From Wasserstein Distances to Wasserstein Barycenters: Prediction-Space Ensembling in Set-Valued Knowledge Distillation

Pengcheng Xu et al.

University of California, Irvine

Abstract

When a student must reproduce a *set* of outputs rather than a single label, say a predicted cell population, a feature map or a distribution over classes, knowledge distillation is pursued almost entirely through *objective design*: optimal transport, kernel MMD, correlation and concordance surrogates, robust and precision-weighted moment matching. On a fixed teacher/student/data triple we show that changing the *shape* of the loss is capped in advance and exhausted in practice, while two first-moment levers stay open and compose. Averaging in *prediction space* across independently seeded students gains $+0.0179$ on the primary metric (t=13.6 over 70 held-out perturbations, $70/70$ improved) against a control emitting the same number of predicted cells, and $+0.022$ against one checkpoint emitting as many *distinct* cells. Ten loss changes, from entropic optimal transport and MMD through Lin's concordance to sliced-Wasserstein and shape-residual legs, nine trained to convergence under one protocol with same-seed pairing, give six ties, three losses and a collapse; of two post-hoc corrections, shrinkage does not help, while a leak-free offset calibration that corrects the predicted mean gains $+0.0145$ and adds to averaging, though by removing a shift shared by every perturbation. Two properties of the evaluation metric say why in advance: it quotients out a two-dimensional gauge, so reweighting has no first-order reason to help, and it reads the predicted set only through its first moment, capping what a change of shape can buy through the mean at ≈ 0.006, a cap the data attain. An ambiguity identity in the metric's own seminorm then splits the remaining error into the *seed variance of the trained function*, 26% measured without a fit, which no tested loss reaches but averaging does, and a part shared by every seed, from which calibration removes a common bias. We give the identity under which pooling and the W_2 barycenter score identically, show that snapshots halve the training cost ($+0.0007$, t=1.2) and that pooling composes with the objective (for every tested loss), and document two benchmark confounds with the controls that remove them: a seed-derived evaluation control block worth up to 0.0177, and a predicted-cell-count effect worth up to 0.34 on differential-expression metrics.

1 Introduction

Knowledge distillation is usually framed as a search for the right discrepancy between teacher and student outputs, natural for a single vector and far more tempting when the output is a *set*, because a rich menu of distributional discrepancies opens up: optimal transport, kernel two-sample statistics, moment matching, correlation surrogates. Much recent set-valued distillation work walks that menu. We report that on a fixed teacher, student and dataset the loss's *shape* can be exhausted while two simpler levers are untouched: train K students and average their *predictions*, and remove the shared bias of the predicted mean using a held-out fold. Averaging gains $+0.0179$ against a control emitting exactly the same number of cells; ten loss changes, spanning loss form, weighting and target estimator, give six ties, three losses and a collapse under one protocol; the offset correction gains $+0.0145$; and the two levers add.

The contribution is not the arithmetic mean but the account of *why* the two axes differ, stated before the experiments rather than after. The strongest published objective here is a Wasserstein *distance* between teacher and student sets; the remaining headroom is reached not by refining it but by moving one level up, to a Wasserstein *barycenter* of many predicted sets, whose first moment averaged predictions share (Proposition 1). Two properties of the evaluation metric (Sec. 3.1) bound the objective axis in advance: it reads the predicted set only through its per-coordinate mean, and it quotients out a two-dimensional gauge. The second gives an expansion (Proposition 2) whose per-coordinate error weight is *uniform* on that quotient, so reweighting objectives have no first-order reason to help; the first gives a *measured* cap of ≈ 0.006, a third of the pooling gain, on the channel through which second- and higher-moment changes reach it (Proposition 3). Both bound only the *direct* channel, to the order stated: what a reweighting or a change of shape (entropic OT, kernel MMD, concordance's scale term) can buy through the predicted mean, not the gradient interference such a leg causes, which we measure.

Neither bounds *averaging*, which does not change what any single run optimizes, nor a correction of the predicted mean itself. The ambiguity decomposition (Krogh and Vedelsby 1994) in the metric's own seminorm (Propositions 4–5, supplementary material) splits the converged student's residual into an averageable and a shared part and puts 26% of it, measured without a fit, in the *seed variance of the trained function*, a quantity a per-step loss contains no term for. The shared part holds a bias common to every seed, which no pool size reaches but a first-moment correction can: leak-free offset calibration removes it and, on the pri-

mary metric, adds to pooling: a calibrated $K{=}5$ pool scores 0.9120 against our 0.9025. That asymmetry, not the mean itself, is the finding, and it is not the *weight*-space one: weight averaging (SWA, model soups) is a within-basin approximation that ties on our converged image bed, whereas the diversity that pays is *between* basins. Neither cost is inherent: five runs $\times$ two snapshots match ten independent runs ($+0.0007$, $t{=}1.2$) for half the training, and distilling the pool into one student recovers $+1.20$ top-1 on CIFAR-100 at $1\times$ inference over the same KD recipe taught by the original teacher.

Contributions. **(i)** prediction-space averaging for set-valued distillation, $+0.0179$ ($t{=}13.6$, $70/70$) against an exactly output-size-matched control, with variants that halve its training cost and part of its inference cost; **(ii)** two structural caps on what changing the loss's shape can buy, derived from the evaluation metric rather than fitted to the results, and a trained-to-convergence test of ten loss changes against them reporting intervals rather than verdicts; **(iii)** a first-moment coincidence, a corollary of Agueh and Carlier (2011) made exact by the metric: pooled predictions and the members' W_2 barycenter score identically here, with a prediction of when they would not; **(iv)** the ambiguity decomposition, taken in the metric's seminorm, which splits the residual into seed variance, which no tested loss reaches but averaging does, and shared error, from which a leak-free first-moment calibration removes a bias that adds to averaging; and **(v)** two benchmark confounds with their controls, worth up to 0.0177 and 0.34.

2 Related Work

Distillation objectives. Classical distillation matches softened logits under KL (Hinton, Vinyals, and Dean 2015); DIST (Huang et al. 2022) replaces KL with an affine-invariant Pearson-correlation distance. WKD (Lv, Yang, and Li 2024) brings optimal transport to distillation, and its feature form WKD-F reduces, under the diagonal covariance its own ablation prefers, to matching per-coordinate means and standard deviations, precisely a set-valued objective in our sense; our baseline is that objective transplanted, with cells as spatial positions, and we claim no novelty for it. Past a diagonal Gaussian one compares the sets as distributions: entropic optimal transport (Cuturi 2013), its debiased Sinkhorn-divergence form (Feydy et al. 2019), already used for distillation (Cui et al. 2024b), and maximum mean discrepancy (Gretton et al. 2012) are present inside our own teacher, and Proposition 3 bounds their direct channel. Lin's coefficient (Lin 1989) factors as $\mathrm{CCC} = \rho \cdot C_b$, so our concordance arm, the single-cell analogue of DIST, is the metric with its affine gauge penalized instead of granted; it is not a reweighting, so we tested rather than predicted it, and an industrial perturbation model reports adopting it at scale (Wang et al. 2026).

Ensembling, weight averaging and evaluation practice. Deep ensembles (Lakshminarayanan, Pritzel, and Blundell 2017) convert seed diversity into accuracy and calibration and our result is an instance of that mechanism, whose necessity is debated (Abe et al. 2022; Theisen et al. 2023); new here are the comparison against objective design at matched inference parameters and the contrast with its *weight*-space relatives, SWA (Izmailov et al. 2018) along one trajectory and model soups (Wortsman et al. 2022) across fine-tuning runs. Snapshot ensembles (Huang et al. 2017) pool cyclic-schedule checkpoints; ours needs no cyclic schedule. Dognin et al. (2019) already ensemble models by a Wasserstein barycenter, over a label-space ground metric for classifiers; what we add is that on a first-moment metric the barycenter and plain pooling coincide, and when they part. Ensemble distillation that tries to keep the members' diversity, by distribution distillation (Malinin, Mlodozeniec, and Gales 2020), multi-head students (Tran et al. 2021) or diversity-preserving perturbations with weight averaging (Nam et al. 2021, 2022), is the natural next step after our Sec. 7, which tests only plain pool distillation. Allen-Zhu and Li (2023) argue ensemble gains *can* be distilled back into one model, which Sec. 7 partly confirms; our claim is that no *loss on the teacher–student pair* reaches the averaged-out variance, a property of the training procedure rather than of the teacher's outputs. Stanton et al. (2021) find students rarely match teacher fidelity even when distillation "works", a mismatch (5) makes quantitative. Our reporting follows the reality-check genre (Musgrave, Belongie, and Lim 2020; Narang et al. 2021; Reimers and Gurevych 2017; Bouthillier et al. 2021; Demšar 2006; Card et al. 2020).

Barycenters, and single-cell perturbation prediction. Agueh and Carlier (2011) characterize the W_2 barycenter as the law of the weighted average under an optimal multi-marginal coupling, so its mean is the weighted mean of the members'; in one dimension, quantile and probability averaging share their mean, the latter being wider (Lichtendahl, Grushka-Cockayne, and Winkler 2013). We use this to relate pooling to barycentric aggregation and to predict when the two come apart (Sec. 4.3). Perturbation-response models range from graph-prior regressors (Roohani, Huang, and Leskovec 2024) through transcriptome foundation models (Cui et al. 2024a) to set-valued models trained on large chemical compendia (Adduri et al. 2025; Zhang et al. 2025; Gandhi et al. 2025). Ahlmann-Eltze, Huber, and Anders (2025) show current deep perturbation predictors do not beat simple linear or additive baselines, which is why every claim here is a paired difference against a fixed baseline on one fixed triple, not a leaderboard position. Mejía et al. (2025) show unweighted delta metrics can reward mode collapse and propose DE-weighted alternatives, several of which Sec. 6.2 finds confounded by predicted-cell count. An independent sweep of an MMD term's weight, on another dataset, reports degradation at larger weights on delta-correlation, MSE and MAE, with small gains on distribution-shape metrics (Jiang et al. 2026), the pattern our shape objectives show.

3 Setting

A perturbation model receives a cell line's control cells and a drug identity and predicts the perturbed cell population over $G = 2000$ highly variable genes, a set-valued output. The teacher is a released STATE checkpoint (Adduri et al. 2025) (108.7M parameters) held out on cell line C32 of Tahoe-100M (Zhang et al. 2025); the student is a 3.21M-parameter

set-transformer of the same family. The split is the teacher authors' own: 348 training, 56 validation and 735 test perturbations, of which 70 are scored on the evaluation plate. No arm sees a validation or test perturbation's real cells. Every arm comparison here is a within-seed paired difference under the single protocol of Sec. 8; the supplementary material locates this regime in the baseline hierarchy.

3.1 The metric, and its two relevant properties

With $\hat{\mu}_p$ the predicted per-gene mean of perturbation p, μ^{ctrl} the control mean and μ_p the real perturbed mean, the primary metric is the Pearson correlation *across the G genes* of the mean-difference vectors, averaged over the $P = 70$ evaluation perturbations:

$$\begin{aligned} \text{pearson}\Delta &= \tfrac{1}{P}\textstyle\sum_p \operatorname{corr}_g\big(\hat{\Delta}_p,\ \Delta_p\big), \\ \hat{\Delta}_p &= \hat{\mu}_p - \mu^{\text{ctrl}}, \qquad \Delta_p = \mu_p - \mu^{\text{ctrl}}. \end{aligned} \tag{1}$$

(M1) The predicted set enters only through its per-gene mean: the evaluator reduces each cloud to a bulk mean before correlating. **(M2)** The metric is invariant to $\hat{\Delta} \mapsto a\hat{\Delta} + b\mathbf{1}$ for $a > 0$, i.e. it quotients out $\operatorname{span}\{\mathbf{1}, \Delta_c\}$, which is two directions, not one.

3.2 The baseline

The strongest objective we could field is WKD-F (Lv, Yang, and Li 2024), matching the teacher cloud's per-gene mean and standard deviation. From

$$\begin{aligned} W_2^2\big(\mathcal{N}(\mu_1,\Sigma_1),\mathcal{N}(\mu_2,\Sigma_2)\big) &= \|\mu_1 - \mu_2\|^2 \\ &\quad + \operatorname{tr}\big(\Sigma_1 + \Sigma_2 - 2(\Sigma_2^{1/2}\Sigma_1\Sigma_2^{1/2})^{1/2}\big), \end{aligned} \tag{2}$$

the Bures term collapses under diagonal Σ to $\sum_g(\sigma_{1g}-\sigma_{2g})^2$ (verified against a full eigendecomposition to 1.8×10^{-15}), giving

$$\mathcal{L}_{\text{WKD-F}} = \rho\,\|\Delta\mu\|^2 + \|\Delta\sigma\|^2. \tag{3}$$

For every $\rho > 0$ this is the *square* of a metric on diagonal Gaussians, so $\sqrt{\mathcal{L}_{\text{WKD-F}}}$ is a distance while $\mathcal{L}_{\text{WKD-F}}$ is not, and it coincides with W_2^2 *only* at $\rho = 1$ against a shipped default of $\rho = 2$, at which it is a legitimate distance on (μ, σ) but not the transport distance (supplementary material). The baseline scores 0.8882 over seeds 0–4 (sd 0.0057, se 0.0025) and 0.8889 over seeds 5–9, above the teacher's own 0.8472.

4 Distill, Then Average

Train K students under the same objective and protocol, differing only in seed; at inference emit each student's predicted cell set and *pool the sets*: no retraining, no distillation of the ensemble, no change to the loss (Figure 1, which also draws the fork between prediction and weight space). An ensemble of K students emits K times as many predicted cells, so the first objection is that any gain is statistical power rather than model quality. Its exact control duplicates a *single* checkpoint K times: identical model, identical predicted mean, $K\times$ the cells. The objection then fails by construction: by (M1) the evaluator reduces each cloud to a bulk mean, so duplication is a no-op. The $K{=}5$ duplicated control matches the single model exactly on all 70 perturbations, while the $K{=}5$ ensemble at the *same* 1280 predicted cells per perturbation scores 0.9025. Duplication holds the predicted mean fixed, so it cannot ask whether pooling is merely more *distinct* samples. The sharper control emits 1000 distinct cells per perturbation from one checkpoint, through five disjoint 200-cell control-input blocks, against five checkpoints at the same 1000 cells. Distinct cells alone gain $+0.0036$ and $+0.0074$ on two disjoint pools' members, and the pools beat that control by a further $+0.0217$ and $+0.0216$ (t=19.4 and 18.4, 70/70 perturbations each). Finite-emission averaging is therefore a small part of the gain, not its explanation.

Metric	single	matched ctrl	Δ ensemble
pearsonΔ	0.8847	0.8847	+0.0179
DE Spearman (LFC, sig.)	0.7712	0.7712	+0.0594
DE direction match	0.9925	0.9925	+0.0025
precision@100	0.1446	0.0456	+0.2224
overlap@100	0.1326	0.0377	+0.2087
discrimination (cosine)	0.9851	0.9851	+0.0008
MAEΔ ↓	0.0087	0.0087	−0.0007
PR-AUC	0.2328	0.3259	+0.0096
ROC-AUC	0.7436	0.7916	+0.0014
clustering agreement	0.4206	0.4206	+0.1075
DE gene recall	0.3126	0.6487	−0.0652
DE Spearman (sig.)	0.0204	−0.1955	−0.1456
discrimination (ℓ^1)	0.9908	0.9908	−0.0035
discrimination (ℓ^2)	0.9927	0.9927	−0.0012

Table 1: The $K{=}5$ prediction ensemble against an *exactly* output-size-matched control: 1280 predicted cells per perturbation in both right-hand columns, the middle one the same checkpoint with $5\times$ the cells, isolating the cell-count confound of Sec. 6.2. The bottom block is all four metrics the ensemble loses on against this seed-0 member: the two DE losses are the confounded ones, DE gene recall gaining $+0.336$ from sample size alone and then *falling* once that is controlled. The two discrimination scores rank bulk mean effects, so they are first-moment metrics; their losses are within noise (per-perturbation 95% CIs $[-0.011, +0.004]$ and $[-0.006, +0.003]$) and reverse against the pool's own five members ($+0.0041$, $+0.0027$), seed 0 being the best of twelve.

4.1 Result

On the primary metric the paired per-perturbation gain is $+0.0179$ (paired se 0.0013, $t = 13.6$ on 69 df, 70/70 perturbations improve; Figure 2). That interval is over perturbations, so it establishes that *this* pool beats *this* model, not that pooling reliably beats its members. For that we cut 12 independently seeded baseline runs into three *disjoint* pools of $K{=}4$, no member appearing in two pools so that the three contrasts are independent, and emit all 15 through one shared control block. Each pool beats the mean of its own members ($+0.0228$, $+0.0229$, $+0.0283$; over pools $+0.0246$, replicate se 0.0018, t=13.5, n=3), and each beats its own *best* member by $+0.0168$, $+0.0172$ and $+0.0162$, a standard deviation of 0.0005 and *tighter* than the comparison against the mean, so the gain is not a lucky member laundered by four mediocre

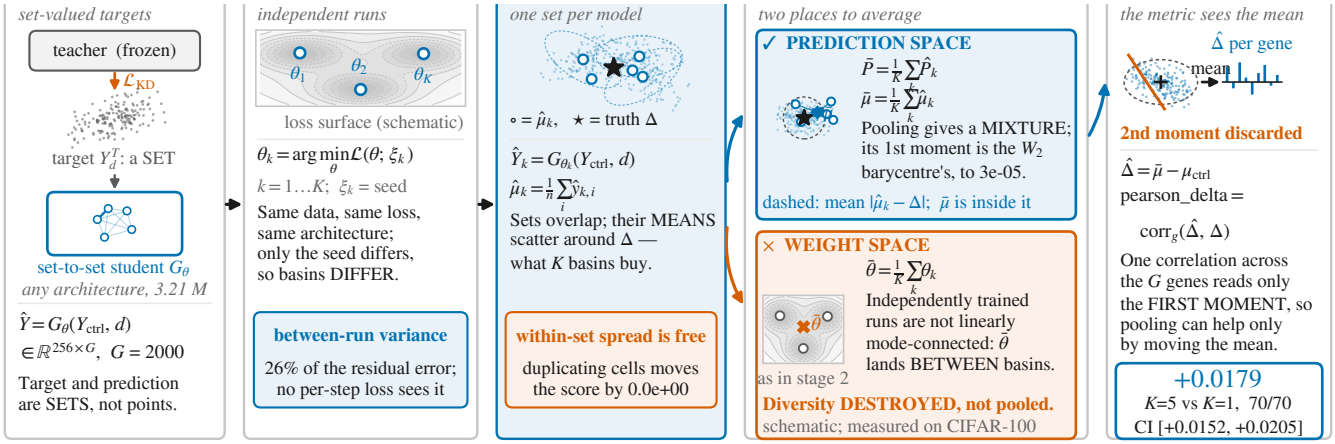

Figure 1: The method, in five stages: **(1)** a frozen teacher supplies *set*-valued targets; **(2)** K students differ only in seed and settle in different basins; **(3)** their sets overlap but their *means* scatter; **(4)** the fork: prediction-space averaging pools the sets, *weight* averaging places $\bar{\theta}$ between basins; **(5)** the *primary* metric reads only the first moment. Both loss-surface panels are schematic and labelled as such; the weight-space arm is measured in Sec. 7.

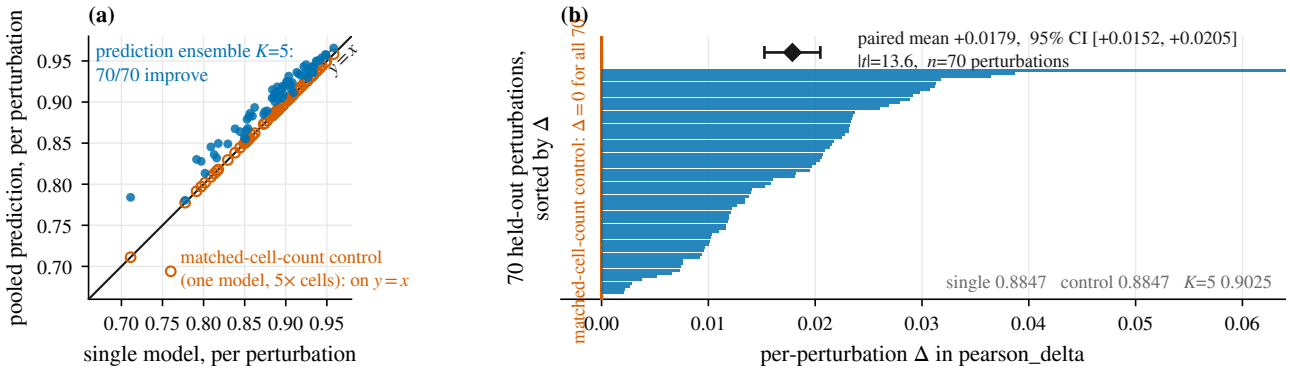

Figure 2: The main result, paired per perturbation on one shared control block. **(a)** All 70 held-out perturbations improve under the K=5 ensemble; the matched-cell-count control (5× the cells, one model) lies exactly on $y = x$. **(b)** Paired differences.

ones. The K=5 ensemble closes 30% of the gap from its seed-0 member to the metric's measured noise ceiling (0.9451, an oracle expectation; supplementary material).

Across all 26 scored metrics 22 improve and four degrade, and Table 1 lists all four rather than only the ones that flatter us: two are the differential-expression metrics most inflated by the cell-count confound, and the other two are rank-based discrimination scores that read only bulk means: their losses against seed 0 are inside per-perturbation noise and become gains against the pool's own members, because seed 0 is the best of the twelve. Against those five members the pool improves every metric with a per-perturbation interval except the count-confounded DE family (supplementary material).

4.2 Snapshots: the same gain at zero extra training

Every checkpoint file already stores two states: the validation-selected iterate and the live iterate at write time, two points on one trajectory typically 100–200k steps apart. Pooling both doubles K for free. At *matched* member and emitted-cell count, ten members drawn as five runs × two snapshots score +0.0007 over ten drawn as ten independent runs (paired se 0.0006, $t = 1.2$, $[-0.0005, +0.0019]$, 34/70 perturbations): indistinguishable at this n, at half the training compute. Nor does one wider model work: five copies of the student beat a same-scale monolith by +0.0185 at matched inference parameters, so the ceiling is one of *variance*, not capacity (supplementary Figure S3).

4.3 Pooling versus the barycenter

Averaging distributions has a canonical answer, the Wasserstein barycenter $\bar{\nu} = \arg\min_\nu \sum_k w_k W_2^2(\nu, \nu_k)$. Is arithmetic pooling a cheap surrogate for it? Not on this metric.

Proposition 1 (First-moment coincidence) *Let $\nu_1, \ldots, \nu_K$ have finite second moments and let $w_k \geq 0$ with $\sum_k w_k = 1$. Then any W_2 barycenter satisfies $\text{mean}(\bar{\nu}) = \sum_k w_k \text{mean}(\nu_k)$, which is also the mean of the pooled mixture $\sum_k w_k \nu_k$. By (M1), (1) depends on the predicted set only through this mean. Hence barycentric*

aggregation and arithmetic pooling receive the same *score in exact arithmetic.*

This is a corollary of the Agueh–Carlier characterization (Agueh and Carlier 2011) (proof in the supplementary material); what is specific here is (M1), which turns a shared mean into a shared score.

On five real members, emitting their diagonal-Gaussian barycenter (mean of means, mean of SDs) instead of the mixture moves the score by -3.2×10^{-5} (paired $t=-2.54$), two orders of magnitude below every effect here and an artifact of the emission's non-negativity clamp (supplementary material); we report it rather than claim exact agreement.

Proposition 1 says nothing about the second moment: the pooled cloud is a *mixture*, its variance inflated by between-member disagreement, so the two generically differ there and only their means coincide. We claim not that our ensemble *is* the optimal-transport aggregation but that on a first-moment metric the two are indistinguishable and generically come apart on metrics with a genuine second-moment component, as the DE family confirms: emitting the barycenter drops precision@100 from 0.293 to 0.004. The supplementary material gives the Gaussian characterization and the counterexample (0.966) showing it needs Gaussianity.

