POPFLOOR: SPLIT-HALF DIAGNOSTICS FOR FINITE-SAMPLE WASSERSTEIN BIAS IN SINGLE-CELL PERTURBATION BENCHMARKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Single-cell perturbation models are evaluated by comparing predicted and measured populations. Yet the plug-in Wasserstein distance has a finite-sample floor and can fall when a prediction shrinks toward its mean, so underdispersed predictions can outscore a target sample. PopFloor estimates this split-half floor from measured cells. By an exact per-draw identity for the unregularised squared 2-Wasserstein distance, a spread-free population beats an equal-size target sample wherever the floor's ratio to the held-out variance exceeds one plus the relative mean error. In an exploratory recomputation, scPerturBench's leader falls to fourth without Wasserstein distance. Our reproduced CPA scores below the floor on 88.7 to 92.6% of the largest screen's perturbations in three seeds, and by the pre-registered rule beats it on three of five screens. The ratio and that share rise together from 10 to 5,000 genes under the exact estimator, as pre-registered, while its relative mean error falls. Pre-registered tests show the ratio's margin over that threshold predicts, beyond gene and cell counts, the preference on disjoint cells of two unused screens and, on the reproductions, entrants' advantage over an independent floor in 11 of 12 runs. PopFloor shows where Wasserstein rankings can reward underdispersion.

1 INTRODUCTION

Single-cell perturbation screens measure how a population responds when a gene is silenced or activated (Dixit et al., 2016; Replogle et al., 2022). Models now predict responses for perturbations that were never measured (Lotfollahi et al., 2023; Roohani et al., 2024; Cui et al., 2024; Hao et al., 2024; Piran et al., 2024; Zhu & Li, 2026), although their advantage over simple baselines remains disputed (Ahlmann-Eltze et al., 2025; Csendes et al., 2025; Miller et al., 2025). Benchmarks compare predicted and measured cell populations under several distributional distances, then aggregate those comparisons into rankings across methods and screens (Wei et al., 2026; Wu et al., 2025; Wenteler et al., 2025). Transport distances also serve to evaluate new methods (Bunne et al., 2022; Klein et al., 2025; Palla et al., 2026; Li et al., 2026). These scores assess heterogeneity beyond the mean, but also depend on the metric's behaviour on finite cell samples.

However, the plug-in Wasserstein distance can prefer populations tighter than the target distribution. Its sample loss has biased gradients that can lead to a degenerate solution (Bellemare et al., 2017). The plug-in Fréchet distance can rank a slightly underdispersed Gaussian above the true one (Bińkowski et al., 2018). The same pattern appears in single-cell evaluation. Variance-shrunk simulations can score better than samples from the target distribution. Models sometimes beat one held-out half scored against the other on real data (Heidari et al., 2026), while the Wasserstein-2 distance only weakly separates a held-out half from the pooled perturbed population (Schäfer et al., 2026). Reported predictions also collapse modes or compress variance (Wu et al., 2025; Li et al., 2026). No evaluation we know of measures this preference from the cells before a model is scored or tests how much of a published ranking rests on it. A low empirical transport score can reflect recovery of the response distribution. It can also reflect a spread that matches the estimator's finite-sample preference. An aggregate leaderboard does not distinguish them.

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

To measure this preference from the cells before any model is scored, we propose PopFloor. Its split-half floor is a same-law reference computed between two halves of the measured population, not a bound that a correct prediction cannot cross. An exact per-draw identity for the unregularised squared 2-Wasserstein distance recovers a shrunk truth sample's preferred spread from this floor and the samples' first two moments. It also states when the sample's repeated mean outscores the sample itself (Figure 1). PopFloor re-ranks the benchmark's candidate genes on a discovery third, computes a model-independent floor-to-variance ratio, and tests below-floor advantages by varying the same predictions' spread. The floor and ratio use measured cells rather than entrant predictions, and comparisons treat the perturbation rather than the cell or training seed as the independent unit. Exploratory measurements motivated the framing. Later tests were pre-registered with dated amendments, and post hoc analyses are labelled where they appear (Appendix B).

The audit connects the diagnostic to model rankings. On two screens whose cells no earlier analysis had used, the margin measured on calibration cells predicts the preference on disjoint cells better than gene count and cell number, as pre-registered. Separately, on saved K562 and RPE1 predictions, it predicts how far reproduced entrants fall below an independent floor in 11 of 12 runs. On scPerturBench's largest screen, a repeated mean beats a truth sample in 97% of splits at 100 scoring genes. From 10 to 5,000 genes, the floor ratio rises from 0.65 to 1.92, while the share of perturbations on which our reproduction of the top-ranked method beats the floor rises from 6.5% to 100%. The reproduced cells are 3.4 times tighter than the truth's (Figure 3), and two further entrants are 15.8 to 40.5 times tighter. An exploratory recomputation leaves the published leader first only when the Wasserstein distance is included. On identical cells in a seven-arm comparison, the three entrants occupy the top three ranks under that distance and the bottom three under the energy distance. On PerturBench, two spaces were added after a pre-registered top-100 test failed. In these spaces, a truth sample ranks last among nine arms where the floor ratio is highest and second in the 30-component space where it is lowest.

Our contributions connect a measured-data diagnostic to benchmark outcomes and reporting.

1. We derive an identity, exact for every draw in any dimension, and a corollary that give the preferred spread from the split-half floor and the samples' first two moments. They also give the condition under which a repeated mean beats a truth sample. We verify both by direct exact solves in every run. Pre-registered tests show the cell-only margin predicts the preference on disjoint cells of two untouched screens beyond gene and cell counts. It predicts reproduced entrants' advantage over an independent floor on previously analysed screens beyond the same counts (Sections 3 and 4.4, Appendices A and K).

2. We use PopFloor to audit two perturbation benchmarks. Across eight gene counts, the floor ratio rises with the below-floor shares of a reproduced leader and a fixed-error surrogate, as pre-registered. Three reproduced entrants occupy the top three Wasserstein ranks and the bottom three energy ranks among seven arms on identical cells, while the same preference follows the floor ratio across three PerturBench spaces (Sections 4.2–4.6, Figure 2, Tables 1 and 2).

3. We turn the audit into reporting changes on the floor and ratio, a two-sided check, an energy-distance comparison and completion rules that route methods differently (Sections 4.7 and 5).

2 Related work

Predicting perturbation responses, and a dispute over metrics. Methods differ in what they condition on and in what they emit. Deep generative models predict a shifted cell distribution (Lotfollahi et al., 2023; Piran et al., 2024; Lotfollahi et al., 2019), graph and attention models a mean change (Roohani et al., 2024; Bai et al., 2024), and foundation models and gene-embedding predictors one or the other (Cui et al., 2024; Hao et al., 2024; Yang et al., 2024; Liu et al., 2026; Chen & Zou, 2025; 2024; Zhu & Li, 2026). Ahlmann-Eltze et al. (2025) report that deep predictors do not beat simple linear baselines. Kernfeld et al. (2025) and Csendes et al. (2025) reach compatible conclusions, while Miller et al. (2025) reply that the mean-level metrics are poorly calibrated. Apparent difficulty also depends on how the held-out set is built and how the signal is diluted (Gao et al., 2024; Mejia et al., 2025; Viñas Torné et al., 2026; Kedzierska et al., 2025). The dispute concerns mean-level metrics, and we do not attempt to resolve it.

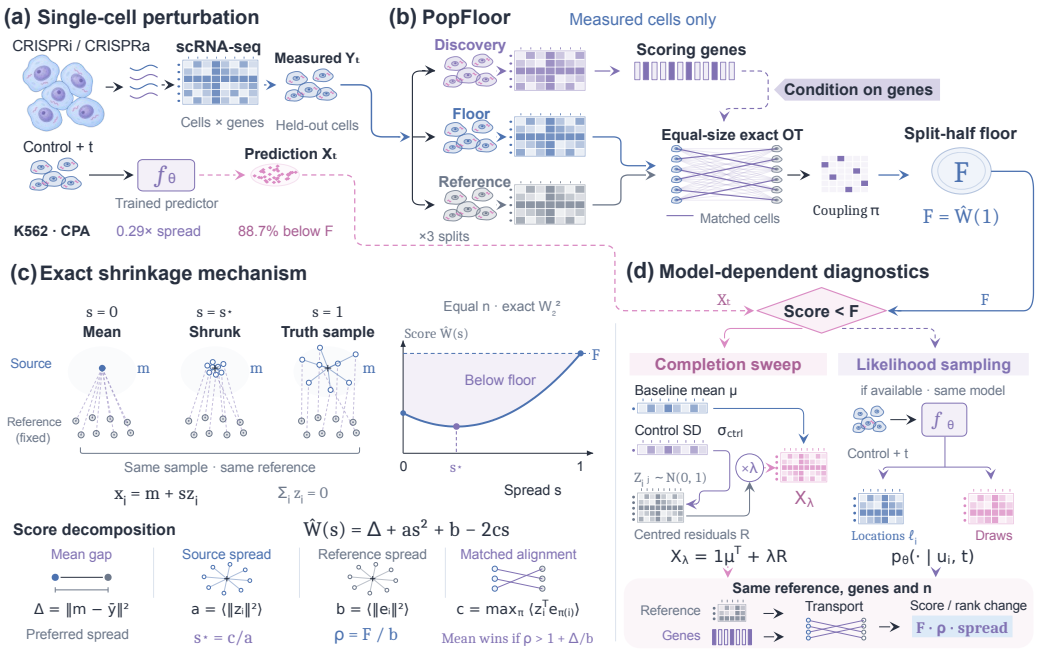


Figure 1: **PopFloor diagnoses shrinkage advantages in single-cell perturbation benchmarks.** **(a)** Measured responses define the floor, while predictions are scored separately. The K562 CPA callouts are exploratory. **(b)** Held-out cells are split into discovery, floor and reference thirds. The discovery third re-ranks the benchmark's candidate genes, and the other thirds define the split-half floor over three splits. **(c)** Exact equal-size transport decomposes the score and identifies its preferred spread. Points, couplings and the curve are schematic. **(d)** Completion and likelihood-sampling probes rescore predictions with the reference, genes and population size fixed within each comparison. Neither changes the measured-only floor.

Metrics for cell populations, and finite-sample bias. Ji et al. (2023) benchmark sixteen single-cell population distances and recommend mean squared error. Mejia et al. (2025) and Miller et al. (2025) calibrate mean-level metrics with a technical duplicate, one half of a perturbation's cells scored against the other. The current Virtual Cell Challenge scales its mean-level and differential-expression metrics so that such a split scores one (Arc Institute, 2026). The closest work is Heidari et al. (2026). They score distributional metrics against a held-out cell split. Baseline and trained models sometimes beat it under the Wasserstein distance on two real datasets. In simulated expression data the variance scaling that minimises that distance falls below one as the dimension grows, while the energy distance is minimised at the match. Schäfer et al. (2026) compare the split with pooled perturbed cells and find that the Wasserstein-2 distance separates them only weakly on seven screens. A dispersed negative reference cannot show which way the distance errs, whereas a population with no spread can. Transport distances score predicted populations on genes (Wei et al., 2026; Bunne et al., 2022; Driessen et al., 2026), in low-dimensional embeddings (Klein et al., 2025; Palla et al., 2026; Atanackovic et al., 2025) and one gene at a time (Lotfollahi et al., 2023; Li et al., 2026), in nineteen works we found (Appendix H). The theory is older. A sample Wasserstein loss has biased gradients that the energy distance removes (Bellemare et al., 2017). Minimising its expectation between samples underestimates the scale of a one-dimensional location-scale model through the same quadratic in the spread (Jang & Noh, 2026). Subtracting the distance between two independent samples of one law centres the plug-in squared distance (Papp & Sherlock, 2025). The empirical measure converges slowly in high dimension (Fournier & Guillin, 2015; Genevay et al., 2019). A model-dependent finite-sample bias can reorder models, as for the Fréchet inception distance (Chong & Forsyth, 2020; Bińkowski et al., 2018). A score whose expectation the true distribution does not minimise is not proper (Gneiting & Raftery, 2007). The contraction warning and split reference are prior work. Our contribution is predictive and interventional. An exact per-draw identity in any dimension gives a cell-only margin. Pre-registered tests show it predicts

the preference on disjoint cells of two untouched screens and reproduced entrants' advantage over independent floors beyond gene and cell counts. Spread and completion interventions establish the mechanism on the same reproduced outputs. The published-rank audit shows metric dependence.

Benchmarks, and rankings that turn out to be artefacts of the protocol. Recent benchmarks evaluate many methods across many screens under a shared split and metric set (Wei et al., 2026; Wenteler et al., 2025; Wu et al., 2025) on harmonised data and tooling (Peidli et al., 2024; Heumos et al., 2026; Wolf et al., 2018; Virshup et al., 2024). Elsewhere in machine learning, tuned simple baselines have matched the published state of the art and reported gaps have fallen within training variance (Ferrari Dacrema et al., 2019; Musgrave et al., 2020; Melis et al., 2018; Bouthillier et al., 2021). In drug-response prediction, a metric choice reverses the ranking (Agarwal & Bisht, 2026). Most of these audits vary the method and hold the evaluation fixed. We hold the predictions fixed and vary scoring and completion, as Schäfer et al. (2026) vary protocols on scPerturBench.

3 What the split-half floor says about the bias

A screen provides unperturbed control cells and, for each perturbation t in a held-out set $\mathcal{T}$, a measured population $\mathbf{Y}_t \in \mathbb{R}^{n_t \times G}$ over G genes. A benchmark scores a method by comparing its predicted population $\mathbf{X}_t$ with $\mathbf{Y}_t$ under a distance between distributions and averaging over $\mathcal{T}$. For a method that emits only a mean response, the benchmark completes the response by drawing cells around that mean (Section 4.7). Transport distances are computed on the two populations' empirical measures using the plug-in estimator.

Proposition 1. *Let $x_i = m + sz_i$ and $y_i = \bar{y} + e_i$ for $i = 1, \dots, n$, with $\sum_i z_i = \sum_i e_i = 0$, $s \geq 0$, $a = \frac{1}{n}\sum_i \|z_i\|^2$, $b = \frac{1}{n}\sum_i \|e_i\|^2$ and $\Delta = \|m - \bar{y}\|^2$. The plug-in squared 2-Wasserstein distance between the two empirical measures is*

$$\widehat{W}(s) \;=\; \Delta + as^2 + b - 2cs, \qquad c = \max_{\pi} \tfrac{1}{n}\textstyle\sum_i \langle z_i, e_{\pi(i)} \rangle,$$

the maximum over permutations, with $0 \leq c \leq \sqrt{ab}$. For $a > 0$ it is minimised at $s^\star = c/a$, and $s^\star$ is below the spread $\sqrt{b/a}$ that matches the held-out cells unless the matched residuals are proportional.

Corollary 1. *If x is a truth sample of the same size as y, then $F = \widehat{W}(1)$ is the split-half floor and $c = (a + b + \Delta - F)/2$. Together with a, b and Δ, the floor therefore gives the preferred spread, $s^\star = (a+b+\Delta-F)/(2a)$. A shrunk copy then beats the sample for every $s \in \big((b+\Delta-F)/a,\, 1\big)$. In particular, the sample's own mean, repeated, scores $\Delta+b$ and beats the sample whenever $F > b+\Delta$.*

Expanding a coupling's cost gives an objective linear in the coupling, so the optimum over the Birkhoff polytope is attained at a permutation independent of s (Appendix A). The statement is exact for every draw, not in expectation. It covers the unregularised estimator at equal sizes, not the benchmark's entropic call, which the same appendix compares with it on identical arrays. The ratio $\rho = F/b$ needs no model. The sample's repeated mean outscores the sample exactly when $\rho > 1 + \Delta/b$, a threshold near one when Δ is small. The preferred spread is near $1 - \rho/2$ of the truth's when $a \approx b$ and $\Delta \approx 0$. The magnitude of ρ depends on how well independent residuals align. In one gene, the sorted coupling aligns them almost perfectly and ρ is near zero. Across many genes of high effective dimension, they are nearly orthogonal and ρ approaches two, consistent with the slow $n^{-1/d}$ convergence of empirical transport (Fournier & Guillin, 2015). These limits are heuristics, and Section 4.4 measures ρ itself. A prediction of the same size with residual norm a', alignment c' and mean error Δ' is below the floor exactly when $\rho - 1 > (\Delta' + a' - 2c')/b$, so ρ is its model-independent part. The floor is thus a same-law reference, not a bound that a correct prediction cannot cross. A sample of the truth falls below it in about half of single comparisons, so a score below the floor does not by itself show that a prediction is wrong. Neither the floor nor ρ is a detector with a controlled false-alarm rate. Section 4.4 tests whether the measured margin $\rho - 1 - \Delta/b$ predicts the preference on separate cells.

Calibrating the floor on the data. Each held-out population is randomly split into thirds. The first ranks genes within the benchmark's candidate list. The others define the floor and reference (Appendix D). With genes selected independently of the scored samples, a same-law prediction is

exchangeable with the floor third. Its candidate-minus-floor score averages zero and, absent ties, is negative half the time in expectation. This is the split reference of Heidari et al. (2026) and technical duplicate of Mejia et al. (2025), with gene re-ranking moved off the scored thirds. We average scores over three random splits per perturbation. Candidate filtering and averaging need not preserve the half-below-floor null, as the full-pipeline simulation shows (Appendix K).

