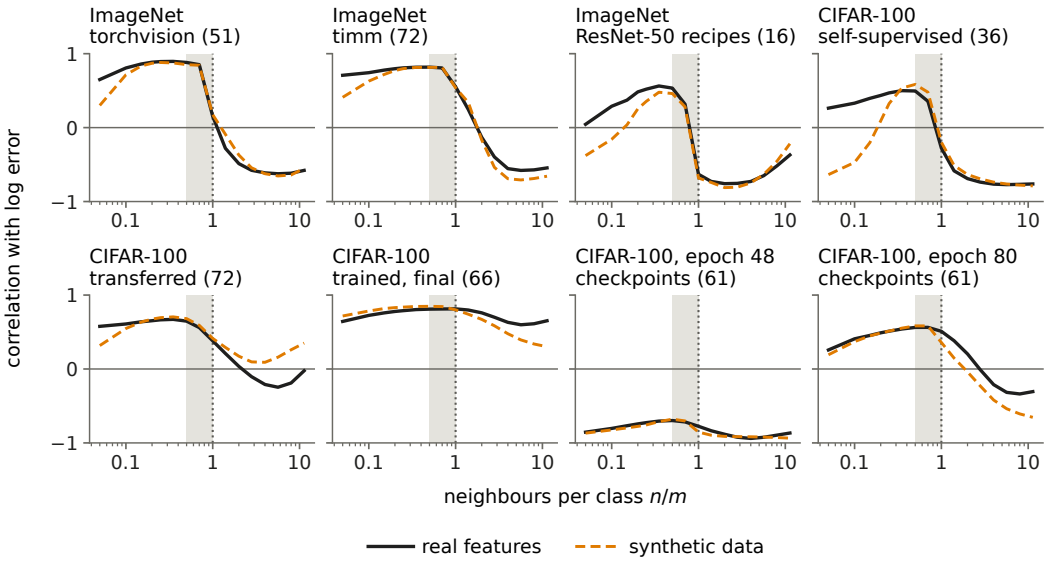

Figure 8: **The synthetic data track the curve in all seven populations.** Correlation of the estimate with log error across networks against n/m, real (black) and synthetic data (orange dashed); grey, the crossover band. The checkpoint population is shown at two epochs.

Table 22: The surrogate against the real curve in seven populations, three on ImageNet validation images ($K = 100$, $m = 20$) and four on CIFAR-100 test features ($m = 20$; pretrained encoders are scored by linear-probe error). The construction, populations, criteria, baselines and controls were fixed in advance. An earlier pre-registered surrogate test on ImageNet validation images failed. Top: correlation between the real and surrogate curves (Pearson over 16 neighbour counts; checkpoints pooled over epochs). Bottom: curve error (root-mean-square, in units of r) against three baselines, and the crossing.

population	N	r	population	N	r
ImageNet, torchvision	51	0.988	CIFAR-100, timm as frozen extractors	72	0.902
ImageNet, timm	72	0.991	CIFAR-100, trained by us (never reverse)	66	0.867
ImageNet, ResNet-50 recipes	16	0.956	CIFAR-100, their checkpoints, pooled	366	0.986
CIFAR-100, self-supervised	36	0.816	checkpoints, epoch 48 / 80	61	0.939 / 0.988

comparison	outcome
surrogate vs mean real curve of the other six populations	better in 6 of 7 (0.10–0.22 against 0.25–0.84; self-supervised 0.35 against 0.34)
surrogate vs itself with networks shuffled	better in 7 of 7
surrogate vs class means only, one radius per network or per class	better in 7 of 7
surrogate crossing within a factor two of the real one	5 of 5 populations that cross

E.2 The interventions

Class means over all of a class's validation images serve as fixed anchors. Each point keeps its residual direction and only its distance to the anchor changes, in four cells: torchvision and timm, each at $K = 1000$, $m = 2$ and at $K = 100$, $m = 20$. The cells, placebos and rules were fixed in advance. Only equalising every distance to the centre removes the negative branch (Table 23; Figure 4E). It met its rule in all four cells, overshooting to positive on timm. Permuted scales and the cap control change almost nothing, so the removal is not an artefact of rescaling. Two other rules failed, equal class scales and, in an earlier test, equal norms within each class. The second kept the branch in all four cells, while the positive branch survived as predicted (4 of 4).

Equalising moves each class's empirical mean off its anchor, by a median 8.9–9.7% of the distance to the nearest other anchor (9–11% with a per-class scale), so part of the effect can act through between-class geometry. Re-fixing the means by re-centring the probe points removed the branch in 2 of 4 cells against a bar of 3, and re-centring whole classes, the secondary rule, in 3 of 4 (Table 23).