5 What Changing the Loss Can and Cannot Reach

Proposition 2 (Uniform weighting on the quotient) *Let $\eta=\hat{\Delta}-\Delta$, let subscript c denote gene-centering, and let $P_\perp$ project onto the complement of Δ_c. Then to second order in η_c,*

$$1-\mathrm{corr}_g(\hat{\Delta},\Delta)\;\approx\;\frac{\|P_\perp\eta_c\|^2}{2\|\Delta_c\|^2}. \tag{4}$$

The proof is in the supplementary material: the along-Δ component of the error cancels in the Pearson quotient and so enters only at higher order. Empirically (4) recovers $0.938\times$ the exact deficit here, an operating-point-specific cancellation of higher-order terms rather than a transferable constant.

The per-gene error weight implied by (4) is $1-\frac{1}{G}-\Delta_{c,g}^2/\|\Delta_c\|^2$: uniform *on the quotient* by $\mathrm{span}\{\mathbf{1},\Delta_c\}$, but in gene coordinates the *complement* of the reference difference's own energy profile. The metric therefore *down*-weights exactly the genes a differential-expression-weighted objective *up*-weights, sharper than uniformity and pointing the same way (supplementary material). The consequence we claim is narrow: an objective that reweights genes (robust losses, Mahalanobis or precision weighting, differential-expression weighting) is *no first-order reason to help*: it optimizes against a weighting the metric does not have. We explicitly do *not* claim it cannot help at all: (4) says nothing about the estimator's bias–variance tradeoff, and its η is measured against the real difference vector while training uses the teacher's. Table 2's ties are the evidence; (4) is the reason to expect them.

Proposition 3 (First-moment cap) *Assume the emitted set is used unmodified, in particular that it is not clipped to a valid expression range. Then under (M1), a change to the objective that acts only through the predicted set's second or higher moments reaches* (1) *only through the sampling variance $\mathbb{E}\|z_c\|^2$ of the predicted per-gene mean, and can therefore* gain *at most $\mathbb{E}\|z_c\|^2/(2\|\Delta_c\|^2)$ in the sense of* (4)*. The bound is one-sided: it does not limit how much such a change can lose.*

The premise (M1) is verified exactly: duplicating a 256-cell block five times changes the metric by $0.000\mathrm{e}{+00}$. The proposition itself is tested by the one operation that removes z without moving the first moment: emitting more *distinct* cells from one checkpoint, through disjoint control-input blocks. Estimated with no independence assumption between cells, as the variance of the emitted mean across disjoint 256-cell input blocks and averaged per perturbation as (1) does, the cap is 0.0064 (95% CI $[0.0058, 0.0071]$ over perturbations, 14 checkpoints); extrapolating each checkpoint's deficit to infinitely many distinct cells gains 0.0068 $[0.0061, 0.0075]$. The cap is reached, as expected of an operation that removes z entirely and so gains the projected term (0.0063), the small excess being of the size of (4)'s second-order error, so we quote ≈ 0.006 (proof and sweep in the supplementary material). Three qualifications keep this honest, the third found by trying to break the proposition. It is *one-sided*: re-drawing from an inflated cloud costs up to -0.62, two orders of magnitude past the cap. It bounds only the *direct* channel, while a real auxiliary loss also perturbs shared parameters. And the no-clipping hypothesis is not cosmetic: our emit path clamps and the converged student is 93% exact zeros, so a *strictly mean-preserving* 5% variance inflation, in which the mean moves by 4×10^{-15}, still costs -0.0103, the clamp rectifying the widened lower tail and shifting the mean after the fact. That second channel is real and negative. So the supportable statement is: *the first-moment-mediated component of any second-moment objective is capped at ≈ 0.006 here; everything else acts through the clamp or the shared parameters*, which we observe to be negative.

5.1 The test

Table 2 is the whole objective axis: among the changes to the loss, six ties, three losses and one collapse, none excluding zero on the positive side; of the post-hoc corrections, shrinkage is refuted and the single gain is the first-moment offset calibration, which neither cap covers; Table S1 and Figure S2 give per-seed values and intervals.

Protocol, and the two objections it anticipates. Every arm runs the protocol of Sec. 8 *paired by seed*: the auxiliary legs draw no additional RNG, so an arm and the baseline at seed s share initialization, batch order and the evaluation control block of Sec. 6. Against undertraining (Beyer et al. 2022): the median arm peaks at 192k steps and runs 200k more, and the reported score is that best-validation checkpoint. Against tuning parity (Musgrave, Belongie, and Lim 2020): each arm is reported at its best configuration over the grids in the supplementary material, except entropic OT, shown at the median-heuristic blur it was first trained at; swept in revision (blur 3, 9, 27; n=3 off the default), it at best ties per seed (-0.0030).

Objective change	n	Δ	95% CI	verdict
Concordance γ=0.3	5	$+0.0015$	$[-0.0043, +0.0073]$	tie
MMD, calibrated bw	5	-0.0005	$[-0.0076, +0.0066]$	tie
MMD, default bw	5	-0.0036	$[-0.0168, +0.0097]$	tie
Leg-ratio ρ=4	5	-0.0008	$[-0.0065, +0.0050]$	tie
Concordance $C_b^{0.5}$	5	-0.0037	$[-0.0074, -0.0001]$	loss
Rao–Blackwell target	5	-0.0029	$[-0.0123, +0.0065]$	tie
Calibration fold only	5	-0.0060	$[-0.0102, -0.0017]$	loss
Entropic OT	5	-0.0143	$[-0.0218, -0.0069]$	**loss**
Anti-collapse	1	-0.2978		– collapse
Dir. shrinkage	–	$\leq +0.0005$, then < 0		refuted
Sliced W (centred)‡	5	-0.0317	$[-0.0731, +0.0096]$	tie†
Shape residual‡	5	-0.0194	$[-0.0265, -0.0123]$	loss
Offset calibration, Eq. (19)	5	$+0.0145$	$[+0.0102, +0.0188]$	**gain**

Table 2: The objective axis, each arm a within-seed paired difference against the WKD-F baseline (0.8882, n=5) with two-sided t intervals on $n-1$ df; the pairing unit is seeds here and 70 perturbations for the ensemble of Sec. 4, so the two are not on one axis. Anti-collapse is a single run and shrinkage a post-hoc sweep, so they carry no interval. We claim no equivalence: no interval lies inside ± 0.005. Under a Bonferroni correction across the eight pre-specified arms with an interval, only entropic OT remains significant ($p = 0.0058$, adjusted 0.046), and it is a *loss*; no loss change is a win after correction. ‡Added in revision under the same protocol. †All five seeds below the baseline (per-perturbation t=-20.3); one seed (0.794) widens the interval. The last row is Eq. (19) implemented, a post-hoc correction of the predicted mean; the fold row is its cost.

How we report ties. We report two-sided t intervals on $n-1$ df and claim non-significance, not *equivalence* (Lakens 2017). What the table supports is a bound on the upside, the quantity a reader wants from a negative result (Card et al. 2020): with every arm now at n=5, by each loss-side arm's *upper* confidence limit no change to the loss is consistent with an improvement larger than $+0.010$, against $+0.018$ for the ensemble of Sec. 4 at an interval width of 0.005. That is per-arm, not simultaneous: Bonferroni over the eight pre-specified loss-side arms raises the limit to $+0.022$ on the seed-paired footing, and to $+0.003$ when each arm's seed-averaged differences are paired over the 70 perturbations, the ensemble's own footing.

Reading the table. The caps bind and the failures are not transport's. Entropic OT's -0.0143 is far past the ≈ 0.006 Proposition 3 allows its direct channel, so it is gradient interference rather than transport geometry: a leg that cannot help can still hurt. Anti-collapse is catastrophic because the magnitude collapse it targets (Wang et al. 2026) does not exist under the baseline's *absolute* moment matching.

5.2 The channel that is left

The seed-variance channel is the one quantity we can both bound from the loss side and *measure* from the ensemble side, and the two must agree; the pool-size curve of Figure S1 measures it. We cut twelve independently seeded students into *disjoint* pools at each size, twelve of one through two of six, so that no member is counted twice at any K and the spread at a given K is replicate error rather than a redescription of one draw. Fitting the metric deficit against pool size,

$$1-\mathrm{corr}(K) \;=\; c_\infty + \frac{v}{K}, \qquad c_\infty = 0.0926, \quad v = 0.0332, \tag{5}$$

with $R^2 = 0.999$, splits the single model's residual into an averageable $v/(c_\infty + v) = 26\%$ and a remaining 74% shared across seeds, made of reference measurement noise and bias common to every run, that no pool size reaches. The curve therefore saturates rather than continuing, and levels off sooner than the $1/K$ decay of independent, equal member errors.

That 26% is the quantity the argument turns on, and the largest single effect in this paper. No per-step loss targets it: a loss is written one example at a time and contains no term for the variance of the learned *function* across training runs, so it can reduce that variance only indirectly; no entry in Table 2 addresses it, not because those objectives are poor but because no object of that type represents the quantity. Averaging *does* reach it, and by construction the measured gain is what (5) predicts. Nor is the form specific to this bed or metric (Lobacheva et al. 2020): empirically it fits three CIFAR-100 beds ($R^2 \geq 0.988$, supplementary material), where the second member again buys about half the total gain, so what transfers is the form and not the size of the share, which rises with the capacity gap. Disjointness matters: a nested curve, whose K=1 point is the best of the twelve members, gave a biased 20% (supplementary material, which also withdraws an earlier five-way decomposition that did not sum to 100%). The fit is now a consistency check rather than the evidence: Propositions 4–5 (supplementary material), the ambiguity decomposition (Krogh and Vedelsby 1994) and the bias–variance–covariance law (Ueda and Nakano 1996) in the seminorm of (4), derive (5) for exchangeable seeds and measure v directly from member disagreement, without a fit, at 26.4% (perturbation bootstrap $[25.5, 27.3]$%); refitting (5) on 400 random disjoint re-partitions gives $[25.1, 28.3]$%.

What the shared part leaves to a correction. c_∞ is not one thing: it holds the reference's measurement noise, the sampling error of the shared control block, and a bias common to every seed. No pool size reaches it and no change of loss shape is first-order helpful against it, but a direct correction of the predicted mean lies outside both caps. Eq. (19) of the supplementary material estimates the shared offset on a fold of training perturbations supervised exactly as the evaluation ones (teacher only) and subtracts it: $+0.0145$ $[+0.0102, +0.0188]$ at n=5 (Table 2), net of the fold's own cost (-0.0060). Because it acts on c_∞ while pooling acts on v/K, the two add: a calibrated K=5 pool scores 0.9120, $+0.0094$ $[+0.0077, +0.0111]$ over our uncalibrated K=5 pool (t=11.0, 64/70). The bias it removes is shared by every perturbation, which control-referenced metrics reward (Viñas Torné et al. 2025): against the perturbed centroid Eq. (19) ties ($+0.0027$ $[-0.0029, +0.0083]$) while pooling still gains $+0.032$ $[+0.025, +0.040]$ (64/70).

CIFAR-100, RN32×4 → RN8×4	top-1	ECE
single student (n=6)	76.06 ± 0.15	0.0389
within-run SWA, epochs 200+240	75.96 ± 0.17	–
prediction pool, K=6	$\mathbf{78.48}$	$\mathbf{0.0197}$
across-run weight soup, K=6	1.05	–
teacher (reference)	79.42	0.0845
KD from the single teacher (n=6)	73.43 ± 0.20	–
KD from the pool (n=6)	$\mathbf{74.63 \pm 0.22}$	–

Table 3: The cross-domain bed, authors' configuration, final-epoch checkpoints. Pooling gains $+2.42$; averaging weights *within* a run gains nothing, while averaging *across* independent runs is not a model at all. Below the first rule: a student taught by the pool gains $+1.20 \pm 0.08$ (t=15.7, 6/6 seeds) at $1\times$ inference, and that pooled teacher is *less* accurate than the one it replaces while $4.3\times$ better calibrated. Chance is 1.00; bold marks the best among comparable rows; the pool is over softmax, but over logits as a distillation target.

6 Benchmark Hygiene

Two confounds here are large enough to change conclusions, and neither is specific to our method.

6.1 The evaluation control block is seed-derived

The held-out control block is built from the run seed and is *both* the model's input *and* the evaluator's difference reference. Re-scoring one checkpoint against another block moves its score by up to 0.0177; resampling the reference alone has standard deviation 0.0052, comparable to the whole seed-to-seed spread. Comparisons across blocks partly measure the block, which is why every objective-axis number here is a within-seed difference and the ensemble comparisons share one block.

6.2 Predicted-cell count moves differential-expression metrics

The matched-count column of Table 1 isolates the effect with the model held fixed. Multiplying emitted cells by five moves DE gene recall by $+0.336$, PR-AUC by $+0.093$ and ROC-AUC by $+0.048$, but precision@100 by -0.099, overlap@100 by -0.095 and DE Spearman by -0.216. The mechanism is direct: the genes the evaluator calls significant rise from 150.9 to 485.5 against a real count of 100.5, purely from statistical power. Comparing methods that emit different numbers of cells is uninterpretable on these metrics without a matched control; the DE-weighted metrics proposed as better calibrated (Mejía et al. 2025) are the ones most sensitive to output-set size.

7 Cross-Domain Evidence

Is any of the above a property of pearsonΔ? We re-run the *method* claim where task, metric and architecture all differ: CIFAR-100 at WKD-F's published pair (ResNet32×4 → ResNet8×4), through the authors' released trainer and untouched configuration, varying only the seed. Our six students score 76.06 ± 0.15, 0.71 below the published 76.77 ± 0.26. We use the *final*-epoch checkpoint throughout: this harness evaluates on the test set, so best-epoch selection would be selection on it.

Table 3's two weight-space rows separate a claim a single soup number cannot: weight averaging is bounded by the basin its members share, whereas prediction pooling is not; that shared basin is the condition Wortsman et al. (2022) require for souping, and independently initialized members violate it.

The bottom block answers the objection that pooling costs $K\times$ inference. With objective, schedule, seed and code path identical, only the teacher differs; every seed improves, and the control is within 0.1 of the published KD figure (73.33). The distilled student recovers part of the pool's benefit, not all of it: it stays below the full pool and below a student trained with the stronger WKD-F recipe (74.63 against 76.06). The objective-side evidence remains the cell-side campaign of Sec. 5; the image bed carries the *method* claim only. At matched inference parameters, whether one model beats the pool depends on *which* model the budget buys: a deeper published architecture wins, a naively widened student loses (supplementary material).

8 Reproducibility

One frozen protocol: 800k-step ceiling, 120k minimum, patience 200, batch 256, $G = 2000$ raw genes, WKD-F weight 1.0 and ratio 2.0 unless swept, Adam at constant learning rate 10^{-3}. The 70 scored perturbations are the teacher's test perturbations present on the evaluation plate with at least 100 cells (a deterministic filter); the control block is 2000 of that plate's DMSO cells drawn and permuted with the run seed, whose second half is the evaluator's reference. Scores are best-validation checkpoints at n=5 seeds (baseline n=10), compared within seed, by the teacher authors' released evaluator (`cell-eval` v0.6.8).

9 Conclusion and Limitations

For set-valued distillation on a fixed teacher, student and dataset, changing the loss's shape was exhausted, as the caps of Sec. 5 predict, while averaging, which acts on the seed variance of the trained function (26% of the residual), was not, and survives a perturbed-centroid reference; a leak-free offset correction adds to it on the primary metric by removing a shift every perturbation shares. On any first-moment metric, pooling is indistinguishable from barycentric aggregation.

Two limits: one teacher, student, cell line and primary (mean-response) metric, with cross-domain evidence for the method half only; and "exhausted" means no detected improvement among the variants tested, not equivalence: on distribution shape a shape objective beats our pool, though pooling its members helps it.

References

Abe, T.; Buchanan, E. K.; Pleiss, G.; Zemel, R.; and Cunningham, J. P. 2022. Deep Ensembles Work, But Are They Necessary? In *Advances in Neural Information Processing Systems*, volume 35.

Adduri, A. K.; Gautam, D.; Bevilacqua, B.; Imran, A.; Shah, R.; Naghipourfar, M.; Teyssier, N.; Ilango, R.; Nagaraj, S.; Dong, M.; Ricci-Tam, C.; et al. 2025. Predicting cellular responses to perturbation across diverse contexts with State. bioRxiv. Preprint, bioRxiv 2025.06.26.661135.

Agueh, M.; and Carlier, G. 2011. Barycenters in the Wasserstein Space. *SIAM Journal on Mathematical Analysis*, 43(2): 904–924.

Ahlmann-Eltze, C.; Huber, W.; and Anders, S. 2025. Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines. *Nature Methods*, 22(8): 1657–1661.

Allen-Zhu, Z.; and Li, Y. 2023. Towards Understanding Ensemble, Knowledge Distillation and Self-Distillation in Deep Learning. In *International Conference on Learning Representations (ICLR)*. OpenReview.net.

Beyer, L.; Zhai, X.; Royer, A.; Markeeva, L.; Anil, R.; and Kolesnikov, A. 2022. Knowledge distillation: A good teacher is patient and consistent. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10915–10924. IEEE.

Bouthillier, X.; Delaunay, P.; Bronzi, M.; Trofimov, A.; Nichyporuk, B.; Szeto, J.; Mohammadi Sepahvand, N.; Raff, E.; Madan, K.; Voleti, V.; Ebrahimi Kahou, S.; Michalski, V.; Arbel, T.; Pal, C.; Varoquaux, G.; and Vincent, P. 2021. Accounting for Variance in Machine Learning Benchmarks. In *Proceedings of Machine Learning and Systems (MLSys)*, volume 3, 747–769.

Card, D.; Henderson, P.; Khandelwal, U.; Jia, R.; Mahowald, K.; and Jurafsky, D. 2020. With Little Power Comes Great Responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 9263–9274. Association for Computational Linguistics.

Cui, H.; Wang, C.; Maan, H.; Pang, K.; Luo, F.; Duan, N.; and Wang, B. 2024a. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21(8): 1470–1480.

Cui, X.; Qin, Y.; Gao, Y.; Zhang, E.; Xu, Z.; Wu, T.; Li, K.; Sun, X.; Zhou, W.; and Li, H. 2024b. Sinkhorn Distance Minimization for Knowledge Distillation. arXiv:2402.17110.

Cuturi, M. 2013. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In *Advances in Neural Information Processing Systems*, volume 26, 2292–2300. Curran Associates, Inc.

Demšar, J. 2006. Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, 7: 1–30.

Dognin, P.; Melnyk, I.; Mroueh, Y.; Ross, J.; Santos, C. D.; and Sercu, T. 2019. Wasserstein Barycenter Model Ensembling. In *International Conference on Learning Representations*.

Feydy, J.; Séjourné, T.; Vialard, F.-X.; Amari, S.-i.; Trouvé, A.; and Peyré, G. 2019. Interpolating between Optimal Transport and MMD using Sinkhorn Divergences. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, 2681–2690. PMLR.

Gandhi, S.; Javadi, F.; Svensson, V.; Khan, U.; Jones, M. G.; Yu, J.; Merico, D.; Goodarzi, H.; and Alidoust, N. 2025. Tahoe-x1: Scaling Perturbation-Trained Single-Cell Foundation Models to 3 Billion Parameters. bioRxiv. Preprint, bioRxiv 2025.10.23.683759.

Gretton, A.; Borgwardt, K. M.; Rasch, M. J.; Schölkopf, B.; and Smola, A. J. 2012. A Kernel Two-Sample Test. *Journal of Machine Learning Research*, 13: 723–773.

Hinton, G. E.; Vinyals, O.; and Dean, J. 2015. Distilling the Knowledge in a Neural Network. NIPS 2014 Deep Learning Workshop, arXiv:1503.02531.

Huang, G.; Li, Y.; Pleiss, G.; Liu, Z.; Hopcroft, J. E.; and Weinberger, K. Q. 2017. Snapshot Ensembles: Train 1, Get M for Free. In *International Conference on Learning Representations (ICLR)*. OpenReview.net.

Huang, T.; You, S.; Wang, F.; Qian, C.; and Xu, C. 2022. Knowledge Distillation from A Stronger Teacher. In *Advances in Neural Information Processing Systems*, volume 35, 33716–33727. Curran Associates, Inc.

Izmailov, P.; Podoprikhin, D.; Garipov, T.; Vetrov, D. P.; and Wilson, A. G. 2018. Averaging Weights Leads to Wider Optima and Better Generalization. In *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence (UAI)*, 876–885. AUAI Press.

Jiang, D.; Yan, Z.; Wang, Q.; An, Z.; Zhao, Y.; and Lai, L. 2026. OCOO-T: A Simple and Scalable Virtual Cell Model for Transcriptional Perturbation Response Prediction. Preprint, arXiv:2606.12838.

Krogh, A.; and Vedelsby, J. 1994. Neural Network Ensembles, Cross Validation, and Active Learning. In Tesauro, G.; Touretzky, D.; and Leen, T., eds., *Advances in Neural Information Processing Systems*, volume 7, 231–238. MIT Press.

Lakens, D. 2017. Equivalence Tests: A Practical Primer for *t* Tests, Correlations, and Meta-Analyses. *Social Psychological and Personality Science*, 8(4): 355–362.

Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2017. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In *Advances in Neural Information Processing Systems*, volume 30, 6402–6413. Curran Associates, Inc.

Lichtendahl, K. C.; Grushka-Cockayne, Y.; and Winkler, R. L. 2013. Is It Better to Average Probabilities or Quantiles? *Management Science*, 59(7): 1594–1611.

Lin, L. I.-K. 1989. A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics*, 45(1): 255–268.

Lobacheva, E.; Chirkova, N.; Kodryan, M.; and Vetrov, D. 2020. On Power Laws in Deep Ensembles. In *Advances in Neural Information Processing Systems*, volume 33.

Lv, J.; Yang, H.; and Li, P. 2024. Wasserstein Distance Rivals Kullback-Leibler Divergence for Knowledge Distillation. In *Advances in Neural Information Processing Systems*, volume 37, 65445–65475. Curran Associates, Inc.

Malinin, A.; Mlodozeniec, B.; and Gales, M. 2020. Ensemble Distribution Distillation. In *International Conference on Learning Representations*.

Mejía, G. M.; Miller, H. E.; Leblanc, F. J. A.; Wang, B.; Swain, B.; and de Lima Camillo, L. P. 2025. Diversity by Design: Addressing Mode Collapse Improves scRNA-seq Perturbation Modeling on Well-Calibrated Metrics. Preprint, arXiv:2506.22641.

Musgrave, K.; Belongie, S. J.; and Lim, S.-N. 2020. A Metric Learning Reality Check. In *Computer Vision – ECCV 2020*, volume 12370 of *Lecture Notes in Computer Science*, 681–699. Springer.

Nam, G.; Lee, H.; Heo, B.; and Lee, J. 2022. Improving Ensemble Distillation With Weight Averaging and Diversifying Perturbation. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvari, C.; Niu, G.; and Sabato, S., eds., *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 16353–16367. PMLR.

Nam, G.; Yoon, J.; Lee, Y.; and Lee, J. 2021. Diversity Matters When Learning From Ensembles. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, volume 34, 8367–8377. Curran Associates, Inc.

Narang, S.; Chung, H. W.; Tay, Y.; Fedus, L.; Fevry, T.; Matena, M.; Malkan, K.; Fiedel, N.; Shazeer, N.; Lan, Z.; Zhou, Y.; Li, W.; Ding, N.; Marcus, J.; Roberts, A.; and Raffel, C. 2021. Do Transformer Modifications Transfer Across Implementations and Applications? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 5758–5773. Association for Computational Linguistics.

Reimers, N.; and Gurevych, I. 2017. Reporting Score Distributions Makes a Difference: Performance Study of LSTM-networks for Sequence Tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 338–348. Association for Computational Linguistics.

Roohani, Y.; Huang, K.; and Leskovec, J. 2024. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nature Biotechnology*, 42(6): 927–935.

Stanton, S.; Izmailov, P.; Kirichenko, P.; Alemi, A. A.; and Wilson, A. G. 2021. Does Knowledge Distillation Really Work? In *Advances in Neural Information Processing Systems*, volume 34, 6906–6919. Curran Associates, Inc.

Theisen, R.; Kim, H.; Yang, Y.; Hodgkinson, L.; and Mahoney, M. W. 2023. When Are Ensembles Really Effective? In *Advances in Neural Information Processing Systems*, volume 36.

Tran, L.; Veeling, B. S.; Roth, K.; Swiatkowski, J.; Dillon, J. V.; Snoek, J.; Mandt, S.; Salimans, T.; Nowozin, S.; and Jenatton, R. 2021. Hydra: Preserving Ensemble Diversity for Model Distillation. arXiv:2001.04694.