PopFloor. The diagnostic first estimates a split-half reference for each metric and held-out population. It then computes the floor-to-variance ratio ρ of the exact transport score in the same scoring space. Only the third step needs a model. For predictions that score below the floor, it tests the contribution of spread through controlled completion and, when available, likelihood sampling. Section 4 applies all three steps, the last only to scPerturBench.

4 EXPERIMENTS

4.1 SETUP

Data, split and metrics. scPerturBench's genetic single-perturbation track ranks 15 methods on 9 datasets (Wei et al., 2026). We use the 5 screens whose published test sets we reproduce exactly with the benchmark's split procedure at seeds 1, 2 and 3. The K562 and RPE1 essential screens, Replogle-8, and Tian activation and inhibition hold out 323, 266, 12, 19 and 36 perturbations. Models see every gene in the released archive (5,822 on the largest screen), and scoring uses each perturbation's top 100 differentially expressed genes as the harness ranks them (Appendix J). The benchmark ranks on six metrics. We reproduce the five population distances, namely the mean squared error, the Pearson distance, the energy distance (Székely & Rizzo, 2013), the symmetric Kullback-Leibler divergence (Kullback & Leibler, 1951) and the Wasserstein distance. We compute them through the harness's call to pertpy's distance module (Heumos et al., 2026) at version 1.0.3 with ott-jax 0.6.0. Its Wasserstein distance is an entropy-regularised transport cost on squared Euclidean distances, returned without its self-terms (Cuturi et al., 2022). We also score the same arrays with the Sinkhorn divergence and the exact unregularised estimator, which Proposition 1 describes and every floor ratio below uses. Table A.1 names the score behind every figure and table.

Statistics and pre-registration. The independent unit is the perturbation, never the cell and never the seed. Paired comparisons are bootstrapped over held-out perturbations within a seed, with 10,000 resamples and percentile intervals, each for its own comparison. The largest screen's CPA runs, the interventions on them and the recomputations on the published table were exploratory. Apart from two sensitivity spaces, later runs were pre-registered with their decision rules before any output existed. Those rules are the only pre-specified checks (Appendix B). Analyses of existing runs are labelled post hoc where they appear.

4.2 FIRST PLACE ON SCPERTURBENCH TURNS ON THE WASSERSTEIN DISTANCE

The benchmark's published table has no single winner when ranked one metric at a time (Schäfer et al., 2026). Exploratory recomputation from its per-metric ranks with its own composite aggregation leaves the top method, CPA (Lotfollahi et al., 2023), first when any other metric is dropped and fourth without the Wasserstein distance. Without that distance it ranks first in none of 2,000 bootstrap resamples of the perturbations. Its pooled lead comes from the two largest screens, which hold 81% of the ranked rows (Appendix F). Eight of the fifteen entrants write their own cells, six hand a mean to the harness, which completes it, and one has no released script for this track. The three best published Wasserstein ranks belong to entrants that write their own cells, and three of those, CPA, scGPT (Cui et al., 2024) and scouter (Zhu & Li, 2026), rank 10.1, 13.5 and 13.2 of fifteen on the energy distance. No entrant's cells are released, so we reproduce these three from the benchmark's released scripts (Appendix C). Our CPA reproduces the published CPA's orderings against the linear baseline on 98.5 to 100% of perturbations under the Wasserstein distance. Under the two mean-level metrics it agrees on only 30.3 to 37.8%, and the substituted embedding can change spread as well as mean (Appendices C and F). The published-rank audit and the reproduced-model interventions therefore support separate claims, and the original CPA's dispersion remains unmeasured.

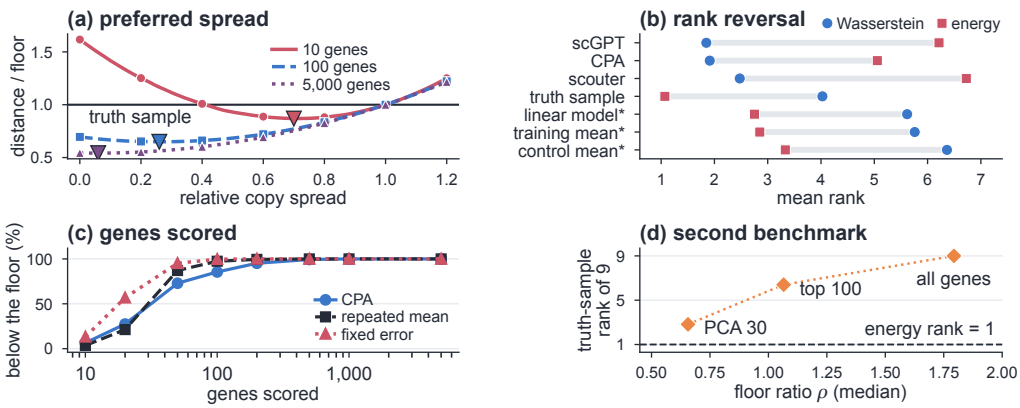


Figure 2: **The floor tracks transport's preference for compressed populations across scoring spaces.** **(a)** Target-median exact W_2^2 curves divided by the floor. **(b)** Seed-averaged seven-arm ranks on identical cells under the benchmark's Wasserstein call and the energy distance. Asterisks mark completed means. **(c)** Gene-count response under exact W_2^2. CPA and fixed-error shares count perturbations. The repeated-mean share counts splits. **(d)** Exact floor ratio and truth-sample rank under the benchmark's call across three spaces. Panels (a)–(c) use the largest screen.

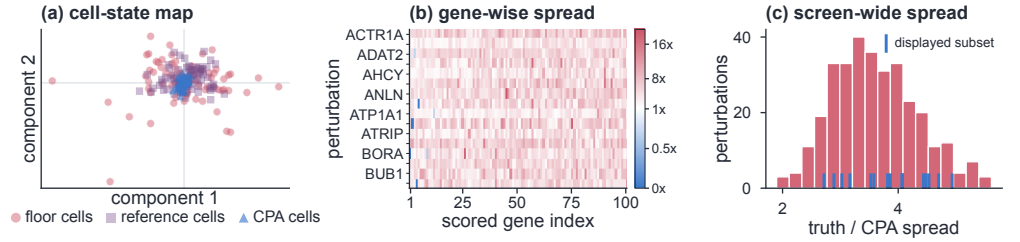


Figure 3: **Single-cell transcriptomic populations show that reproduced CPA responses are systematically tighter.** **(a)** Measured and predicted cells for the first predeclared perturbation. The projection is fitted on floor cells. **(b)** Reference-to-CPA gene-wise spread ratios over three splits for the first 16 held-out perturbations alphabetically, with alternate rows labelled. No outcome orders rows or genes. **(c)** Ratios for all 323 perturbations. Ticks mark the displayed subset.

4.3 The ranked entrants are too tight and beat a sample of the truth

Too tight, everywhere. In the exploratory first-seed runs, CPA's reproduced population has a mean per-gene standard deviation of 0.117 against the truth's 0.402, a ratio of 0.29. It is tighter than the truth on *every one* of the 323 held-out perturbations, by a factor of 2.0 to 5.6 (Figure 3). After centring every population, the two distances retain that ordering (Appendix E). The reproduced scouter is 15.8 to 18.3 and scGPT 21.5 to 40.5 times tighter than the truth across three seeds, again on every held-out perturbation.

Below the floor under the Wasserstein distance, never under the energy distance. On the largest screen, CPA is below the split-half floor on 88.7% of perturbations, paired -4.45 $[-5.05, -3.75]$ over 310. The shares are 92.6% and 88.7% on the other two seeds, each with its own trained CPA. Two further model seeds on the first split give 88.4% and 88.7%. The scouter reproduction is below the floor on 94.2 to 95.5% and the scGPT reproduction on 92.6 to 94.9%. None of the three is below the energy-distance floor on any perturbation in any seed (Table D.1). Post hoc, the floor third beats the harness-completed prediction of no change on 91.3 to 94.6% of perturbations under the Wasserstein distance. The metric on which these entrants lead therefore passes a check of the kind Schäfer et al. (2026) run, a split against a dispersed negative (Appendix D). The debiased Sinkhorn divergence and the exact estimator show the same preference, on 79.3% and 78.3% of perturbations on the released genes. The regularisation in the benchmark-era ott-jax release strengthens

it to 88.9% (Table A.2). Across the five screens CPA is tighter on every perturbation of four and on 94 to 96% of the fifth, and beats the floor on three. Full-axis re-ranking retains K562's below-floor exact-W_2^2 mean, while RPE1 intervals include zero (Appendix L).

The lead goes with the spread, and with how the cells are written. In the exploratory first-split intervention, at the control's spread, the harness-completed linear baseline trails CPA on the Wasserstein distance by 10.55 $[10.26, 10.82]$. Completing the baseline at CPA's own ratio, 0.29 of the control's spread, reduces the gap to -0.06 $[-0.26, 0.12]$. Rescaled to the truth's spread, CPA loses to the completed baseline by 2.92 $[2.48, 3.37]$ (Appendix A). With the same draws recentred to the predicted mean, as pre-registered, the gap at CPA's ratio remains -0.08 $[-0.28, 0.09]$. The rescaling preserves CPA's sample mean. The benchmark's CPA script writes the location of the decoder's likelihood as each cell. Drawing each cell from that likelihood instead, with the same trained model and inputs, multiplies the per-gene spread by five. It puts CPA above the floor on 99.7% of perturbations, paired $+7.88$ $[7.09, 8.73]$, and behind the harness-completed linear baseline (Appendix C). Recentred to the mean of the decoder's locations it stays above the floor on 99.7%, though the draw also changes the covariance (Appendix A).

4.4 The floor ratio tracks where the preference is strong

Corollary 1 turns each population's floor into a condition. We pre-registered three tests before computing any floor ratio on real data, replaying the thirds, genes and entrant subsamples of the floor runs above under the exact estimator (Appendix B). Every checked replay matched the stored floor, and the identity held on every direct solve to a relative error of 4×10^{-14}.

On the largest screen, at the benchmark's 100 scoring genes, the median floor ratio is 1.50. The repeated mean of a sample of the truth beats the sample itself in 97.4% of splits, and a shrunk copy scores best at a median of 0.27 of the sample's spread, close to CPA's own ratio of 0.29. Varying only the number of scoring genes, from 10 to 5,000, moves the median ratio from 0.65 to 1.92. The share of splits in which the repeated mean beats the sample rises from 3.5% to 100%, and the share of perturbations on which CPA beats the floor from 6.5% to 100% (Figure 2(c)). Both pre-registered rules hold, with a Spearman correlation of 1.0 between gene count and ratio and of 0.99 between ratio and CPA's share. Both hold again on the other two seeds, each with its own trained CPA. CPA's mean error relative to the held-out variance also falls as genes are added, from 0.33 to 0.04. A second pre-registered test holds CPA's spread and mean error fixed, in a surrogate rather than in CPA. The surrogate uses the floor third's residuals at CPA's 100-gene values and is scored exactly by the identity. Its below-floor share rises from 13.2% to 100% over the same gene counts, with a Spearman correlation of 0.94 with the ratio. Across the five screens at 100 genes the median ratio runs from 0.51 to 1.50. Post hoc, Replogle-8, the one screen where CPA never beats the floor, is also the one where the ratio is below one and even the repeated mean does not (Appendix B).

Prediction on disjoint cells. Pre-registration P predicts the repeated-mean margin on validation cells. Per gene count, frozen regressions on the log half size are compared with and without the margin $q = \rho - 1 - \Delta/b$ measured on separate calibration cells. On 72 Adamson and 104 Replogle-7 targets with cells unused by earlier analyses, adding q lowers the mean absolute error over 10, 100 and 1,000 genes by 0.082 $[0.068, 0.095]$ and 0.453 $[0.439, 0.466]$, as pre-registered. The reductions are 0.052 $[0.039, 0.064]$ and 0.059 $[0.047, 0.071]$ with both regressions refitted within each screen. The unfitted q has the lowest error in all six settings (Appendix K.4). Pre-registration Q instead predicts the margin $y = [\widehat{W}(X, D) - \widehat{W}(C, D)]/b_D$ of saved entrant predictions X. The evaluation halves C and D are disjoint from the calibration cells, and a negative y places the entrant below their floor. Over five target folds per run, adding q lowers the out-of-fold error in all 12 runs, by 0.014 to 0.056, with an interval above zero in 11, at 100 and 1,000 genes but not ten. The all-seed rule holds for CPA on RPE1, scGPT and scouter but not CPA on K562. In R, registered after Q's outcome, adding q to a quadratic cell-count model lowers the pooled error by 10.4% $[6.8, 12.4]$, above its 5% registered threshold (Appendix K.5). Q reuses analysed screens, and neither test establishes biological model selection.

Table 1: **Among seven original arms, the three entrants occupy the top three Wasserstein ranks and the bottom three energy distance ranks.** Post hoc, from the existing floor runs. Largest screen, 310 to 312 held-out perturbations per seed. Each arm is scored with the benchmark's Wasserstein call and energy distance against the reference third on genes chosen on the discovery third, ranked per perturbation (1 is the smallest distance) and the ranks averaged, then averaged over three seeds, which never differ by more than 0.21. D, defined in Section 5, subtracts both self-terms and is a descriptive ordering, not a corrected ranking. These runs did not compute the truth sample's own self-term, so it has no entry there.

Arm	Wasserstein	Energy	D
A sample of the truth	4.03	1.07	–
scGPT	1.85	6.21	4.75
CPA	1.91	5.06	4.30
scouter	2.48	6.73	5.18
Linear model, completed	5.62	2.75	1.86
Training mean, completed	5.76	2.85	2.03
Control mean, completed	6.36	3.33	2.87

4.5 On reproduced cells, the seven-arm ordering reverses

The seven original arms of the floor runs are scored against the same reference thirds on the same genes. This allows a post hoc ranking per perturbation. Under the Wasserstein distance the three entrants take the first three places, a sample of the truth the fourth and the harness-completed baselines the rest. Under the energy distance the order reverses (Figure 2(b)). The truth sample is first, followed by the completed baselines and then the three entrants (Table 1). Two added raw means rank last among nine arms under both energy estimators (Appendix L). On each of the other four screens CPA's mean rank is first or second of five under the Wasserstein distance and last under the energy distance, in every seed (Appendix D).

4.6 A second benchmark, trained by its own code

PerturBench (Wu et al., 2025) uses an energy statistic for its distributional ranking, not a transport distance. Its models therefore show what a transport distance would have done there. We trained its eight tuned configurations for Norman19 (Norman et al., 2019) with its code, one seed each as released, kept their per-cell predictions, and scored them with a sample of the truth on 45 held-out combinations. Seven of the eight score close to PerturBench's published means on its own metrics, and our linear additive model is worse (Appendix G). Every model is tighter than the measured cells on all genes. The least tight is 3.1 times tighter, and one has only identical cells.

On all genes, where PerturBench computes its energy statistic, a sample of the truth ranks last of nine under the Wasserstein distance. Under PerturBench's own statistic it ranks first, with the eight models following in exact order of tightness (Table 2). In PerturBench's 30-component PCA the same sample ranks second, also with the projection fitted on training cells alone. On the pre-registered top-100 genes it ranks eighth. The pre-registered test that the Wasserstein distance orders the models by tightness narrowly fails, at a Spearman correlation of -0.48 against a registered threshold of -0.5. That gene rule degenerates on these small populations, where on average 93.7% of the genes chosen on a discovery third are constant. The all-gene and PCA spaces were therefore added as sensitivity analyses after that result was read and before they ran. The floor ratio, registered after these ranks had been read, orders the three spaces as the truth sample's rank does, at 1.79 on all genes, where the repeated mean beats the sample on every combination, 1.07 on the top-100 genes and 0.66 in the PCA, where it beats it on none.

4.7 The completion rule routes methods differently

scPerturBench completes a mean-emitting method with independent Gaussian draws at the control's per-gene spread, but never completes a method that writes cells. Giving CPA the linear baseline's completion reverses the Wasserstein comparison from about ten ahead to about one behind in every seed, also under mean-exact completion, every bootstrap interval excluding zero (Appendix I).

Table 2: **A sample of the truth ranks last of nine on all genes and second in the 30-component embedding.** PerturBench's Norman19 models, one run each, 45 held-out combinations. Times tighter divides the truth's mean per-gene standard deviation by the model's, over all genes. Mean rank among the nine arms per combination, 1 the smallest distance, under the benchmark's Wasserstein call in three spaces and under PerturBench's own energy statistic. The top-100 genes are the pre-registered space, and the two others were added after its result was read (Appendix G). The floor ratio is the exact split-half floor over the held-out variance, registered after the ranks were read (Section 4.4). *Run with the constant covariate column its released code requires.

		Wasserstein, mean rank			Energy
Arm	Times tighter	Top 100	All genes	PCA 30	mean rank
A sample of the truth	–	6.40	9.00	2.82	1.00
Floor ratio ρ, median	–	1.07	1.79	0.66	–
Linear additive	3.1	7.56	7.98	7.18	2.13
biolord	3.3	6.36	7.02	7.96	3.09
SAMS-VAE, modified	4.9	2.69	2.80	2.49	3.96
SAMS-VAE	5.0	5.44	5.60	6.62	4.96
CPA, no adversary	10.9	3.76	2.29	3.04	6.04
CPA	13.0	4.02	2.31	3.11	6.82
Latent additive	1,055	3.44	3.62	5.71	8.11
Decoder only*	∞	5.33	4.38	6.07	8.89

5 What a benchmark should report

P1. Report references matched to the estimator. Report the scorer's floor, cell counts and gene rule (Mejia et al., 2025; Heidari et al., 2026). Compute ρ and the margin $\rho - 1 - \Delta/b$ with exact W_2^2 on genes chosen on separate cells. A positive margin says the repeated mean wins, not that a model is wrong. The two-sided score D, the score minus half of each self-term (Papp & Sherlock, 2025), prefers the completed linear baseline to CPA in Table 1 but is not proper (Appendix A).