With the probe points re-centred, the removal holds in the $m = 20$ cells but not in the $m = 2$ ones, where re-centring two points restores between-class differences in radius (8.8–9.3% between-class variation). The spread of distances to class centres is necessary for the branch only jointly with the class-mean displacement it induces. Changing that spread in either direction does not move the branch predictably. Re-weighting which points are sampled, across a wide range of the spread, moved the branch in only 3 of 18 cells, short of its rule. A regulariser that raised the spread in training lowered the estimate at large n.

Real distances with isotropic directions keep 108–112% of the negative branch in the four supervised cells, Gaussian residuals without real distances only 62% on torchvision and 8% on timm, and a placebo permuting distances across points 89–99%, so equalisation does not act by degrading the estimator; in self-supervised encoders no single component is necessary. The dispersion of residual norms itself correlates with log error at -0.81, -0.74 and -0.81 in the torchvision, timm and held-out collections; that of raw feature norms does not ($+0.04$ on timm).

Table 23: **Only equalising every distance to the centre removes the negative branch.** Entries are the fraction $\phi = r_{\text{after}}/r_{\text{before}}$ of the negative branch kept, at the neighbour count where the real curve is most negative ($n = 80$ for $m = 2$, $n = 113$ for $m = 20$; r_{before} between -0.51 and -0.56). Re-centred rows re-fix the means on the anchors; the bar fixed in advance was $\phi \leq 0.2$ in at least 3 of 4 cells, and the probe-point test failed. The single-component replacements in the text belong to the pre-registered surrogate protocol, whose spectrum-only prediction for self-supervised encoders failed.

	torchvision		timm	
intervention	$m = 2$	$m = 20$	$m = 2$	$m = 20$
all distances to centre equal	$+0.14$	$+0.04$	-0.54	-0.68
probe points re-centred on the anchor (fixed in advance; **failed**)	$+0.50$	$+0.07$	$+0.43$	-0.59
whole classes re-centred (secondary rule)	$+0.25$	$+0.12$	-0.44	-0.59
class scales equal (rule fixed in advance **failed**, 2 of 4)	0.71	0.65	0.50	0.43
norms equal within each class	0.85	0.79	0.63	0.70
ℓ_2-normalised features	0.75	0.79	0.92	0.95
largest radii capped	1.00	1.00	0.84	0.90
matched control for the cap	1.00	1.01	0.99	1.00
class scales, permuted classes	1.01	0.98	1.00	1.03

F LAYER PROFILES AND LANGUAGE MODELS

What follows are the full layer profiles, language-model results and layer-choice tests behind Section 4, with the plateau rule and the published designs.

F.1 LAYER PROFILES

The thirteen torchvision classifiers are AlexNet, VGG-11/13/16/19 without batch normalisation, ResNet-18/34/50/101/152 and DenseNet-121/161/169. Each was read at six to eight layers, with two draws, on two probes. The pooled probe takes two images from each of 1,000 classes and is read at $n = 1$. The stratified probe has $K = 100$ and $m = 20$. A profile's fall is $T = (\text{peak} - \text{last})/\text{peak}$, from its highest value to the last hidden layer.

Read on the pooled probe, or at $n = m/2$ on the stratified one, the profile falls by more than half in every network (Table 24). The deleted estimates fall much less, and so does the full estimate in the layout regime (Figure 5B). Every reading shows the rise. The extra fall of the pooled reading coincides with the late layers where classes form (Table 24, bottom block). The pooled and per-class readings are different estimands, so we do not split the fall between them.

The probes, the layers and three rules were fixed in advance, before the features were read (Table 31). The primary rule required the fall at $n = m/2$ on the stratified probe to exceed the falls at $n = 1$ and at $n = 4m$ in at least 11 of 13 networks. It **failed**, holding against $n = 1$ in only 10 of 13. The other two rules passed. The per-layer deleted estimates were added afterwards.