Ueda, N.; and Nakano, R. 1996. Generalization Error of Ensemble Estimators. In *Proceedings of International Conference on Neural Networks (ICNN'96)*, volume 1, 90–95. IEEE.

Viñas Torné, R.; Wiatrak, M.; Piran, Z.; Fan, S.; Jiang, L.; Teichmann, S. A.; Nitzan, M.; and Brbić, M. 2025. Systema: a Framework for Evaluating Genetic Perturbation Response Prediction beyond Systematic Variation. *Nature Biotechnology*, 44(6): 1050–1059.

Wang, C.; Karimzadeh, M.; Ravindra, N. G.; Bounds, L. R.; Alerasool, N.; Huang, A. C.; Ma, S.; Gulbranson, D. R.; Cui, H.; Lee, Y.; et al. 2026. X-Cell: Scaling Causal Perturbation Prediction Across Diverse Cellular Contexts via Diffusion Language Models. bioRxiv. Preprint, bioRxiv 2026.03.18.712807.

Wortsman, M.; Ilharco, G.; Gadre, S. Y.; Roelofs, R.; Gontijo Lopes, R.; Morcos, A. S.; Namkoong, H.; Farhadi, A.; Carmon, Y.; Kornblith, S.; and Schmidt, L. 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 23965–23998. PMLR.

Zhang, J.; Ubas, A. A.; de Borja, R.; Svensson, V.; Thomas, N.; Thakar, N.; Lai, I.; Winters, A.; Khan, U.; Jones, M. G.; et al. 2025. Tahoe-100M: A Giga-Scale Single-Cell Perturbation Atlas for Context-Dependent Gene Function and Cellular Modeling. bioRxiv. Preprint, bioRxiv 2025.02.20.639398.

Supplementary Material
From Wasserstein Distances to Wasserstein Barycenters: Prediction-Space Ensembling in Set-Valued Knowledge Distillation

Pengcheng Xu et al.

University of California, Irvine

1 Setting, Hardware and Provenance

Baseline hierarchy. For orientation, the hierarchy on this teacher/student/data triple runs: mean-prediction floor -0.009, real cells alone 0.448, teacher alone 0.607, teacher-plus-real 0.724, and the WKD-F objective 0.886, against a measured metric ceiling of 0.9451. These reference points were measured by a collaborating site under a 500k-step variant of our protocol; they locate the operating regime only, and every arm comparison in the main paper is a within-seed paired difference under the single protocol it specifies.

The metric ceiling. That 0.9451 is a split-half reliability estimate of the reference, Spearman–Brown corrected. It is an *oracle expectation* for the infinite-sample mean difference and captures sampling noise only; it is not a mathematical upper bound, and a predictor that reproduced the 500-cell real mean exactly would score 1.0. It is measured against the same seed-derived control block the arms are scored on; the full-plate reference gives 0.9469. We use it only as a scale against which to read the ensemble's gain.

A trap in re-scoring. The seed-derived evaluation control block of the main paper is also a tooling hazard: a re-scoring utility that hard-codes one seed will silently produce block-mismatched comparisons, which is exactly the error an audit of an earlier draft of this paper caught.

The baseline loss is a squared metric, and only at ρ=1 is it W_2^2. The WKD-F loss of the main paper, $\mathcal{L}_{\text{WKD-F}} = \rho\|\Delta\mu\|^2 + \|\Delta\sigma\|^2$, is for every $\rho > 0$ the squared weighted ℓ^2 norm of $(\sqrt{\rho}\,\Delta\mu, \Delta\sigma)$, hence the *square* of a metric on diagonal Gaussians: $\sqrt{\mathcal{L}_{\text{WKD-F}}}$ is a distance while $\mathcal{L}_{\text{WKD-F}}$ itself, like any squared metric, violates the triangle inequality, as three diagonal Gaussians on the line at $\mu = 0, 1, 2$ with equal σ show. The square coincides with W_2^2 only at $\rho = 1$, whereas the shipped default is $\rho = 2$ (on on converged emissions the mean leg carries 10.2% of the objective at ρ=1 and 18.6% at ρ=2); at $\rho = 2$ it is W_2^2 up to the reweighting of the mean leg, a legitimate distance on (μ, σ) but not the transport distance.

Mechanisms behind the related work the main paper cites. DIST replaces the KL objective because an exact point-wise match is harmful when the teacher is much stronger than the student. WKD's feature form models each feature map as a Gaussian over spatial positions under the closed-form 2-Wasserstein distance, and its own ablation prefers a diagonal covariance to the full Bures term, which is why the deployed objective reduces to matching per-coordinate means and standard deviations. Entropic optimal transport and its debiased Sinkhorn-divergence form interpolate between optimal transport and maximum mean discrepancy.

Hardware and throughput. All runs are on one $8\times$B200 node, two to three training processes per GPU (about 9.8 GB each), which does not affect numerics because each process has its own CUDA context, seeded initialization and batch sequence. Throughput at that packing is roughly 65 steps/s for the baseline and 30 steps/s for the entropic-OT arm, the latter dispatch-bound rather than FLOP-bound at ≈ 300-cell clouds. The median arm reaches its best validation score at 192k steps and stops at roughly 392k.

The teacher's simulated target is regenerable from public artifacts alone: the released checkpoint plus the public ground-truth file reproduce it to the documented distributional signature in about one minute of GPU time, and pushing a certified checkpoint through the regenerate-then-score path reproduces its score to 3×10^{-10}.

2 Proofs

The six headline results are formalized for finite emitted clouds and their vector errors in Lean 4 with Mathlib (`rebuttal/lean/project`), with no unproved steps and only Lean's standard axioms.

Throughout, gene vectors lie in $\mathbb{R}^G$, $G \geq 2$, $\mathbf{1}$ is the all-ones vector, $x_c = x - G^{-1}(\mathbf{1}^\top x)\mathbf{1}$, Δ and $\hat{\Delta}$ are the real and predicted mean changes from the same control mean μ_{ctrl}, and $\Delta_c \neq 0$. Set $A = \|\Delta_c\|^2$ and let $P_\perp$ be the orthogonal projector onto $\operatorname{span}\{\mathbf{1}, \Delta_c\}^\perp$; on centered vectors it projects onto the complement of Δ_c. Write $Q(e) = \|P_\perp e\|^2/(2A)$ for the quadratic deficit in (4) of the main paper. Expectations are on a probability space, conditional on the fixed reference and checkpoint where appropriate; all random vectors below have finite second moments. Pearson scores are evaluated on nonzero centered vectors.

Proposition 1 of the main paper (first-moment coincidence). Pooling and the Wasserstein barycenter have the same mean, so a score that reads only that mean treats them equally.

Statement. Let $\nu_1, \dots, \nu_K$ be probability measures with finite second moments and let $w_k \geq 0$ with $\sum_k w_k = 1$. Then any W_2 barycenter $\bar{\nu}$ satisfies $\mathrm{mean}(\bar{\nu}) = \sum_k w_k \,\mathrm{mean}(\nu_k)$, also the mean of $\sum_k w_k \nu_k$. By (M1), (1) of the main paper depends only on this mean, so the two aggregations receive the same score in exact arithmetic.

Intuition. Translating a cloud toward the weighted mean of the members lowers its transport cost until the two means coincide. This is the first-moment consequence of Agueh–Carlier (Agueh and Carlier 2011); in one dimension it is the agreement of probability and quantile averages (Lichtendahl, Grushka-Cockayne, and Winkler 2013).

Proof. Write $\mu(\nu) = \int x \, d\nu$, $F(\nu) = \sum_k w_k W_2^2(\nu, \nu_k)$ and $\mu_* = \sum_k w_k \mu(\nu_k)$. Translation by t bijects the transport plans and changes every plan's cost by $2\langle t, \mu(\nu) - \mu(\nu_k)\rangle + \|t\|^2$, by expanding $\|x + t - y\|^2$ and using the marginals. Taking infima and summing gives $F(\nu^t) = F(\nu) + 2\langle t, \mu(\nu) - \mu_*\rangle + \|t\|^2$. At a minimizer, $t = \mu_* - \mu(\bar{\nu})$ would lower F by $\|t\|^2$, so $t = 0$. Linearity gives the mixture's mean, and (M1) gives equality of scores.

Paper connection. This supports the pooling–barycenter comparison, tested by the implementation's diagonal moment rescale with a reported score difference of -3.2×10^{-5} after the emission clamp.

Remark (spread and emission). For diagonal Gaussian members with means μ_k and coordinate standard deviations σ_{kg}, the Gaussian barycenter has mean μ_* and standard deviation $\sum_k w_k \sigma_{kg}$: coordinatewise optimal coupling reduces its objective to weighted squared distances between means and standard deviations. The mixture instead has coordinate variance $\sum_k w_k[\sigma_{kg}^2 + (\mu_{kg} - \mu_{*g})^2]$, by expanding around μ_*. For general one-dimensional laws, the barycenter's quantile is the weighted quantile average (Lichtendahl, Grushka-Cockayne, and Winkler 2013): the centered, unit-variance uniform and exponential quantiles $\sqrt{12}(s - 1/2)$ and $-\log(1 - s) - 1$, $0 < s < 1$, have inner product $\sqrt{3}/2$, so their equal-weight barycenter has standard deviation $\sqrt{1/2 + \sqrt{3}/4} \approx 0.966$. In the shipped emission the per-gene rescale to the mean member SD can exceed one by up to $\sqrt{1279/1275}$, because pooled and member SDs use different degrees of freedom, and the non-negativity clamp that follows shifts the mean slightly.

Proposition 2 of the main paper (uniform weighting on the quotient). To second order, Pearson deficit measures the squared error perpendicular to the true change.

Statement. Let $\eta = \hat{\Delta} - \Delta$, let subscript c denote gene-centering, and let $P_\perp$ project onto the complement of Δ_c in the centered space. Then, to second order in η_c, $1 - \mathrm{corr}_g(\hat{\Delta}, \Delta) \approx \|P_\perp \eta_c\|^2 / (2\|\Delta_c\|^2)$, as in (4) of the main paper.

Intuition. Correlation measures an angle: a small sideways displacement turns the prediction, whereas a small displacement along the reference changes its length. Gene-centering first removes the constant direction.

Proof. Put $\alpha = \langle \eta_c, \Delta_c\rangle / A$ and $e = P_\perp \eta_c$, so $\Delta_c + \eta_c = (1 + \alpha)\Delta_c + e$ is an orthogonal sum. For $1 + \alpha > 0$, set $r^2 = \|e\|^2 / [(1 + \alpha)^2 A]$; Pythagoras gives $\mathrm{corr}_g = 1/\sqrt{1 + r^2}$ exactly. For $f(x) = 1 - (1 + x)^{-1/2}$, $f(0) = 0$, $f'(0) = 1/2$ and $-3/4 \leq f''(x) \leq 0$ on $x \geq 0$; integrating twice gives $x/2 - 3x^2/8 \leq f(x) \leq x/2$. Thus $r^2/2 - 3r^4/8 \leq 1 - \mathrm{corr}_g \leq r^2/2$. With $\varepsilon = \|\eta_c\|/\|\Delta_c\| \to 0$, $|\alpha| \leq \varepsilon$ and $r^2 = O(\varepsilon^2)$, so $(1 + \alpha)^{-2} = 1 + O(\varepsilon)$ gives deficit $Q(\eta_c) + O(\varepsilon^3)$. Centering and cancellation of a positive scalar in the cosine also give (M2).

Paper connection. This supports the metric's uniform quadratic weighting on the quotient; (4) of the main paper recovers $0.938\times$ the measured exact deficit at this operating point.

Remark (gene coordinates). The orthonormal vectors $\mathbf{1}/\sqrt{G}$ and $\Delta_c/\sqrt{A}$ give $P_\perp = I - \mathbf{1}\mathbf{1}^\top/G - \Delta_c\Delta_c^\top/A$. For coordinate unit vector e_g, symmetry and idempotence give $(P_\perp)_{gg} = \|P_\perp e_g\|^2 = 1 - 1/G - \Delta_{c,g}^2/A \in [0, 1 - 1/G]$; summing gives trace $G - 2$. The full form is $x^\top P_\perp x = \sum_g (P_\perp)_{gg} x_g^2 + \sum_{g \neq h} (P_\perp)_{gh} x_g x_h$, so the diagonal measures an isolated gene error's sensitivity; its measured range is 0.758 to 0.9995.

Proposition 3 of the main paper (first-moment cap). At a fixed emitted mean, removing sampling noise can recover only its current contribution to the quadratic deficit.

For this result write $\hat{\mu} = m + z$, where $m = \mathbb{E}\hat{\mu}$ is fixed, $\mathbb{E}z = 0$, and $b = m - \mu_{\mathrm{ctrl}} - \Delta$; a change acting only through higher moments keeps m fixed. *Statement.* Assume the emitted set is used unmodified, in particular that it is not clipped to a valid expression range. Then under (M1), a change to the objective that acts only through the predicted set's second or higher moments reaches (1) of the main paper only through the sampling variance $\mathbb{E}\|z_c\|^2$ of the predicted per-gene mean, and can therefore gain at most $\mathbb{E}\|z_c\|^2/(2\|\Delta_c\|^2)$ in the sense of (4) of the main paper. The bound is one-sided: it does not limit how much such a change can lose.

Intuition. The error is a fixed displacement plus a cloud of mean-zero fluctuations. Collapsing that cloud to its center removes its spread while retaining the displacement.

Proof. Linearity of $P_\perp$ and $\mathbb{E}z_c = 0$ remove the cross term: $\mathbb{E}Q(b_c + z_c) = Q(b_c) + \mathbb{E}\|P_\perp z_c\|^2/(2A)$. For two such emissions z_1, z_2, put $V_j = \mathbb{E}\|P_\perp z_{j,c}\|^2/(2A)$. Their gain is exactly $V_1 - V_2 \leq V_1 \leq \mathbb{E}\|z_{1,c}\|^2/(2A)$, using nonnegativity and contraction of orthogonal projection. This is the claimed cap on the quadratic deficit.

Paper connection. This supports the measured finite-emission cap: at 256-cell blocks the projected and unprojected estimates are 0.0063 and 0.0064, compared with the exact-score extrapolated gain 0.0068.

Remark (attaining the cap). The first inequality is an equality exactly when $P_\perp z_{2,c} = 0$ almost surely; setting $z_2 = 0$ achieves it. The orthogonal sum $z_c = P_\perp z_c + (\langle z_c, \Delta_c\rangle/A)\Delta_c$ gives $\mathbb{E}\|z_c\|^2 - \mathbb{E}\|P_\perp z_c\|^2 = \mathbb{E}\langle z_c, \Delta_c\rangle^2/A$. Thus the second inequality is an equality exactly when $\langle z_{1,c}, \Delta_c\rangle = 0$ almost surely, since a nonnegative integrable function has zero expectation exactly when it vanishes almost surely.

Remark (exact deficit). For the same fixed b, assume z_1, z_2 have finite fourth moments and $0 < \ell \leq t_j = 1 + \langle b_c + z_{j,c}, \Delta_c\rangle/A \leq u$ almost surely. Set $q_j = Q(b_c + z_{j,c})$, $d_j =$

$1-(1+2q_j/t_j^2)^{-1/2}$, $\beta = Q(b_c)$ and $R_2 = \mathbb{E}q_2^2$. The scalar bounds in Proposition 2 of the main paper give $q_j/u^2 - 3q_j^2/(2\ell^4) \le d_j \le q_j/\ell^2$; integrating, using $\mathbb{E}q_j = \beta + V_j$ and $V_2 \ge 0$, yields $\mathbb{E}d_1 - \mathbb{E}d_2 \le V_1/\ell^2 + \beta(\ell^{-2} - u^{-2}) + 3R_2/(2\ell^4)$. If $\|b_c + z_{j,c}\| \le \varepsilon\|\Delta_c\|$ almost surely as $\varepsilon \to 0$, take $\ell = 1-\varepsilon$, $u = 1+\varepsilon$: $\beta, V_j = O(\varepsilon^2)$ and $R_2 = O(\varepsilon^4)$ recover $V_1 + O(\varepsilon^3)$.

Proposition 4 (pooling gain is disagreement). Pooling removes the part of the squared error on which the members disagree.

Use quotient errors $u_k = P_\perp(\hat{\mu}_k - \mu_{\mathrm{ctrl}} - \Delta)_c$ and $\bar{u} = K^{-1}\sum_k u_k$, for $K \ge 1$. *Statement.* (i) $K^{-1}\sum_k \|u_k\|^2 - \|\bar{u}\|^2 = K^{-1}\sum_k \|u_k - \bar{u}\|^2$. (ii) For random members with common moments $\mathbb{E}\|u_k\|^2 = a$ and $\mathbb{E}\langle u_k, u_l\rangle = c$ for $k \ne l$, $\mathbb{E}\|\bar{u}\|^2 = c + (a-c)/K$; for $K \ge 2$, $\mathbb{E}\|u_1 - u_2\|^2 = 2(a-c)$.

Intuition. Draw each error as an arrow: the pool keeps the centroid of the arrows and removes their spread around it. Common second moments, supplied for example by exchangeable members, allow that spread to be summarized by one number.

Proof. Expand $\|u_k - \bar{u}\|^2$ and sum; the cross term is $2K\|\bar{u}\|^2$, proving (i), the Krogh–Vedelsby ambiguity identity (Krogh and Vedelsby 1994). For (ii), expand the squared sum: its expectation has K diagonal terms a and $K(K-1)$ terms c, so division by K^2 gives the claim; expanding $\|u_1 - u_2\|^2$ gives the last identity. All cross terms are integrable by Cauchy–Schwarz. This is the Ueda–Nakano bias–variance–covariance calculation (Ueda and Nakano 1996). Dividing by $2A$ gives $\mathbb{E}Q(\bar{e}) = c_\infty + v/K$, where $\bar{e} = K^{-1}\sum_k(\hat{\mu}_k - \mu_{\mathrm{ctrl}} - \Delta)_c$, $c_\infty = c/(2A)$ and $v = (a-c)/(2A) = \mathbb{E}\|u_1 - u_2\|^2/(4A)$ for $K \ge 2$.

Paper connection. This gives the quadratic form behind (5) of the main paper, whose fit to exact Pearson deficits has $R^2 = 0.999$.

Remark (shared error). For a sequence of exchangeable seeds with the same moments at every pool size, nonnegativity of $c + (a-c)/K$ gives $c \ge 0$ as $K \to \infty$, while pairwise disagreement gives $a - c \ge 0$. When $a > 0$, $v/(c_\infty + v) = (a-c)/a$ is therefore the fraction of quadratic error removable by pooling. The shared term includes common bias and correlated reference, control-block and seed errors.

Proposition 5 (disagreement measures the finite pooling gain). The same measurement predicts the gain from pooling seeds or from pooling more emitted blocks.

Statement. For $K \ge 2$ square-integrable vectors U_k with common moments $\mathbb{E}\|U_k\|^2 = a$ and $\mathbb{E}\langle U_k, U_l\rangle = c$ for $k \ne l$, put $\bar{U} = K^{-1}\sum_k U_k$ and $S^2 = (K-1)^{-1}\sum_k \|U_k - \bar{U}\|^2$. Then $\mathbb{E}S^2 = a - c$ and $\mathbb{E}\|U_1\|^2 - \mathbb{E}\|\bar{U}\|^2 = (1-1/K)\mathbb{E}S^2$.

Intuition. The members' spread around their own mean reveals precisely the error that averaging removes. The factor $1 - 1/K$ accounts for the spread still present in a finite pool.

Proof. Proposition 4(i) gives $\sum_k \|U_k - \bar{U}\|^2 = \sum_k \|U_k\|^2 - K\|\bar{U}\|^2$ on each realization. Taking expectations and using Proposition 4(ii) gives $(K-1)(a-c)$; divide by $K-1$. Subtracting $c + (a-c)/K$ from a gives the second formula. For seeds, take $U_k = u_k/\sqrt{2A}$ to measure v. For $B \ge 2$ exchangeable block means M_j from a fixed checkpoint, take $U_j = P_\perp(M_j - \mu_{\mathrm{ctrl}} - \Delta)_c/\sqrt{2A}$ and $K = B$: the gain is $(1-1/B)\mathbb{E}S^2$.

Paper connection. This supports the fit-free disagreement measurement (28.5% on the quadratic scale, reported as 26.4% after the measured conversion) and predicts $0.0072 \times 4/5 \approx 0.0058$ for $200 \to 1000$ cells, against 0.0057 observed.

Remark (block variance). Writing $M_j = m + z_j$, the measured quantity is $\mathbb{E}S^2 = [\mathbb{E}\|P_\perp z_{1,c}\|^2 - \mathbb{E}\langle P_\perp z_{1,c}, P_\perp z_{2,c}\rangle]/(2A)$, by cancellation of the common bias. It equals the projected emission variance when the displayed cross moment is zero; under that condition, pooling B such blocks leaves variance $1/B$ times the single-block variance.

Proposition 6 (shared-offset correction). A well-aligned correction removes shared error, preserves pooling's gain, and cancels from residuals measured about their own centroid.

Statement. Fix centered b, h and mean-zero z, and set $C(h) = \mathbb{E}Q(b+z) - \mathbb{E}Q(b+z-h)$. Then $C(h) = [2\langle P_\perp b, P_\perp h\rangle - \|P_\perp h\|^2]/(2A)$, so $C(h) \ge 0$ exactly when $2\langle P_\perp b, P_\perp h\rangle \ge \|P_\perp h\|^2$. For any $K \ge 1$ vectors u_k sharing correction h, $\overline{u-h} = \bar{u} - h$ and $(u_k - h) - \overline{u-h} = u_k - \bar{u}$.

Intuition. Averaging contracts a cloud around its center; a common correction translates the whole cloud. Translation helps when it moves the center toward the target, while leaving distances from the cloud's own center unchanged.

Proof. Expand both projected squares and use $\mathbb{E}z = 0$ to obtain $C(h)$; $A > 0$ gives its sign criterion. Linearity of averaging gives $\overline{u-h} = \bar{u} - h$; subtraction gives the residual identity. Apply Proposition 4(i) before and after translation: the ambiguity terms agree, so $K^{-1}\sum_k[\|u_k\|^2 - \|u_k - h\|^2] = \|\bar{u}\|^2 - \|\bar{u} - h\|^2$. Applying this to $P_\perp u_k/\sqrt{2A}$ and $P_\perp h/\sqrt{2A}$ gives equal average-member and pool correction gains in Q. With $u_p = m_p$ the predicted means of $N \ge 1$ perturbations, the same residual identity proves own-centroid invariance exactly.

Paper connection. This explains the composition and reference dependence of (19) of this supplement. The equal-gain identity is for one common h; the implemented offset is estimated per run, so it also holds that run's seed-specific error, which pooling already removes. Consistent with this, the correction gains $+0.0145$ on members and a smaller but positive $+0.0094$ on the $K = 5$ pool. Its own-centroid gain, predicted to vanish, is $+0.0027$ ($t = 0.96$), a tie.

Remark (estimating and applying the offset). For $\hat{b} = b + \xi$ with finite second moment, $\mathbb{E}\xi = 0$ and $\mathbb{E}\langle P_\perp z, P_\perp \xi\rangle = 0$, expansion gives expected gain $[\|P_\perp b\|^2 - \mathbb{E}\|P_\perp \xi\|^2]/(2A)$: nonnegative gain is equivalent to projected estimation variance at most the squared projected bias. For a gene with cell mean $m_g \ne 0$, the unclipped rescale $x_g \mapsto x_g(m_g - h_g)/m_g$ has mean $m_g - h_g$ by linearity. For nonnegative cells this applies directly when $m_g > 0$ and $m_g - h_g \ge 0$; the implemented rule also clips and uses run-specific fold estimates. The own-centroid result applies to a correction common

across perturbations (Viñas Torné et al. 2025); the reported gain against the fixed real centroid is $+0.0344$.

3 The Objective-Axis Arms, Arm by Arm

The ten arms, their Δ, 95% CI, n and verdict are Table 2 of the main paper; Figure S2 below plots the same intervals and Table S1 gives the per-seed values behind them. This section annotates the arms the main paper's "Reading the table" passage does not reach.