P2. Report an energy-distance comparison. It is the divergence of a strictly proper scoring rule (Gneiting & Raftery, 2007), and U removes V's $1/n$ bias (Székely & Rizzo, 2013). Both detect the specified Gaussian errors and real-cell collapse, though not every permutation or wrong-target control. Neither shows that energy selects biologically better models (Appendices K and L).

P3. Make the completion visible. Score mean-only metrics on the predicted mean. Compare differently routed methods under both routings, exclude undefined arms, and publish the raw-distance-to-rank transform, which scPerturBench does not release (Appendices I and J).

6 Limitations

Some entrants remain unreproduced, and our CPA matches the published one only in distributional orderings (Appendix F). Three screens hold 12 to 36 perturbations. PerturBench's transport ranking is counterfactual, with one run per configuration, and its Norman19 screen is also K562. A weaker linear baseline and the failed gene rule limit the post hoc spaces and ratio comparison. Completion, published-table recomputation, first-screen CPA runs and interventions preceded registration, and Table 1 and the five-screen comparison are post hoc. The below-floor share is not a calibrated 50% null. Ten-gene flags are sparse and, where estimable, adverse. Q reuses analysed screens and fails for CPA on K562 (Appendix K). The floor mixes biological and technical variation (Appendix A).

7 Conclusion

Before any model is scored, PopFloor measures when finite-sample transport favours contraction over a truth sample, and its margin helps predict how far reproduced entrants score below an independent floor.

Ethics statement

We re-analyse publicly released single-cell CRISPR screens from scPerturBench and PerturBench and collect no new human or animal data. Trained models are our reproductions from public releases, and the results describe transport scores on finite samples, not the merit of any method. We know of no dual-use risk and follow the ICLR Code of Ethics.

Reproducibility statement

The supplementary code recomputes the reported numbers from the released per-perturbation results with one script and rebuilds the figures. It includes a record of the pre-registered analyses with their dates, rules and amendments, two pinned verification environments and scripts that prepare cell matrices from public archives. The test on the two untouched screens fixes its fitted predictors in the released code and refuses development inputs that do not reproduce them. Both out-of-sample tests are recomputed from target-level rows (Appendices K.4 and K.5), and the later analyses keep per-run observations, repeats, exclusions and intervals (Appendix L). Biological scores need cell matrices and model outputs, and the synthetic controls (Appendix K) need neither. The checks of the split, the mean-only metrics and the Wasserstein call are in Section 4.1 and Appendices I and A, and Appendix J lists what we could not reproduce.

AI use statement

Large language models contributed to code development, manuscript revision and polishing, and the design, implementation and interpretation of the experiments in Appendices K and L. Synthetic data in Appendix K come from its stated samplers. Generative AI drew the icons in Figure 1(a), which are not experimental images. The authors take full responsibility for the manuscript and code.

References

B Adamson, T M Norman, M Jost, M Y Cho, J K Nuñez, Y Chen, J E Villalta, L A Gilbert, M A Horlbeck, M Y Hein, R A Pak, A N Gray, C A Gross, A Dixit, O Parnas, A Regev, and J S Weissman. A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell*, 167(7):1867–1882.e21, 2016. doi: 10.1016/j.cell.2016.11.048.

Dhruv Agarwal and Riya Bisht. The metric picks the winner: Evaluation choice flips model rankings for drug-response prediction in unseen chemistry. *arXiv preprint arXiv:2606.12639*, 2026.

Constantin Ahlmann-Eltze, Wolfgang Huber, and Simon Anders. Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines. *Nature Methods*, 22: 1657–1661, 2025. doi: 10.1038/s41592-025-02772-6.

Arc Institute. The vcc2026 metric suite, 2026. URL `https://github.com/ArcInstitute/cell-eval2/blob/main/docs/vcc2026_metrics/vcc2026-metrics.md`. Metric specification of the 2026 Virtual Cell Challenge, dated 2026-08-19.

Lazar Atanackovic, Xi Zhang, Brandon Amos, Mathieu Blanchette, Leo J. Lee, Yoshua Bengio, Alexander Tong, and Kirill Neklyudov. Meta flow matching: Integrating vector fields on the Wasserstein manifold. In *International Conference on Learning Representations*, 2025.

Ding Bai, Caleb N. Ellington, Shentong Mo, Le Song, and Eric P. Xing. AttentionPert: accurately modeling multiplexed genetic perturbations with multi-scale effects. *Bioinformatics*, 40 (Supplement 1):i453–i461, 2024. doi: 10.1093/bioinformatics/btae244.

Marc G. Bellemare, Ivo Danihelka, Will Dabney, Shakir Mohamed, Balaji Lakshminarayanan, Stephan Hoyer, and Rémi Munos. The Cramér distance as a solution to biased Wasserstein gradients. *arXiv preprint arXiv:1705.10743*, 2017.

Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Tal Arbel, Chris Pal, Gael Varoquaux, and Pascal Vincent. Accounting for variance in machine learning benchmarks. In *Proceedings of Machine Learning and Systems*, volume 3, 2021.

Charlotte Bunne, Andreas Krause, and Marco Cuturi. Supervised training of conditional Monge maps. In *Advances in Neural Information Processing Systems*, 2022.

Charlotte Bunne, Stefan G. Stark, Gabriele Gut, Jacobo Sarabia del Castillo, Mitch Levesque, Kjong-Van Lehmann, Lucas Pelkmans, Andreas Krause, and Gunnar Rätsch. Learning single-cell perturbation responses using neural optimal transport. *Nature Methods*, 20:1759–1768, 2023. doi: 10.1038/s41592-023-01969-x.

Yiqun Chen and James Zou. GenePert: Leveraging GenePT embeddings for gene perturbation prediction. *bioRxiv*, 2024. doi: 10.1101/2024.10.27.620513.

Yiqun Chen and James Zou. Simple and effective embedding model for single-cell biology built from chatgpt. *Nature Biomedical Engineering*, 9:483–493, 2025. doi: 10.1038/s41551-024-01284-6.

Changxi Chi, Yufei Huang, Jun Xia, Jiangbin Zheng, Yunfan Liu, Zelin Zang, and Stan Z. Li. Departures: Distributional transport for single-cell perturbation prediction with neural Schrödinger bridges. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 20544–20552, 2026a. doi: 10.1609/aaai.v40i25.39190.

Changxi Chi, Jun Xia, Yufei Huang, Zhuoli Ouyang, Cheng Tan, Yunfan Liu, Jingbo Zhou, Chang Yu, Liangyu Yuan, Siyuan Li, Zelin Zang, and Stan Z. Li. Doloris: Dual conditional diffusion implicit bridges with sparsity masking strategy for unpaired single-cell perturbation estimation. In *International Conference on Learning Representations*, 2026b.

Min Jin Chong and David Forsyth. Effectively unbiased FID and inception score and where to find them. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6069–6078, 2020. doi: 10.1109/CVPR42600.2020.00611.

Gerold Csendes, Gema Sanz, Kristóf Z. Szalay, and Bence Szalai. Benchmarking foundation cell models for post-perturbation RNA-seq prediction. *BMC Genomics*, 26:393, 2025. doi: 10.1186/s12864-025-11600-2.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21:1470–1480, 2024. doi: 10.1038/s41592-024-02201-0.

Marco Cuturi, Laetitia Meng-Papaxanthos, Yingtao Tian, Charlotte Bunne, Geoff Davis, and Olivier Teboul. Optimal transport tools (OTT): A JAX toolbox for all things Wasserstein. *arXiv preprint arXiv:2201.12324*, 2022.

Paul Datlinger, André F. Rendeiro, Christian Schmidl, Thomas Krausgruber, Peter Traxler, Johanna Klughammer, Linda C. Schuster, Amelie Kuchler, Donat Alpar, and Christoph Bock. Pooled CRISPR screening with single-cell transcriptome readout. *Nature Methods*, 14:297–301, 2017. doi: 10.1038/nmeth.4177.

Andrea Giuseppe Di Francesco, Andrea Rubbi, and Pietro Liò. Retrieval-augmented generation for predicting cellular responses to gene perturbation. In *ICLR 2026 Workshop on Generative AI in Genomics (Gen^2): Barriers and Frontiers*, 2026.

A Dixit, O Parnas, B Li, J Chen, C P Fulco, L Jerby-Arnon, N D Marjanovic, D Dionne, T Burks, R Raychowdhury, B Adamson, T M Norman, E S Lander, J S Weissman, N Friedman, and A Regev. Perturb-seq: Dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell*, 167(7):1853–1866.e17, 2016. doi: 10.1016/j.cell.2016.11.038.

Alice Driessen, Dhruva Abhijit Rajwade, Benedek Harsanyi, Marianna Rapsomaniki, and Jannis Born. Conditional Monge gap enables generalizable single-cell perturbation modelling. *Nature Machine Intelligence*, 8:984–996, 2026. doi: 10.1038/s42256-026-01242-8.

Luca Eyring, Dominik Klein, Théo Uscidda, Giovanni Palla, Niki Kilbertus, Zeynep Akata, and Fabian Theis. Unbalancedness in neural Monge maps improves unpaired domain translation. In *International Conference on Learning Representations*, 2024.

Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. Are we really making much progress? a worrying analysis of recent neural recommendation approaches. In *Proceedings of the 13th ACM Conference on Recommender Systems*, 2019. doi: 10.1145/3298689.3347058.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. 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*, pp. 2681–2690, 2019.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3–4):707–738, 2015. doi: 10.1007/s00440-014-0583-7.

Chris J. Frangieh, Johannes C. Melms, Pratiksha I. Thakore, Kathryn R. Geiger-Schuller, Patricia Ho, Adrienne M. Luoma, Brian Cleary, Livnat Jerby-Arnon, Shruti Malu, Michael S. Cuoco, Maryann Zhao, Casey R. Ager, Meri Rogava, Lila Hovey, Asaf Rotem, Chantale Bernatchez, Kai W. Wucherpfennig, Bruce E. Johnson, Orit Rozenblatt-Rosen, Dirk Schadendorf, Aviv Regev, and Benjamin Izar. Multimodal pooled Perturb-CITE-seq screens in patient models define mechanisms of cancer immune evasion. *Nature Genetics*, 53:332–341, 2021. doi: 10.1038/s41588-021-00779-1.

Yicheng Gao, Zhiting Wei, Kejing Dong, Ke Chen, Jingya Yang, Guohui Chuai, and Qi Liu. Toward subtask-decomposition-based learning and benchmarking for predicting genetic perturbation outcomes and beyond. *Nature Computational Science*, 4:773–785, 2024. doi: 10.1038/s43588-024-00698-1.

Aude Genevay, Lénaïc Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of Sinkhorn divergences. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1574–1583, 2019.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma, Xuegong Zhang, and Le Song. Large-scale foundation model on single-cell transcriptomics. *Nature Methods*, 21:1481–1491, 2024. doi: 10.1038/s41592-024-02305-7.

Mahshid Heidari, Mina Karimpour, Sumana Srivatsa, and Hesam Montazeri. Evaluating single-cell perturbation response models is far from straightforward. *bioRxiv*, 2026. doi: 10.64898/2026.02.14.705879.

Lukas Heumos, Yuge Ji, Lilly May, Tessa D. Green, Stefan Peidli, Xinyue Zhang, Xichen Wu, Johannes Ostner, Antonia Schumacher, Karin Hrovatin, Michaela Müller, Faye Chong, Gregor Sturm, Alejandro Tejada, Emma Dann, Mingze Dong, Gonçalo Pinto, Mojtaba Bahrami, Ilan Gold, Sergei Rybakov, Altana Namsaraeva, Amir Ali Moinfar, Zihe Zheng, Eljas Roellin, Isra Mekki, Chris Sander, Mohammad Lotfollahi, Herbert B. Schiller, and Fabian J. Theis. Pertpy: an end-to-end framework for perturbation analysis. *Nature Methods*, 23:350–359, 2026. doi: 10.1038/s41592-025-02909-7.

Cheongjae Jang and Yung-Kyun Noh. On the finite-sample bias of minimizing expected Wasserstein loss between empirical distributions. In *Proceedings of The 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 4420–4428. PMLR, 2026. URL `https://proceedings.mlr.press/v300/jang26b.html`.

Yuge Ji, Tessa D. Green, Stefan Peidli, Mojtaba Bahrami, Meiqi Liu, Luke Zappia, Karin Hrovatin, Chris Sander, and Fabian J. Theis. Optimal distance metrics for single-cell RNA-seq populations. *bioRxiv*, 2023. doi: 10.1101/2023.12.26.572833.

Qun Jiang, Shengquan Chen, Xiaoyang Chen, and Rui Jiang. scPRAM accurately predicts single-cell gene expression perturbation response based on attention mechanism. *Bioinformatics*, 40(5): btae265, 2024. doi: 10.1093/bioinformatics/btae265.

Kasia Z. Kedzierska, Lorin Crawford, Ava P. Amini, and Alex X. Lu. Zero-shot evaluation reveals limitations of single-cell foundation models. *Genome Biology*, 26:101, 2025. doi: 10.1186/s13059-025-03574-x.

Eric Kernfeld, Yunxiao Yang, Joshua S. Weinstock, Alexis Battle, and Patrick Cahan. A comparison of computational methods for expression forecasting. *Genome Biology*, 26:388, 2025. doi: 10.1186/s13059-025-03840-y.

Dominik Klein, Jonas Simon Fleck, Daniil Bobrovskiy, Lea Zimmermann, Sören Becker, Alessandro Palma, Leander Dony, Alejandro Tejada-Lapuerta, Guillaume Huguet, Hsiu-Chuan Lin, Nadezhda Azbukina, Fátima Sanchís-Calleja, Theo Uscidda, Artur Szalata, Manuel Gander, Aviv Regev, Barbara Treutlein, J. Gray Camp, and Fabian J. Theis. CellFlow enables generative single-cell phenotype modeling with flow matching. *bioRxiv*, 2025. doi: 10.1101/2025.04.11.648220.

S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951. doi: 10.1214/aoms/1177729694.

Lanxiang Li, Yue You, Yunlin Fu, Wenyu Liao, Xueying Fan, Shihong Lu, Ye Cao, Bo Li, Wenle Ren, Jiaming Kong, Shuangjia Zheng, Jizheng Chen, Xiaodong Liu, and Luyi Tian. A systematic comparison of single-cell perturbation response prediction models. *Science Advances*, 12(37): eaed3414, 2026. doi: 10.1126/sciadv.aed3414.

Tianyu Liu, Tianqi Chen, Wangjie Zheng, Xiao Luo, Yiqun Chen, and Hongyu Zhao. Embeddings from language models are good learners for single-cell data analysis. *Patterns*, 7:101431, 2026. doi: 10.1016/j.patter.2025.101431.

Mohammad Lotfollahi, F. Alexander Wolf, and Fabian J. Theis. scGen predicts single-cell perturbation responses. *Nature Methods*, 16:715–721, 2019. doi: 10.1038/s41592-019-0494-8.

Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Leon Hetzel, Yuge Ji, Ignacio L. Ibarra, Sanjay R. Srivatsan, Mohsen Naghipourfar, Riza M. Daza, Beth Martin, Jay Shendure, Jose L. McFaline-Figueroa, Pierre Boyeau, F. Alexander Wolf, Nafissa Yakubova, Stephan Günnemann, Cole Trapnell, David Lopez-Paz, and Fabian J. Theis. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology*, 19:e11517, 2023. doi: 10.15252/msb.202211517.

Xinjie Mao, Songming Zhang, Qianhong Wen, Xiangyu Wen, Kedu Jin, Hao Wu, Shuizhou Chen, Yuqiang Li, Lei Bai, Qi Liu, Ning Ding, Siqi Sun, and Zhangyang Gao. Benchmarking virtual cell models for in-the-wild perturbation response. *arXiv preprint arXiv:2604.27646*, 2026.

Gabriel M. Mejia, Henry E. Miller, Francis J. A. Leblanc, Bo Wang, Brendan Swain, and Lucas Paulo de Lima Camillo. Diversity by design: Addressing mode collapse improves scRNA-seq perturbation modeling on well-calibrated metrics. In *ICML 2025 Generative AI and Biology (GenBio) Workshop*, 2025. doi: 10.48550/arXiv.2506.22641.

Gábor Melis, Chris Dyer, and Phil Blunsom. On the state of the art of evaluation in neural language models. In *International Conference on Learning Representations*, 2018.

Henry E. Miller, Gabriel M. Mejia, Francis J. A. Leblanc, Bo Wang, Brendan Swain, and Lucas Paulo de Lima Camillo. Deep learning-based genetic perturbation models do outperform uninformative baselines on well-calibrated metrics. *bioRxiv*, 2025. doi: 10.1101/2025.10.20.683304.

Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. In *Computer Vision – ECCV 2020*, pp. 681–699, 2020. doi: 10.1007/978-3-030-58595-2_41.

T M Norman, M A Horlbeck, J M Replogle, A Y Ge, A Xu, M Jost, L A Gilbert, and J S Weissman. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science*, 365(6455):786–793, 2019. doi: 10.1126/science.aax4438.

Giovanni Palla, Sudarshan Babu, Payam Dibaeinia, James D. Pearce, Donghui Li, Aly A. Khan, Theofanis Karaletsos, and Jakub M. Tomczak. Scalable single-cell gene expression generation with latent diffusion models. In *International Conference on Machine Learning*, 2026.

Efthymia Papalexi, Eleni P. Mimitou, Andrew W. Butler, Samantha Foster, Bernadette Bracken, William M. Mauck, III, Hans-Hermann Wessels, Yuhan Hao, Bertrand Z. Yeung, Peter Smibert, and Rahul Satija. Characterizing the molecular regulation of inhibitory immune checkpoints with multimodal single-cell screens. *Nature Genetics*, 53:322–331, 2021. doi: 10.1038/s41588-021-00778-2.

Tamás P. Papp and Chris Sherlock. Centered plug-in estimation of Wasserstein distances. *arXiv preprint arXiv:2203.11627*, 2025.

Matteo Pariset, Ya-Ping Hsieh, Charlotte Bunne, Andreas Krause, and Valentin De Bortoli. Unbalanced diffusion Schrödinger bridge. *arXiv preprint arXiv:2306.09099*, 2023.

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

Zoe Piran, Niv Cohen, Yedid Hoshen, and Mor Nitzan. Disentanglement of single-cell data with biolord. *Nature Biotechnology*, 42:1678–1683, 2024. doi: 10.1038/s41587-023-02079-x.

J M Replogle, R A Saunders, A N Pogson, J A Hussmann, A Lenail, A Guna, L Mascibroda, E J Wagner, K Adelman, G Lithwick-Yanai, N Iremadze, F Oberstrass, D Lipson, J L Bonnar, M Jost, T M Norman, and J S Weissman. Mapping information-rich genotype-phenotype landscapes with genome-scale perturb-seq. *Cell*, 185(14):2559–2575.e28, 2022. doi: 10.1016/j.cell.2022.05.013.

Joseph M. Replogle, Thomas M. Norman, Albert Xu, Jeffrey A. Hussmann, Jin Chen, J. Zachery Cogan, Elliott J. Meer, Jessica M. Terry, Daniel P. Riordan, Niranjan Srinivas, Ian T. Fiddes, Joseph G. Arthur, Luigi J. Alvarado, Katherine A. Pfeiffer, Tarjei S. Mikkelsen, Jonathan S. Weissman, and Britt Adamson. Combinatorial single-cell CRISPR screens by direct guide RNA capture and targeted sequencing. *Nature Biotechnology*, 38:954–961, 2020. doi: 10.1038/s41587-020-0470-y.

Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nature Biotechnology*, 42:927–935, 2024. doi: 10.1038/s41587-023-01905-6.

Philipp S. L. Schäfer, Kendall A. Reid, Zach Boldyga, Ekin D. Aksu, Hugo Hakem, and Julio Saez-Rodriguez. Towards principled evaluation of single-cell perturbation prediction models. *bioRxiv*, 2026. doi: 10.64898/2026.07.23.740433.

G Schiebinger, J Shu, M Tabaka, B Cleary, V Subramanian, A Solomon, J Gould, S Liu, S Lin, P Berube, L Lee, J Chen, J Brumbaugh, P Rigollet, K Hochedlinger, R Jaenisch, A Regev, and E S Lander. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943.e22, 2019. doi: 10.1016/j.cell.2019.01.006.

Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5): 2263–2291, 2013. doi: 10.1214/13-AOS1140.

Peiting Shi, Ningfeng Que, Xianzhe Huang, Xiaofei Wang, and Jianzhong Jeff Xi. StateXDiff: Cell state-contextualized multimodal diffusion for single-cell perturbation prediction. *arXiv preprint arXiv:2605.16104*, 2026.

Bingchi Sun, Haibin Zheng, Jinyin Chen, and Jingjing Jin. A neural-network-free calibration matches or beats deep single-cell perturbation response models across four datasets. *Genes*, 17 (7):816, 2026. doi: 10.3390/genes17070816.

Gábor J. Székely and Maria L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143:1249–1272, 2013. doi: 10.1016/j.jspi.2013.03.018.

R Tian, M A Gachechiladze, C H Ludwig, M T Laurie, J Y Hong, D Nathaniel, A V Prabhu, M S Fernandopulle, R Patel, M Abshari, M E Ward, and M Kampmann. CRISPR interference-based platform for multimodal genetic screens in human iPSC-derived neurons. *Neuron*, 104(2):239–255.e12, 2019. doi: 10.1016/j.neuron.2019.07.014.

Ruilin Tian, Anthony Abarientos, Jason Hong, Sayed Hadi Hashemi, Rui Yan, Nina Dräger, Kun Leng, Mike A. Nalls, Andrew B. Singleton, Ke Xu, Faraz Faghri, and Martin Kampmann. Genome-wide CRISPRi/a screens in human neurons link lysosomal failure to ferroptosis. *Nature Neuroscience*, 24:1020–1034, 2021. doi: 10.1038/s41593-021-00862-0.

Théo Uscidda and Marco Cuturi. The Monge gap: A regularizer to learn all transport maps. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *PMLR*, pp. 34709–34733, 2023.

Ramon Viñas Torné, Maciej Wiatrak, Zoe Piran, Shuyang Fan, Liangze Jiang, Sarah A. Teichmann, Mor Nitzan, and Maria Brbić. Systema: a framework for evaluating genetic perturbation response prediction beyond systematic variation. *Nature Biotechnology*, 44:1050–1059, 2026. doi: 10.1038/s41587-025-02777-8.

Isaac Virshup, Sergei Rybakov, Fabian J. Theis, Philipp Angerer, and F. Alexander Wolf. anndata: Access and store annotated data matrices. *Journal of Open Source Software*, 9(101):4371, 2024. doi: 10.21105/joss.04371.

Zhiting Wei, Yiheng Wang, Yicheng Gao, Shuguang Wang, Ping Li, Duanmiao Si, Yuli Gao, Siqi Wu, Danlu Li, Kejing Dong, Xingbo Yang, Chen Tang, Shaliu Fu, Xiaohan Chen, Wannian Li, Yuzhou You, Chen Zhang, Aibin Liang, Guohui Chuai, and Qi Liu. Benchmarking algorithms for generalizable single-cell perturbation response prediction. *Nature Methods*, 23:451–464, 2026. doi: 10.1038/s41592-025-02980-0.

Aaron Wenteler, Martina Occhetta, Nikhil Branson, Victor Curean, Magdalena Huebner, William Dee, William Connell, Siu Pui Chung, Alex Hawkins-Hooker, Yasha Ektefaie, César Miguel Valdez Córdova, and Amaya Gallagher-Syed. PertEval-scFM: Benchmarking single-cell foundation models for perturbation effect prediction. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 66633–66677. PMLR, 2025.

F. Alexander Wolf, Philipp Angerer, and Fabian J. Theis. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, 19:15, 2018. doi: 10.1186/s13059-017-1382-0.

Yan Wu, Esther Wershof, Sebastian Schmon, Marcel Nassar, Błażej Osiński, Ridvan Eksi, Zichao Yan, Rory Stark, Kun Zhang, and Thore Graepel. PerturBench: Benchmarking machine learning models for cellular perturbation analysis. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3225.

Xiaodong Yang, Guole Liu, Guihai Feng, Dechao Bu, Pengfei Wang, Jie Jiang, Shubai Chen, Qinmeng Yang, Hefan Miao, Yiyang Zhang, Zhenpeng Man, Zhongming Liang, Zichen Wang, Yaning Li, Zheng Li, Yana Liu, Yao Tian, Wenhao Liu, Cong Li, Ao Li, Jingxi Dong, Zhilong Hu, Chen Fang, Lina Cui, Zixu Deng, Haiping Jiang, Wentao Cui, Jiahao Zhang, Zhaohui Yang, Handong Li, Xingjian He, Liqun Zhong, Jiaheng Zhou, Zijian Wang, Qingqing Long, Ping Xu, The X-Compass Consortium, Hongmei Wang, Zhen Meng, Xuezhi Wang, Yangang Wang, Yong Wang, Shihua Zhang, Jingtao Guo, Yi Zhao, Yuanchun Zhou, Fei Li, Jing Liu, Yiqiang Chen, Ge Yang, and Xin Li. GeneCompass: deciphering universal gene regulatory mechanisms with a knowledge-informed cross-species foundation model. *Cell Research*, 34:830–845, 2024. doi: 10.1038/s41422-024-01034-y.

Ouyang Zhu and Jun Li. Scouter predicts transcriptional responses to genetic perturbations with large language model embeddings. *Nature Computational Science*, 6:21–28, 2026. doi: 10.1038/s43588-025-00912-8.

Appendix

Appendix A proves the identity of Section 3 and maps every result to its estimator. Appendix B holds the floor-ratio tests of Section 4.4 and every pre-registered rule with its outcome. Appendices C to F support the entrant reproductions and the published ranking of Sections 4.2 and 4.3, Appendix G the second benchmark, Appendix H the survey of transport scores and Appendix I the completion analysis. Appendix J lists what could not be reproduced. Appendix K gives the calibration, sensitivity and disjoint-cell tests, including the two out-of-sample tests of Section 4.4, and Appendix L the saved-prediction rescoring and measured-cell controls. Unless a caption says otherwise, Wasserstein in the appendix is the benchmark's entropic call and energy is the Euclidean V-statistic the harness computes, while floor ratios, preferred spreads and every quantity marked exact use the unregularised squared distance (Table A.1).

Contents

A Estimators and proofs
B Floor-ratio tests
C Predictor implementations
D Reproduced entrants
E Centred dispersion
F Published-rank recomputations
G Second benchmark
H Use of transport scores
I Completion analysis
J Reproduction limits
K Calibration and separate cells
L Saved predictions and controls

A Transport estimators and proofs

This appendix proves Proposition 1 and Corollary 1 of Section 3 and shows on identical arrays that the benchmark's entropic call carries the same preference as the exact estimator (Section 4.3).

Metric background and library parity. The energy distance is equivalent to a kernel two-sample statistic (Gretton et al., 2012; Sejdinovic et al., 2013), and optimal transport has been used to map single-cell populations across developmental time and perturbation (Schiebinger et al., 2019; Bunne et al., 2023). Our independent reimplementations of the mean squared error, the Pearson distance and the energy distance agree exactly with pertpy 1.0.3 on twelve random cases across four group-size and dimension combinations. The library's energy distance has two code paths. Up to version 0.10 its direct call uses squared Euclidean distances, for which the statistic reduces exactly to twice the squared distance between the group means. The harness's one-sided call instead uses precomputed Euclidean distances. The benchmark's energy column, like ours, is therefore the Euclidean energy distance, which sees spread.

The metrics as the harness calls them. The mean squared error and the Pearson distance depend only on the two group means, with the Pearson distance computed on their change from the control mean. The energy distance is $2\,\mathbb{E}\|x-y\| - \mathbb{E}\|x-x'\| - \mathbb{E}\|y-y'\|$ on Euclidean distances. Each average includes all pairs, including a point with itself, giving the V-statistic. The symmetric Kullback-Leibler divergence is reported as $\log_2(1+v)$ of the library's value. Every value is rounded to four decimals, and a group of more than 2,000 cells is first subsampled with a generator seeded 42, as the harness does. Every other random stream in this paper uses numpy's default generator with a fixed, recorded seed. Table A.4 gives each metric's split-half floor on every screen, beside what the harness-completed mean of one half scores against the other.

The benchmark's Wasserstein column is an entropy-regularised transport cost computed with ott-jax's default regularisation. In our version, this is a twentieth of the cost matrix's standard deviation. The cost is returned without the two self-terms that turn it into a Sinkhorn divergence. To separate the regularisation from the finite-sample bias, the same arrays are scored three ways. The first is the benchmark's own call, reproduced here to within its four-decimal rounding. The second is the Sinkhorn divergence from ott-jax with the regularisation shared across its three terms, which removes the entropic bias (Feydy et al., 2019). The third is the exact unregularised empirical squared 2-Wasserstein distance, which has no regulariser at all and is solved as an assignment problem. The held-out half is trimmed by at most one row to give equal sizes.

All three prefer the tight candidate on the synthetic Gaussian of Table A.2, and under all three the ranked entrant scores below the split-half floor on the largest screen, with every bootstrap interval excluding zero. Debiasing removes about 5% of the raw floor on the synthetic Gaussian and 11%

Table A.1: **Proposition 1 and Corollary 1 hold exactly only for the exact estimator, and every result names the score it uses.** The benchmark's call uses pertpy's distance module to return an ott-jax entropic transport cost at the library's default regularisation, without its self-terms. Table A.2 compares it with the exact estimator on identical arrays. The energy distance is not a transport score and has no floor ratio. Its V-statistic keeps the self-pairs and is biased by order $1/n$, and its U-statistic drops them.

	Exact unregularised W_2^2	Benchmark's call	Energy distance, V and U
Computed as	assignment on squared Euclidean cost, equal sizes	entropic cost, self-terms not subtracted	$2\,\mathbb{E}\|x-y\| - \mathbb{E}\|x-x'\| - \mathbb{E}\|y-y'\|$
Proposition 1	exact for every draw	not covered, same preference measured	not applicable
Main text	ρ, preferred spread and repeated mean (Section 4.4); Figure 2(a), (c) and the ratio axis of (d); the ratio row of Table 2; the two out-of-sample tests	below-floor shares and interventions (Section 4.3); Wasserstein columns of Tables 1 and 2; Figure 2(b) and the rank axis of (d); completion (Section 4.7)	energy floors (Section 4.3); energy columns of Tables 1 and 2, the latter PerturBench's own statistic; Figure 2(b)
Appendices	B, K, L (full-axis genes)	D, F, G, I	V in D; V and U in K and L

on the largest screen, and on neither does it remove the preference. The exact estimator's floor is the two-sample bias of the plug-in estimator, to which the benchmark's call adds its regularisation. At 100 genes, we measure a large floor at every population size these screens provide. This is consistent with the slow $n^{-1/d}$ convergence of unregularised empirical transport in high dimension (Fournier & Guillin, 2015). It is a property of the cells as measured, so it carries whatever technical variation the assay and the preprocessing leave in them, and we do not separate that from biological variation. Reconstructed arrays match the floor measurement of Section 4.3 on all 969 perturbation-and-split values to 10^{-13}.

Proof of Proposition 1 and Corollary 1. A coupling of two uniform measures on n points is P/n with P doubly stochastic, and it costs $\frac{1}{n}\sum_{i,j} P_{ij}\|m-\bar{y}+sz_i-e_j\|^2$. Expanding, the terms linear in $m-\bar{y}$ vanish, because every row and column of P sums to one and z and e are centred. The remaining cost is $\|m-\bar{y}\|^2 + as^2 + b - \frac{2s}{n}\sum_{i,j} P_{ij}\langle z_i, e_j\rangle$. The last sum is linear in P, so its maximum over the Birkhoff polytope is attained at a permutation. For $s>0$, the cheapest coupling maximises this sum regardless of s, giving the identity with c as stated. At $s=0$ every coupling costs $\Delta+b$, which the formula also gives. The average of $\frac{1}{n}\sum_i \langle z_i, e_{\pi(i)}\rangle$ over all permutations is $\langle \bar{z}, \bar{e}\rangle = 0$, so $c \geq 0$. Applying the Cauchy–Schwarz inequality twice gives $c \leq \sqrt{ab}$, with equality only if the matched residuals are proportional. The quadratic is minimised at $s^\star = c/a$, and $c < \sqrt{ab}$ puts $s^\star$ below $\sqrt{b/a}$. For the corollary, $s=1$ gives $F = \Delta + a + b - 2c$, which solves for c. Then $\widehat{W}(s) - F = (s-1)\big(a(s+1)-2c\big)$, which for $s<1$ is negative exactly when $s > 2c/a - 1 = (b+\Delta-F)/a$.

Checks of the identity, and the scale of the floor. The identity is checked against the exact solver in the released tests and on every run of Appendix B. On the synthetic Gaussian, identical copies of the correct mean score $\operatorname{tr}\Sigma = 100\times 0.40^2 = 16.0$ in expectation under the benchmark's call and the exact estimator alike, since every coupling of identical copies costs the same. Taking $a=b=\operatorname{tr}\Sigma$ and fixing c from the matched candidate's exact score predicts 14.84 at 0.29 of the spread against 14.83 measured, with a minimiser near 0.27. Between two samples of the same law at 300 cells the benchmark's call returns 2.2 at a per-gene standard deviation of 0.12 and 24.2 at 0.40 on 100 genes, and 111.5 at 0.40 on 400 genes.

The benchmark's numerical environment. The benchmark used pertpy 0.7 but did not pin ott-jax. Its Wasserstein call uses the same four lines as the version we tested, so the ott-jax release in the benchmark's environment remains unknown. The ott-jax release current with pertpy 0.7 was 0.4.5. Its default regularisation is one twentieth of the mean cost, about three times the setting used here, which is one twentieth of the cost's standard deviation. We scored the same arrays with those four lines in an environment containing only that ott-jax release and jax. Under this regularisation

the preference is stronger (last row of Table A.2). This direction is expected because increasing regularisation moves the cost toward the expected squared distance under independent coupling, which a point mass at the target mean minimises.