Table 24: **How far a layer profile falls depends on how it is read.** 13 torchvision classifiers, not the original fourteen (three DenseNets added, the four VGG-bn left out). All rows but the first use the stratified probe. The bottom block shows where classes form (78 layer-to-layer transitions).

reading	median fall T	$T > \frac{1}{2}$
pooled probe, two images per class, $n = 1$	0.82	13 of 13
crossover band, $n = m/2$	0.80	13 of 13
layout regime, $n = 2m$	0.43	6 of 13
layout regime, $n = 4m$	0.29	4 of 13
residuals-only, $n = 3$	0.46	6 of 13
class-means-only, $n = 2m$	0.37	2 of 13
residuals-only peak within one layer of the pooled peak		10 of 13
median classmate share at $n = 1$, stratified: first two thirds of depth; penultimate; last		0.06; 0.59; 0.80
Spearman: change in log pooled estimate vs in its classmate share		-0.832

F.2 Language models

The 26 pretrained models are GPT-2 (four sizes), Pythia (seven, 70M–6.9B), GPT-Neo 1.3B, OPT 1.3B, OLMo 1B and 7B, Qwen2.5 (four, 0.5B–7B), SmolLM2 360M and 1.7B, and the encoders BERT-base, DistilBERT, RoBERTa-base, all-mpnet-base-v2 and a BERT fine-tuned on banking77 (Devlin et al., 2019). Every layer is read, 588 units in all, from the last token in decoders and the token mean in encoders. The data are the test sets of banking77 (77 intents of 40 examples) and clinc (150 intents of 30, out-of-scope dropped). A unit's error is that of a linear intent probe on a stratified 80/20 split. Probes draw $m \in \{1, 2, 5, 20\}$ per intent, with three draws, and are read up to $n = 80$.

Across models at the last layer both datasets show the vision pattern (Table 25; Figure 5C). The correlation is positive below m and negative above, and for $m \geq 5$ it changes sign at n/m between 2 and 3. The deleted estimates carry over from vision.

The signs within one model do not carry over. Deeper clinc layers have residual clouds of higher dimension, class means further from equidistant, and lower probe error, so across the layers of one clinc model both branches invert. The branch above m inverts only far above it, at $16m$ for $m = 5$ and weakly at $4m$ for $m = 20$; for $m = 5$ the median at $4m$ is -0.14. Counted per model on clinc ($m = 5$), the class-regime sign inverts in 19 of 26 models, the layout sign is positive at $16m$ in 20, and both hold in 14 (4 with the layout sign read at $4m$). The same three counts on banking77 are 19, only 10 and 5. Within models the class-means term is *positively* related to probe error, whereas in every vision group it is negative (Table 19). The negative relation is a property of the vision networks and does not come from the estimator. The clinc class means are also far less regular than those of any vision network.

Table 25: The estimate against probe error in language models. Top: correlation of $\log \widehat{d}_n$ with log probe error, across the 26 models at their last layer (Pearson) and within each model across its layers (median Spearman). $n = 80$ is $16m$ at $m = 5$ and $4m$ at $m = 20$. The last column gives the first sign change, n^*/m, at $m = 5$ and 20. Bottom: the deleted estimates against the full estimate, the class-means term of Table 19 against probe error within models (median r), and the regularity of the class means.

		$m = 5$		$m = 20$		
dataset	view	$n = 1$	$n = 80$	$n = 1$	$n = 80$	n^*/m
banking77	across models, last layer	$+0.832$	-0.389	$+0.893$	-0.120	2.7, 2.8
clinc	across models, last layer	$+0.801$	-0.322	$+0.804$	-0.148	2.2, 2.4
banking77	within model, across layers	-0.586	-0.149	-0.358	$+0.058$	1.4, 1.4
clinc	within model, across layers	-0.937	$+0.365$	-0.932	$+0.089$	4.4, 3.9

	banking77	clinc
residuals-only vs estimate at $n = 1$, all units, $m = 20$	$+0.697$	$+0.977$
class-means-only vs estimate at $n = 3m$, last layer	$+0.851$	$+0.853$
class-means term vs probe error, within model, $4m$	$+0.232$	$+0.535$
CV of class-mean distances (vision, 5th–95th pct.: 0.06–0.13)		up to 3.41

F.3 A neighbour count chosen by a plateau

The plateau rule is our own, not the scale-selection procedure of any published estimator. For each network it takes the adjacent pair of grid neighbour counts over which $\log \widehat{d}_n$ changes least, on the

published design (2,000 images uniform over 1,000 classes, about two per class) and on five further designs (Table 26). On every design most networks' flattest pair lies at or above the class size, and there, as at each network's maximum, the correlation with error is negative.

Table 26: The plateau rule on the published design (178 networks, top) and on five designs (222 networks, 224-pixel matched errors, bottom).

published design: networks whose flattest pair lies within $n = 15$ to 70; at $n = 1, 2$	135; 23
networks whose pair starts at $n \geq 2$, the images per class; median smaller n of the pair	155; 20
r with log error at the pair: all; torchvision, timm, held-out	-0.44; -0.39, -0.42, -0.58
r at each network's maximum	-0.63

design $K \times m$	median n/m at the pair	pair at $n \geq m$	r at the pair	r at the maximum
1000×1	10	100%	-0.64	-0.65
500×2	7.5	100%	-0.66	-0.66
250×4	5	99%	-0.66	-0.65
50×20	2	91%	-0.47	-0.62
100×20, stratified	3	90%	-0.44	-0.62

F.4 Published designs on the n/m axis

Table 27 places eleven published neighbour-based dimension results on the n/m axis.