Arm-by-arm notes. MMD had only ever been run at the median heuristic; calibrating the bandwidth at convergence moves the arm from -0.0036 to -0.0005, and the calibrated setting also has the smallest seed variance of the three bandwidths. The Rao–Blackwellized target replaces the resampled teacher statistics with full-cloud ones, removing 49% of the mean leg's gradient noise and changing nothing measurable. The directional shrinkage pulls each predicted difference toward the cross-perturbation mean by α, scoring, block-matched on two beds, $+0.0005$, $+0.0005$, -0.0009, -0.0044 and $+0.0002$, $+0.0001$, -0.0016, -0.0051 at $\alpha \in \{0.05, 0.10, 0.20, 0.30\}$: positive but an order of magnitude below resolution at $\alpha \leq 0.1$, and negative and monotone beyond it (an earlier block-mismatched curve is withdrawn; see the method specification). The anti-collapse penalty is imported from the large industrial perturbation model cited in the main paper, where magnitude collapse is the documented failure mode; that same model reports adopting a concordance objective after instrumenting per-loss gradient magnitudes, which is the prior positive evidence at scale the main paper credits our concordance arm with. Lin's accuracy factor C_b, which the main paper's Related Work names, penalizes departure of the best-fit line from the identity; the concordance leg therefore acts on the mean-difference vector's direction rather than on a per-gene weighting. The offset calibration needs a leak-free estimate of the shared-across-perturbations offset it corrects, i.e. a calibration fold supervised exactly as the evaluation perturbations are (teacher only). The fold alone withholds 35 of 347 real supervision signals and costs -0.0060 at n=5; the submitted row (-0.0074, n=3) measured only this cost, because Eq. (19) was not yet applied. With Eq. (19) applied the arm gains $+0.0145$, net of that cost (Section 8).

4 Per-Seed Values

5 Protocol

Every arm runs: 800000-step ceiling, validation fraction 0.15, evaluation every 1000 steps, strict validation policy, early-stop patience 200, minimum delta 10^{-4}, early-stop EMA 0.3, minimum steps 120000, batch 256, 300 context cells, 500 evaluation cells, $G = 2000$ raw gene space, and WKD-F leg weight 1.0 and ratio 2.0 except where swept.

Tuning budget. MMD's bandwidth was calibrated at convergence over $\{0.33, 1.0, 3.0\}$ and its weight over the same grid; concordance over seven leg weights $\{0.03, \ldots, 30\}$, with four accuracy exponents $\{0.25, \ldots, 1.0\}$ explored at reduced replication. Table 2 of the main paper reports each arm at its best configuration. Entropic OT is the exception noted in the main paper: its blur is fixed at the median heuristic and only that single configuration was trained to convergence.

Arm	s0	s1	s2	s3	s4
WKD-F baseline	0.8847	0.8902	0.8843	0.8974	0.8847
Concordance $\gamma 0.3$	0.8802	0.8893	0.8874	0.8991	0.8928
MMD bw 3.0	0.8766	0.8919	0.8884	0.8925	0.8893
MMD bw 1.0	0.8635	0.8867	0.8871	0.8948	0.8913
MMD bw 0.33	0.8648	0.8862	0.8870	–	–
Leg-ratio ρ=4	0.8772	0.8896	0.8819	0.8991	0.8896
Rao–Blackwell	0.8689	0.8890	0.8857	0.8952	0.8881
Concordance $C_b^{0.5}$	0.8784	0.8872	0.8813	0.8905	0.8851
Calibration fold only	0.8774	0.8795	0.8800	0.8959	0.8786
Entropic OT	0.8633	0.8762	0.8706	0.8800	0.8794
Sliced W (centred)‡	0.8647	0.8714	0.8715	0.8812	0.7936
Shape residual‡	0.8647	0.8742	0.8596	0.8725	0.8731
Offset calibration, Eq. (19)	0.8977	0.9018	0.9005	0.9092	0.9044
Anti-collapse	0.5869	–	–	–	–

Table S1: Per-seed test pearsonΔ from the released evaluator. Columns are paired: arm seed s and baseline seed s share initialization, batch order and evaluation control block. Every row was re-emitted from its checkpoint at its own seed's control block and re-scored, reproducing the submitted aggregates; seeds 3–4 of the former three-seed arms were trained under the same frozen protocol. Seeds 5–9 of the baseline score 0.8884, 0.8844, 0.8866, 0.8934, 0.8917. ‡Added in revision under the same frozen protocol.

Undertraining. Three arms in this campaign are recorded cases of a short screen over-promising relative to the converged run. The starkest is a sharpness-aware arm that screened at $+0.034$ and converged to $+0.002$. That is why the post-peak tail described in the main paper is non-negotiable.

Why t intervals, and why not equivalence. At n=3 the two-sided 5% critical ratio is 4.30, not 1.96, so a ratio of 3.8 is *not* significant; treating it as significant is a common and consequential error, which is why the main paper reports two-sided t intervals on $n-1$ degrees of freedom rather than z-scores. A claim of statistical *equivalence* would additionally require a pre-specified margin and a two-one-sided-test procedure, and our per-arm power does not support one at any margin the field would find interesting. "Tie" in the main paper therefore means a non-significant difference, which is a weaker statement, and we keep it weak.

Default-inertness. Each auxiliary leg is separately verified *default-inert*: at weight zero the resulting parameters are bit-identical to the unmodified baseline, not merely close, so the λ=0 boundary of the leg-weight sweeps provably recovers the baseline. The accuracy-exponent grid is the one sweep whose boundary does *not* recover WKD-F, since at exponent zero the concordance leg reduces to plain Pearson, and that grid accordingly starts at 0.25.

6 The Capacity Ladder

Spending the same parameters on one wider model does not work. A capacity ladder run to convergence with a swept per-

Arm	Δ	t	improved	Bonferroni p
Concordance $\gamma 0.3$	$+0.0015$	2.5	46/70	0.123
MMD, calibrated bw	-0.0005	-0.6	28/70	1.000
MMD, default bw	-0.0036	-3.7	25/70	0.004
Leg-ratio ρ=4	-0.0008	-0.9	35/70	1.000
Concordance $C_b^{0.5}$	-0.0037	-2.8	22/70	0.051
Rao–Blackwell	-0.0029	-2.4	29/70	0.165
Calibration fold only	-0.0060	-6.4	9/70	< 0.001
Entropic OT	-0.0143	-13.6	1/70	< 0.001
Sliced W (centred)‡	-0.0317	-20.3	0/70	< 0.001
Shape residual‡	-0.0194	-17.8	0/70	< 0.001
Offset calibration, Eq. (19)	$+0.0145$	12.1	67/70	

Table S2: Every arm on the ensemble's footing: seed-averaged differences against the baseline, paired over the 70 perturbations (t on 69 df), the perturbations improved, and the Bonferroni-adjusted two-sided p over the eight pre-specified arms, or over all ten for the two arms added in revision (‡). The pooling gain on this footing is $+0.0179$ (t=13.6, 70/70); the largest Bonferroni upper limit is $+0.003$, over the eight arms and over all ten.

Ensemble configuration	members	pearsonΔ
Single member (seed 0)	1	0.8847
Duplicated control (5$\times$ cells)	1	0.8847
K=5 runs	5	0.9025
K=5 runs + snapshots	10	0.9054
K=10 runs	10	0.9046
K=10 runs + snapshots	20	0.9067

Table S3: Aggregation configurations, all scored by the released evaluator on one shared control block. The matched snapshot comparison of the main paper is *rows 4 and 5*, and only those: both emit 2560 cells per perturbation from ten members, differing solely in whether the members come from ten independent runs or from five runs $\times$ two snapshots (0.9054 vs 0.9046; the paired value is $+0.0007$, and the difference of the rounded aggregates shown here is 0.0008). Rows 3 and 4 are *not* matched, differing in both member and cell count, and are listed for completeness rather than as a comparison.

size learning rate is flat, with 3.2M scoring 0.886, 5.2M 0.879 and 15.9M 0.884, so a 5$\times$ larger monolith buys nothing, while five copies of the student total 16.06M parameters and score 0.9025, beating the same-scale monolith by $+0.0185$ (Figure S3). What reads as a capacity ceiling is therefore a *variance* ceiling: it does not break when the parameters go into one wider model, and it does break when they go into independent members.

Which model the parameter budget buys. Whether one model beats the pool at matched inference parameters depends on *which* model the budget is spent on, and two beds disagree. Spending it on a deeper published architecture wins: an RN32$\times$4 student at 7.43M parameters, a 6.02$\times$ ratio, scores 80.06 ± 0.07 and beats our K=6 pool by $+1.58$. Spending it on the student one already has, widened 5$\times$ to 6.16M at a 5.00$\times$ ratio, loses: that monolith scores 75.85 against our second replication's K=5 pool of the same 6.16M at 78.35, a deficit of 2.50. The two together say more than either alone. A better-designed deeper architecture can beat the pool, while naively widening the existing student does not, so the matched-parameter comparison turns on architecture choice and not on capacity alone. This is why the main paper's limitation is stated in those terms rather than only as a saturated-ladder assumption.

Second-replication ensemble distillation. Our own second replication, on separate hardware under a logit-KD recipe, ran the within-recipe control. A single-teacher student scores 72.75 (n=1); the ensemble-distilled students average 74.80, 74.84 and 74.88 at $K = 5$, 11 and 15 (n=3 each), i.e. $+2.05$ to $+2.13$ over that one control, while trailing a member (76.07) by 1.2–1.3. Our $+1.20 \pm 0.08$ at K=6 agrees in sign on a different recipe, and the replication adds that the gain grows slowly with pool size.

Three caveats keep the matched-parameter comparison of the main paper honest. The parameter match is approximate and tilted slightly *toward* the ensemble (16.06 versus 15.86M). It matches *inference* parameters only, since the ensemble still costs $K\times$ the training compute, which the snapshot variant halves but does not remove. And the ladder was measured at a collaborating site: the protocol is ours field for field (800k-step ceiling, patience 200, minimum 120k steps, validation fraction 0.15, batch 256, the same split, teacher and evaluation), with the learning rate swept per size so that the wider model is not handicapped by a setting tuned for a smaller one, so the monolith is at its own plateau and not merely undertrained; it shares no seed with our arms, so we quote it as a same-scale reference. We have since reproduced the comparison on this machine, under the frozen protocol of every arm and at the ladder's two converged learning rates (10^{-4} and 3×10^{-4}, three seeds each, all early-stopped at 272–332k steps). On the shared seed-0 block, `state_m` (15.86M) scores 0.8689 and 0.8699, below the average `state_xs` member on that block (0.8783), and the K=5 pool beats it by $+0.0337$ and $+0.0326$ (t=13.6 and 12.7, 70/70 perturbations each), and by $+0.0236$ over the best of the six runs (66/70). Our earlier attempt at this size, reported in the submission (0.7593–0.7895, one seed undertrained, Figure S5), is superseded by these runs.

7 The CIFAR-100 Beds

The main paper's Sec. 7 carries the cross-domain argument; this section records what it states compactly.

The fitted constants behind Table S6. Fitting $1 - \mathrm{score}(K) = c_\infty + v/K$ bed by bed gives $0.0926 + 0.0332/K$ at $R^2 = 0.999$ for the cell bed, $22.730 + 3.838/K$ at $R^2 = 0.9925$ for CIFAR in-shim, $21.082 + 2.894/K$ at $R^2 = 0.9974$ for CIFAR authors', and $26.950 + 5.116/K$ at $R^2 = 0.9885$ for the 29$\times$ gap. For CIFAR-100 the score is top-1 in percent, so the deficit is $100 -$ top-1; each image bed is six independently seeded students scored at the *final* epoch and pooled over every $\binom{6}{K}$ subset, the fourth a 29$\times$ gap to a

MobileNetV2 student. The cell bed's five points are disjoint pools of twelve independently seeded students: twelve of one, six of two, four of three, three of four and two of six, so that no member is counted twice at any K. Figure S7 rescales all four by their own constants.

Which bed is which. Table S6 fits the pool-size law to three image beds, and they are different runs. The *in-shim* bed, six students at 73.50 ± 0.34 pooling to 76.79 with a weight soup of 0.97, is the one drawn in Figure S6 and quoted in the main paper's prose. The *authors'-configuration* bed is a later, better-tuned run of the same six-seed design: 76.06 ± 0.15 individually, 78.48 pooled, weight soup 1.05, against the same 79.42 teacher. It is the third row of Table 3 and the pool used as a teacher below.

The $29\times$ bed. The fourth row is a different pair and a much wider capacity gap: ResNet50 (23.71M) distilled to MobileNetV2 (0.81M), a $29\times$ ratio and a change of architecture family, against the $6\times$ same-family pair of the two rows above it. Six independently seeded students were trained under mdistiller's own KD configuration at their MobileNet learning rate. They score 68.06 ± 0.76 individually, just above the published KD figure of 67.35 for this pair, and the $K{=}6$ pool reaches 72.42, a gain of $+4.36$ against $+2.42$ at the $6\times$ gap. This is the bed behind the main paper's statement that the averageable share rises with the capacity gap.

Ensemble distillation, and why it works. Training a fresh student against the six-member prediction pool instead of against the original single teacher, with objective, schedule, seed and code path identical and only the teacher supplying the targets changed, gains $+1.20 \pm 0.08$ top-1: 74.63 ± 0.22 against a control of 73.43 ± 0.20, $n{=}6$ each, paired $t{=}15.7$ with all $6/6$ seeds positive (per-seed differences $+1.31$, $+1.26$, $+1.04$, $+0.96$, $+1.18$ and $+1.48$). The control reproduces the published KD figure for this pair, 73.33, so the difference is attributable to the teacher and not to our harness. The pooled teacher is *less* accurate than the one it replaces (78.48 against 79.42) and $4.3\times$ better calibrated (ECE 0.0197 against 0.0845): what transfers is calibration, not accuracy.

8 Rebuttal-Phase Analyses

All numbers below come from re-emitting the trained checkpoints and scoring them with the released evaluator. Our reimplementation of pearsonΔ from per-perturbation means matches the evaluator to $\leq 2.6 \times 10^{-7}$ per perturbation, and the code added for these analyses is bitwise inert when off.

Matched-count control with distinct cells. At 1000 predicted cells per perturbation (five disjoint 200-cell control-input blocks), one checkpoint gains $+0.0036$ and $+0.0074$ over its own 200-cell emission (members of two disjoint pools). Five checkpoints beat that distinct-cell control by $+0.0217$ and $+0.0216$ ($t{=}19.4$, 18.4; $70/70$).

Proposition 3, measured. Emitting more distinct cells from one checkpoint removes the finite-emission error z without moving the first moment. Across 14 checkpoints at the trained 256-cell block size, the cap $\mathbb{E}\|z_c\|^2/(2\|\Delta_c\|^2)$ is 0.0064 $[0.0058, 0.0071]$. Here $\mathbb{E}\|z_c\|^2$ is estimated without any independence assumption between cells, as the variance of the emitted mean across disjoint, exchangeable input blocks. The extrapolated $n \to \infty$ gain is 0.0068 $[0.0061, 0.0075]$; the intervals are a perturbation bootstrap. At 200-cell blocks the figures are 0.0074 and 0.0074, and the observed $200 \to 1000$ gain is 0.0057 against 0.0058 predicted. The cap is reached: removing z gains the projected term $\mathbb{E}\|P_\perp z_c\|^2/(2\|\Delta_c\|^2)$ (0.0063 at 256-cell blocks, 0.0072 at 200), which Proposition 3's proof guarantees and which the projection lowers only slightly. By Proposition 5, the $200 \to 1000$ prediction is that projected term times $1 - 1/B$ for $B = 5$ blocks, which assumes exchangeable blocks.

Propositions 4–5, measured. From 14 members' disagreement on the shared block, with no fit, the averageable share is 26.4%: 28.5% on the quadratic scale of Proposition 2, times its measured second-order factor 0.93 (below), against 26.4% from the fit of (5) in the main paper. The intervals are $[25.5, 27.3]\%$ (the quadratic-scale perturbation bootstrap $[27.5, 29.4]\%$ times the same factor), a member-jackknife standard error of 1.0 points, and $[25.1, 28.3]\%$ when (5) is refitted over 400 random disjoint re-partitions of the twelve K-curve members. The identity predicts both disjoint $K{=}5$ pools' gains to within 7% (the second-order factor 0.93 of Proposition 2), with per-perturbation correlations of 0.92 and 0.94. Per-perturbation gains rank with member disagreement (Spearman 0.91) and with the members' own deficit (0.88), and inversely with effect size $\|\Delta_c\|$ (-0.85). Grouped by mechanism of action (the drug annotations of Tahoe-100M; the 70 perturbations are 70 drugs), every class gains: all three broad and all 17 fine classes improve on every perturbation, by similar amounts (Kruskal–Wallis $p = 0.45$ and 0.97), so the gain follows effect size and member error rather than drug class.

Offset calibration, Eq. (19) implemented. Each run's 35-perturbation teacher-only fold is reproduced exactly; it equals the fold the training run logged and is disjoint from the 70 evaluation perturbations. b_0 is estimated from the run's own fold predictions and subtracted from the emitted per-gene mean by the non-negative rescale $x \mapsto x(m - b_0)/m$, which clips at most 1.4% of gene–perturbation means at 0 (genes whose predicted mean is already 0 are left there). The result is $+0.0145$ $[+0.0102, +0.0188]$ at $n{=}5$ (per perturbation $t{=}12.1$), while the fold alone costs -0.0060. On the shared block a calibrated $K{=}3$ pool scores 0.9087: $+0.0093$ over a $K{=}3$ pool of baseline members ($t{=}9.2$), and $+0.0139$ over the same three members without the correction. With all five calibrated members the $K{=}5$ pool scores 0.9120, $+0.0094$ $[+0.0077, +0.0111]$ over our uncalibrated $K{=}5$ pool ($t{=}11.0$, $64/70$). That is consistent with Proposition 4: calibration removes the bias shared by every seed, which lies in c_∞, while pooling removes v/K.

Both levers against a perturbed-centroid reference. Control-referenced scores reward variation shared by every perturbation (Viñas Torné et al. 2025). We therefore rescored the same emissions with the control mean replaced by

the centroid of the 70 evaluation perturbations, either common to both sides (the real centroid) or each side's own, under which a shift shared by every predicted perturbation cancels exactly (`rebuttal/analysis.py systema`; every emission is first matched to its cell-eval score, $\leq 3 \times 10^{-7}$). Pooling survives both: the $K{=}5$ pool beats its seed-0 member by $+0.045$ $[+0.037, +0.054]$ (common) and $+0.032$ $[+0.025, +0.040]$ (own; $t{=}8.8$, $64/70$). Eq. (19) gains $+0.034$ $[+0.027, +0.042]$ on the common reference but ties on its own ($+0.0027$ $[-0.0029, +0.0083]$, $n{=}5$ seeds averaged per perturbation), and so does the calibrated pool over the uncalibrated one ($+0.0047$ $[-0.0007, +0.0100]$). So the offset calibration is a correction of the shared shift, which the primary metric scores, and not of perturbation-specific error; the seed variance that pooling removes is perturbation-specific.

Secondary and distribution-sensitive metrics. The evaluator's discrimination scores rank bulk mean effects, so they are first-moment metrics. Against seed 0 the pool's discrimination changes are within noise (ℓ^1 -0.0035 $[-0.011, +0.004]$), and against the pool's own five members they are gains ($+0.0041$ $[+0.0021, +0.0061]$). On distribution shape the conclusion reverses. Entropic OT, the worst arm on pearsonΔ, raises the evaluator's energy-distance correlation by $+0.180$ (seed-paired), while pooling barely moves it: $+0.002$ over the pool's own five members on the same block ($+0.066$ over seed 0, the weakest of them on this metric). It also cuts the unbiased energy distance to the real cells from 1.75 to 0.02 on 20 perturbations, while losing the differential-expression family (precision@100 -0.156, ROC-AUC -0.115).

The image-bed replication, per seed. Our second replication, on separate hardware with a logit-KD recipe at $T{=}1$ (the submission called it an independent site), distils the pool into a single student (Table S4), against a member mean of 76.07 and a single-teacher control of 72.75 ($n{=}1$). No student of either replication lies inside the published 76.77 ± 0.26 (maximum 76.39).

K	distilled student, seeds $0/1/2$	mean	pool top-1
5	74.74/74.70/74.97	74.80	78.35
11	74.77/74.84/74.91	74.84	78.64
15	74.95/74.86/74.82	74.88	78.86

Table S4: The second replication: CIFAR-100 top-1 of the pool-distilled single student per seed, and of the pool it distils. Member mean 76.07; single-teacher KD control 72.75 ($n{=}1$).

Pooling composes with the objective. Pooling helps every loss we tested, as it helps the baseline. Each arm's $K{=}5$ pool of its own five frozen-protocol seeds beats the average of those five members, all emitted on one shared control block and scored by the released evaluator, by $+0.0224$ to $+0.0502$ ($t \geq 18.1$; $70/70$ perturbations for every arm; Table S5). After pooling the loss choice no longer matters: no arm's pool beats the baseline pool (the closest, calibrated MMD, by $+0.0016$ $[-0.0002, +0.0035]$). For the shape objectives pooling keeps their energy-distance lead while it lifts pearsonΔ: the OT pool scores 0.910 against 0.709 for the baseline pool, its members 0.902. A $K{=}3$ pool of OT members at blur 27 ties a $K{=}3$ baseline pool on pearsonΔ ($+0.0008$ $[-0.0013, +0.0029]$) while scoring 0.852 against 0.694 on the energy-distance correlation; it stays below the baseline pool on the DE family (precision@100 0.225 against 0.269). A mixed pool of five baseline and five OT members scores 0.9034.

arm	members	pool	gain	t	energy distance
Baseline (WKD-F)	0.8783	0.9025	$+0.0243$	19.8	$0.707 \to 0.709$
Concordance $\gamma{=}0.3$	0.8756	0.9011	$+0.0255$	19.3	$0.693 \to 0.705$
MMD, calibrated bw	0.8782	0.9042	$+0.0260$	18.9	$0.691 \to 0.697$
MMD, default bw	0.8731	0.9000	$+0.0269$	18.1	$0.755 \to 0.767$
Leg-ratio $\rho{=}4$	0.8769	0.9006	$+0.0237$	20.1	$0.680 \to 0.689$
Concordance $C_b^{0.5}$	0.8756	0.9020	$+0.0263$	19.1	$0.705 \to 0.715$
Rao–Blackwell target	0.8761	0.8985	$+0.0224$	19.2	$0.721 \to 0.738$
Entropic OT	0.8612	0.8950	$+0.0338$	20.9	$0.902 \to 0.910$
Sliced W (centred)	0.8445	0.8947	$+0.0502$	22.4	$0.905 \to 0.912$
Shape residual	0.8551	0.8934	$+0.0383$	20.3	$0.911 \to 0.915$

Table S5: Pooling helps every loss. Each arm's $K{=}5$ pool of its own five seeds against the average of those five members, on one shared control block (seed 0), scored by the released evaluator: pearsonΔ of the members and of the pool, the gain paired over the 70 perturbations (t on 69 df; every pool improves $70/70$), and the energy-distance correlation from the member average to the pool. Each member scored alone matches its score inside the pool's emission to 4×10^{-7}, and the two re-emitted controls reproduce exactly.

The remaining objective arms. Under the frozen protocol, at the pre-registered gradient-parity weights (1329 centred, 268 standardized; re-measured exactly), the shape leg added to the baseline loses: sliced Wasserstein -0.0317 $[-0.0731, +0.0096]$ (all five seeds below the baseline, per-perturbation $t{=}-20.3$; one seed hit the step ceiling, was resumed and early-stopped at its step-783k best, scoring 0.794), shape residual -0.0194 $[-0.0265, -0.0123]$ ($n{=}5$). Both raise the energy-distance correlation ($+0.185$, $+0.189$) and lose the DE family, as entropic OT does. Entropic OT's blur sweep is monotone: -0.0362 at blur 3, -0.0143 at the median-heuristic 9, -0.0030 $[-0.0110, +0.0050]$ at 27, so its best setting ties and none gains.

Pool distillation under the authors' strongest recipe. On the image bed ($n{=}3$, seeds disjoint from the pool's), WKD-F scores 76.10 (reproducing the 76.06 of Table 3 of the main paper), WKD-LF with the single teacher 76.63, and WKD-LF with the pool's mean logits as the logit target 76.70. The pool-distilled student exceeds the plain single student, but as a logit target the pool only ties the single teacher ($+0.07 \pm 0.20$): the $+1.20$ of Table 3 of the main paper is specific to plain KD. Pooling itself composes with that recipe: six WKD-LF students (seeds 100–105, disjoint from Table 3's) score 76.48 ± 0.23 and pool to 78.84, $+2.36$ over their mean and above every member (best 76.82) and

the WKD-F pool's 78.48; each member re-scores to its own training log.