Spread minima and calibration. We sweep the candidate's spread from zero to twice the truth's at 100, 300 and 1,000 cells, using fresh draws rather than the three of Table A.2. The benchmark call's minimum is at 0.20, 0.25 and 0.29 of the truth's spread, and the exact estimator's at 0.25, 0.25 and 0.29. Identical copies of the correct mean score 16.0 at 300 cells against 24.4 for the matched candidate. On 1,000 synthetic populations, a candidate drawn from the truth's own law lands below the split-half floor on 50.6% with a paired mean of -0.001 (standard error 0.008), consistent with single-comparison exchangeability. A candidate at 0.29 of the spread lands below it on 100%. This calibration draws the samples independently in a fixed scoring space. Under those conditions, averaging repeats preserves the zero expected margin, but need not preserve the half-below-floor probability. This check does not establish calibration after data-dependent candidate filtering. Appendix K tests the full pipeline. The two-sided score of P1 is $D = W(X, Y) - \frac{1}{2}W(X, X') - \frac{1}{2}W(Y, Y')$. Here, X' is a second prediction sample of the same size (a second completion, or a disjoint half of an entrant's cells), and $W(Y, Y')$ is the split-half floor, all under the benchmark's call. Over the same range of spreads, with the candidate and the truth each drawn as two halves of 150 cells, it is minimised at the truth's spread, while the benchmark's call between the same halves is minimised at a fifth of it. Its failure of propriety can be computed exactly and is small. Let P put equal mass on $(\pm 1, 0)$ and $(0, \pm 1)$ and Q on $(\pm 1, \pm 1)/\sqrt{2}$, which share mean and covariance. Averaging the exact distance over every equally likely pair of two-point samples gives Q an expected D against P of $(7 - 5\sqrt{2})/16 \approx -0.0044$, while P scores zero. The released tests check that value.

Dispersion at a fixed predicted mean on the real screen. First seed throughout, three random half-splits per perturbation. CPA emits one cell per held-out cell, and its scored half is a random subset of those, of the floor half's size. Held-out populations on the two largest screens hold 50 to 648 cells on K562, median 140, and 50 to 1,699 on RPE1, median 90. The linear baseline's mean is completed at $\lambda \in \{0, 0.29, 0.5, 1\}$ times the control's per-gene spread, with a second, independent draw for its self-term. The draws are not recentred, so each completion's sample mean differs from the predicted mean by noise of per-gene variance λ^2 times the control's over the half's size. CPA's cells are rescaled about their own mean by the ratio of the floor half's mean per-gene spread to theirs on the 100 scoring genes. The factor is 1.9 to 5.9 with median 3.5 over half-splits, the half-size counterpart of the 2.0 to 5.6 of Section 4.3. The target spread comes from the half that is not scored. Arms that also appear in the floor measurement reproduce it on all 1,938 shared perturbation-and-split values. Paired over 323 perturbations, the Wasserstein distance of the linear baseline minus CPA's is 10.55 at $\lambda = 1$, -0.06 at 0.29 and -0.30 $[-0.51, -0.11]$ at 0, and on the two-sided score of P1 it is -2.09 $[-2.30, -1.91]$ at $\lambda = 1$. The energy distance prefers the baseline completed at the control's spread to CPA's own cells by 1.28 $[1.23, 1.33]$. CPA's likelihood draw trails that baseline on the Wasserstein distance by 0.59 $[0.40, 0.81]$.

The same completions recentred. Pre-registration G recentred every completion to the predicted mean with the same draws. The gap to CPA becomes 10.28 $[9.99, 10.54]$ at $\lambda = 1$ and -0.08 $[-0.28, 0.09]$ at 0.29. The exact estimator's median a' and c' do not change. The median relative mean error at $\lambda = 1$ falls from 0.193 to 0.177, the value for identical copies. CPA's likelihood draw, recentred to the mean of the decoder's locations at the same rows, stays above the floor on 99.7% of perturbations, by 7.63 $[6.84, 8.48]$ on average, and behind the recentred linear baseline by 0.62 $[0.42, 0.83]$. Recentring lowers its score by 0.25 $[0.22, 0.27]$ and its median relative mean error from 0.219 to 0.207.

Gene selection and the symmetric KL divergence. The scoring genes are the benchmark's released top-100 lists, ranked on the whole held-out population, so the floor and every arm are scored on genes chosen with both halves. That choice is the benchmark's and applies to every arm alike. Re-ranking the benchmark's candidate genes on a discovery third makes the entrant's lead over the floor larger, and every trivial baseline still beats CPA on the energy distance and on D (Table A.3). On the symmetric KL divergence the entrant is below the floor on 43.7% of perturbations. Its paired

Table A.2: **The preference for a tight population belongs to the transport estimator, not to the solver the benchmark uses.** Exploratory comparison of three estimators on identical arrays. The benchmark's call is an entropy-regularised cost without the self-terms, at our library version. The Sinkhorn divergence subtracts them with a shared regularisation. The exact estimator is the unregularised empirical squared 2-Wasserstein distance, solved as an assignment problem. The last row uses the same call with the regularisation of ott-jax 0.4.5, a twentieth of the mean cost. That release was current with the benchmark's pertpy but is not a pinned dependency of it. The synthetic columns compare a Gaussian truth at 100 genes and 300 cells with candidates at the truth's spread and at 0.29 of it, averaging three draws. They also show which of five spreads, 0.29, 0.5, 1, 2 and 3.4 times the truth's, each estimator scores lowest. It is a grid minimum, not the continuous one, which a finer sweep in the text locates. The three draws differ by at most 0.54 at the matched spread and 0.13 at 0.29 of it, under any estimator, against gaps of 8.5 to 9.1 between the two. The largest-screen columns give the split-half floor (SHF) and the ranked entrant, both at the same half size, over 323 held-out perturbations, with the paired difference and its bootstrap interval. The first row reproduces Table A.4. Section 4.3.

	synthetic			largest screen			
estimator	matched	0.29×	minimum	SHF	entrant	below (%)	entrant − SHF
call, our version	24.52	15.41	0.29×	27.76	24.51	81.7	-3.26 $[-3.98, -2.49]$
Sinkhorn divergence	23.25	14.74	0.29×	24.66	22.14	79.3	-2.52 $[-3.27, -1.72]$
exact W_2^2	23.35	14.83	0.29×	25.24	22.91	78.3	-2.33 $[-3.09, -1.50]$
call, ott-jax 0.4.5	28.95	16.83	0.29×	30.76	25.29	88.9	-5.47 $[-6.16, -4.73]$

Table A.3: **Discovery re-ranking enlarges CPA's measured lead over the floor. Completed baselines still beat CPA on energy and the two-sided score.** Largest screen, first seed unless stated, 310 held-out perturbations with at least 60 cells, each split at random into thirds three times. A discovery third re-ranks genes within the released whole-population candidate list. The other thirds define the floor and reference. Applied to the whole population, the same ranking rule reproduces the released top-100 lists exactly on all 60 perturbations checked. Each baseline is completed by the harness, with a second completion for its self-term. Paired means with 95% bootstrap intervals, and the share of perturbations on which the difference is negative. Section 4.3.

difference	score	paired mean	negative (%)
CPA − floor, genes from the discovery third	Wass.	-4.45 $[-5.05, -3.75]$	88.7
the same, second seed and its own CPA	Wass.	-4.98 $[-5.47, -4.40]$	92.6
the same, third seed and its own CPA	Wass.	-4.52 $[-5.09, -3.91]$	88.7
CPA − floor, released genes, same thirds	Wass.	-3.72 $[-4.45, -2.90]$	85.2
CPA − floor, genes from the discovery third	energy	$+1.66$ $[+1.56, +1.77]$	0.0
linear model − CPA	Wass.	$+8.38$ $[+8.12, +8.64]$	0.6
linear model − CPA	energy	-1.21 $[-1.26, -1.17]$	100.0
linear model − CPA	D	-1.91 $[-2.10, -1.75]$	98.7
training mean − CPA	Wass.	$+8.63$ $[+8.41, +8.87]$	0.0
training mean − CPA	energy	-1.19 $[-1.22, -1.15]$	100.0
training mean − CPA	D	-1.67 $[-1.77, -1.57]$	98.1
control mean − CPA	Wass.	$+9.26$ $[+8.95, +9.58]$	0.0
control mean − CPA	energy	-1.12 $[-1.15, -1.09]$	100.0
control mean − CPA	D	-1.04 $[-1.15, -0.92]$	86.8

mean is below the floor, -1.49 $[-2.51, -0.48]$, while its median is above it, $+2.36$, because a minority of perturbations with very large floors dominate that metric's mean.

B THE FLOOR RATIO IN FULL

This appendix gives the floor-ratio tests behind Section 4.4, dates every registration and lists every pre-specified rule with its outcome.

Table A.4: **A perfect mean completed by the harness scores worse than the split-half floor on every screen under the Wasserstein distance.** Each cell reads *SHF / perfect mean*. The SHF is one half of the held-out population scored against the other. The perfect mean is the mean of that first half, completed by the harness, so it never sees the half it is scored on. Both are at the same size. Held-out counts in the header, seed 1. Section 4.3.

metric	K562 (323)	RPE1 (266)	Replogle-8 (12)	Tian act. (19)	Tian inh. (36)
MSE	0.006 / 0.008	0.008 / 0.013	0.002 / 0.003	0.004 / 0.006	0.002 / 0.003
Pearson	0.095 / 0.136	0.066 / 0.103	0.004 / 0.009	0.163 / 0.234	0.178 / 0.255
energy	0.183 / 0.277	0.259 / 0.462	0.049 / 0.427	0.129 / 0.215	0.087 / 0.132
sym. KL	16.92 / 14.10	31.65 / 29.97	29.25 / 41.02	34.31 / 39.61	40.09 / 43.54
Wasserstein	27.76 / 30.11	25.54 / 35.30	27.78 / 58.96	20.23 / 23.21	9.83 / 11.51

Registration timeline. Table B.1 dates each pre-registered rule and each exploratory, sensitivity and post hoc analysis. The exploratory analyses that shaped the framing all precede the first pre-registration. The second benchmark's two sensitivity spaces were specified after its registered top-100-gene result had been read and before they were run. One check was amended after a test run of its code and before any comparison was computed. G's replay excludes the self-term of identical copies from the stored values it must reproduce because their entropic score is undefined on a zero cost matrix.

Decision rules. Each rule is a threshold on a descriptive statistic, fixed with the statements its outcomes would support. None is a hypothesis test with a controlled error rate, and the paper's intervals are per comparison. A rule required of every seed is a conjunction, which more seeds make harder to pass. Table B.2 lists every rule and its outcome, failures included.

The three tests and their checks. The three tests of Section 4.4 were pre-registered, with their decision rules and the statement each outcome would support, before any floor ratio had been computed on real data. The runs replay the floor runs' random stream, so every population has already been scored in those runs. The thirds, discovery genes and CPA's subsample are drawn again in the same order. Three checks were completed before any outcome was computed. First, the replayed discovery genes reproduce the stored overlap with the released genes for every perturbation and split, together with the stored cell counts and third sizes. Second, the benchmark's own call on the replayed thirds reproduces the stored floor exactly on the first five populations of every dataset and space. Third, the exact distance of the floor third shrunk about its own mean, solved directly at four spreads, equals the identity everywhere it was checked. All three passed in every run. Every quantity here uses the exact estimator. A population whose reference third has no spread, $b = 0$, would have no ratio. One whose floor third has no spread, $a = 0$, would have no preferred spread. Either would be excluded and counted, and neither occurred.

Test outcomes. Both rules on the number of genes hold (Table B.3). The rule on the second benchmark's spaces also holds. The top-100 space, which enters no rule, falls between the other two in the same order as the truth sample's rank (Table B.5). CPA's mean error relative to the held-out variance also falls as genes are added, so the gene-count design shows the two terms of the identity moving together and does not separate them. In a post hoc comparison, the five screens line up with the corollary for a very tight prediction, whose own spread and alignment are small, so that it scores close to $\Delta + b$ and beats the floor roughly where its relative mean error is below $\rho - 1$ (Table B.4). Replogle-8 is the one screen with ρ below one and the one where CPA never beats the floor. Among the other four, RPE1 is where CPA beats it least often and where its relative mean error is largest. With five screens that is a consistent ordering rather than a demonstration.

The surrogate with the mean error held fixed. Pre-registration E asks how much of the gene-count rise the ratio produces without CPA's falling mean error. On each split the identity scores the floor third's residuals shrunk to spread s, about a mean whose squared distance to the reference third's mean is δb, exactly as $\delta b + as^2 + b - 2cs$, with a, b and c those of that split's floor and reference thirds. Averaging over a perturbation's three splits gives $\delta\bar{b} + \bar{a}s^2 + \bar{b} - 2\bar{c}s$. The arm counts as below the floor when this is below the split-averaged floor, following the rule for CPA's

Table B.1: **Registration timeline.** Dates of the pre-registration's entries and of the first saved result of each exploratory or post hoc analysis, all in 2026. A rule counts as pre-registered only when it was fixed before any output it judges existed.

when	what	status
Sep 8	the completion analysis on five screens	exploratory
Sep 11	CPA's floor on the largest screen, in half-splits and then in thirds over three seeds, the centred and estimator comparisons, the leave-one-metric-out ranking, the spread sweep and the likelihood draw	exploratory
Sep 12	rules for scGPT and scouter on the largest screen and for CPA on RPE1, before any of their outputs	pre-registered
Sep 12	rules for CPA on further screens, amended from seven to three before any ran because four splits do not reproduce	pre-registered
Sep 17	CPA with two further model seeds, and PerturBench's CPA verification run	pre-registered
Sep 17	PerturBench's eight models, amended twice before any output was read	pre-registered
Sep 17	the ranking of every floor-run arm in Table 1, from existing runs	post hoc
Sep 17	the all-gene and PCA spaces, after the pre-registered gene result was read and before they ran	sensitivity
Sep 17	the floor ratio's three rules, before any ratio on real data, with the eight models' ranks already read	pre-registered
Sep 17	the relative mean error held fixed, and the floor ratio on seeds 2 and 3, before either was computed, the second amended once before it ran	pre-registered
Sep 17	the completions and the likelihood draw recentred, and the PCA fitted without held-out cells, before any code for either existed, one check of the first amended after a test run of its code and before any comparison	pre-registered
Sep 17	a split against the completed baselines, and each perturbation dropped in turn, from the existing floor runs	post hoc
Sep 19	full-pipeline null calibration, controlled metric sensitivity and disjoint-cell diagnostic transfer, fixed before outcomes on previously inspected data	pre-registered
Sep 20	energy rescoring of saved predictions, full-axis gene selection and measured-cell error controls, before outcomes	pre-registered
Sep 20	the continuous association of calibration and validation margins, motivated by the disjoint-cell outcomes	post hoc
Sep 24	the calibration margin against gene count and cell number on two screens whose cells no analysis had used, with the development fit frozen	pre-registered
Sep 24	the same margin against a reproduced entrant's advantage over an independent floor, on twelve saved-prediction runs	pre-registered
Sep 24	a quadratic cell-count comparator for Q, specified after Q's outcome and before its own	pre-registered

column. The pre-registered arm takes s^2 as CPA's median variance ratio at 100 genes and δ as its median relative mean error there, $s = 0.2616$ and $\delta = 0.1247$. Before the surrogate's outcome was computed, the arm at $s = 1$ and $\delta = \bar{\Delta}/\bar{b}$ reproduced every split-averaged floor to a relative error of 1.7×10^{-15}. Its eight shares appear in the surrogate column of Table B.3. Its rule holds, with a Spearman correlation of 0.94 with the median ratio against the registered threshold of 0.8 and a rise of 86.8 points against half of CPA's 93.5. The arm has the truth's residuals, not CPA's, so it shows that the ratio can carry the rise at CPA's spread and mean error, not how the two terms divide CPA's own rise.

Seeds 2 and 3. The same replay uses each seed's own trained CPA, with a further check requiring the replayed CPA subsample to reproduce the stored entrant values. On the largest screen, the median ratio runs from 0.69 at 10 genes to 1.93 at 5,000 in the second seed and from 0.66 to 1.93 in the third. CPA's share below the floor rises from 7.7% and 8.4% to 100%. Both rules hold in both seeds at the same correlations as the first. At 100 genes on the other four screens, the ratios in the second and third seeds are 1.46 and 1.45 on RPE1, 0.34 and 0.37 on Replogle-8, 1.19 and 1.19 on Tian activation and 1.24 and 1.23 on Tian inhibition. Replogle-8 is again the only screen with ρ below one, where neither CPA nor the repeated mean beats the floor.