Table 27: Eleven published designs on the n/m axis. Each is placed by arithmetic on the stated design, and no result is re-analysed. Most read in the class regime and describe the shape of classes. Three read at or above the class size and describe the layout of classes, two of them only under an inferred class size (†). The inner neighbour count is n, and TwoNN reads $n = 1$. The number of probe points per class is m. "inf." marks a value inferred from the stated design. †Neither paper states a class grouping; m counts points per token position and output class. Ruppik et al. (2025): about 3,000 last-layer token states pooled over the four positions of 'a o b =', $p = 197$, so $m \approx 3{,}000/(4 \cdot 197)$. Tang et al. (2026): 2,000 states pooled over two token positions, $p \in \{113, 149, 197\}$, so $m = 2{,}000/(2p)$, or 10 to 18 per output class alone.

paper	representation, probe	n	m	n/m	side of m
Ansuini et al. (2019), error	last layer; 2,000 ImageNet images uniform over 1,000 classes	1	2 (N/K; own class ≈ 3)	$\frac{1}{2}$ ($\frac{1}{3}$)	below
Ansuini et al. (2019), depth	each layer; seven classes, one at a time, 450 images per subsample	1	450	0.002	one class only
Recanatesi et al. (2019)	per-class manifolds, Fashion-MNIST and CIFAR-10; local TwoNN-type estimator (a global geodesic one uses $k = 20$)	1 (local)	not stated	$\ll 1$ if the full class	below
Doimo et al. (2020)	ImageNet, 300 classes of 300 images	30	300	0.1	below
Valeriani et al. (2023)	iGPT, 300 classes of 300 images, decimated fourfold	1	75	0.013	below
Doimo et al. (2024)	MMLU, 57 subjects, up to 200 each	16	≈ 160 (inf.)	≈ 0.1	below
Konz & Mazurowski (2024)	final layer, binary tasks, 500 to 1,750 images	1	250–875	≤ 0.004	below
Ma et al. (2018)	minibatches of 100 from 10-class data	20–30	≈ 10	2–3	at or above
Ruppik et al. (2025)	modular addition, local TwoNN in 64-point neighbourhoods	64	≈ 4 (inf.)†	≈ 16	above
Tang et al. (2026)	modular addition, 2,000 points, 64-point neighbourhoods	64	5–9 (inf.)†	7–13	above
Osman et al. (2026)	two 8B language models, 10,000 WikiText-103 samples grouped by last word	8–256 (their $k = 2n$)	10,000 / number of words	varied	both, by design; their $k = m$ is $n/m = \frac{1}{2}$

F.5 Choosing layers by the estimate

The vision encoders are 32 timm encoders of at least twelve blocks (DINO, DINOv2 with and without registers, MAE, CLIP and SigLIP variants, EVA-02, two supervised ViTs), read on 30,000 ImageNet training crops (100 classes of $m = 300$, blocks mean-pooled). Our operationalisation of the observation of Valeriani et al. (2023), that relative minima after the first peak are the most semantic layers, takes the block where the estimate is lowest after its global maximum (relative depth ≤ 0.8), with the published TwoNN fit or with GRIDE at $n/m \approx 0.003$, $\frac{1}{2}$ and 2. The loss is the best block's linear-probe top-1 accuracy (80/20 split) minus the chosen block's, in points. The language models are seven Pythia and OLMo models (160M to 7B), with the peak of Cheng et al. (2025),

which marks the onset of their abstraction phase, on last-token states. Models, rules, scales and probes were fixed in advance, before any feature was extracted.

Reading above m picks early blocks (relative depth near 0.2) in most vision encoders and costs far more than reading below it (Table 28). Read with the published fit, the rule already costs points, as predicted. Reading at $n/m = \frac{1}{2}$ does not beat it, and that comparison **failed**. In language models the published peak finds good sentiment layers. On topic text the peak moves with n/m, but reading in the band does not choose better layers; this comparison **failed** too. Two of the three comparisons were met in each part.