Every arm paired per perturbation. Reviewer 5MXf asked that the objective arms be put on the ensemble's footing. Averaging each arm-minus-baseline difference over seeds and pairing over the 70 perturbations gives (intervals conditional on the seeds drawn, as the ensemble's are):

arm	Δ	95% CI	t_{69}
Concordance γ=0.3	$+0.0015$	$[+0.0003, +0.0027]$	2.5
MMD, calibrated bw	-0.0005	$[-0.0022, +0.0012]$	-0.6
MMD, default bw	-0.0036	$[-0.0055, -0.0016]$	-3.7
Leg-ratio ρ=4	-0.0008	$[-0.0024, +0.0009]$	-0.9
Concordance $C_b^{0.5}$	-0.0037	$[-0.0064, -0.0011]$	-2.8
Rao–Blackwell	-0.0029	$[-0.0053, -0.0005]$	-2.4
Calibration fold only	-0.0060	$[-0.0078, -0.0041]$	-6.4
Entropic OT	-0.0143	$[-0.0165, -0.0122]$	-13.6
Sliced W (revision)	-0.0317	$[-0.0349, -0.0286]$	-20.3
Shape residual (revision)	-0.0194	$[-0.0216, -0.0172]$	-17.8
Offset calibration, Eq. (19)	$+0.0145$	$[+0.0121, +0.0169]$	12.1
Prediction pool, K=5	$+0.0179$	$[+0.0152, +0.0205]$	13.6

With the eight pre-specified interval arms, the Bonferroni (m=8) upper limit of any loss change is $+0.003$ on this footing and $+0.022$ on the seed-paired one.

Per-perturbation intervals for every metric. The K=5 pool against the mean of its own five members, paired over the 70 perturbations on the shared block (the released evaluator's per-perturbation output). Metrics the evaluator computes once over all perturbations (DE Spearman on significant genes, the energy-distance correlation, clustering agreement) have no per-perturbation interval. *The pool emits five times as many cells as a member, which moves these DE-family metrics by itself (Sec. 6.2 of the main paper); Table 1 gives the count-matched comparison.

metric	pool − members	95% CI
pearsonΔ	$+0.0243$	$[+0.0218, +0.0267]$
MSE (lower better)	-0.0001	$[-0.0001, -0.0001]$
MAE (lower better)	-0.0009	$[-0.0010, -0.0009]$
MSEΔ (lower better)	-0.0001	$[-0.0001, -0.0001]$
MAEΔ (lower better)	-0.0009	$[-0.0010, -0.0009]$
discrimination ℓ^1	$+0.0041$	$[+0.0021, +0.0061]$
discrimination ℓ^2	$+0.0027$	$[+0.0012, +0.0042]$
discrimination cosine	$+0.0048$	$[+0.0023, +0.0073]$
overlap@100*	$+0.1071$	$[+0.0746, +0.1396]$
precision@100*	$+0.1400$	$[+0.1087, +0.1713]$
DE direction match*	$+0.0054$	$[+0.0028, +0.0080]$
DE Spearman (LFC, sig.)*	$+0.0318$	$[+0.0252, +0.0384]$
DE gene recall*	$+0.2734$	$[+0.2511, +0.2957]$
PR-AUC*	$+0.1122$	$[+0.1009, +0.1234]$
ROC-AUC*	$+0.0377$	$[+0.0292, +0.0463]$

Cost, measured. On one H100 NVL, emitting 256 cells for all 70 evaluation perturbations takes 151 ms for one student and 754 ms for a K=5 pool (5.0×); on CIFAR-100, one RN8x4 classifies the 10,000 test images in 76 ms and six in 458 ms, against 317 ms for the RN32x4 teacher. One member trains in a median 2.1 h of wall-clock with two to three runs sharing a B200 (about one GPU-hour); a K-member pool costs K times that, and two snapshots per run halve the runs needed for a given K. The student has 3.21M trainable parameters; its backbone also carries a frozen, unused 6.14M token-embedding table, which no forward pass reads.

bed	avg. share	2nd member
cells, disjoint	26.4%	58%
CIFAR, in-shim	14.4%	53%
CIFAR, authors'	12.1%	56%
CIFAR, 29× gap	16.0%	50%

Table S6: The same pool-size law on four beds: two tasks, three architecture families, two metrics and three training recipes are each described to $R^2 \geq 0.988$ by Eq. (5) of the main paper, *a shared error averaging cannot remove plus a seed-variance term falling as* $1/K$, and on all four the second member alone buys about half the gain. The invariant is that form, *not* the averageable share, which rises with the capacity gap: $+4.36$ top-1 at 29× against $+2.42$ at 6×, so pooling pays most where distillation is hardest. Each bed's fitted constants and its R^2, and how the beds are built, are in the supplementary material.

Moved from the main text: the nested curve and a withdrawn decomposition. We first estimated the averageable share from a *nested* curve, taking K=1 as the first seed, K=2 as the first two, and so on, which gave 20%. That construction is biased, because its K=1 point is one particular member (here the best of the twelve, 0.8847 against a twelve-seed mean of 0.8743), so the curve starts too high and looks flatter than it is. We also withdraw a previously reported five-way decomposition of the residual: its shares summed to 101%, so they were estimated separately rather than as one decomposition, and we could not reproduce them from the artifacts.

9 Supplementary Figures

Figure S1 carries the pool-size measurement the main paper argues from directly and refers to by number; Figure S2 plots the intervals of its Table 2, and Figures S3–S7 are corroborating evidence for claims the main paper states in prose. Figure S7 in particular shows the pool-size law of the main paper's Table 3 collapsing onto one parameter-free line. The method schematic is Figure 1 of the main paper.

10 Formula-Exact Specification of the Objective-Axis Arms

This section specifies each objective-axis arm at the level of detail required to reimplement it. Every expression is transcribed from the code that produced the reported numbers; where the main paper's body prints an idealized form, the shipped form is printed here and the difference is stated.

Reading the implementation pointers. A bare line reference such as `:2358` continues the file named immediately before it. Paths are relative to the repository root; `m7_fewshot_distill.py` lives in `experiments/` and is the trainer, which assembles the per-step objective from the leg functions in `set_distill/`.

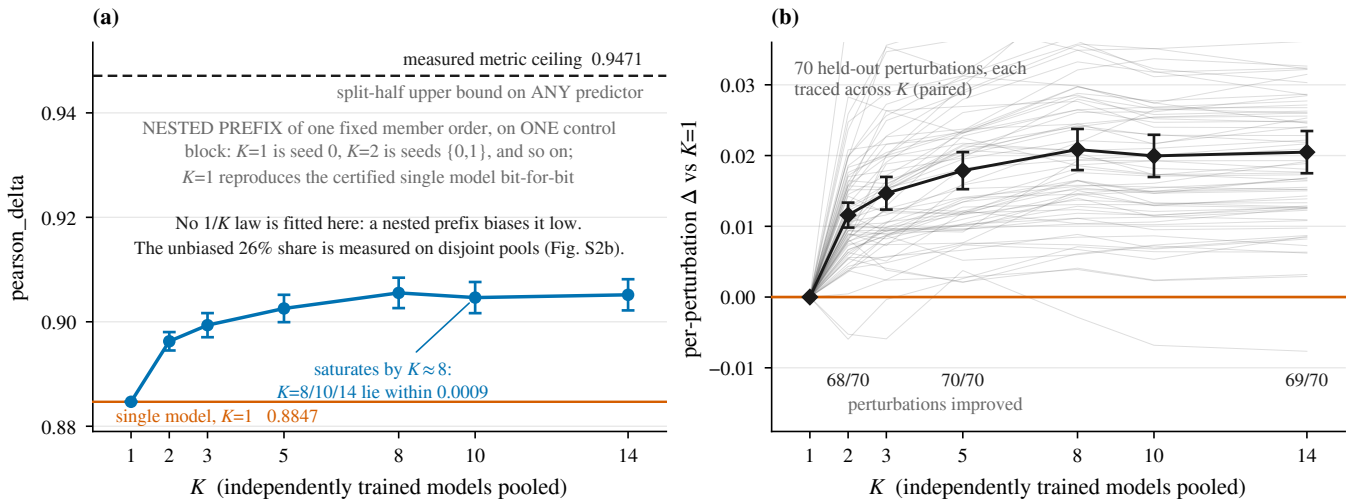

Figure S1: How the gain scales with the pool size K, on a **fixed evaluation control block**. Because the evaluation control block is seed-derived, every point is regenerated on-path through the model's own emit path at seed 0 and scored by the same evaluator, so all K share one control block and one set of 70 held-out perturbations; the stored control-block and gene-selection provenance is byte-identical across all seven points, $K{=}1$ reproduces the certified single-model score bit for bit, and $K{=}5$ reproduces the paired headline exactly. **(a)** The level, with the measured metric ceiling for scale; the curve saturates by $K \approx 8$, with $K = 8, 10, 14$ lying within 0.0009 of one another. **(b)** The same data paired: each of the 70 perturbations traced across K, with the mean paired difference against $K{=}1$ and its 95% t interval. These seven points are the ones Eq. (5) of the main paper is fitted to. This curve supersedes an earlier one that pooled control *blocks* as well as cells; that version was withdrawn in audit and no number from it appears anywhere in this paper.

10.1 Notation and Shared Conventions

A *cloud* is a matrix of cells by genes. At each training step the sampler draws one supervising cloud (d, Y, src, k): a perturbation index d, a target cloud Y, a source tag $\mathrm{src} \in \{\textsc{teacher}, \textsc{real}\}$, and a control-pool key k. It then draws, with replacement, $b = \min(B, |Y|)$ rows of Y to form $\tilde{Y} \in \mathbb{R}^{b\times G}$ and b rows of the control pool to form $C \in \mathbb{R}^{b\times G}$, and runs the student once, $\hat{X} = G_\theta(C, d) \in \mathbb{R}^{b\times G}$. Throughout, $B = 256$ is the batch size, $G = 2000$ is the number of highly variable genes, and $T = 800{,}000$ is the step ceiling. Per-gene statistics of a cloud $X \in \mathbb{R}^{n\times G}$ are

$$\mu_g(X) = \frac{1}{n}\sum_{i=1}^{n} X_{ig}, \qquad s_g^2(X) = \frac{1}{n-1}\sum_{i=1}^{n}\big(X_{ig} - \mu_g(X)\big)^2, \tag{1}$$

i.e. s^2 is the *unbiased* ($\mathrm{ddof} = 1$) sample variance unless a term explicitly says otherwise. Write $\mu(X) \in \mathbb{R}^G$ for the vector of per-gene means. The *mean-delta* of a cloud against the step's control block is $\delta(X) = \mu(X) - \mu(C)$; the student's is $\hat{\delta} = \mu(\hat{X}) - \mu(C)$ and the target's is $\delta^{\mathrm{tgt}} = \mu(\tilde{Y}) - \mu(C)$. Gradients flow through $\hat{X}$ only: every target is detached, and C enters as a constant reference.

Which cloud each leg is computed against, and on what fraction of steps. The `distilled_plus_wkdf` supervision list is the concatenation of one teacher cloud per perturbation with one real cloud per *training* perturbation. The sampler is uniform over perturbations and then uniform over the sources available for that perturbation, with probability $w_{(d,\mathrm{src})} \propto \big(D \cdot |\mathcal{S}(d)| \cdot n_{d,\mathrm{src}}\big)^{-1}$ over $D = 417$ perturbations. Because 347 perturbations carry both a teacher and a real cloud while the 70 transductive evaluation perturbations carry a teacher cloud only,

$$\Pr[\mathrm{src} = \textsc{real}] = \frac{347}{2\cdot 417} = 0.4161, \qquad \Pr[\mathrm{src} = \textsc{teacher}] = \frac{70}{417} + \frac{347}{2\cdot 417} = 0.5839. \tag{2}$$

The WKD-F base and the entropic-OT and MMD legs are computed against *whichever* cloud was sampled, so they fire on every step. The concordance, anti-collapse and metric-alignment legs are gated on $\mathrm{src} = \textsc{teacher}$ and take the **teacher's** mean-delta as their target; they therefore fire on 58.4% of steps and are absent from the remaining 41.6%. This gate is deliberate—it is what keeps those arms' target the teacher's delta rather than the real one—but it means the arms are not what a description of an always-on leg would suggest.

Default-inertness, stated once. Every knob in this section is *default-inert*: at its default value the trained parameters are bit-identical to the unmodified WKD-F baseline's. This was verified by re-running the trainer on the synthetic fixture (30 steps, CPU, seed 0, `state_xs` student) and comparing full `state_dict` tensor equality—not a decimal-truncated training trace—for every *added-leg* knob in Table S7, and additionally for the robust-Huber and sliced-shape leg weights.

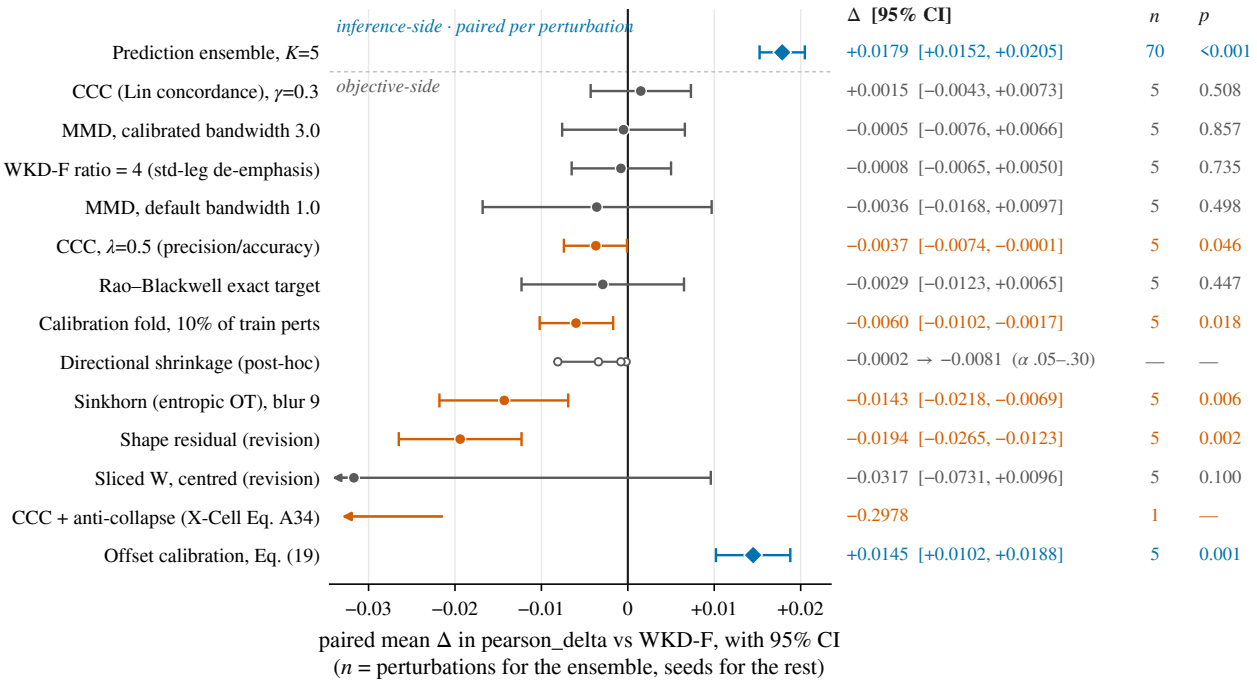

Figure S2: The closed loss axis. Every change on the objective axis as a within-seed paired difference against the WKD-F baseline (0.8882, n=5), with two-sided t intervals on $n-1$ df, with all eight interval arms at n=5. Among the loss changes, six ties, three losses and one collapse; none excludes zero on the positive side. The one gain is the post-hoc offset calibration, Eq. (19). The pairing unit differs by row and the axis label says so: the inference-side row is paired over 70 perturbations, the objective-side rows over seeds. We deliberately draw no "resolvable effect" band: we claim no equivalence, and no interval lies inside ± 0.005. The prediction ensemble (top row) is on the same axis for scale. Exact values are in Table 2 of the main paper and per-seed values in Table S1 above.

(The baseline's own ρ and ε are excepted: they define the baseline rather than add to it, so inertness at their defaults is a tautology.) All are byte-identical when off, and each is non-decorative when on (turning it up changes the student). Only the two beyond-Wasserstein weights are pinned in continuous integration, by `test_wiring_is_byte_identical_when_off` in `set_distill/tests/test_beyondw_arms.py`. The concordance and ensembling families have no test coverage, so their inertness is measured but not protected.

10.2 The WKD-F Baseline

What it fixes. It supplies the arms' common base: the strongest published single-teacher *distributional* KD objective, obtained as the closed-form squared 2-Wasserstein distance between two diagonal Gaussians fitted to the two clouds, so that the student matches the teacher's per-gene location *and* dispersion rather than only its mean.

Objective, as shipped.

$$\mathcal{L}_{\text{WKD-F}}(\hat{X},\tilde{Y}) = \rho \sum_{g=1}^{G} \big(\mu_g(\hat{X}) - \mu_g(\tilde{Y})\big)^2 + \sum_{g=1}^{G} \Big(\sqrt{s_g^2(\hat{X}) + \varepsilon} - \sqrt{s_g^2(\tilde{Y}) + \varepsilon}\Big)^2, \tag{3}$$

entering the total loss as $\gamma_w \mathcal{L}_{\text{WKD-F}}$. Constants: $\rho = 2.0$ is the mean/covariance leg ratio (`-wkdf_ratio`, the source's `cfg.WKD.MEAN_COV_RATIO`); $\varepsilon = 10^{-5}$ is the variance stabiliser (`-wkdf_eps`, the source's `cfg.WKD.EPS`); $\gamma_w = 1.0$ (`-wkdf_gamma`). Both reductions over genes are *sums*, not means. s_g^2 is the unbiased variance of (1). There is no averaging over perturbations: one cloud is sampled per step, so the loss is per-perturbation. In the teacher-only arms there is no cross-entropy term—no real labels are consumed—so (3) is the whole loss and γ_w is a bare rescale to which Adam is invariant.

Hyperparameters swept. $\rho \in \{2, 4\}$ (2 is the source repository's default and its ImageNet configuration, 4 its CIFAR-100 configuration); ε and γ_w were not swept.

Implementation. `set_distill/wkd_baseline.py:105-131`

Table S7: Every objective-axis knob: default, swept grid, and the seeds actually *scored*. Knobs are default-inert at the value in the second column.

Knob	Default	Grid swept	Scored
`wkdf_ratio` ρ	2.0	$\{2, 4\}$	n=5 at ρ=4 (revision)
`wkdf_eps` ε	10^{-5}	—	—
`wkdf_exact_target`	off	{off, on}	n=5 (revision)
`sinkhorn_blur`	0.05	$\{3, 9, 27\}$ nominal	only 9, n=5
`sinkhorn_gamma`	0	$\{0.1, 0.3, 1.0\}$ nominal	only 1.0, n=5
`mmd_bw_scale` β	1.0	$\{0.33, 1.0, 3.0\}$ (+$\{10, 30\}$)	all, n=3–5
`mmd_gamma`	0	$\{0.1, 0.3, 1.0\}$	all; 1.0 at n=5
`ccc_gamma`	0	$\{0.03, 0.1, 0.3, 1, 3, 10, 30\}$	0.3 at n=5; see text
`ccc_cb_power` λ	1.0	$\{0.25, 0.5, 0.75, 1.0\}$	0.5 at n=5 (revision); rest n=1
`ccc_decay_frac` f	0.0	$\{0.3, 0.6\}$	n=3, null (see text)
`dnorm_gamma`	0	$\{0.3, 3\}$	n=1
`leg_autobalance`	off	shares $\{0.1, 0.3, 1.0\}$	0.3 at n=3; rest n=1
`metric_align_gamma`	0	—	never launched
`calib_frac` f	0.0	$\{0.10\}$	n=5 (revision)
`swa_frac` ϕ	0.0	—	no on-box run
shrinkage α (post hoc)	0	{0, .05, .1, .2, .3, .5}	curve, no retraining
`ensemble_from` K	none	pool-size curve	K=1, 5, 10
`ensemble_snapshots`	off	{off, on}	$2K$=10
`ensemble_barycenter`	off	{off, on}	K=5

(`wkd_feature_loss_cloud`) and `:134-145` (`wkdf_loss`); called at `m7_fewshot_distill.py:2130`.

Citation. Lv, Yang, and Li (2024), feature branch, `grid==1`.

Deviation from the main paper's printed form. The body prints $\mathcal{L} = \rho\|\Delta\mu\|^2 + \|\Delta\sigma\|^2$. Two details of the shipped function are absent from it and are stated here instead. First, ε **sits inside the square root**: the code computes $\left(\sqrt{\mathrm{Var} + \varepsilon} - \sqrt{\mathrm{Var}' + \varepsilon}\right)^2$, not $(\sigma - \sigma')^2$. The Bures-term identity the body verifies against a full eigendecomposition to 1.8×10^{-15} holds for the ε-free form; the shipped leg is the ε-regularised version, so the verified identity is not literally the shipped objective. Second, **the standard deviation uses** $\mathrm{ddof} = 1$: `torch.var` is called at its default `unbiased=True`. Switching to $\mathrm{ddof} = 0$ moves the covariance leg by $\approx 0.4\%$ on a test batch. Neither detail changes a reported number—every reported run used this same code—but (3), not the body's equation, is what ran.

Deviation from the source. Only the `grid==1` branch is ported; the source's `adaptive_avg_std_pool2d` has no cell analogue, there being no spatial axis to subdivide. The source's `ConvReg` projector is not ported: it exists to align a student's channel count to the teacher's, and here both clouds already live in the same 2000-gene space. The source's cross-entropy term is dropped for the reason given above. Their (batch, channel, H, W) tensor maps onto a cloud $A \in \mathbb{R}^{n \times d}$ as batch $= 1$, channel $= d$ genes, spatial $= (n, 1)$ cells; under that reshape the ported function is bit-for-bit theirs, and unlike their four-dimensional version it additionally admits $n \neq m$ because the statistics are per gene.

10.3 Entropic Optimal Transport (Sinkhorn Divergence)

What it fixes. WKD-F truncates each cloud to a diagonal Gaussian. This leg distils what that truncation discards, by putting a genuine transport-based discrepancy on the *actual* clouds.

Objective. With ground cost $c(x, y) = \|x - y\|^p/p$ (the library's convention, so $\frac{1}{2}\|x - y\|^2$ at the shipped p=2) and entropic strength $\varepsilon_{\mathrm{OT}} = \mathrm{blur}^p$, let

$$\mathrm{OT}_{\varepsilon_{\mathrm{OT}}}(\alpha, \beta) = \min_{\pi \in \Pi(\alpha, \beta)} \int c \, \mathrm{d}\pi \; + \; \varepsilon_{\mathrm{OT}} \, \mathrm{KL}\big(\pi \,\|\, \alpha \otimes \beta\big). \tag{4}$$

The leg is the *debiased* Sinkhorn divergence

$$\begin{aligned} \mathcal{S}(\hat{X}, \tilde{Y}) = \mathrm{OT}_{\varepsilon_{\mathrm{OT}}}(\hat{X}, \tilde{Y}) \\ - \tfrac{1}{2}\mathrm{OT}_{\varepsilon_{\mathrm{OT}}}(\hat{X}, \hat{X}) \\ - \tfrac{1}{2}\mathrm{OT}_{\varepsilon_{\mathrm{OT}}}(\tilde{Y}, \tilde{Y}), \end{aligned} \tag{5}$$

both clouds carrying uniform weights $1/b$. The total loss is $\gamma_w \mathcal{L}_{\text{WKD-F}} + \gamma_S \, \mathcal{S}$. As $\mathrm{blur} \to 0$, (5) approaches W_2^p/p, i.e. $\frac{1}{2}W_2^2$ here; as $\mathrm{blur} \to \infty$ it approaches an MMD, so this arm and the next are the two ends of one family.

The call, verbatim in its arguments. The library is used directly, not reimplemented; the constructor is `SamplesLoss` with arguments `"sinkhorn"`, `p=2`, `blur=`b, `scaling=0.9`, `backend= "tensorized"` and `debias=True`. Constants: $p = 2$ (squared-Euclidean ground cost, matching WKD-F's W_2); $b = 9$; scaling $= 0.9$; debiasing on.