Table B.2: **Every pre-specified rule holds except the below-floor rule for CPA on two screens and on PerturBench's registered genes, the tightness correlation, the ten-gene transfer and Q for CPA on K562.** Each rule is a threshold on a descriptive statistic, fixed with the statements its outcomes would support before the output it judges existed (Table B.1). None is a test with a controlled error rate, and a rule required of every seed is a conjunction over seeds.

registration	rule	outcome
entrants	reproduction check: ordering agreement with the published entrant against the linear model at least 80%, mean of seeds	passes for every entrant and screen
	under-dispersed: truth's spread over the entrant's above 1.5, tighter on at least 90%	yes for scouter, scGPT and CPA on all five screens
	beats the floor: below it on more than half, paired interval below zero, on every seed	yes for scouter, scGPT and CPA on K562 and both Tian screens, no for CPA on RPE1 and Replogle-8
A	both labels unchanged across model seeds on the first split	holds
B	PerturBench's CPA under-dispersed, and beats the floor	yes, and no on the pre-registered top-100 genes (interval includes zero)
C	Spearman of tightness with Wasserstein rank at most -0.5; the truth sample first under both energy statistics	fails (-0.48); holds
D	Spearman of gene count with ρ at least 0.9; of ρ with CPA's share at least 0.8; ρ and the repeated mean's share larger on all genes than in the PCA	holds (1.00); holds (0.99); holds
E	the surrogate's Spearman with ρ at least 0.8, and its rise at least half of CPA's	holds (0.94; 86.8 against 46.8)
F	D's two gene-count rules on seeds 2 and 3 separately	both hold on both (1.00 and 0.99)
G	the completion recentred to the predicted mean leaves an interval including zero against CPA at CPA's spread and still trails it at the control's; the likelihood draw recentred to its locations' mean above the floor on at least 90%, interval above zero	both hold
H	D's space rule with the PCA fitted on the training partition	holds
K	a flagged subset of at least ten targets with its validation interval above zero; a gain over the dimension-only rule only with its interval above zero	holds in all six contexts at 100 and 1,000 genes, not at ten; gain in 1 of 18 settings
P	the calibration margin lowers a frozen gene-count and cell-number model's error, interval above zero, on each of two screens	holds on both
Q	the calibration margin lowers the out-of-fold error of an entrant's margin below the floor, interval above zero, on all three seeds of an entrant and screen	holds for CPA on RPE1, scGPT and scouter; not for CPA on K562 (third seed inconclusive)
R	against a quadratic cell-count model, the margin lowers the pooled out-of-fold error by more than 5%, interval lower bound	holds (10.4%, lower bound 6.8%)

Two ways to count the repeated mean. The pre-registration counts the repeated mean split by split, because Corollary 1 holds for each draw, while CPA's column averages each perturbation's three splits first. Using CPA's averaging rule, the repeated mean beats the floor on 0.6, 16.5, 87.7, 97.4, 99.7, 100, 100 and 100% of perturbations over the eight gene counts. The shares are 97.4, 93.1, 0.0, 94.7 and 100% on the five screens, and 55.6, 100 and 0% of combinations in the second benchmark's three spaces. The order over gene counts and over spaces is the same either way, and Replogle-8 is the one screen at zero under both.

C PREDICTORS IN FULL

This appendix describes the datasets and every predictor behind Sections 4.2, 4.3 and 4.7, including where our reproductions depart from the benchmark's scripts.

Table B.3: **The floor ratio and CPA's share below the floor rise together as more genes are scored.** Largest screen. Seed 1, 310 held-out perturbations, exact estimator. The floor ratio ρ is the split-half floor over the held-out variance, and $s^\star$ is the spread at which a shrunk copy of the floor third scores best. Both are formed per split, averaged over the three splits, then summarised by the median over perturbations. The repeated mean column, as pre-registered, is the share of splits, over every perturbation and its three splits, in which the floor third's own mean, repeated, scores below the floor. By Corollary 1 this occurs where that split's floor exceeds $b + \Delta$. CPA's column is the share of perturbations whose split-averaged distance is below the split-averaged floor. The surrogate column is pre-registration E's arm, the floor third's residuals at spread $s = 0.2616$ about a mean at relative error $\delta = 0.1247$, CPA's 100-gene values. It is scored as $\delta\bar{b} + \bar{a}s^2 + \bar{b} - 2\bar{c}s$ against the split-averaged floor and counted as CPA's column is. CPA's mean error is its squared mean distance to the reference third over b. The last row scores every gene of the 5,000 the benchmark's released ranking keeps for each perturbation. The Spearman correlation is 1.00 between gene count and ρ and 0.99 between ρ and CPA's share.

genes	median ρ	repeated mean below (%)	median $s^\star$	CPA below (%)	surrogate below (%)	CPA mean error / b
10	0.65	3.5	0.71	6.5	13.2	0.327
20	0.92	21.3	0.56	27.4	56.8	0.282
50	1.29	87.1	0.38	72.9	95.2	0.186
100	1.50	97.4	0.27	85.5	99.7	0.125
200	1.65	99.5	0.19	95.2	100.0	0.085
500	1.78	100.0	0.13	99.4	100.0	0.062
1,000	1.85	100.0	0.10	100.0	100.0	0.050
5,000 (all ranked)	1.92	100.0	0.06	100.0	100.0	0.038

Table B.4: **Replogle-8 is the one screen whose median floor ratio is below one, and the one where CPA never beats the floor.** The five screens at the benchmark's 100 genes. Seed 1, exact estimator, columns as in Table B.3. This comparison is post hoc and has no decision rule.

screen	n	median ρ	repeated mean below (%)	median $s^\star$	CPA below (%)	CPA mean error / b
K562	310	1.50	97.4	0.27	85.5	0.125
RPE1	232	1.48	93.1	0.29	69.8	0.236
Replogle-8	12	0.51	0.0	0.75	0.0	0.792
Tian act.	19	1.21	93.0	0.41	94.7	0.024
Tian inh.	36	1.19	89.8	0.42	100.0	0.028

Dataset sources and scope. Of the five screens, two come from a genome-scale Perturb-seq study (Replogle et al., 2022), one from an earlier study with direct guide capture (Replogle et al., 2020), and two from CRISPR interference and activation screens in human neurons (Tian et al., 2019; 2021). The genetic track's other four screens, from Adamson et al. (2016), the same earlier study and two multimodal screens (Frangieh et al., 2021; Papalexi et al., 2021), are not used because their splits do not reproduce (Appendix J). These screens build on pooled single-cell CRISPR platforms (Dixit et al., 2016; Datlinger et al., 2017), and the second benchmark's Norman19 is the combinatorial screen of Norman et al. (2019).

CPA, the cell-emitting entrant. CPA was trained and run on the same split using the benchmark's released configuration, with every hyperparameter and the training schedule transcribed from its released implementation. Its prediction route replaces each held-out cell by a bootstrapped control cell, encodes that cell's basal state, adds the perturbation's embedding and decodes. The benchmark writes the location of the decoder's Gaussian likelihood as the predicted cell. One input is substituted. The released implementation expects external gene embeddings that were not released, so we extract them from the public scGPT whole-human checkpoint (Cui et al., 2024). They have 512 dimensions, as in the benchmark's run, with 1,615 of 1,616 perturbation keys found where its log reports 1,617. The values cannot be identical to theirs. The embedding enters through the perturbation's shift, which every cell of that perturbation shares. The decoder is nonlinear, so a different shift can change the spread as well as the mean. The substitution's effect cannot be tested without

Table B.5: **The three spaces order the floor ratio as the truth sample's Wasserstein rank orders them.** PerturBench's Norman19, 45 held-out combinations, no combination excluded. The ratio and the preferred spread use the exact estimator. The last column is the truth sample's mean rank among nine arms under the benchmark's Wasserstein call, from Table 2. The decision rule compares all genes with the PCA, two spaces added after the pre-registered top-100 result was read, and the top-100 space is reported beside them. The largest relative error of the identity on any direct solve, in any run, is 4.1×10^{-14}.

space	n	median ρ	repeated mean below (%)	median $s^\star$	truth sample's rank
top-100 discovery genes	45	1.07	57.8	0.48	6.40
all genes	45	1.79	100.0	0.12	9.00
30-component PCA	45	0.66	0.0	0.69	2.82

the original file. Two observations bear on it. Drawing from the same trained model's likelihood multiplies the spread by five, so most of the tightness comes from writing the location, and PerturBench's own CPA, trained with its own code and inputs, is 84.9 times tighter than the truth on all genes (Appendix G). For the counterfactual of Section 4.3 the trained model is reloaded and first run through the unchanged path, which reproduces the saved predictions to 3.3×10^{-6} over all 313,776,690 values. The one change is a draw from the decoder's Gaussian in place of its location, under a fixed seed, with the same bootstrapped control cells.

scouter, a second cell-emitting entrant. The scouter model was trained and run from the benchmark's released script on the same split, with the public package (Zhu & Li, 2026) and the public release of the language-model gene embeddings named in the script (Chen & Zou, 2025). Training runs forty epochs under the package's early stopping, and prediction then emits 500 cells per held-out perturbation, each a control cell drawn with replacement and passed through a network conditioned on the perturbed gene's embedding. Six perturbed genes have no embedding and receive the script's uniform random vector. There are three departures from the released script. Each seed runs in its own process from the script's seed, where the script runs the three in one loop. The random vectors are assigned in sorted gene order, because the script's order depends on the interpreter's string hashing and differs from run to run. The first seed's final forward pass is taken 5,000 pairs at a time with the same control draws, which reproduces the package's single pass to zero difference. Against the linear model under the Wasserstein distance, our reproduction orders itself as the published scouter does on 99.1 to 100% of held-out perturbations over the three seeds, and under the energy distance and the symmetric KL divergence on all of them.

scGPT. We fine-tuned the public whole-human checkpoint on the same split, following the benchmark script after its perturbation tutorial. Training used fifteen epochs, an initial learning rate of 10^{-4} multiplied by 0.9 after each epoch, batches of 64, 1,536 sampled genes per training cell, and mixed precision. Prediction used all genes. We selected the checkpoint by validation Pearson correlation. Each prediction is the network's output for a control cell paired by the data loader with one held-out cell. Predictions and truth therefore have the same number of cells, but no held-out expression enters the prediction. The script requests a fast attention library that our environment lacks, so the model's constructor used standard attention, which computes the same operation with a different kernel. This fallback changes the names of pretrained attention weights. We applied the package's own renaming rule for this case and required every pretrained transformer weight to load. We supplied the frozen split from our GEARS run. The older GEARS release bundled with scGPT retains two perturbations without ontology entries and would otherwise draw a different split.

We also departed from the released script in three implementation details. We ran each seed in a separate process initialised with that seed. The released script runs all three in one loop, so later seeds inherit the previous run's random state. At prediction, we reused the control pairing from training. The released script rebuilds the dataset and draws a new pairing. Both are random draws of control inputs. We read each predicted cell's perturbation label from its batch rather than inferring it from row order. We checked the labels against the truth cell by cell.

The five mean predictors. The linear model, the training mean, the control mean, the regression baseline and retrieval each emit one mean expression vector per held-out perturbation, which the harness then completes and scores as it does every mean. Each is fitted on the training perturbations only. No held-out expression enters any of them. Gene embeddings come from a public pretrained checkpoint, descriptors from the gene ontology, and every fitted quantity from training perturbations. The scoring genes are ranked on held-out cells, as the harness ranks them, which defines the evaluation rather than feeding a predictor.

Linear model. The benchmark's own linear method, ported from its released code and checked against it. The first ten principal components of the training conditions' pseudobulk give each gene a ten-dimensional embedding. A perturbation is represented by the embedding of the gene it targets. The map from that representation to the response is solved as a two-sided ridge with penalty 0.1 on each side. A perturbation whose gene is not among the measured genes returns a missing prediction, which the harness's own handler converts to the control mean at scoring time. The reported scores keep this behaviour.

Training mean. The mean expression over every cell of the training perturbations, so a perturbation with more cells weighs more. No fitting.

Control mean. The mean of the unperturbed control population, that is, a prediction of no change. This is the arm whose Pearson distance is undefined (Appendix I).

Regression baseline. The benchmark's regression baseline, which regresses each training perturbation's mean expression on the embedding of its target gene with ridge regression at scikit-learn's default penalty, and gives a gene without an embedding the mean of all embeddings. The embedding file the authors used is not released, so we extract ours from the public scGPT whole-human checkpoint, as for CPA.

Retrieval. Training-free. Each perturbation is described by its gene-ontology term membership, and similarity between two perturbations is the cosine between their descriptor vectors. For a held-out perturbation we take its five most similar training perturbations and weight them by a softmax over their similarities at temperature 0.05, then predict the weighted mean of their measured responses. Both values were fixed before the first run and are the same in every run reported here. The same five neighbours and weights supply the neighbour-residual generation rule of Appendix I, so that rule is computed alongside this predictor and then applied identically to every predictor in the crossing.

D EVERY REPRODUCED ENTRANT, SCREEN BY SCREEN

This appendix gives every reproduced entrant's tightness and below-floor share, screen by screen and seed by seed, behind Section 4.3 and Table 1. Each entrant is scored against the same thirds of the same held-out populations as the entrant beside it, so the floor rows are shared and only the entrant arm differs. The times-tighter column is the truth's mean per-gene standard deviation over the entrant's, on each perturbation's released scoring genes. The last column is the reproduction check, the share of perturbations on which our run orders itself against the linear baseline under the Wasserstein distance as the published entrant does.

Thirds and eligible perturbations. Every held-out population of at least 60 cells is permuted three times by one generator seeded once per run, and each permutation is cut into three parts of $\lfloor n/3 \rfloor$ cells, a discovery third, the floor third and the reference third. The discovery third ranks the same candidate genes the benchmark ranks for that perturbation, by the absolute Welch t statistic against the control cells, and keeps the top 100. Ties hold the archive's gene order, and a gene whose pooled standard error is zero takes a statistic of zero, so it ranks below every gene that varies. That candidate list is the benchmark's own released ranking for the perturbation, computed on the whole population and truncated at 5,000 genes. The discovery third therefore separates the choice of the top 100 within the list from the scored cells, but not the choice of the list itself. The list is a symmetric function of every cell of the population, so the floor third, the reference third and a truth

sample drawn as another third of it remain exchangeable. An independent sample of the same law need not be exchangeable with the floor after this filtering, which Appendix K measures. Scoring the same thirds on the released top 100 instead, which is the fully dependent choice, puts CPA below the floor on 85.2% rather than 88.7%, and the two sets of 100 genes share a median of 25. An entrant that writes one cell per held-out cell is subsampled to a third by a permutation from the same stream, and one that writes a fixed number of cells by a second generator. Each harness-completed baseline is drawn twice, once to be scored and once for its self-term, and every score is averaged over a perturbation's three splits before it is compared with the floor. The 60-cell rule removes 13, 11 and 12 of the largest screen's 323 held-out perturbations in the three seeds, 34, 34 and 28 of RPE1's 266 and none on the three small screens, which accounts for every difference between those counts and the n of Table D.2. scouter writes 500 cells per perturbation and is scored only where two thirds fit in them. This excludes one perturbation of 992 cells from its third seed, so its n there is 310 where CPA's and scGPT's is 311. Table 1 ranks the arms on the perturbations all of them have.

Table D.1: **Every reproduced entrant is tighter than the truth. Each is below the Wasserstein floor on most perturbations of every screen but Replogle-8, and below the energy floor on none.** Every entrant that writes its own cells, on the same thirds. Times tighter divides the truth's mean per-gene standard deviation by the entrant's, each averaged over held-out perturbations on their released top-100 genes, and the column gives its range over three seeds. Tighter on is the share of held-out perturbations on which the entrant's spread is smaller. Below the floor is the share of perturbations, per seed, on which the entrant's Wasserstein distance is lower than a split-half sample of the truth's with scoring genes chosen on the discovery third, and the paired difference beside it is its range over seeds. A dagger marks a seed whose 95% bootstrap interval for that difference includes zero, so CPA on RPE1 fails the pre-registered rule. The last column is the share of perturbations on which our reproduction orders itself against the linear model under the Wasserstein distance as the published entrant does.

Entrant	Screen	Times tighter	Tighter on (%)	Below floor, W (%)	W − floor	Below floor, energy (%)	Published order (%)
CPA	K562	3.36–3.43	100	88.7 / 92.6 / 88.7	-4.98–-4.45	0.0	98.5–100.0
CPA	RPE1	1.94–2.09	94–96	71.6† / 75.9 / 71.0	-1.16–-0.61	0.0	84.6–89.1
CPA	Replogle-8	5.36–8.17	100	0.0 / 0.0 / 0.0	$+18.45$–$+21.77$	0.0	91.7–100.0
CPA	Tian act.	5.98–6.76	100	94.7 / 89.5 / 94.7	-1.93–-1.64	0.0	100.0
CPA	Tian inh.	5.99–6.95	100	100.0 / 100.0 / 100.0	-1.64–-1.53	0.0	100.0
scGPT	K562	21.51–40.51	100	93.2 / 94.9 / 92.6	-5.43–-5.10	0.0	100.0
scouter	K562	15.79–18.26	100	94.2 / 95.5 / 94.5	-5.32–-4.94	0.0	99.1–100.0

Reading the two tables. Table D.1 summarises each entrant over its seeds and Table D.2 gives every seed's paired differences with their intervals. The seeds are separate estimates on the same perturbation units and are never pooled. scGPT orders itself against the linear baseline as the published scGPT does on all 323 perturbations in every seed, and on the 310, 312 and 311 that enter the floor analysis it is below the floor under the Wasserstein distance and above it under the energy distance, with every interval excluding zero. The reproduction check is agreement in that pairwise ordering, not equality with the published scores. CPA's first seed on RPE1 is the one case whose Wasserstein interval includes zero, so that screen fails the pre-registered requirement that every seed beat the floor. Ranked per perturbation as in Table 1, among five arms (a sample of the truth, CPA and the three completed baselines), CPA's mean rank over three seeds under the Wasserstein and energy distances is 1.30 and 4.96 on RPE1, 2.00 and 5.00 on Replogle-8, 1.07 and 5.00 on Tian activation and 1.00 and 5.00 on Tian inhibition.