Table 28: **Choosing a layer by the estimate read above the class size costs far more than reading it below.** Every comparison but the last row was fixed in advance, before any feature was extracted. Two comparisons **failed**.

vision encoders, selection read at	median loss [IQR]	comparison fixed in advance
published fit ($n = 1$)	2.21 [1.36, 3.77]	median ≥ 2 points: met
GRIDE, $n/m \approx 0.003$	2.27 [1.63, 7.73]	
GRIDE, $n/m = \frac{1}{2}$	2.41 [1.85, 7.74]	better than published in 2, worse in 4, tied 26: **failed**
GRIDE, $n/m = 2$	36.22 [20.52, 40.74]	worse than $n/m = \frac{1}{2}$ in 24, better in 0, tied 8: met
language models		**outcome**
published peak within two points of the best IMDb sentiment layer		5 of 7: met
topic text (DBpedia-14): peak moves > 0.1 in depth, $n/m \approx 0.003$ to $\frac{1}{2}$		6 of 7: met
band peak chooses better topic layers than the published one		**failed**
post hoc: sign of the across-layer correlation with topic accuracy		$+$ at the published scale, $-$ at $n/m = \frac{1}{2}$, all seven

G Reflections on real networks

The reflections of Section 3.3 were run on the stratified probe ($K = 100$, $m = 20$), and every change reported here is a signed change in $z_n = \log \widehat{d}_n$. Theorem 2 and Proposition 2 are proved in Appendix A.5.

G.1 Two reflections and the floor

Two reflections were applied to real networks (Table 29). The first reflects each point through its class mean, $x_i \mapsto \mu_{y_i} - r_i$. It keeps every class mean, every within-class distance and every class covariance, so NC1 and NC2 do not change. Only cross-class squared distances change, by $-4\langle \mu_{y_i} - \mu_{y_j}, r_i - r_j \rangle$. The second, the head-null reflection $r_i \mapsto (I - 2P_{\ker W})r_i$, keeps every logit, class mean and within-class distance, but NC1 generally changes. It needs $D > K$, and on the 44 torchvision networks with $D \geq 1024$ it keeps the logits to 3×10^{-14}. For the Gaussian null, sampled and recentred clouds with each class's covariance go through the same code.

The odd part $o = \frac{1}{2}[z(X) - z(X^-)]$ of Theorem 2 is half the shift that the class-mean reflection makes, and its root-mean-square over the 222 networks is the floor. At $4m$ the floor is 30% of the estimate's spread across the original networks (Table 29, bottom). There the summaries can account for at most about 91% of the estimate's variance on the paired clouds.

Reflecting through the class means moves the estimate up below the class size and down above it, so a sign measured at one n cannot be carried to another. The change exceeds the Gaussian null on both the development and the held-out networks, though by less on the held-out ones. There the gate of 0.10 on the median $|\Delta z|$ at $4m$ was **missed** (0.079), and the signed median is smaller still. The head-null reflection moves the estimate at $n = m$ in the same direction in every network (Figure 9), although four networks with almost no residual energy in the null space (the three DenseNets and ViT-L/16) barely move. Gaussian clouds do not reproduce the odd part (median $|o| = 0.045$ against 0.013), and for Gaussian classes the expected shift is zero (Proposition 2).

G.2 What the reflection leaves fixed

We computed every statistic of the horse race (Table 21) that is defined on the stratified probe, a few close variants of them, hubness (Radovanović et al., 2010) and nearest-class-mean accuracy, on the original and reflected clouds of the development and held-out networks (Table 30). No criterion was

Table 29: Median changes in z_n under the two reflections, signed unless marked $|\cdot|$. Brackets give the 95% interval of the excess over the Gaussian null, or the IQR. The held-out networks (timm) and the gate were fixed before the effect was seen. Bottom: the floor, as the root-mean-square of o and as a share of the across-network standard deviation of z_n (in parentheses, families weighted equally).

reflection (networks)	statistic	$n = m/2$	$n = m$	$n = 4m$
class means, development (24)	median Δz	$+0.137$	-0.022	-0.106
	share moving up	96%	—	29%
	median $\lvert\Delta z\rvert$; gate ≥ 0.10	—	—	0.113: met
	excess over the null	—	—	$+0.111$ [$+0.064$, $+0.163$]
class means, held-out (55)	median Δz	$+0.157$	—	-0.033
	median $\lvert\Delta z\rvert$; gate ≥ 0.10	—	—	0.079: **missed**
	excess over the null	—	—	$+0.073$ [$+0.048$, $+0.099$]
Gaussian null (development)	median $\lvert\Delta z\rvert$	0.012	—	0.025
head-null, torchvision (44)	median Δz [IQR]	—	$+0.057$ [$+0.010$, $+0.134$]	—
	same direction; $\lvert\Delta z\rvert > 0.05$	—	100%; 52%	—
odd part o (the floor of Theorem 2)		$n = 1$	$n = 9$	$n = 4m$
class means (222)	root-mean-square of o	0.014	0.082	0.090
	share of the spread of z_n	10%	13%	30% (27%)