Hyperparameters swept. $\mathrm{blur} \in \{3, 9, 27\}$ nominally and $\gamma_S \in \{0.1, 0.3, 1.0\}$ nominally, with $\mathrm{blur} =$

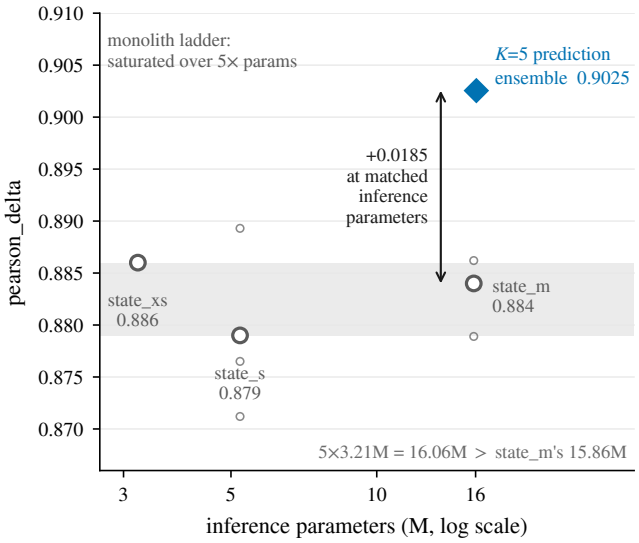

Figure S3: A variance ceiling, not a capacity ceiling. The monolith ladder is shown as a saturation band with its per-seed points, not as a trend line: over a $5\times$ range of parameters the fully-trained ladder moves within ± 0.004. The $K{=}5$ prediction ensemble sits $+0.0185$ above `state_m` at, if anything, a slightly *unfavourable* parameter match: $5 \times 3.21\,\mathrm{M} = 16.06\,\mathrm{M}$ exceeds `state_m`'s $15.86\,\mathrm{M}$, which the panel states in place of claiming an exact match. Ladder scores come from the hardened patience-200 capacity sweep, at the same protocol as the main paper's arms but without a shared seed, so they are a same-scale reference and not a same-seed paired arm.

9 chosen because the feature-normalised clouds' median RBF scale is 8.88 (median over the 70 evaluation perturbations of `energy.median_bandwidth` applied to the teacher and real clouds after the shipped per-gene gain of `energy.fit_gene_gain`, fitted on the pooled teacher cells of the 348 training perturbations; squared-distance median 157.7; re-derived by `paper/verify/closeout_constants.py`). *On this machine only* blur $= 9$ *and* $\gamma_S = 1.0$ *were ever trained to completion and scored* ($n = 5$ seeds). Every launcher hard-codes `-sinkhorn_blur 9`; blur 3 and 27 exist only as an off-machine, $n = 1$, 60k-step screen that the laboratory record itself labels undertrained; and the six $\gamma_S \in \{0.1, 0.3\}$ runs were terminated before completion, leaving directories with neither a summary nor scored output. The submitted entropic-OT result was therefore a single configuration, not the argmax of a grid, which is why the revision swept the blur under the frozen protocol ($n{=}3$): blur 3 loses -0.0362, and blur 27, the best, ties per seed at -0.0030 $[-0.0110, +0.0050]$ (per perturbation $[-0.0053, -0.0006]$). The γ_S grid remains unscored.

Implementation. `set_distill/energy.py:601-631` (`sinkhorn_divergence`); called at `m7_fewshot_distill.py:2286`.

Citation. Feydy et al. (2019) for the debiased divergence and the library; Cuturi (2013) for the entropic regularisation (4).

Deviation from the source. Two arguments depart from the library's defaults. scaling $= 0.9$ against a library default of 0.5: this is the geometric ratio of the ε-scaling annealing descent, whose documentation describes it as the speed/accuracy trade-off with < 0.4 fast and > 0.9 accurate, so 0.9 is the accurate end and makes the returned value a tighter approximation of (5) at more inner iterations. The value is the wrapper's own default and the trainer never overrides it, so every entropic-OT run used 0.9. And backend $=$ `"tensorized"` is forced against a default of `"auto"`, because the `pykeops` kernel backend is not installed in this environment; the tensorized path is a pure-`torch`, fully differentiable computation of the same quantity and differs only in memory traffic.

10.4 Kernel MMD

What it fixes. The kernel counterpart of the previous arm: a full-distribution discrepancy that sees *all* moments, where WKD-F's diagonal Gaussian keeps only the first two.

The kernel and its bandwidths. Let $Z = \begin{bmatrix}\hat{X}\\ \tilde{Y}\end{bmatrix}$ be the pooled cloud of $N = 2b$ points, detached. The median heuristic is

$$\sigma_{\mathrm{med}} = \sqrt{\tfrac{1}{2}\,\mathrm{median}\big\{\|z_i - z_j\|^2 \;:\; i \neq j\big\}}, \qquad (6)$$

the median being taken over the $N^2 - N$ strictly off-diagonal squared distances, with the argument clamped below at 10^{-12}. The bandwidth *set* is then built by multiplying (6) by the calibrated scale β *first* and expanding to five scales afterwards:

$$\sigma_s = c_s\,\beta\,\sigma_{\mathrm{med}}, \qquad c_s \in \{0.25,\ 0.5,\ 1,\ 2,\ 4\}, \quad S = 5, \qquad (7)$$

so β (`-mmd_bw_scale`) shifts all five bandwidths together rather than selecting among them; at $\beta = 0.33$ the set is $\{0.0825, 0.165, 0.33, 0.66, 1.32\} \cdot \sigma_{\mathrm{med}}$. The kernel is the equal-weight mixture

$$k(x, y) = \frac{1}{S}\sum_{s=1}^{S} \exp\!\left(-\frac{\|x - y\|^2}{2\sigma_s^2}\right) \in (0, 1]. \qquad (8)$$

Objective. The estimator is the standard unbiased MMD_u^2, with both self-terms excluding their diagonals and the cross term a full $b \times b$ mean:

$$\begin{aligned}\mathrm{MMD}_u^2(\hat{X}, \tilde{Y}) = {} & \frac{1}{b(b-1)}\sum_{i\neq j} k(\hat{x}_i, \hat{x}_j) \\ & + \frac{1}{b(b-1)}\sum_{i\neq j} k(\tilde{y}_i, \tilde{y}_j) \\ & - \frac{2}{b^2}\sum_{i,j} k(\hat{x}_i, \tilde{y}_j).\end{aligned} \qquad (9)$$

Dropping the diagonals is exactly what removes the finite-sample bias, so that $\mathbb{E}[\mathrm{MMD}_u^2] = 0$ under the null; the cross

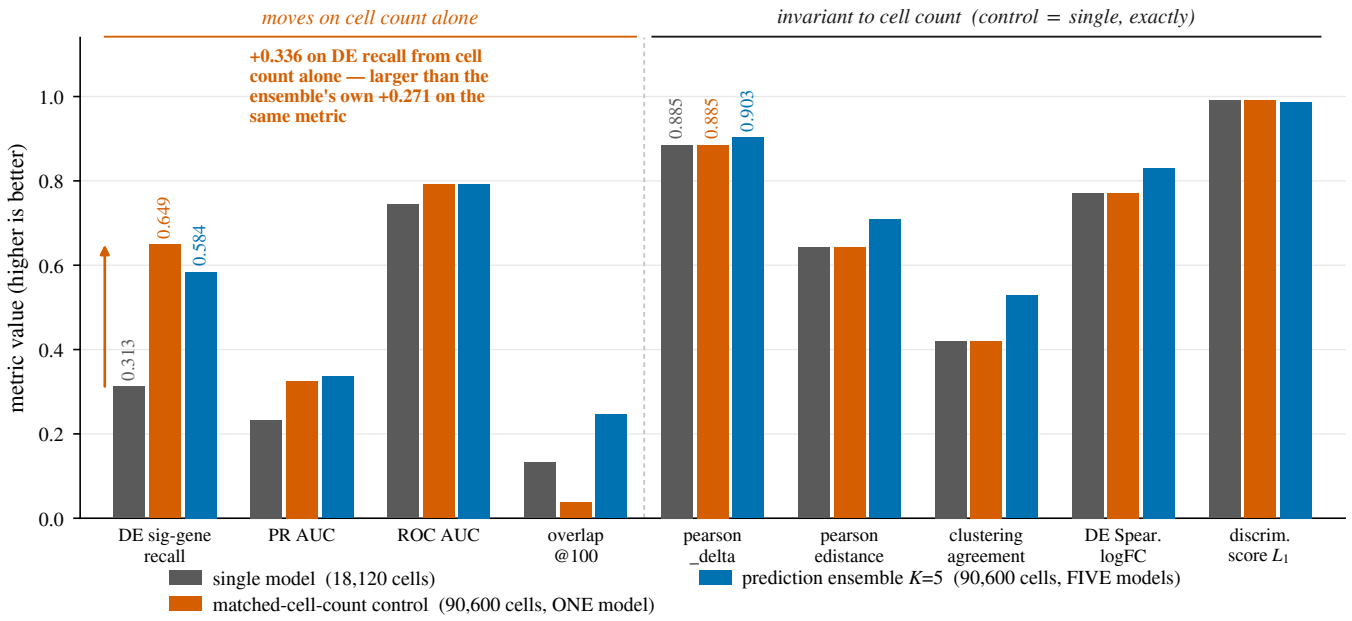

Figure S4: Nine evaluator metrics under three arms that differ in a controlled way: one model (18,120 predicted cells), the *same* model's cells duplicated to the ensemble's count (90,600 cells, one model), and the K=5 prediction ensemble (90,600 cells, five models). Four metrics move on predicted-cell count alone. DE significant-gene recall gains +0.336 from duplication, larger than the ensemble's own +0.271 on that same metric, so a table of these numbers cannot be read as a model comparison unless cell counts match. The five metrics on the right are invariant: the control reproduces the single model *exactly* ($\Delta = 0.0\text{e}{+00}$), which is why pearson$\Delta$ can carry the main claim. Note that the ensemble is very slightly *worse* on the L_1 discrimination score.

term pairs independent samples and carries no diagonal to drop. The total loss is $\gamma_w \mathcal{L}_{\text{WKD-F}} + \gamma_M \, \text{MMD}_u^2$. The general form of (9) for $n \neq m$ replaces the two b's independently; the trainer always calls it at $n = m = b$.

Hyperparameters swept. $\beta \in \{0.33, 1.0, 3.0\}$ at $\gamma_M = 1.0$, plus an off-grid edge probe at $\beta \in \{10, 30\}$; and $\gamma_M \in \{0.1, 0.3, 1.0\}$ at $\beta = 1.0$. All were scored. $\beta = 1.0$ and $\beta = 3.0$ ran at $n = 5$ seeds ($\beta = 3.0$ escalated from 3 to 5 after it tied), the rest at $n = 3$. The default $\beta = 1.0$ is the pure median heuristic; calibrating β was worth +0.0066 relative to it.

Implementation. `set_distill/energy.py:531-547` ((6)), `:550-556` ((7)), `:559-598` ((8)–(9)); called at `m7_fewshot_distill.py:2294-2295`.

Citation. Gretton et al. (2012) for the estimator and the median heuristic.

Deviation from the source. The source states the estimator for a single characteristic kernel; the mixture (8) over five bandwidths is the deep-MMD convention rather than theirs, and a finite mixture of Gaussian kernels remains characteristic, so MMD_u^2 remains a metric. The bandwidth is recomputed *per batch* from (6) and detached, so it is a fixed hyperparameter within a step but a step-varying quantity across training—a departure from treating σ as a constant fitted once. Finally, the median in (6) is taken over strictly off-diagonal pairs, which excludes the N structural zeros that a naive pooled median would include.

10.5 Lin's Concordance (CCC) and Its Generalization

What it fixes. The scored metric is an across-genes Pearson correlation, which grants each perturbation a free affine gauge; a scale-invariant loss therefore has $\partial \mathcal{L} / \partial \|\hat{\delta}\| = 0$ whenever $\hat{\delta} \parallel \delta$ and supplies no magnitude gradient at all. Lin's concordance is the object the metric would be if it did not grant that gauge: the same correlation, times a factor penalising departure of the best-fit line from the identity.

Objective. All statistics are taken *across the G genes* of the two mean-delta vectors, and all moments are *population* ($1/G$) moments, which is Lin's definition and what makes the factorisation exact. Let $\hat{\delta}$ be the student's mean-delta and let the target be the **teacher's** mean-delta $\delta^t = \mu(\tilde{Y}) - \mu(C)$, detached. With $\bar{a} = \frac{1}{G} \sum_g a_g$, write

$$\begin{aligned} v_p &= \overline{(\hat{\delta} - \bar{\hat{\delta}})^2}, & v_t &= \overline{(\delta^t - \bar{\delta}^t)^2}, \\ q &= \overline{(\hat{\delta} - \bar{\hat{\delta}})(\delta^t - \bar{\delta}^t)}, & m &= \left(\bar{\hat{\delta}} - \bar{\delta}^t\right)^2. \end{aligned} \tag{10}$$

For $\lambda = 1$ the code takes a dedicated fast path that is Lin's coefficient literally:

$$\mathcal{L}_{\text{CCC}}^{\lambda=1} = 1 - \frac{2q}{v_p + v_t + m + \epsilon}. \tag{11}$$

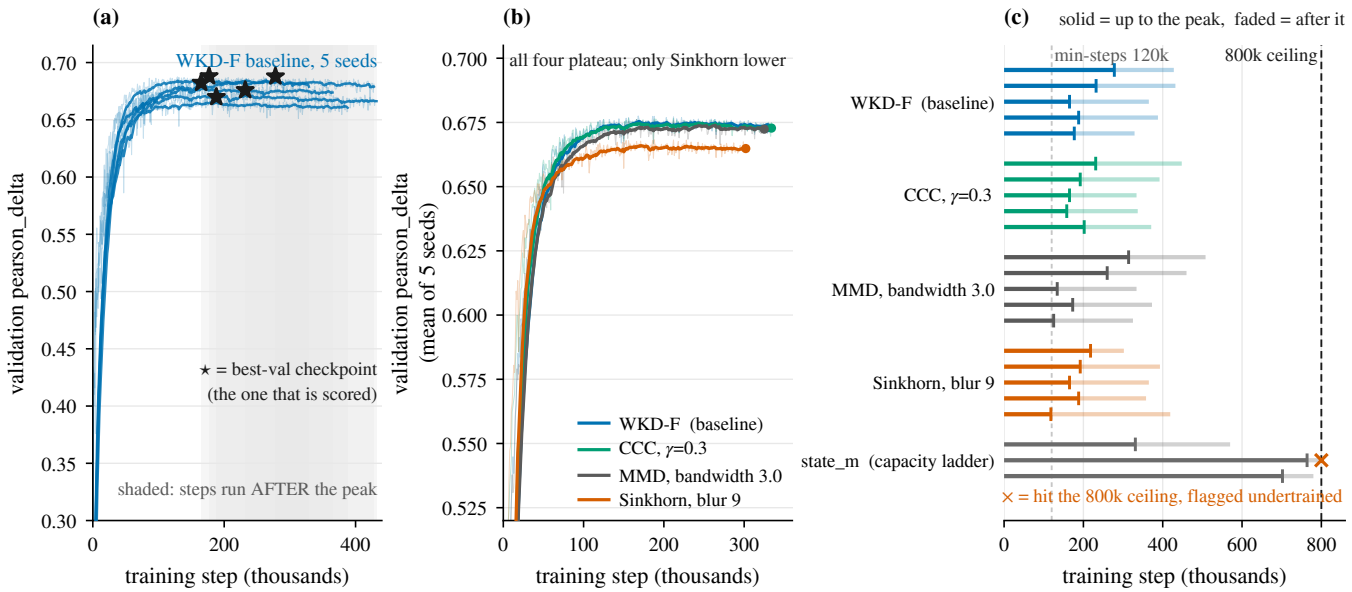

Figure S5: Every arm is trained to convergence; none is stopped early. **(a)** The five WKD-F baseline seeds' complete per-evaluation validation histories (one evaluation every 1000 steps). The star marks the best-validation checkpoint, which is the iterate that is actually scored, and the shaded span is the training that continued *after* that peak. **(b)** All four arms on one axis, averaged over five seeds each: they plateau at the same level and only the entropic-OT arm plateaus lower, which is exactly the ordering Table 2 of the main paper reports. **(c)** A convergence certificate for every run: the bar spans the steps actually run, the tick marks the best-validation step, and the faded segment is the post-peak tail. Early stopping is patience 200 evaluations (`min-delta` 10^{-4} on an EMA-smoothed validation score) with a 120k-step floor and an 800k-step ceiling; the median run continued 200 evaluations, or 200k steps and about half its wall-clock, past its own peak purely to certify the plateau. Exactly one run on this machine reached the ceiling without early-stopping: it is a capacity-ladder `state_m` seed, the protocol guard flagged it `undertrained`, and it is shown here rather than dropped.

For $\lambda \neq 1$ it computes the factorised form and raises only the accuracy factor to the power λ:

$$s_p = \sqrt{v_p + \epsilon}, \qquad s_t = \sqrt{v_t + \epsilon},$$
$$\varrho = \frac{q}{s_p s_t + \epsilon}, \qquad C_b = \frac{2 s_p s_t}{v_p + v_t + m + \epsilon}, \tag{12}$$
$$\mathcal{L}^{\lambda}_{\mathrm{CCC}} = 1 - \varrho \cdot \big(\max(C_b, \epsilon)\big)^{\lambda}.$$

The guard constant is $\epsilon = 10^{-8}$ (`eps`) and λ is `-ccc_cb_power`. The leg enters as $\gamma_c \mathcal{L}^{\lambda}_{\mathrm{CCC}}$, optionally through the auto-balance and cosine-decay operators of the following subsections. It fires only when src = TEACHER, i.e. on 58.4% of steps by (2).

What λ does, and why the grid starts at 0.25. Lin's identity is $\mathrm{CCC} = \varrho \cdot C_b$, with ϱ the *precision* factor (scatter about the best-fit line, literally the scored quantity) and C_b the *accuracy* factor (how far that line sits from the 45° identity). Full CCC assigns C_b weight exactly 1; λ makes that weight tunable, and $\lambda \in (0, 1)$ buys some scale-sensitivity without the collapse a pure accuracy penalty causes. Since $C_b > 0$ always, the power is well defined; ϱ may be negative, which is correct, since anti-correlation must stay penalised. **At $\lambda = 0$ the leg reduces to $1 - \varrho$, i.e. plain Pearson—which is not the baseline.** The baseline is the WKD-F objective, recovered at $\gamma_c = 0$, not at $\lambda = 0$; $\lambda = 0$ merely swaps concordance for correlation and leaves an extra, differently-behaved leg attached to the base. Sweeping down to $\lambda = 0$ would therefore not have produced a baseline anchor, and the grid starts at $\lambda = 0.25$ for that reason.

Hyperparameters swept. $\gamma_c \in \{0.03, 0.1, 0.3, 1, 3, 10, 30\}$ (seven weights; $\gamma_c = 0.3$ at $n = 5$, $\gamma_c = 3$ and $\gamma_c \in \{0.03, 0.1\}$ at $n = 3$, and $\gamma_c \in \{1, 10, 30\}$ at $n = 1$); $\lambda \in \{0.25, 0.5, 0.75, 1.0\}$ (four exponents, all at $\gamma_c = 0.3$; $\lambda = 0.5$ at $n = 5$ in revision, $\lambda \in \{0.25, 0.75\}$ at $n = 1$, and $\lambda = 1.0$ being the default path). The $n = 1$ points cannot separate a configuration from seed noise—the baseline alone spans 0.6699–0.6881 in best-validation across seeds, wider than every gap in the concordance grid—so they orient the curve rather than rank it.

Implementation. `set_distill/energy.py:634-695` (`delta_ccc_loss`: fast path at `:678-680`, general branch at `:692-695`); gated and weighted at `m7_fewshot_distill.py:2333` and `:2358-2361`.

Citation. Lin (1989) for the coefficient; Wang et al. (2026) for its adoption as a perturbation-model distillation leg.

Deviations. Three, all stated plainly. (i) The source's printed accuracy term is ill-typed for vectors; the shipped

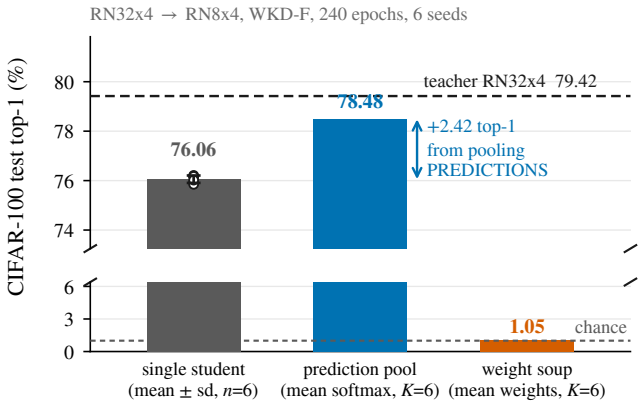

Figure S6: The mechanism is not an artefact of pearsonΔ. Six WKD-F students (ResNet32×4 → ResNet8×4, CIFAR-100, the published 240-epoch schedule, seeds 0–5) differing only in seed, evaluated three ways on the 10,000-image test set at the *final*-epoch checkpoint: individually (73.50 ± 0.34 top-1), as a prediction-space pool of their softmax outputs (76.79), and as a weight-space soup, the elementwise mean of the same six `state_dicts` (0.97, against a chance rate of 1.00). Pooling *predictions* gains +3.29 top-1; averaging *weights* over the identical checkpoints destroys the model outright. The soup is the clean negative control for "the useful diversity lives *between* basins," not a strawman: model souping requires a shared pre-trained initialisation, and these six runs are independently initialised, so its collapse is the predicted outcome. Note the broken axis; the teacher (79.42) is shown for scale. The pool's 76.79 is close to, but is not, WKD's published single-model 76.77 quoted in the main paper; they are different quantities.

code uses the squared difference of the two *gene-means*, $\left(\overline{\hat{\delta}} - \overline{\delta^t}\right)^2$, and *not* the mean of squared per-gene differences $\overline{(\hat{\delta}_g - \delta_g^t)^2}$, which is an MSE and would destroy the $\varrho \cdot C_b$ identity. (ii) The λ generalization (12) is ours; no published form exposes it. The two branches agree to 8.1×10^{-9} in float64 at $\lambda = 1$, so the fast path is equivalent but not bit-identical, and they are not the same *expression*: (11) uses ϵ once whereas (12) uses it five times, so they are not interchangeable at small v_p. (iii) `delta_ccc_loss` carries **no** dead-student tolerance guard—that guard exists only in the Pearson leg specified later—and this matters differently in the two branches. At $\lambda = 1$ it is harmless: (11)'s denominator retains the teacher's variance v_t, and the gradient norm was measured flat at 1.4×10^{-1} as the student's delta scale shrinks over six orders of magnitude. At $\lambda \neq 1$ it is not: (12) reintroduces a $1/s_p$ factor through ϱ, and the gradient norm was measured at 9.8 at delta scale 10^{-4} and 14 at exactly zero. The $\lambda \in \{0.25, 0.5, 0.75\}$ arms therefore ran unguarded, with gradient clipping disabled (`-max_grad_norm 0.0`) as the production protocol defaults. The blow-up is five orders of magnitude milder than the unguarded Pearson leg's (5.4×10^6 at zero), but it is real, and it is an omission of ours rather than a property inherited from the source.

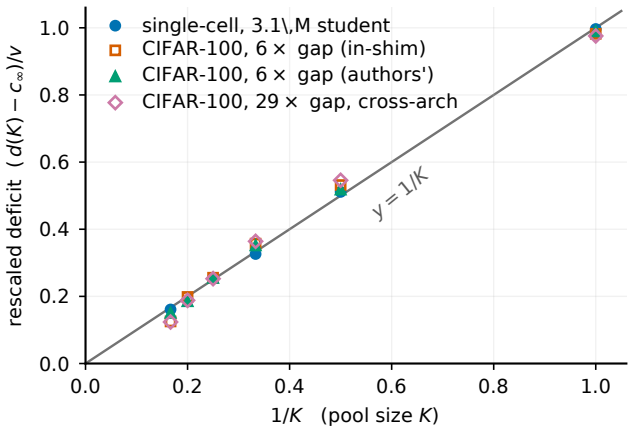

Figure S7: Four beds, one law. Each bed's deficit is rescaled by its own fitted c_∞ and v as $y(K) = (d(K) - c_\infty)/v$, so Eq. (5) of the main paper predicts $y = 1/K$ with no freedom left. Two tasks, three architecture families and three training recipes fall on the same line. The shared object is the *form*: the averageable share itself runs from 12.1% to 26.4% and grows with the capacity gap, which is why pooling helps most where distillation is hardest.