The split against a dispersed negative, and one perturbation at a time. Both analyses are post hoc and reuse the floor runs. Under the Wasserstein distance the floor third beats the harness-completed prediction of no change on 91.3 to 94.6% of the largest screen's perturbations over the three seeds and on 72.2 to 100% of the other four screens', and under the energy distance on 89.5 to 100% of every screen's. A check that sets a split of the measured cells against a dispersed negative, as Schäfer et al. (2026) do with a pooled population of perturbed cells, therefore passes both distances, while CPA, scouter and scGPT beat the same split under the Wasserstein distance on 88.7 to 95.5% of the largest screen's perturbations and never under the energy distance. Dropping

Table D.2: **The Wasserstein difference favours the entrant in every combination but the three on Replogle-8, while the energy difference favours the floor in all of them.** Every seed's paired difference from the floor, with its interval. Entrant minus the discovery-third floor under the Wasserstein distance and the energy distance, with 95% percentile bootstrap intervals over the n perturbations that enter, which the two distances share. Negative values favour the entrant. Seeds are not pooled. All 21 combinations of entrant, screen and seed are shown.

Entrant	Screen	Seed	n	Wasserstein difference [95% CI]	Energy difference [95% CI]
CPA	K562	1	310	-4.449 [-5.050, -3.754]	$+1.662$ [$+1.562$, $+1.773$]
CPA	K562	2	312	-4.981 [-5.469, -4.405]	$+1.572$ [$+1.487$, $+1.663$]
CPA	K562	3	311	-4.522 [-5.094, -3.910]	$+1.669$ [$+1.564$, $+1.780$]
scGPT	K562	1	310	-5.121 [-5.631, -4.538]	$+2.400$ [$+2.290$, $+2.521$]
scGPT	K562	2	312	-5.433 [-5.850, -4.934]	$+2.334$ [$+2.237$, $+2.436$]
scGPT	K562	3	311	-5.102 [-5.568, -4.602]	$+2.343$ [$+2.228$, $+2.461$]
scouter	K562	1	310	-4.935 [-5.429, -4.371]	$+2.526$ [$+2.419$, $+2.641$]
scouter	K562	2	312	-5.324 [-5.748, -4.814]	$+2.388$ [$+2.291$, $+2.492$]
scouter	K562	3	310	-5.026 [-5.516, -4.491]	$+2.480$ [$+2.365$, $+2.601$]
CPA	RPE1	1	232	-0.613 [-1.645, $+0.468$]	$+1.677$ [$+1.544$, $+1.815$]
CPA	RPE1	2	232	-1.163 [-2.159, -0.062]	$+1.669$ [$+1.533$, $+1.819$]
CPA	RPE1	3	238	-1.164 [-2.077, -0.211]	$+1.599$ [$+1.482$, $+1.721$]
CPA	Replogle-8	1	12	$+18.453$ [$+15.840$, $+21.558$]	$+4.593$ [$+3.889$, $+5.502$]
CPA	Replogle-8	2	12	$+19.711$ [$+17.302$, $+22.620$]	$+4.616$ [$+3.967$, $+5.485$]
CPA	Replogle-8	3	12	$+21.771$ [$+19.013$, $+24.528$]	$+4.902$ [$+4.090$, $+5.817$]
CPA	Tian act.	1	19	-1.926 [-2.431, -1.416]	$+1.073$ [$+0.913$, $+1.289$]
CPA	Tian act.	2	19	-1.640 [-2.684, -0.269]	$+1.216$ [$+0.906$, $+1.666$]
CPA	Tian act.	3	19	-1.661 [-2.037, -1.260]	$+1.017$ [$+0.942$, $+1.101$]
CPA	Tian inh.	1	36	-1.534 [-1.696, -1.374]	$+0.719$ [$+0.679$, $+0.760$]
CPA	Tian inh.	2	36	-1.568 [-1.742, -1.398]	$+0.808$ [$+0.770$, $+0.849$]
CPA	Tian inh.	3	36	-1.644 [-1.823, -1.480]	$+0.772$ [$+0.717$, $+0.847$]

any one perturbation in turn leaves CPA's pre-registered label, beating the floor or not, unchanged on every screen and seed, the three small ones included, and the median paired difference has the sign of the mean in every one of the fifteen cases.

Retraining on one split. The three seeds above change the split and the model together. On the first split alone, CPA retrained with its original model seed reproduces the reported predictions exactly, so its GPU training is deterministic here. Model seeds 1118 and 1119 give populations 3.40 and 3.33 times tighter than the truth, tighter on every perturbation. They are below the floor on 88.4% and 88.7% with intervals $[-5.11, -3.82]$ and $[-5.06, -3.77]$, and below the energy floor on none. Training moves these numbers less than the split does.

E Dispersion with the mean removed

This appendix supports the statement of Section 4.3 that the entrants' tightness survives removing every population's mean.

Every population is centred on its own mean before scoring, the held-out truth on the truth's mean, the entrant's cells on the entrant's mean, and each residual block on its own mean, so the harness's Gaussian and the two residual rules enter as zero-mean blocks. Every metric then sees dispersion and higher moments only, and the two mean-only metrics are identically zero. The spread column divides each arm's mean per-gene standard deviation by the truth's for each perturbation, then averages that ratio over the 323 perturbations of the first seed (Table E.1).

F Recomputations on the published table

This appendix supports Section 4.2 with recomputations of the benchmark's published ranks and the agreement of our CPA with the published one.

Every analysis here uses the benchmark's released per-perturbation results for the genetic single-perturbation track, nine datasets and fifteen methods, and its own published per-metric ranks. None needs the score transform that Appendix J says is missing, because ranks are published directly.

Leave one metric out. Table F.2 recomputes the ranking with each of the six ranked metrics removed in turn, using the benchmark's own aggregation, which is also the one its own published analysis uses, a pooled mean over every perturbation-by-seed row of each method's rank. Without the Wasserstein distance CPA falls to fourth, and a plain average of the metric ranks without re-ranking also puts it fourth. The pooled first place comes from the two largest screens, which hold 81% of the rows. CPA is first on each of them, and without the Wasserstein distance it falls to third on K562 and holds first on RPE1 (Table F.1). On the other seven datasets it ranks between third and eleventh with all six metrics.

Output representation against Wasserstein rank. Methods are classified by whether their released implementation writes a per-cell matrix or hands a mean to the harness's synthesis routine. The six that hand over a mean hold Wasserstein ranks 8.9, 9.4, 9.8, 10.7, 11.6 and 12.6. The three best ranks, CPA's 1.6, scGPT's 2.2 and scouter's 2.5, belong to methods that write cells, eight in all. The fifteenth, biolord, has no genetic-track script in the released repository and is left unclassified. The grouping is not clean in the other direction. One method writing its own cells ranks 12.3, and on the energy distance the two groups invert, 9.3 against 6.7. The grouping is descriptive and does not identify an effect of the output representation.

Metric ranks of the three entrants. CPA, scGPT and scouter rank 10.1, 13.5 and 13.2 of fifteen on the energy distance. That is the inversion Section 4.3 measures on CPA's own populations, where the Wasserstein distance prefers it and the energy distance puts it last. No entrant's cell-level output is released, so the published ranks alone cannot establish that mechanism in the original predictions, and Appendix D measures dispersion in our reproductions of all three.

Agreement with the published CPA. The benchmark's per-perturbation rank columns let the reproduction be checked without its missing score transform, because a transform applied within a perturbation keeps the order of two methods there. On the largest screen, our CPA orders itself against the linear model as the published CPA does on 98.5 to 100% of perturbations under the Wasserstein distance, 99.7 to 100% under the symmetric KL divergence and 84.8 to 89.5% under the energy distance, across the three seeds. Under the two mean-level metrics it agrees on only 30.3 to 37.8%. The published CPA beats the linear model's mean on 96.9 to 99.4% of perturbations and ours on 28.5 to 37.5%. The dispersion-driven orderings are reproduced, but the mean-level orderings are not. That is agreement in order, not a measurement of the published CPA's dispersion, whose cells were never released. Our CPA's mean-level orderings are worse than the published one's, plausibly through the substituted embedding. By Proposition 1 the mean enters the exact score as an added $\|m - \bar{y}\|^2$, so a better mean with the same spread and alignment would lower CPA's score and put it below the floor more often. The published cells are unavailable, so this is an inference for the exact estimator rather than a measurement. Across all nine datasets the published CPA beats the

Table E.1: **After centring, the entrant that wins on the Wasserstein distance is the worst of the four on the other two distributional metrics.** Centred populations scored against the centred truth on the largest screen, 323 held-out perturbations. The three metric columns are means over three seeds, without intervals, and the entrant's row is in bold. A mean-only metric is identically zero here, so only dispersion and higher moments remain. The final column gives each arm's per-gene standard deviation as a fraction of the truth's, per perturbation, averaged over the first seed. The entrant is the only arm that does not reproduce the spread, and the only one the Wasserstein distance prefers. Section 4.3.

centred population	energy dist.	sym. KL	Wasserstein	spread vs truth
the ranked entrant	**1.106**	**8.111**	**17.908**	**0.29**
harness Gaussian	0.091	6.331	28.636	1.02
real control cells	0.091	6.333	28.018	1.02
real neighbour cells	0.077	6.279	28.790	1.07

Table F.1: **Dropping the Wasserstein distance costs CPA first place on K562 and not on RPE1.** Its place on each dataset, with all six metrics and without that distance. The benchmark's aggregation within each dataset. Rows are perturbation-by-seed entries in the released file. The pooled first place of Table F.2 comes from the two largest screens, which hold most of the rows.

		CPA's place		winner			CPA's place		winner
dataset	rows	all six	no Wass.	no Wass.	dataset	rows	all six	no Wass.	no Wass.
K562	969	1	3	GenePert	Tian act.	57	4	8	training mean
RPE1	798	1	1	CPA	Adamson	45	11	12	GenePert
Tian inh.	108	9	11	GenePert	Replogle-8	36	9	13	training mean
Frangieh	93	3	8	scGPT	Papalexi	15	4	7	scELMo
Replogle-7	60	11	15	GenePert					

Table F.2: **CPA holds pooled first place unless the Wasserstein distance is dropped, when it falls to fourth.** Exploratory recomputation of the benchmark's own published ranks, leaving one ranked metric out. The benchmark's aggregation, reproduced exactly on 32,651 of its 32,715 rows. Per perturbation it averages the metric ranks and ranks the fifteen methods on that average, then averages over perturbations. The other 64 rows, all on K562 and RPE1, are second-to-last and last places that the published table records as 27th and 28th of fifteen. DE, differentially expressed. Lower is better. The last column is CPA's position. In 2,000 resamples of the 1,781 perturbations within each dataset, 780 behind K562's 969 rows and 655 behind RPE1's 798, each carrying its rows from every seed and a tie taking the better place, CPA is first with all six metrics in 100% and, with the Wasserstein distance dropped, first in none, second in 7.1%, third in 31.0% and fourth in 61.9%.

metric dropped	first	second	third	CPA
none	CPA (4.20)	GenePert (5.30)	biolord (6.04)	1
Pearson	CPA (4.96)	GenePert (5.16)	biolord (5.87)	1
energy distance	CPA (2.92)	scGPT (4.95)	scouter (5.71)	1
mean squared error	CPA (4.72)	GenePert (5.20)	biolord (5.94)	1
DE-gene overlap	CPA (3.97)	GenePert (5.50)	biolord (6.19)	1
symmetric KL	CPA (3.68)	GenePert (5.58)	scGPT (6.06)	1
Wasserstein	GenePert (4.60)	biolord (5.41)	linear model (5.46)	4

linear model under the Wasserstein distance on 71.7 to 100% of perturbations and under the energy distance on 0 to 31.2%.

G THE SECOND BENCHMARK IN FULL

This appendix gives the runs, the agreement with published scores and the failure of the pre-registered gene rule behind Section 4.6 and Table 2.

Runs on Norman19. PerturBench's processed Norman19 holds 91,168 cells over 5,575 genes in 237 conditions, 131 combinations, 105 single perturbations and the control. Its combination-prediction splitter, as configured, holds out 46 combinations in the test partition, 45 of them with at least sixty cells, and every run saved the same split. Each of the eight tuned configurations was trained once with the configured seed, 245, by PerturBench's own code, and its per-cell predictions were kept. For each held-out combination the model maps 1,000 sampled control cells. There are two departures, both recorded before any output was read. First, the released decoder-only model cannot run with its covariate switched off because its input width still counts the covariate. It was therefore run with that covariate on. On a screen with one cell type, this adds a constant column to a first layer that already has a bias. Second, five models were evaluated with three of PerturBench's pipelines rather than all of them, omitting all-gene pairwise rankings that no analysis here uses.

Agreement with published scores. Table G.1 sets our runs beside PerturBench's published means on its own metrics. Seven models score close to them. Our linear additive model is worse on logFC

Table G.1: **Every model is below the all-gene Wasserstein floor on every combination and below the all-gene energy floor on none.** Share of held-out combinations below the split-half floor under the Wasserstein distance, with the 95% bootstrap interval of the paired difference, in the three spaces, and under PerturBench's energy statistic on all genes. The last three columns compare our single run with PerturBench's published means over five seeds on its own metrics, logFC cosine, its PCA energy score and mean RMSE. The asterisk marks the departure noted in Table 2.

	Below the Wasserstein floor (%) [paired interval]			Energy	Ours / published		
Model	Top 100	All genes	PCA 30	(%)	cosine	PCA energy	RMSE
Linear additive	53 $[+0.3, +2.2]$	100 $[-320.3, -295.0]$	7 $[+26.8, +46.7]$	0	0.30 / 0.60	1.91 / 1.20	0.088 / 0.057
biolord	53 $[+0.1, +1.9]$	100 $[-353.5, -329.5]$	4 $[+30.9, +54.8]$	0	0.39 / 0.41	1.66 / 1.60	0.086 / 0.086
SAMS-VAE, modified	78 $[-1.2, -0.5]$	100 $[-438.5, -402.5]$	62 $[-3.7, +6.5]$	0	0.79 / 0.78	0.88 / 0.74	0.053 / 0.047
SAMS-VAE	67 $[-0.4, +1.5]$	100 $[-394.4, -370.7]$	9 $[+27.2, +52.1]$	0	0.49 / 0.45	1.94 / 1.90	0.086 / 0.084
CPA, no adversary	73 $[-0.9, -0.1]$	100 $[-436.5, -406.5]$	51 $[+1.7, +12.7]$	0	0.77 / 0.77	2.06 / 2.20	0.050 / 0.046
CPA	76 $[-0.9, -0.2]$	100 $[-438.7, -406.7]$	44 $[+1.3, +10.3]$	0	0.77 / 0.76	2.16 / 2.20	0.046 / 0.046
Latent additive	71 $[-0.9, -0.2]$	100 $[-431.9, -399.8]$	2 $[+12.3, +20.6]$	0	0.81 / 0.79	3.17 / 3.20	0.041 / 0.043
Decoder only*	69 $[-0.9, -0.1]$	100 $[-429.7, -397.5]$	2 $[+12.8, +22.2]$	0	0.68 / 0.73	3.22 / 3.20	0.046 / 0.043

cosine, 0.30 against 0.60, and on mean RMSE, 0.088 against 0.057, so its row in Table 2 describes our run more than the published model. Dropping that row changes neither reading, because on all genes the truth sample is ranked ninth of nine on every combination, and in the PCA the only arm ahead of it is the modified SAMS-VAE.

PerturBench's distributional score. Its configuration calls the score MMD, and its code computes geomloss's energy loss, half the energy distance with the diagonal included. A reimplementation agrees with PerturBench's own function to a relative error of 1.1×10^{-4} at three dimensionalities, and the analysis uses it on all genes and in the PCA.

The PCA, and a refit without held-out cells. The PCA is PerturBench's own projection, 30 components fitted on every cell of the test partition and centred on their mean, with the truth and every prediction projected alike. Both scored thirds of every combination are among the cells it is fitted on and the predictions are not, so the two thirds stay exchangeable with each other but not with cells outside the fit. Each third is at most 2.5% of the fitted cells. Pre-registered as H, the projection was refitted on the training partition, which holds no held-out cell. The median floor ratio is 0.65 against 0.66, and the repeated mean beats the floor on no combination in either. The truth sample ranks second of nine in both, at 2.89 against 2.82, so the PCA result does not depend on the scored cells entering the projection.

Failure of the pre-registered genes. Ranking genes by absolute Welch t against 9,702 control cells favours a gene the held-out cells never express and a few control cells do, whose standard error is tiny. On these populations, on average 46% of a combination's top 100 genes are constant in its measured cells, all 100 are on five of the 45, and on a discovery third of about thirty cells an average 93.7% of the discovery genes are constant. We report the pre-registered numbers as they are, and the two spaces without gene selection as sensitivity analyses, added after the pre-registered result was read and before they ran. The rules come out differently by space. The Spearman correlation between tightness, measured in each space, and Wasserstein mean rank is -0.48 on the top-100 genes, -0.52 on all genes and -0.36 in the PCA. The tightness rule, with a registered threshold of -0.5, fails on the top-100 genes and in the PCA and passes on all genes. The truth sample is first under the energy statistic in all three.