Figure 9: **A reflection that keeps every logit still moves the estimate.** Change in $\log \widehat{d}_n$ against n/m under the head-null reflection, which fixes every logit, class mean and within-class distance. The networks are the 44 torchvision networks with $D \geq 1024$. Horizontal band: the Gaussian null of the reflection through the class means (no null was run for this reflection), as in Figure 3C, the median over its 24 development and 55 held-out networks (Table 29 gives the development value alone); vertical band: the crossover band.

set for this analysis. Fifteen collapse, spectral and class-mean statistics are unchanged to rounding in every network, the largest change being 2×10^{-14} of the spread. They are NC1 $= \operatorname{tr}(\Sigma_W \Sigma_B^{+})/K$, NC2, CDNV, class separation R^2 in its squared Euclidean form, the within-class spread ratio τ, the within-class spread $\operatorname{tr} \Sigma_W$, the class-mean-distance coefficient of variation, the participation ratio, spectral decay α, RankMe (centred and uncentred), the gap statistics G_{2m} and G_{4m}, and both deleted estimates, the class-means-only $\log \widehat{d}^{\,0}_{4m}$ and the residuals-only $\log \widehat{d}^{\,\mathrm{res}}_{9}$.

Every neighbour statistic of the full cloud moves, among them the estimate at each neighbour count (most at $4m$ relative to its spread), the foreign fraction, hubness and class separation in its cosine form. So do 1-NN and nearest-class-mean accuracy, which the reflection raises in almost every network. The collapse summaries fix neither the estimate nor the probe's own neighbour accuracies.

Table 30: **The reflection through the class means moves every neighbour statistic of the full cloud and leaves the collapse, spectral and class-mean statistics unchanged.** $|\Delta|$ is the size of the change, on the log scale where marked, and /spread divides its median by the statistic's across-network standard deviation. The fifteen fixed statistics are listed in Appendix G.2. †Not in the horse race.

		development (24)		held-out (55)	
statistic	scale	median $\lvert\Delta\rvert$ [IQR]	/spread	median $\lvert\Delta\rvert$ [IQR]	/spread
moved by the reflection					
1-NN accuracy (leave-one-out)	lin	0.059 $[0.045, 0.072]$	0.54	0.069 $[0.048, 0.084]$	0.80
5-NN accuracy (leave-one-out)	lin	0.034 $[0.018, 0.043]$	0.29	0.036 $[0.018, 0.047]$	0.39
foreign fraction, $m/2$ neighbours	lin	0.020 $[0.012, 0.031]$	0.13	0.018 $[0.0078, 0.032]$	0.13
hubness (skewness, $k = 10$)†	lin	1.4 $[0.88, 2.8]$	0.38	1.8 $[0.80, 2.6]$	0.33
nearest-class-mean accuracy†	lin	0.032 $[0.028, 0.040]$	1.0	0.042 $[0.034, 0.053]$	1.3
class separation R^2 (cosine)	lin	0.0094 $[0.0062, 0.018]$	0.084	0.013 $[0.0064, 0.019]$	0.11
$\log \widehat{d}_1$	log	0.011 $[0.0051, 0.033]$	0.081	0.022 $[0.0059, 0.037]$	0.12
$\log \widehat{d}_{10}$ $(n = m/2)$	log	0.14 $[0.10, 0.18]$	0.17	0.16 $[0.11, 0.21]$	0.20
$\log \widehat{d}_{20}$ $(n = m)$	log	0.093 $[0.051, 0.14]$	0.32	0.091 $[0.033, 0.15]$	0.22
$\log \widehat{d}_{40}$ $(n = 2m)$	log	0.066 $[0.022, 0.19]$	0.24	0.076 $[0.032, 0.13]$	0.26
$\log \widehat{d}_{80}$ $(n = 4m)$	log	0.11 $[0.028, 0.19]$	0.41	0.079 $[0.036, 0.13]$	0.34

H THE LEDGER

For every main-text result, Table 31 gives what was fixed in a protocol file before its outcome existed and the outcome, with any analysis added afterwards named in the outcome column.