10.6 The Delta-Norm-Ratio (Anti-Collapse) Leg

What it fixes. Conservative collapse: a student that gets the perturbation *direction* right but systematically undershoots its *magnitude*. An affine-invariant metric silently permits this, so nothing in a correlation-only objective pushes the delta norm back up.

Objective. With uncentred Euclidean norms and the same detached teacher target,

$$\mathcal{L}_{\text{norm}} = \left(1 - \frac{\left\|\hat{\delta}\right\|_2}{\left\|\delta^t\right\|_2 + \epsilon}\right)^2, \qquad \epsilon = 10^{-8}, \tag{13}$$

entering as $\gamma_n \mathcal{L}_{\text{norm}}$, optionally through the auto-balance operator. It is non-negative and vanishes exactly when the two delta norms agree. Like the concordance leg it fires only on teacher-sourced steps. It is complementary to, not redundant with, (11): C_b penalises a location/scale mismatch of the *centred* gene profile, whereas (13) constrains the *overall* delta norm.

Hyperparameters swept. $\gamma_n \in \{0.3, 3\}$, paired with $\gamma_c \in \{0.3, 3\}$; the reported row is $\gamma_c = \gamma_n = 0.3$ at $n = 1$.

Implementation. `set_distill/energy.py:698-717` (`delta_norm_ratio_loss`); `m7_fewshot_distill.py:2362-2363`.

Citation. Wang et al. (2026).

Deviation. The formula matches the source term for term. The *arm* does not: the reported anti-collapse row was run with `-leg_autobalance` enabled, i.e. with the share-normalised leg of the next subsection rather than the absolute weight every other row in the objective-axis table uses. It is

therefore not weight-comparable with its neighbours, and its magnitude should not be read as though it were.

10.7 The Rao–Blackwellized Target

What it fixes. The resampling noise that the trainer's own sampler injects into the WKD-F target statistics, at no cost to the objective.

What is being replaced by what. The step's target cloud $\tilde{Y}$ is a *with-replacement resample* of $b \approx 256$ rows from the full teacher cloud Y of $|Y| \approx 300$ cells, and (3) depends on $\tilde{Y}$ only through the per-gene statistic pair $T(\tilde{Y}) = \big(\mu(\tilde{Y}),\ \sqrt{s^2(\tilde{Y}) + \varepsilon}\big)$. The flag `-wkdf_exact_target` replaces the resampled statistic by the full-cloud one,

$$\mathcal{L}_{\text{WKD-F}}\big(\hat{X}, \tilde{Y}\big) \ \longmapsto\ \mathcal{L}_{\text{WKD-F}}\big(\hat{X}, Y\big), \tag{14}$$

i.e. $T(\tilde{Y}) \mapsto T(Y)$, computed on all $|Y|$ cells.

Why this is Rao–Blackwellization and not merely a better estimate. The randomness being removed is *the loss's own auxiliary randomization*, not randomness in the data. Conditioning on Y—which is sufficient for the resampling distribution, the resample being i.i.d. from the empirical measure of Y—leaves the population objective identical and can only reduce the variance of its gradient estimate; that is exactly the Rao–Blackwell operator, applied to the loss's estimator of its own target rather than to a data estimator. The mean leg's conditioning is *exact*, since $\mathbb{E}[\mu(\tilde{Y}) \mid Y] = \mu(Y)$ for with-replacement resampling. The standard-deviation leg's is the plug-in conditional statistic rather than the literal conditional expectation, since $\sqrt{\cdot}$ is concave and $\mathbb{E}[s^2(\tilde{Y}) \mid Y]$ is the $1/|Y|$-normalised variance, so the two agree only to $O(1/b)$. This is distinct from target *denoising*, which attacks the cloud's estimate of the population; (14) attacks the loss's estimate of the cloud, and therefore changes no target.

Magnitude of what is removed. At convergence $\mathbb{E}\|\mu(\tilde{Y}) - \mu(Y)\|^2 = \sum_g v_{t,g}/256 = 0.5917$ against a systematic mean-leg term $\|\mu(Y) - \mu(\hat{X})\|^2 = 0.6100$, so the resampling share of the mean leg's *value* is $0.5917/(0.5917 + 0.6100) = 49.2\%$. Being student-independent it biases no gradient, but it roughly doubles the mean-leg gradient variance. One scope caveat belongs with the number: the leg fires on *training* perturbations, whereas this ratio was measured on the 70 *evaluation* perturbations' clouds.

Hyperparameters swept. Boolean, default off; the on-arm ran at $n = 3$ seeds, $n = 5$ in revision.

Implementation. `m7_fewshot_distill.py:2070` (the ternary that selects the target) and `:2130` (its only consumer); flag declared at `:3198`.

Deviation. The substitution reaches *only* the plain WKD-F branch. The robust-Huber branch (`m7_fewshot_distill.py:2122`) and the logit-standardised branch (`:2128`) both consume the resampled target, so the flag is silently inert when either is active; and the entropic-OT and MMD legs deliberately keep the resampled target because they require equal-sized clouds.

10.8 Leg Auto-Balance, Metric Alignment, Post-Hoc Shrinkage, and Offset Calibration

Leg auto-balance

What it fixes. WKD-F's legs are gene *sums* over $G = 2000$ whose magnitude falls by roughly two orders of magnitude between initialization and convergence, so no fixed leg weight holds a constant share of the objective: a weight that gives parity at step 1 is $O(10^3)$ too large near convergence.

Operator. Let L_{base} be the value of the WKD-F base at that step, captured by detaching the running loss *before* any auxiliary leg is added. With $\mathrm{sg}[\cdot]$ the stop-gradient, each balanced leg is

$$\widetilde{\ell} = \ell \cdot \frac{\mathrm{sg}\big[L_{\text{base}}\big]}{\mathrm{sg}\big[|\ell|\big] + 10^{-12}}, \tag{15}$$

so that $\gamma\widetilde{\ell}$ contributes a magnitude of exactly $\gamma\, L_{\text{base}}$: the weight becomes a *target share* of the base rather than an absolute coefficient, with $\gamma = 1$ making the leg equal the base and $\gamma = 0.1$ making it a tenth of it. The scale is detached, so it reweights the gradient without contributing one.

Hyperparameters swept. Boolean, default off, at target shares $\gamma \in \{0.1,\ 0.3,\ 1.0\}$; share 0.3 at $n = 3$, the other two at $n = 1$. Applied to the concordance and anti-collapse legs only; no other leg in the trainer is routed through (15). Balanced and absolute variants were run head to head at $\gamma_c = 0.3$. The balanced variant scored -0.0026 against the plain absolute weight at n=3; no standard error or interval was ever computed for it, so it should be read as a direction rather than an estimate.

Implementation. `m7_fewshot_distill.py:2136` (capture of L_{base}), `:2344-2347` (the `_bal` closure), `:2360` and `:2363` (its two call sites); flag at `:3165`.

Citation. Equation (15) is the two-term form of a published self-normalising multi-task balance rule, $\alpha_a = \mathrm{sg}\big[(L_a + L_b)/L_a\big]$, introduced by a concurrent perturbation-model adaptation method. Its bibliography entry is not yet in this paper's reference list.

Metric-alignment leg

What it fixes. Every other leg matches the cell-cloud *distribution* with genes treated independently. This one matches the cross-gene *pattern* of the mean-delta, which is the object the evaluation metric actually scores, and is affine-invariant by construction, so it never spends student capacity on the per-perturbation scale and shift the metric discards.

Objective. With $a^c = a - \bar{a}$ denoting across-gene centring,

$$\mathcal{L}_{\text{align}} = 1 - \frac{\big\langle \hat{\delta}^c,\ \delta^{t,c} \big\rangle}{\big\|\hat{\delta}^c\big\|\ \big\|\delta^{t,c}\big\| + \epsilon}, \qquad \epsilon = 10^{-8}, \tag{16}$$

entering as $\gamma_a \mathcal{L}_{\text{align}}$ and gated on $\mathrm{src} = \textsc{teacher}$, so it too fires on 58.4% of steps and takes the **teacher's** mean-delta as target. It returns a value in $[0, 2]$.

The dead-student guard, which is part of the objective. If either centred delta is negligible, $\|\hat{\delta}^c\| < \mathrm{tol}$ or $\|\delta^{t,c}\| < \mathrm{tol}$ with $\mathrm{tol} = 10^{-9}$, the function returns the constant 1.0 *with no gradient*, matching the evaluation metric's own dead-student convention. This is not cosmetic: $\partial \mathcal{L}_{\mathrm{align}}/\partial \hat{X}$ carries a $1/\|\hat{\delta}^c\|$ factor, and the gradient norm was measured to rise from 2.3×10^{-2} at delta scale 1.0 to 5.4×10^{6} at delta scale exactly 0—maximal precisely where an unguarded implementation would look safest—while gradient clipping is off by default in the production protocol.

Hyperparameters swept. None: this arm was never launched on this machine. The leg is implemented, wired to an arm name and default-inert, but no launcher sets `-metric_align_gamma` and no run directory for it exists. The only figures that exist are from an off-machine screen—-0.040 for the leg alone (collapse) and $+0.0075$ when added to the WKD-F base (dilution), the latter roughly $3.5\times$ below this protocol's resolution at $n = 5$—and the screen script itself is absent from the repository, so even the γ_a that produced them is unrecoverable. The arm is specified here for completeness and must not be reported as measured.

Implementation. `set_distill/energy.py:431-475` (`delta_corr_loss`, guard at `:471`); `m7_fewshot_distill.py:2318-2320`; flag at `:3232`.

Citation. Huang et al. (2022), whose correlation-based relation loss is the mechanism.

Deviation from the source. The source correlates teacher and student *logit* matrices and uses both an inter-class and an intra-class relation; here only the inter-coordinate relation is ported, and it is specialised to the per-perturbation mean-delta across genes rather than to logits across classes. The dead-student guard is ours, added after the audit; it is not in the source.

Post-hoc directional shrinkage

What it fixes. All the preceding arms attack the direction the loss converges to. This one attacks the *variance* of the emitted delta instead, post hoc, with zero retraining, by pulling each perturbation's predicted direction toward the cross-perturbation mean direction—the reason "perturbation mean" baselines score well in these benchmarks.

Operator. Over the P evaluation perturbations, with u_p the unit-normalised predicted delta,

$$u_p = \frac{\hat{\delta}_p}{\|\hat{\delta}_p\| + 10^{-12}}, \qquad \bar{u} = \frac{1}{P}\sum_{p=1}^{P} u_p, \tag{17}$$

$$\hat{\delta}_p \;\longmapsto\; (1-\alpha)\, u_p + \alpha\, \bar{u}.$$

Scalar shrinkage $\hat{\delta} \mapsto c\hat{\delta}$ would be exactly a no-op, because the metric is invariant to $a\hat{\delta} + b\mathbf{1}$ for $a > 0$; the James–Stein idea therefore has content here only because (17) moves *direction*. Correspondingly $\alpha = 0$ reproduces the unshrunk score exactly, since unit-norming alone is a no-op. The operator reads only student predictions—no ground truth anywhere—so it is leak-free by construction.

Hyperparameters swept. $\alpha \in \{0, 0.05, 0.1, 0.2, 0.3, 0.5\}$ in the probe. The curve originally reported here (-0.0002, -0.0008, -0.0034, -0.0081 at $\alpha \in \{0.05, 0.10, 0.20, 0.30\}$) came from an emission whose `adata/` has since been deleted, and it carried a defect: the probe scores every seed's predicted delta against the *first* run's real delta (`posthoc_variance_probe.py:138`) while each run emits its own control block, so four of five contrasts were block-mismatched—the tooling hazard the main paper warns about. Re-derived block-matched on the two surviving beds (five members on one shared control block; four independently seeded runs each on its own), the curve reads $+0.0005$, $+0.0005$, -0.0009, -0.0044 and $+0.0002$, $+0.0001$, -0.0016, -0.0051. Both agree: at $\alpha \le 0.1$ the effect is positive but an order of magnitude below this protocol's resolution, and beyond it the curve is negative and monotone. There is no interior optimum worth having. Re-derived by `paper/verify/closeout_constants.py`.

Implementation. `experiments/posthoc_variance_probe.py:148-165`, with the correlation at `:55-58` and the emitted-mean loader at `:61-81`. It operates on the `adata_pred.h5ad` / `adata_real.h5ad` pairs finished runs already wrote, on CPU.

Deviation. The probe scans α *on the evaluation set* and prints the whole curve rather than an argmax; the script says so itself. A clean claim would require refitting α on training or validation perturbations and re-scoring once. The reported numbers are therefore a curve, not a selected result, and it is the size of the best point—below resolution—rather than strict monotonicity that makes the negative read safe.

Leak-free offset calibration

What it fixes. A shared-across-perturbations offset in the student's output, measured at 36% of its systematic bias ($\|b_0\|^2 = 0.154$ of ≈ 0.43). Removing it scored $+0.0191$ under leave-one-out *within* the evaluation pool, which is not a claim; estimating it without leakage requires a calibration fold whose supervision matches the evaluation perturbations' exactly.

The fold, which is what actually ships. Let P_{tr} be the training perturbations ($|P_{\mathrm{tr}}| = 347$) and $f =$ `-calib_frac`. The trainer draws

$$k = \max\big(1,\ \mathrm{round}(f \cdot |P_{\mathrm{tr}}|)\big), \qquad F \subset P_{\mathrm{tr}}, \quad |F| = k, \tag{18}$$

with a dedicated stream `default_rng(seed + 90210)` that does not perturb the sampler, and then removes F's *real* clouds from the supervision list while keeping F's *teacher* clouds. At $f = 0.10$ this withholds 35 of 347 real supervision signals. Perturbations in F are then supervised exactly as the evaluation perturbations are—teacher cloud only, no real cells—which is the precondition for the offset

$$b_0 = \frac{1}{|F|}\sum_{p \in F}\Big(\hat{\delta}_p - \delta_p^{\mathrm{real}}\Big) \tag{19}$$

estimated on F to be unbiased for the evaluation perturbations' offset. The fold is drawn from training perturbations only, so no held-out perturbation's real cells are ever read; the fold's real cells are simply unused.

Hyperparameters swept. $f \in \{0.10\}$; in revision at $n = 5$ seeds, against the baseline at the same seeds.

Implementation. The fold is drawn at `m7_fewshot_distill.py:3794-3803` and recorded (`calib_perts`, in `summary.json` and the run's W&B configuration). The correction is `-offset_calib` (`:4884-4940`): it reads only the *recorded* fold, refuses a fold that is missing or overlaps the evaluation perturbations, estimates b_0 by (19) from the run's own fold predictions, and removes it from the emitted per-gene mean by the non-negative rescale $x \mapsto x(m - b_0)/m$ (target mean clipped at 0; a gene whose predicted mean is already 0 stays there, and the rescale also scales that gene's spread).

Deviation in the submission, now corrected. The submitted version had no code that estimated or subtracted b_0, so its row (-0.0074, $n = 3$) measured only the *cost of the precondition*, i.e. what withholding 35 of 347 real supervision signals costs. The revision reports that row as "calibration fold only" (-0.0060 at $n = 5$) and the correction separately ($+0.0145$ $[+0.0102, +0.0188]$; Section 8). Two supporting facts of the submission remain prose-only, with no artifact on disk: the 36% share and the $+0.0191$ leave-one-out figure. A related negative that *is* recorded: correcting the student with the *teacher's* offset scored 0.8606, worse than uncorrected, the two offsets correlating at only 0.368.

10.9 The Cosine Leg-Decay Schedule

What it fixes. A leg's gradient share swings by two orders of magnitude inside the first 1000 steps, so a weight fixed at step 1 governs nothing afterwards. On the combined-leg diagnosis traces (`results/combined_diag/fixes_screen/base_s{0,1,2}.json`) the shape leg's gradient norm as a share of WKD-F's runs $0.56 \to 88.6$, $1.54 \to 105$ and $0.76 \to 78.8$ over the three seeds—swings of $159\times$, $68\times$ and $103\times$. The parent method does not gradient-balance either; what it ships is a schedule, and this is that schedule transcribed.

Schedule. With nominal weight γ, current step t, ceiling T and start fraction f,

$$\begin{aligned} \gamma(t) &= \gamma && \text{for } t \le fT, \\ \gamma(t) &= \frac{\gamma}{2}\left(1 + \cos\frac{\pi(t - fT)}{T - fT}\right) && \text{for } t > fT, \end{aligned} \tag{20}$$

with the guard that $f \le 0$ or $T \le 0$ short-circuits to the first case, so the schedule is off and the caller's arithmetic is bit-identical; $f \ge 1$ likewise never fires. The strict inequality and the $\frac{1}{2}(1+\cos)$ shape are the parent method's, unchanged; the only translation is epoch $\to$ step, so $f = 150/240 = 0.625$ reproduces its published CIFAR schedule exactly.

f is a start point, not a duration. `decay_start_frac` names the fraction of the run that elapses *before* decay begins. Decay then runs from fT to T and reaches zero exactly at T. Larger f means the leg is decayed *later* and therefore carries *more* total weight, not less: $f = 0.6$ holds the leg at full strength for $480{,}000$ steps whereas $f = 0.3$ holds it for $240{,}000$. Reading f as a decay duration inverts the comparison.

EXP-17 was a null experiment, not a tie. The schedule was swept at $f \in \{0.3, 0.6\}$ with $\gamma_c = 0.3$, $T = 800{,}000$, $n = 3$ seeds, giving decay onsets at step $240{,}000$ and $480{,}000$. The validation-selected checkpoints landed at steps $231{,}000$, $192{,}000$ and $165{,}000$ for seeds 0, 1 and 2—*all three before either onset*. The two configurations are therefore the same computation up to the selected iterate, and this is visible in the artifacts three ways: `best_step` is identical across the two f values at every seed; `best_val` is bit-identical (0.6859659611413677, 0.6774079055316968, 0.6815248605402673); and the two arms emit the same per-perturbation scores for all 70 perturbations at every seed (0.8802, 0.8893, 0.8874). The checkpoint files differ only in the live iterate and the optimizer state, which is why their checksums differ while the scores do not—and that discrepancy is what exposed the null. The experiment compared a schedule against itself and carries no information about leg decay; it must not be reported as evidence that the two schedules perform equally. Because early stopping fires at 20–30% of the step ceiling here whereas the source method runs a fixed 240 epochs, a fraction chosen by analogy with the source is systematically too late: an informative re-run needs $f \in \{0.05, 0.15\}$, i.e. onsets at 40k and 120k. The general lesson is that on a run with early stopping, any schedule expressed as a fraction of the step *ceiling* is a fraction of a horizon the run never reaches.

Implementation. `set_distill/wkd_baseline.py:65-102` (`cosine_decay_weight`); applied to the shape leg at `m7_fewshot_distill.py:2175-2176` and to the concordance leg at `:2358-2359`; flags `-wkdf_shape_decay_frac` (declared at `:3297`) and `-ccc_decay_frac` (at `:3204`), both defaulting to 0.0.

Citation. Lv, Yang, and Li (2024).

Deviation from the source. The source decays the leg it *adds* to a cross-entropy base. Here the WKD-F moment loss is the base and the auxiliary is the addition, so (20) is applied to the *added* leg; decaying the base instead would grow the auxiliary's share, the direction measured to be harmful.

10.10 The Aggregation Methods

Prediction-space pooling

What it fixes. The seed variance of the trained function, which no per-step loss can see.

Operator. Train K students independently under the same objective, protocol, split and teacher, differing only in seed. At inference load each member's *validation-selected* state—`best["state"]` from its `resume_<arm>.pt`, i.e. the

reported model, not the final iterate—emit its predicted cells on the **same** control block c_{te}, and pool the sets:

$$\hat{\nu}_p^{\text{pool}} = \frac{1}{K}\sum_{k=1}^{K} \hat{\nu}_p^{(k)}, \tag{21}$$

realised as the row-concatenation of the K emitted $n \times G$ matrices into one $Kn \times G$ matrix. No retraining, no distillation of the ensemble, no change to the loss. Because members are equal-sized and the metric reduces each cloud to its per-gene mean before correlating, $\mu(\hat{\nu}_p^{\text{pool}}) = \frac{1}{K}\sum_k \mu(\hat{\nu}_p^{(k)})$ exactly. Sharing c_{te} across members is deliberate: varying the control block per member would also cut control noise, a second effect not conflated here.

Hyperparameters swept. K, over the pool-size curve; the headline comparison is $K = 5$ against a $K = 5$ duplicated-checkpoint control at matched emitted-cell count (1280 cells per perturbation).

Implementation. `m7_fewshot_distill.py:4781-4835`; flag `-ensemble_from` at `:3191`. Heterogeneous pools are permitted—members trained under different arms save under different checkpoint filenames, and the loader falls back to the unique `resume_*.pt` in a member directory—but it refuses rather than guesses when a directory holds more than one.

Citation. Lakshminarayanan, Pritzel, and Blundell (2017).

The emitted W_2 barycenter

What it fixes. Pooling yields a *mixture*, whose per-gene variance carries the between-member disagreement on top of the within-member spread. The barycenter is the correct central object.

Operator. For diagonal Gaussians $\mathcal{N}(\mu_k, \mathrm{diag}(s_k^2))$ with uniform weights, the Agueh–Carlier barycenter is $\mathcal{N}(\bar{\mu}, \mathrm{diag}(\bar{s}^2))$ with

$$\bar{\mu} = \frac{1}{K}\sum_k \mu_k, \qquad \bar{s} = \frac{1}{K}\sum_k s_k, \tag{22}$$

the second identity being the diagonal solution of the fixed point $\Sigma = \sum_k w_k \big(\Sigma^{1/2}\Sigma_k\Sigma^{1/2}\big)^{1/2}$. It is realised *deterministically*, by an affine per-gene rescale of the pooled cloud rather than by parametric resampling—so no new randomness and no distributional assumption enters at emit time. With μ^{pool} and s^{pool} the pooled cloud's per-gene mean and $\mathrm{ddof} = 1$ standard deviation, both computed in float64, and $\bar{s}_g = \frac{1}{K}\sum_k s_g\big(\hat{\nu}_p^{(k)}\big)$ the mean of the *member* standard deviations,

$$\tau_g = \begin{cases} \bar{s}_g / s_g^{\text{pool}}, & s_g^{\text{pool}} > 10^{-12}, \\ 1, & \text{otherwise}, \end{cases}$$
$$x'_{ig} = \max\Big(\mu_g^{\text{pool}} + \big(x_{ig} - \mu_g^{\text{pool}}\big)\tau_g,\ 0\Big), \tag{23}$$

the result being cast to float32. Since $\mu^{\text{pool}} = \bar{\mu}$ already, (23) changes only the second moment in exact arithmetic.

The non-negativity clamp, and why it is not masking an error. The student is ReLU-terminated, so its cells are non-negative and sparse. Jensen gives $\bar{s} = \mathrm{mean}(s_k) \le \sqrt{\mathrm{mean}(s_k^2)}$, but the second step to s^{pool} does *not* hold: both deviations use $\mathrm{ddof}{=}1$, the mixture over $Kn-1 = 1279$ degrees of freedom and the members over $K(n-1) = 1275$, so $\tau_g \le \sqrt{1279/1275} = 1.0015674$ rather than 1. That bound is attained — measured $\max_g \tau_g = 1.001553$, with 11.2% of gene-by-perturbation pairs exceeding 1 and pre-clamp values reaching -3.0×10^{-3}. So the clamp genuinely fires (7.35% of values on the first attempt, which the range check correctly refused to score), and it, not floating-point rounding, is the whole of the residual gap: decomposing the shipped emission gives 1.5×10^{-14} in float64 without the clamp, 5.2×10^{-9} in float32 without it, and 6.27×10^{-4} with it. The clamp is a small, signed, deliberate departure from the exact barycenter, not a masked error.

Hyperparameters swept. Boolean, default off; run at $K = 5$ against the pooled $K = 5$ member set.

Implementation. `m7_fewshot_distill.py:4836-4869`; flag `-ensemble_barycenter` at `:3172`.

Citation. Agueh and Carlier (2011).