Table 31: **Every main-text result has a row in the ledger, failures included (in bold).** What was fixed in advance, and the outcome. The protocol files named in the supplementary ledger are included with it. Holding out models and fixing an endpoint in advance are distinct, and a row can have one without the other.

result	what was fixed in advance	outcome
reversal, torchvision and timm (51, 72)	image draws of the published design	both branches, Holm $p \le 6 \times 10^{-4}$; sweep and primary test added with the correlations known
reversal, held-out collection (55)	membership and pass rule, before any of these networks was read	both branches pass
within architecture families (18 fine, 24 broad)	fine-family rule and statistic; broad family map	$+0.80 \to -0.63$; every leave-one-out keeps both branches (added afterwards, with timm sibling families)
ResNet recipe groups (14, 14)	groups, neighbour counts, pass rule, before the features were read	both branches in both; stratified sweep (step at $n = m$) added afterwards
original fourteen (four architecture types)	reversal criterion; a forecast from two class statistics	never reverses; **the forecast failed**; training images, random subsets, class-means-only ordering added afterwards
boundaries: a ViT family; ImageNet at 64 px; error range per collection	a pass rule for each	**ViT fails**; **64 px never reverses**; error range: torchvision marginal, **timm fails** (first-row errors; 224-pixel: passes unadjusted); pooled range added afterwards (Appendix C)
training change (20 runs)	runs, checkpoints, grid, cells, exclusions, four rules, before training	compression $14/14$ cells, expansion 8/8, matching 8/8; exponent 1.04 inside $[0.6, 1.4]$; per-run exponent (1.02) added afterwards
training change, replications (20 runs re-read; 10 new; Appendix C.7)	new class sizes and images; switch criterion (runs already studied)	$60/60$ per allocation; $30/30$ in two further settings, counted afterwards
fixed-N designs collapse on n/m	the designs	R^2 0.94 on n/m, 0.60 on n; measure added afterwards
crossing line (19 designs; Table 20)	alignment of twelve designs on n/m	**rule held in 9 of 12: failed**; affine fit ($R^2 = 0.98$) added afterwards
rescaling and reshaping residuals (Appendix D)	a five-draw repeat, with the one-image test; crossing read afterwards	crossing moves towards $n = m$ ($23.6 \to 17.2$ at $K = 100$, $m = 20$; also $m \le 2$); added afterwards: rescaling takes r $+0.52 \to +0.40$ at $n = 1$, $+0.79 \to +0.77$ at $n = 10$; reshaping (sharper; flatter) moves median $\log \widehat{d}_n$ by -0.73; $+0.68$ at $n \le 7$, -0.08; -0.01 at $n \ge 30$
one image per class (222; no classmates)	design; equalising tightness must bring r to $\le +0.20$	$r = +0.580$; **bar failed** ($+0.35$); **means of $k = 49$ images failed** (predicted < 0, $+0.27$); **the bar under other tightness targets failed** ($+0.47$ at the 25th percentile); $k = 25$, 10 added afterwards
what moves the estimate, Theorem 1 (194)	the identity (a proof)	within 0.008 per network at $n = m$, occupancy term alone 0.066; 10% contraction and split added afterwards
deleted estimates (222; 18 families)	class-means-only estimate against error within families	within families $p \le 10^{-4}$; each reproduces the full estimate on its side of m (added afterwards, with k-means clusters)
surrogate (7 populations; Appendix E)	construction, populations, criteria; baselines; controls	beats baselines 6/7, 7/7; ties a constant on reversal status; **failed**: an earlier test; spectrum-only prediction for self-supervised encoders
distances to class centres (4 designs)	designs, placebos, three rules; later a re-centred variant, bar 3 of 4; a re-weighting of sampled points should move the branch	equal distances remove the negative branch 4/4 (positive on timm); **primary rule, equal class scales, failed** (2/4); **equal norms per class failed** (0/4); **re-centred failed** (2/4; secondary, whole classes: 3/4); **re-weighting failed** (3/18 cells)