Deviation from the main paper's claim. The body states that pooling and the barycenter score "identically—exactly, not approximately" and that the two first moments agree to $0.000\mathrm{e}{+00}$. The purely arithmetic identity $\mu^{\text{pool}} = \bar{\mu}$ does recompute to exactly 0.0, but the *emitted artifact* does not inherit it: the float32 rescale and the clamp in (23) preserve neither the mean nor the score. Recomputed from the two runs' emitted files, the maximum per-gene absolute first-moment difference is 5.96×10^{-4} and the metric gap is -3.216×10^{-5}, paired $t = -2.54$, $p = 0.0135$, improving on 21 of 70 perturbations—small, but statistically resolvable. The correct statement is "indistinguishable at the resolution of this benchmark".

Snapshot pooling

What it fixes. It doubles K at zero training cost.

Operator. Every `resume_<arm>.pt` already stores two states: `best`, the validation-selected iterate, and `model`, the live iterate at checkpoint-write time. They are different points on the *same* trajectory, typically 100–200k steps apart. With `-ensemble_snapshots`, each member contributes emissions from both states, so (21) runs over $2K$ clouds instead of K.

Hyperparameters swept. Boolean, default off. Against best-validation-only pooling at the same number of *directories* it adds $+0.0028$ at $K = 5$ (10 members) and $+0.0021$ at $K = 10$ (20 members). Against a pool of the same *member count* built from independent seeds—the comparison that isolates whether same-trajectory diversity substitutes for between-seed diversity—it is $+0.000722$ (se 0.000594, $t = 1.2157$, $\mathrm{df} = 69$, $p = 0.228$, improving on 34 of 70 perturbations), i.e. indistinguishable from the independent-seed pool at matched cost of inference.

Implementation. `m7_fewshot_distill.py:4826-4828`; flag at `:3179`.

Citation. Huang et al. (2017).

Deviation from the source. The source collects its snapshots from a *cyclic* learning-rate schedule designed to visit distinct basins. Ours uses the two states an ordinary checkpoint file already carries, with no cyclic schedule and no modification to the optimizer—the same idea, obtained for free, and with correspondingly less member diversity.

Weight-space averaging

What it fixes. It is the weight-space alternative to (21), included as the contrast that makes the prediction-space claim specific.

Operator. A CPU clone of the student's `state_dict` is buffered at every validation step (every 1000 steps). At the end of training, with S the buffer and $\phi =$ `-swa_frac`,

$$\bar{\theta} = \frac{1}{|S_\tau|} \sum_{\theta \in S_\tau} \theta, \qquad S_\tau = \text{the last } \lceil \phi\,|S| \rceil \text{ snapshots}, \tag{24}$$

floating-point tensors being accumulated in float32 and cast back, while non-float buffers (integer and boolean) are copied unchanged from the validation-selected reference state, since they cannot be meaningfully averaged. The averaged state is *returned only if it beats* the best-validation single point on the same validation criterion, so weight averaging can never lose to best-val selection on validation.

Hyperparameters swept. ϕ, default 0.0 (off).

Implementation. `m7_fewshot_distill.py:658-677` (`swa_average_states`), `:2554-2559` (the buffer tail and the acceptance test); flag at `:3047`.

Citation. Izmailov et al. (2018); Wortsman et al. (2022) for the across-run relative.

Provenance caveat. No run with $\phi > 0$ exists in the run tree on this machine. The $+0.009$ quoted for weight averaging comes from an 80k-step short screen at $n = 3$ that its own record marks as not converged (and whose per-seed column in fact lists only two values), and is quoted alongside converged 800k numbers; it should carry that caveat wherever it appears, or be replaced by the qualitative claim that weight averaging was screened and did not help.

The weight-space control that does have an artifact. On the cross-domain replication—CIFAR-100, 6 independently initialised students, final-epoch checkpoints—pooling predictions gives 76.79 top-1 against a per-seed mean of 73.50 ± 0.34, whereas averaging the six *weight* vectors gives 0.97, where chance is 1.00 for 100 classes. The weight-space arm does not merely underperform; it produces something that is not a trained network. This is the expected outcome rather than a strawman: souping requires a shared pre-trained initialization, a condition independently initialised members deliberately violate, and the contrast is what "the diversity lives *between* basins" predicts. It also carries the qualitative claim on its own, without depending on the unsupported $+0.009$.

11 Optimal-Transport Ensembling: The Closest Prior Art

Barycentric aggregation of model outputs is not new, and we do not claim it. Cuturi and Doucet (2014) made Wasserstein barycenters computationally practical, and Benamou et al. (2015) gave the iterative Bregman projection scheme that is still the standard solver; Dognin et al. (2019) ensemble the *outputs* of several models under a Wasserstein barycenter, which is the construction closest to ours, and Singh and Jaggi (2020) fuse model *weights* by optimal transport. Relative to that line our contribution is not the aggregation operator. It is, first, the identity of Proposition 1 of the main paper—on a metric that reads only the first moment of the predicted set, the W_2 barycenter and simple prediction pooling are not merely similar but indistinguishable, so on this class of metric the geometry buys exactly nothing, and we state the second-moment conditions under which it would stop being free. It is, second, the measurement of the channel the aggregation actually exploits: the averageable share of the converged model's residual, 26%, measured on disjoint member pools rather than assumed.

We therefore make a deliberately narrow claim. Barycentric ensembling is prior art; what is new here is knowing *when it cannot pay*, and how large the effect it does buy is on a benchmark where the aggregation is provably equivalent to the trivial alternative.

12 An Independent Capacity Ladder at Per-Size-Optimal Learning Rate

The capacity ladder of the preceding section is a three-point ladder over a $5\times$ parameter range at one swept learning rate. A second, independent ladder exists, run at a collaborating site on a different selector and budget, and it answers the same objection over a wider range and under stronger conditions: four sizes spanning $11\times$, each trained at *its own* optimal learning rate, all under the teacher's own published regularization recipe (AdamW with weight decay 0.01, gradient clipping 0.8, 5000 warmup steps then cosine). It is flat (Table S8).

Per-size tuning is not a formality here. The optimum is size-dependent and *decreasing*: the three smaller students peak at $1\mathrm{e}{-4}$, while the 36M student needs $3\mathrm{e}{-5}$ and scores only 0.749 at $1\mathrm{e}{-4}$—so a single fixed learning rate would have manufactured a bigger-is-worse result from tuning alone. At each size's own optimum the curve is flat within seed noise across the full $11\times$ range, while beating plain energy-only KD by $+0.150$ to $+0.216$ at every size and clearing the linear predict-the-mean bar (0.684) at every size.

A causal control for teacher quality. The natural follow-up—"a flat ladder just means your teacher is the ceiling, and a better teacher would restore bigger-is-better"—was tested directly rather than argued. Mean-correcting the teacher, which lowers its energy-to-truth from 0.130 to 0.0835 (-36%), lifts

student	params	LR	distilled	energy-KD	n
state_xs	3.13M	1e−4	0.782 ± 0.061	0.610	5
state_s	5.22M	1e−4	0.797 ± 0.010	0.581	3
state_m	15.9M	1e−4	0.788 ± 0.030	0.638	5
state_l	36M	3e−5	0.808 ± 0.013	0.619	3

Table S8: A second, independent capacity ladder, each size at its own optimal learning rate under the teacher's regularized recipe (weight decay 0.01, gradient clipping 0.8, 5000-step warmup then cosine). Flat within seed noise across an $11\times$ parameter range, so the flatness reported in the main paper is not an artifact of a learning rate tuned for the small student. The last column but one is a plain energy-only KD arm at the same size; the linear predict-the-mean bar is 0.684 and is cleared at every size. **Protocol note, load-bearing:** this ladder is selected on the teacher's own energy objective at a 200–250k budget, not on pearsonΔ at our 800k one. Its absolute level (≈ 0.79) is therefore *not* comparable with any other number in this paper and the two ladders must never be placed on one axis. Only the flatness transfers.

both measured sizes substantially and still leaves the 36M student at 0.828 against the 3.13M student's 0.829, a difference of -0.001 at n=3 (Table S9). Raising the teacher raises everyone and reorders nobody: student capacity is not the lever in this regime, and teacher quality gates absolute performance for every measured size alike.

student	orig. teacher	corrected teacher	n
state_xs	0.782	0.829 ± 0.013	3
state_l	0.800	0.828 ± 0.023	3
Δ (l−xs)		-0.001	

Table S9: The oracle-teacher control, at the ladder's endpoints state_xs (3.13M) and state_l (36M). Only those two *endpoints* were run under the corrected teacher (6 runs, n=3 each); state_s and state_m were not, so we claim the endpoints, not "every size". The lifts are $+0.047$ and $+0.028$ as recorded; the 36M lift is quoted against the interim original-teacher value 0.800 that was current when the control was run, and against the final n=3 value 0.808 of Table S8 it is $+0.020$. **The corrected teacher is leaky by construction**—it peeks at real per-perturbation means on evaluation perturbations—so this is a mechanism diagnostic and an upper bound, never a deployable method, and it is reported only as a control on the confound.

13 The Metric Cannot See the Second Moment: A Constructive Demonstration

Property (M1) of the main paper—that the primary metric is a functional of the predicted set's first moment only—is load-bearing: it is the premise of the structural caps, of the matched-cell-count control and of the pooling/barycenter equivalence. The main paper supports it with a *destructive* check, that duplicating a block of predicted cells moves the score by exactly 0.0e+00. A constructive check is available and is considerably sharper (Table S10).

Construct two synthetic predictions of the same perturbation, both with the *correct* per-gene mean: one a degenerate point mass with zero within-cloud variance, the other a correctly dispersed cloud with matched diagonal noise. Both score pearsonΔ exactly 1.0 and DE overlap@k exactly 1.0. A two-sample classifier, however, separates the point mass from the real cells at AUC 0.99988—essentially perfectly—while separating the dispersed cloud at 0.527, which is chance. The metric awards a perfect score to a prediction that is trivially detectable as fake, and the differential-expression overlap metric awards it a perfect score too. This is why an objective acting on the predicted set's second moment has so little to reach here: not because second moments are unimportant, but because this evaluator does not read them. Note also that the point mass is the *worst* of the four arms on every dispersion-sensitive statistic in the table—energy excess 4.132 against the copy-the-control baseline's 1.880—so the blindness is specific to the metrics the benchmark scores on, not general.

prediction	pearsonΔ	C2ST	energy	spread
right mean, no spread	**1.000**	**0.99988**	4.132	**0.000**
right mean, real spread	**1.000**	0.527	−0.207	0.993
the teacher's output	0.417	0.814	0.577	1.169
control cells copied	0.000	0.813	1.880	1.000

Table S10: Two predictors with the correct first moment and opposite second moments receive *identical*, perfect scores. Row 2 is a cloud with matched diagonal noise. C2ST is a two-sample classifier's AUC for separating the prediction from the real cells (0.5 is chance); the last two columns are the energy excess over the measurement floor and the ratio of predicted to real per-gene spread, both dispersion-sensitive, reported to show that the point mass is detectable by every route except the one the benchmark scores. The degenerate arm's mean is correct to $\|\Delta\mu\|_2 = 5.5 \times 10^{-7}$, so its perfect score is not a rounding artifact. **Scope caveat:** measured on Replogle RPE1 over 150 held-out perturbations, not on the Tahoe-100M cell line used everywhere else in this paper. That is the right scope, and we state it plainly: the claim is a property of the metric *functional*, which is identical in both settings, not a property of either dataset. As a provenance check on the harness that produced this table, its pearsonΔ correlates at 0.991 with the teacher authors' own released per-perturbation values over the same 150 perturbations.

14 Six Further Objective-Axis Arms

The objective-axis result of the main paper rests on ten changes. Six more exist, run at a collaborating site under the same paired, trained-to-convergence design at n=5, and they cover mechanism classes the ten do not: teacher-target denoising at both supervision levels, intermediate-feature alignment, a robust mean leg, a bias-aware Bayesian fusion of the teacher and real targets, and a graph-geometric penalty on the gene manifold. Three tie and three do not beat the base-

line; none is positive (Table S11). The count of independent objective-side changes that fail to beat WKD-F is therefore sixteen, not ten.

Two of these six had passed a cheap pre-training gate—the bias-aware fusion improved a target-quality proxy by $+0.0128$, and the Laplacian penalty improved a manifold-smoothness statistic at $p = 0.001$—and both then regressed once a student was actually trained on them. That is the sharpest evidence in this campaign that a target-quality or structure-quality proxy does not predict trained-student generalization, and it is why we report no arm on the strength of a proxy.

Objective change	Δ vs WKD-F	verdict
Corr-teacher denoise, t+real	-0.0026 ± 0.0053	tie
Corr-teacher denoise, teacher	$+0.0021 \pm 0.0066$	tie
Feature align (FitNets hint)	-0.0016 ± 0.0173	tie
Huber robust mean leg	-0.0093 ± 0.0042	did not beat
Bias-aware fusion `ba_dn`	-0.0337 ± 0.0029	did not beat
Gene-manifold Laplacian	-0.0217 ± 0.0210	did not beat

Table S11: Six further objective-, target- and geometry-side changes, each a within-seed paired difference against the WKD-F baseline, trained to convergence, all at n=5. Uncertainties are paired standard errors over seeds. Rows 1 and 2 are the same teacher-target denoiser at the two supervision levels, teacher-plus-real and teacher-only. **Cross-site protocol note:** these ran under the collaborating site's 500k-step ceiling and patience 45, not our 800k ceiling and patience 200; the teacher, student (`state_xs`), split, metric and same-seed pairing are identical to ours. They are therefore reported as a separate cross-site block and their rows are deliberately *not* merged into the main objective-axis table. **On the two large negative rows:** the campaign that produced them declines to read a negative as a verdict on the method, because these arms were implemented from the published formulas rather than taken from the authors' released code; we report them as "did not beat the baseline", and the likeliest reading of a large negative is that the added term destabilized training rather than that the method is unsound. The ties are the load-bearing rows, and ties are robust to implementation differences by construction.

15 The Optimizer and Generalization Axis

The inference-side result of the main paper is not a restatement of a cheaper optimizer-side one. A third axis was closed independently: ten optimizer- and regularizer-side changes were screened on the same base at 80k steps, together with the best stacked pair and three further cross-field arms, and the single best was escalated to convergence (Table S12). Nothing on the axis beats the baseline once trained out.

The escalation is worth stating in its own right, because it is the campaign's clearest instance of a short screen overpromising. Fisher-preconditioned SAM screened at $+0.0341$ with all three seeds positive, roughly six screen-level standard errors, which made it the first consistent positive signal on any axis. Trained to convergence at n=4 it settled at $+0.0021$ with a paired standard error of 0.0038: a tie. Its validation traces show why—the sharpness-aware arm converges *faster*, is ahead at 80k, and the plain baseline catches it by convergence. It converged faster, not better. The Huber mean leg of Table S11 is the same story on the objective axis, screening at $+0.0357$ as a CPU battery's standout and converging to -0.0093. Two further arms, the bias-aware fusion and the gene-manifold Laplacian, reversed from cheap pre-training *proxy* gates rather than from a training screen. Naming all four is strictly better than asserting a count, and we do not conflate the two kinds of reversal.

Change	Δ vs WKD-F	mode
Trained to convergence, paired		
Fisher-SAM	$+0.0021 \pm 0.0038$	**tie**
80*k screens,* n=3 *— directional*		
Fisher-SAM + SWA (stack)	$+0.0322$	screen
Fisher-SAM	$+0.0341$	screen
SWA (fraction 0.5)	$+0.0091$	screen
Surrogate-gap SAM (GSAM)	$+0.0086$	screen
IRM invariance penalty	$+0.0085$	screen
Adaptive SAM (ASAM)	$+0.0077$	screen
Vicinal mixup	$+0.0010$	screen
SAM (Euclidean)	-0.0140	screen†
R-Drop	-0.0168	screen
Logit standardization	-0.0973	screen
Relational KD (RKD)	-0.1235	screen
VICReg	-0.1305	screen
Contrastive KD (CRD)	-0.1894	screen
Flow-matching student	≈ -0.27	screen

Table S12: The optimizer and generalization axis. Both plain SAM variants use ρ=0.05; † marks an unstable screen, one of whose three seeds collapsed. Only the first row is trained to convergence, at n=4 paired; *every other row is an* 80*k screen*, and the lesson of this campaign is precisely that screens do not predict converged outcomes—the top screen on this table lost its entire edge when trained out. The screens are reported for coverage, not as results. The mixup and IRM rows are each the best configuration of a small grid (6 and 3 points respectively), neither of which showed a consistent per-seed gain. An EMA(0.999) arm was launched and never returned a number; it is omitted rather than scored. The Fisher-SAM + SWA stack was escalated to a full n=5 run that crashed on a filesystem fault and produced no number—it is **pending**, not refuted. **On the large negatives:** these arms were implemented from the published formulas and mutation-tested against independent oracles, not copied from the authors' released code, and the campaign that produced them states explicitly that values this far below the baseline most likely reflect the added term destabilizing training or being incompatible with a first-moment metric, and are *not* a claim that the method fails. We report them as "did not beat the baseline". The robust, load-bearing claim is the negative one—no fair change on this axis beats WKD-F—and the ties, which are what that claim needs, are by construction insensitive to implementation detail.

16 Null-Teacher Knowledge Controls

The project standard under which these experiments were run states that no number is real until its null-teacher control and its linear baseline exist. Both exist, and we report them here rather than assume the transferred signal is knowledge.

Scrambling the teacher's perturbation-to-response map does not merely remove the gain, it inverts it: the distilled student collapses to 0.19–0.31, far *below* the 0.507 of a student trained on the same real cells with no teacher at all, and below the linear predict-the-mean bar of 0.684 (Table S13). Distilling from a scrambled teacher is worse than not distilling. The same control run at the operating point of this paper reproduces it: the converged WKD-F arms at 0.802 (teacher-only) and 0.886 (teacher-plus-real) collapse to 0.391 and 0.525 under a deranged teacher at $n{=}2$, and a cluster-side run of the same control at $n{=}3$ reports the teacher-plus-real arm falling from 0.888 to 0.53 under derangement and to 0.61 under pooling. The two derangement figures are close enough that they are plausibly the same control at two stages of completion, so we count them as one replication rather than two.

Arm	pearsonΔ	vs. bar
Teacher (108.7M)	0.849	$+0.165$
Linear baseline (mean)	0.684 ± 0.012	— the bar
Distilled (3.13M), best-val	0.670 ± 0.018	$\approx$ ties
From scratch, real cells only	0.507 ± 0.088	-0.177
Distilled, **null** (pool)	0.31	-0.374
Distilled, **null** (derange)	0.19	-0.494
Mean-prediction floor	-0.01	

Table S13: The knowledge control, $n{=}3$ on `state_xs`. A scrambled teacher does not return the student to the no-teacher level; it drives it far below, which is what certifies that a real perturbation-to-response map is being transferred rather than a teacher-shaped regularizer. **Operating-point caveat:** this table is at a different operating point from the rest of the paper—a few-shot split under early-stopped best-validation selection with an energy-only distillation objective—so its absolute levels are not our 0.886, and the 0.670 row rises to 0.692 when the same student is trained to convergence. Quote the *collapse*, not the levels. The replications at our own operating point are quoted in the text.

A null teacher is not a zero-knowledge control. The obvious version of this control is invalid, and we would rather say so than claim more than we ran. A knowledge-free teacher is not the zero point of knowledge, because its effect depends on the softening temperature. At $T{=}1$ a random teacher's distillation gradient is *anti*-aligned with the label gradient on roughly 93% of steps—median cosine between -0.45 and -0.66, against $+0.83$ to $+0.90$ for the real teacher—so the null teacher carries anti-knowledge and actively harms the student. Raising the temperature drives both teachers toward the same near-uniform distribution, at which point the null teacher stops hurting and can appear to help. A null-teacher control run at the wrong temperature cannot fail to tie, which indicts a class of published negative results and, symmetrically, would have indicted ours had it come out the other way. The valid zero point is a budget-matched *no-teacher* arm, which is the row labelled "from scratch" above. **Caveat:** the temperature measurement is from the annotation side of the prior study, not from this benchmark, and its own source flags the four-decimal values as unstable and claims only the mechanism. We claim only the mechanism too.

References

Agueh, M.; and Carlier, G. 2011. Barycenters in the Wasserstein Space. *SIAM Journal on Mathematical Analysis*, 43(2): 904–924.

Benamou, J.-D.; Carlier, G.; Cuturi, M.; Nenna, L.; and Peyré, G. 2015. Iterative Bregman Projections for Regularized Transportation Problems. *SIAM Journal on Scientific Computing*, 37(2): A1111–A1138.

Cuturi, M. 2013. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In *Advances in Neural Information Processing Systems*, volume 26, 2292–2300. Curran Associates, Inc.

Cuturi, M.; and Doucet, A. 2014. Fast Computation of Wasserstein Barycenters. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, 685–693. PMLR.

Dognin, P. L.; Melnyk, I.; Mroueh, Y.; Ross, J.; Nogueira dos Santos, C.; and Sercu, T. 2019. Wasserstein Barycenter Model Ensembling. In *International Conference on Learning Representations (ICLR)*. OpenReview.net.

Feydy, J.; Séjourné, T.; Vialard, F.-X.; Amari, S.-i.; Trouvé, A.; and Peyré, G. 2019. Interpolating between Optimal Transport and MMD using Sinkhorn Divergences. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, 2681–2690. PMLR.

Gretton, A.; Borgwardt, K. M.; Rasch, M. J.; Schölkopf, B.; and Smola, A. J. 2012. A Kernel Two-Sample Test. *Journal of Machine Learning Research*, 13: 723–773.

Huang, G.; Li, Y.; Pleiss, G.; Liu, Z.; Hopcroft, J. E.; and Weinberger, K. Q. 2017. Snapshot Ensembles: Train 1, Get M for Free. In *International Conference on Learning Representations (ICLR)*. OpenReview.net.

Huang, T.; You, S.; Wang, F.; Qian, C.; and Xu, C. 2022. Knowledge Distillation from A Stronger Teacher. In *Advances in Neural Information Processing Systems*, volume 35, 33716–33727. Curran Associates, Inc.

Izmailov, P.; Podoprikhin, D.; Garipov, T.; Vetrov, D. P.; and Wilson, A. G. 2018. Averaging Weights Leads to Wider Optima and Better Generalization. In *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence (UAI)*, 876–885. AUAI Press.

Krogh, A.; and Vedelsby, J. 1994. Neural Network Ensembles, Cross Validation, and Active Learning. In Tesauro, G.; Touretzky, D.; and Leen, T., eds., *Advances in Neural Information Processing Systems*, volume 7, 231–238. MIT Press.

Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2017. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In *Advances in Neural Information Processing Systems*, volume 30, 6402–6413. Curran Associates, Inc.

Lichtendahl, K. C.; Grushka-Cockayne, Y.; and Winkler, R. L. 2013. Is It Better to Average Probabilities or Quantiles? *Management Science*, 59(7): 1594–1611.

Lin, L. I.-K. 1989. A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics*, 45(1): 255–268.

Lv, J.; Yang, H.; and Li, P. 2024. Wasserstein Distance Rivals Kullback-Leibler Divergence for Knowledge Distillation. In *Advances in Neural Information Processing Systems*, volume 37, 65445–65475. Curran Associates, Inc.

Singh, S. P.; and Jaggi, M. 2020. Model Fusion via Optimal Transport. In *Advances in Neural Information Processing Systems*, volume 33, 22045–22055. Curran Associates, Inc.

Ueda, N.; and Nakano, R. 1996. Generalization Error of Ensemble Estimators. In *Proceedings of International Conference on Neural Networks (ICNN'96)*, volume 1, 90–95. IEEE.

Viñas Torné, R.; Wiatrak, M.; Piran, Z.; Fan, S.; Jiang, L.; Teichmann, S. A.; Nitzan, M.; and Brbić, M. 2025. Systema: a Framework for Evaluating Genetic Perturbation Response Prediction beyond Systematic Variation. *Nature Biotechnology*, 44(6): 1050–1059.

Wang, C.; Karimzadeh, M.; Ravindra, N. G.; Bounds, L. R.; Alerasool, N.; Huang, A. C.; Ma, S.; Gulbranson, D. R.; Cui, H.; Lee, Y.; et al. 2026. X-Cell: Scaling Causal Perturbation Prediction Across Diverse Cellular Contexts via Diffusion Language Models. bioRxiv. Preprint, bioRxiv 2026.03.18.712807.

Wortsman, M.; Ilharco, G.; Gadre, S. Y.; Roelofs, R.; Gontijo Lopes, R.; Morcos, A. S.; Namkoong, H.; Farhadi, A.; Carmon, Y.; Kornblith, S.; and Schmidt, L. 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 23965–23998. PMLR.