Table 31, continued

result	what was fixed in advance	outcome
reflection through class means (24; 55 held out)	threshold 0.10 at $4m$	met on development (0.11), **missed held out** (0.08); head-null reflection (44) added afterwards
class-mean gap bound, NC2 (194)	bound (a proof); NC2 null families, seed, statistic	bound loose; raw NC2 correlation matched by the null; partial correlation added afterwards
split into class-means term and remainder (4 cohorts)	nothing (an identity)	descriptive; cohort-by-cohort reading (Appendix D) added afterwards
shell occupancy at $4m$, two CIFAR-100 regulariser sweeps (21; 20; Appendix D)	$\geq 2\times$ the ImageNet value and $\lvert r\rvert \geq 0.7$ in both	**failed**: share held in both; $r = +0.97$ in the weight-decay sweep, $+0.25$ in the label-smoothing and weight-decay sweep
layer profiles (13)	designs, layers, three rules	**primary rule failed** (10 of 13); other two pass; per-layer deleted estimates added afterwards
language models (26)	nothing	switch at the last layer; inversion on clinc (all post hoc)
the reversal within single runs (Appendix C.8)	sign pair $(+, -)$ of log error with the estimate at $m/2$ and $4m$ across each run's checkpoints, in $\geq \frac{3}{4}$ of the cohort and all 20 pre-registered runs	**failed**: 28 of 51 cohort runs (the $4m$ half holds in every run); 20 of 20 pre-registered runs
band additivity along the path to a simplex (remark in Appendix A.3)	L_n in the crossover band rises with residuals contracted to 0.1, in ≥ 18 of 24 networks; 0.3 and 1 secondary	**passed** (24 and 24 of 24 at $n = 10, 15$); at 0.3 also 24 and 24; fades at 1, the networks' own scale (20 and 7)
near-duplicate classes and the crossing (Appendix D)	three tests: on average 1.2–2.2 other class means within a class's median residual radius; class-means-only $r < 0$ from $n = 2m$ at $K = 1000$; the count's crossing shift, sibling-heavy against sibling-light	**all three failed**: count two orders of magnitude too large (230); negative only from $n = 4m$; shift of opposite sign (ratios to the prediction -0.070, -0.041, -0.11); a WordNet sibling count tried afterwards gives the direction, but the shift is three to six times smaller than it implies
two class summaries against the log gap: L_n regressed on log within-class spread, log dispersion of class-mean distances and their product (reported only here)	nothing (descriptive; model and reading rule fixed first)	R^2 0.92 at $n = m/2$, 0.39–0.73 above m; tightness slope largest in the band
one image per class, residual-norm dispersion (Appendix D)	equalising the dispersion of residual norms brings r at $n = 1$ to $\leq +0.20$	**failed**: $+0.57$ (the estimate barely moves, $+0.99$ with the original)
out-of-distribution error (Appendix C.4)	four shifted test sets and ReaL; the class-means-only estimate carries information beyond in-distribution error on at least two	switch holds on all five; **failed** the new-finding bar (only ImageNet-A, -0.35)
first-generation versus modern-recipe torchvision weights (Appendix B)	nothing (post hoc)	first-generation stay positive at $4m$ ($+0.31$), modern flip (-0.86)
what the reflection leaves fixed (Appendix G.2)	nothing (post hoc, no criterion)	fifteen collapse, spectral and class-mean statistics unchanged; every neighbour statistic moves, 1-NN and nearest-class-mean accuracy included
collapse statistics along training (Appendix C.8)	within-run tracking of NC1 at $m/2$ and NC2 at $4m$; the $n = 1$ sign late in training	NC2: met (-1.00); **NC1 failed** ($+0.49$, bar $+0.70$); **late positive branch failed** ($+0.24$, bar $+0.40$)
network scale (Appendix C.3)	partial correlations given log parameter count, criterion on both designs	**passed**: both branches keep their sign ($+0.71$, -0.63 published; $+0.56$, -0.70 stratified)
plateau rule for the neighbour count (Appendix F.3)	nothing (post hoc)	published design: the flattest pair starts at or above m in 155 of 178 networks; $r = -0.44$ there
horse race (Table 21)	nothing; the neighbour count of $\widehat{d}_{10}$ was chosen on the outcomes	nearest-class-mean and 1-NN accuracy track error most closely (0.95, 0.93; Table 18); the estimate's held-out gain beyond it unresolved
two-count check, ten groups (Figure 6)	nothing; the ResNet-50 recipes are not a pre-fixed group	seven flip, the original fourteen unresolved, the two CIFAR-100 sweeps keep their sign (all post hoc)
other estimators (torchvision and timm, 4 cells)	which estimators reverse, before computation	GRIDE and three ratio estimators 4/4: met; **moment-form MLE pattern failed**; **local PCA reverses in 0/4: failed**
layer selection (32 encoders)	encoders, rule, scales, probes, three comparisons, before extraction	the rule read with the published fit costs ≥ 2 points: met; above m worse than $n/m = \frac{1}{2}$: met; $n/m = \frac{1}{2}$ **better than published: failed**
layer selection, language models (7)	models, rule, probes, three comparisons, before extraction	published peak near best sentiment layer, and peak moving with n/m on topic text: met; **band peak picks better topic layers: failed**; estimate–probe correlation sign added afterwards