PerturBench's CPA verification run. PerturBench ships a configuration for checking its CPA against the original implementation, on the curated Norman19 release and its fixed split with 15 held-out conditions. Run with seed 245, its best checkpoint is at epoch 398. On its own metrics it scores a mean RMSE of 0.0255 and a logFC cosine of 0.771. Its cells are 61.7 times tighter than the truth on the top-100 genes, 84.9 times on all genes and 61.1 times in the PCA, tighter on every condition in each. Under the Wasserstein distance, it is below the split-half floor on 53.3% of conditions on the pre-registered top-100 genes, with an interval that includes zero, $[-1.01, +0.07]$. It is below the floor on all 15 on all genes, paired -134.1 $[-146.7, -123.5]$, and on 33.3% in the PCA, $[-1.55, +3.66]$. Under the energy distance and PerturBench's energy statistic it is below the

Table H.1: **Nineteen works that score predicted perturbation populations with a transport distance**, grouped by the space the distance is computed in. Each is counted only where its text or code applies the distance to predicted against measured cells. A Sinkhorn divergence subtracts the two self-terms, and an entropic optimal-transport (OT) cost does not. DE, differentially expressed. PCs, principal components.

Work	Kind	Transport quantity	Space
Multivariate, on genes			
Wei et al. (2026)	benchmark	entropic OT cost	top 100 DE genes
Heidari et al. (2026)	evaluation study	exact W_1	5,000 variable genes
Mao et al. (2026)	benchmark	Sinkhorn divergence	genes, and 50 PCs
Bunne et al. (2022)	method	entropic OT cost	50 marker genes
Uscidda & Cuturi (2023)	method	Sinkhorn divergence	50 marker genes
Driessen et al. (2026)	method	entropic OT cost	top 50 DE genes
Shi et al. (2026)	method	W_1, scPerturBench's protocol	top 100 DE genes
Multivariate, in a low-dimensional embedding or marker panel			
Eyring et al. (2024)	method	Sinkhorn divergence	30 PCs
Klein et al. (2025)	method	Sinkhorn divergence	10 to 100 PCs
Palla et al. (2026)	method	W_2	30 PCs
Atanackovic et al. (2025)	method	exact W_1 and W_2	43 markers
Pariset et al. (2023)	method	entropic OT cost	50 markers
Di Francesco et al. (2026)	method	W_1 and W_2	50 PCs
One dimension at a time, a gene or a random projection			
Sun et al. (2026)	method	sliced Wasserstein	6,000 variable genes
Li et al. (2026)	benchmark	per-gene W_1, averaged	all and top 20 DE genes
Lotfollahi et al. (2023)	method	per-gene Wasserstein	all and DE genes
Jiang et al. (2024)	method	per-gene W_1, summed	top 100 DE genes
Chi et al. (2026b)	method	per-gene W_1, averaged	all and top 20 or 40 DE genes
Chi et al. (2026a)	method	gene-level transport distance	all and top 20 or 40 DE genes

floor on none, in every space. Shrinking the floor third about its own mean by CPA's measured ratio puts it below the floor on 66.7% of conditions on the top-100 genes, $[-1.66, -0.12]$, so on this split spread alone moves the transport score further than CPA's own cells do. PerturBench's authors report that its CPA suffers mode collapse that its mean-level ranks miss (Wu et al., 2025). They measure it as a similarity between the responses predicted for different perturbations, where we measure the spread within one, so the two are related failures rather than one observation.

H Where transport distances already score predicted populations

This appendix lists the works counted in Section 2 in which a transport distance scores predicted populations.

A literature search in September 2026 over methods that predict single-cell perturbation responses and the benchmarks that rank them found nineteen works in which a Wasserstein-type distance scores predicted cells against measured ones (Table H.1). A work counts only where its text or code applies the distance to predictions, not where it is only a training loss, and works we could not open are left out. They sit in three regimes. Proposition 1 covers the unregularised squared distance. Read heuristically, it suggests a different exposure for each regime. The other estimators in the table would need their own checks to confirm this. Multivariate distances on genes are the regime of scPerturBench and of the all-gene analysis of Section 4.6, where independent residuals are nearly orthogonal, the floor ratio is 1.79 and the preference for tight predictions is strongest. Distances in tens of principal components or markers are the regime of that section's PCA analysis, where the floor ratio is 0.66 and the preference is weaker. A distance computed one gene, or one random projection, at a time pairs sorted values, so its alignment term is near its bound and the identity suggests little gain from shrinking. This is consistent with one benchmark in that regime reporting that compressed predictions score worse (Li et al., 2026). We make no claim about any listed method's results, only about which regime its evaluation sits in.

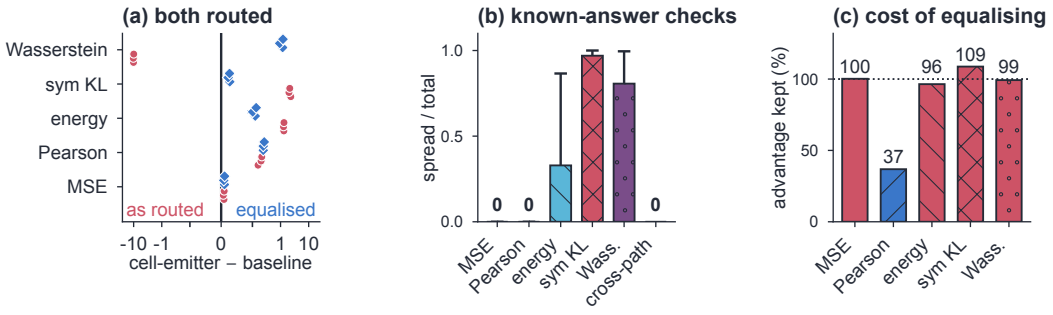


Figure I.1: **Routing a mean through the completion reverses one comparison, and the mean-exact rules leave the mean-only metrics almost unchanged.** Largest screen. **(a)** CPA minus a mean-emitting baseline, as routed and with both completed, per seed with bootstrap intervals. **(b)** Three rules emit the prediction's mean exactly, so a mean-only metric must not change across them. Bars are the median disagreement among the three, as a share of the metric's range over all arms, and whiskers its maximum. **(c)** Share of our own predictor's advantage over the linear baseline left when both get real cells.

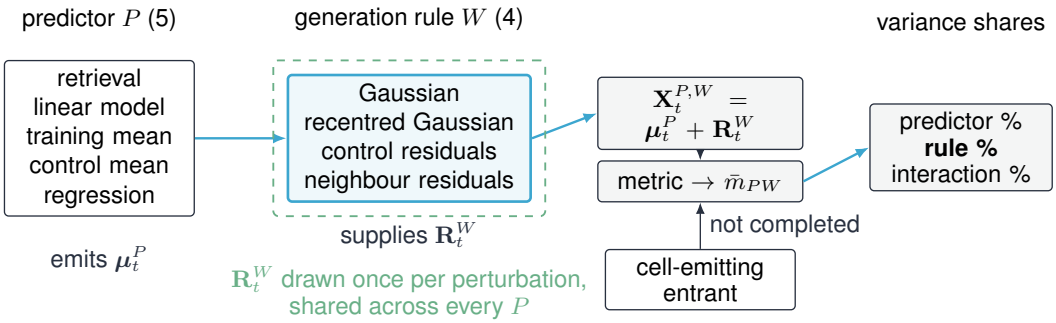


Figure I.2: **Sharing each residual draw across predictors prevents inter-sample noise from producing differences between predictors under one rule.** Twenty arms per screen and seed. $\mathbf{R}_t^W$ is drawn once per perturbation and shared across every predictor, so a difference between predictors under one rule cannot be inter-sample noise. This sharing does not extend across rules, and only the Gaussian and its recentred version share their draws exactly. The cell-emitting entrant reaches the metric directly, without completion, which is the asymmetry Section 4.7 measures.

I THE COMPLETION ANALYSIS IN FULL

This appendix gives the design, decomposition and every reordering behind the completion analysis of Section 4.7.

The four generation rules. Each rule supplies $\mathbf{R}_t^W$ and nothing else (Figure I.2). With $\mathbf{Z}_t \in \mathbb{R}^{n_t \times G}$ a standard normal block and $\boldsymbol{\sigma}$ the per-gene standard deviation of $\mathbf{C}$, the Gaussian rule sets $\mathbf{R}_t = \mathbf{Z}_t \operatorname{diag}(\boldsymbol{\sigma})$. This is the harness's own rule, and it makes three assumptions. Its spread is the unperturbed control's whatever the perturbation did, its genes are independent, and it is Gaussian on log-normalised counts. The recentred Gaussian uses *the same draws*, $\mathbf{R}_t = (\mathbf{Z}_t - \bar{\mathbf{Z}}_t)\operatorname{diag}(\boldsymbol{\sigma})$, so the emitted mean equals $\boldsymbol{\mu}_t^P$ exactly rather than up to an error of order $\boldsymbol{\sigma}/\sqrt{n_t}$. Control residuals bootstrap real control cells with their mean removed, giving real covariance but of unperturbed cells. Neighbour residuals use the real cells of the retrieval predictor's neighbours, each block centred and rescaled by $\sqrt{d/(d-1)}$ for d cells, then centred as a whole. That is real dispersion of *perturbed* cells.

Two design points make the crossing paired. $\mathbf{Z}_t$ and the two residual blocks are drawn *once per perturbation* and reused across every predictor, so a difference between predictors under one rule

is not inter-sample noise, though a rule's main effect still contains its one draw. And the Gaussian rule and its recentred version share their draws exactly, so a difference between them is the mean alignment and nothing else.

The five predictors, and the decomposition. The predictors differ only in $\boldsymbol{\mu}_t$. They are the benchmark's linear model, training mean, control mean and regression baseline, reimplemented from its released code, and a training-free retrieval predictor of our own that takes a weighted mean over neighbouring perturbations in a gene-ontology descriptor space (Appendix C). Four of the five are the benchmark's published baselines, so a reordering among them is a reordering of published methods. The fifth is our own predictor, scored under the same rules.

For a fixed metric and dataset, the twenty arm means $\bar{m}_{PW}$ form a two-way table. Writing $\bar{m}$ for the grand mean, α_P and β_W for the main effects and γ_{PW} for the interaction, the total spread splits as

$$\sum_{P,W} (\bar{m}_{PW} - \bar{m})^2 \;=\; |\mathcal{W}| \sum_{P} \alpha_P^2 \;+\; |\mathcal{P}| \sum_{W} \beta_W^2 \;+\; \sum_{P,W} \gamma_{PW}^2, \tag{1}$$

and we report each term as a share of the total, averaged over three seeds, with the interaction in full in Appendix I.6. The middle term, the completion's footprint, is not by itself a claim about rankings, for the reason given below.

A known-answer check that runs on the real data. Three of the four rules emit a population whose mean is exactly $\boldsymbol{\mu}_t^P$, so any metric that is a function of the group means alone must return the *same number* under all three, for every predictor, dataset and seed. That is checkable without a ground truth. The mean squared error satisfies it with no disagreement in 15 of 15 screen-and-seed combinations and the Pearson distance in 60 of 60 combinations of predictor, screen and seed, the control-mean arm excluded for the reason given below (Figure I.1(b)).

Limits of the footprint. The completion's share of the between-arm variance, from Equation 1, is a footprint, not a statement about any ordering. Under a rule applied uniformly to both arms of a comparison the rule's main effect cancels exactly, so a large share is compatible with a perfectly stable order, and our interaction term is at most 2.4% in any screen, metric and seed. The reason the footprint matters on this benchmark is that the rule is *not* applied uniformly, which is what Section 4.7 measures directly.

The rule exceeds the predictor in 8 of 25 screen-and-metric combinations. On the two large screens the rule's share on the symmetric Kullback-Leibler divergence is 62.6% and 76.5%. Without the neighbour-residual rule, which is ours rather than the benchmark's, those shares become 43.2% and 44.8% and both screens are predictor-dominated, so that reading rests on a rule we introduced. The absolute scale matters too. On the screen holding out twelve perturbations every arm scores between 40.39 and 40.90 on that divergence in the first seed, and the predictor differences sit in the fourth decimal, so a share there is a share of a spread that is itself small against the metric's level. The mean squared error and the energy distance leave the rule a small share on the two large screens, 0.6% and 2.9% on the largest. Which metrics measure the predictor depends on the screen, and the independent reimplementation of Appendix A covers three of the five, not the symmetric Kullback-Leibler divergence.

The share also depends on which arms are compared. Adding the benchmark's control-mean baseline collapses the rule's share on the Pearson distance on every screen, and on two of the five, from 66.6% and 71.3%, it carries them across from rule-dominated to predictor-dominated. The other three were predictor-dominated already, at 30.8%, 24.1% and 4.3%, so the reduction is large on four screens and two change which factor dominates. The cause is specific. The Pearson distance is computed on the change from the control mean. An arm predicting no change has a change of exactly zero, and its correlation with the truth is undefined. The library returns a value near 1.0 for it while every predicting arm sits near 0.5, and that one distant point inflates the predictor main effect. It is also the only arm on which the three mean-exact rules disagree, by up to 0.108, and it disagrees because there is no well-defined correlation for the shared mean to pin down. An undefined quantity is being averaged into a published comparison and given a rank.

Table I.1: **Equalising the routing reverses which entrant wins on the metric that decides first place in this benchmark's own table.** A cell-emitting entrant against a mean-emitting baseline on the largest screen, paired over 323 held-out perturbations. Ranges are over three seeds and negative favours the cell-emitting entrant. *As routed* is what the benchmark scores, its own cells against a baseline the harness completed. *Equalised* puts its mean through that same completion. The predictors are identical in both columns. * marks a 95% bootstrap interval excluding zero in every seed. Only the Wasserstein distance changes sign. Per-seed values and intervals are in Appendix I.5. Section 4.7.

metric	as routed	equalised
mean squared error	+0.007 to +0.010*	+0.008 to +0.011*
Pearson distance	+0.159 to +0.215*	+0.242 to +0.263*
energy distance	+1.277 to +1.324*	+0.103 to +0.136*
symmetric KL divergence	+1.990 to +2.312*	+0.022 to +0.028*
Wasserstein distance	**-10.270 to -9.914***	**+0.934 to +1.221***

Table I.2: **Our own predictor's advantage has an interval excluding zero in all three seeds under the harness's Gaussian and in one when both methods receive the same real cells.** Its Pearson-distance advantage over the linear model on the largest screen, per seed. Retrieval minus linear model, paired over 323 held-out perturbations, with 95% bootstrap intervals from 10,000 resamples. Negative favours retrieval. * marks an interval excluding zero. Under the harness's Gaussian the interval excludes zero in all three seeds. With both methods given the same real cells, the neighbour residuals, it does in one. Section 4.7.

seed	harness's Gaussian	same real cells
1	$-0.0856\ [-0.1167, -0.0548]^*$	$-0.0237\ [-0.0571, +0.0093]$
2	$-0.0789\ [-0.1102, -0.0467]^*$	$-0.0234\ [-0.0579, +0.0118]$
3	$-0.1061\ [-0.1357, -0.0762]^*$	$-0.0526\ [-0.0847, -0.0204]^*$

I.1 The decomposition has no fitted parameter, and the arm set is a choice

For a fixed set of predictors and generation rules the decomposition has no threshold and nothing to tune. Which factor dominates does depend on which arms enter the crossing, and the two paragraphs above report both changes it produces, from the neighbour-residual rule and from the control-mean baseline.

I.2 Per-seed tables

On the Tian inhibition screen our retrieval predictor loses to the linear model under both rules, and the rule changes the size of the loss. The Pearson-distance gap is $+0.0585$, $+0.0698$ and $+0.0539$ across three seeds under the harness's Gaussian and $+0.1594$, $+0.1777$ and $+0.1460$ once both methods receive the same real cells, the neighbour residuals. The harness's Gaussian therefore reduces the measured deficit by roughly two thirds (Figure I.1(c) gives the surviving share on the largest screen). On the largest screen, Table I.2 gives the per-seed intervals behind the advantage that survives equalising in one seed of three.

I.3 The reorderings in full

Each ordering is restricted to the four methods published in the benchmark's table, dropping our own predictor, and we ask whether the generation rule changes their order. It does in 6 of the 25 screen-and-metric combinations, on 4 of the 5 screens. Three of the six involve the regression baseline, whose embedding file the benchmark does not release and which we reconstructed, so the clean subset is the three on the Pearson distance. Measured as the same pair across the same rule transition, the change in the gap is 1.3 to 3.5 times the gap it reverses. Table I.3 lists them all. The 323-perturbation screen shows no reordering and the four that do reorder hold out 266, 36, 19 and 12, but five screens differing in assay, cell type, signal and perturbation distribution do not identify held-out size as the cause.

Table I.3: **Every reordering of four reference baselines under a change of generation rule**, 6 of 25 screen-and-metric combinations across 4 of 5 screens. $A > B$ means A ranks ahead of B, at a smaller distance. The three Pearson-distance rows occur between two rules that share their random draws, which identifies the cause. Section I.3.

screen	metric	the order changes from, to
RPE1	Pearson distance	linear > training mean, training mean > linear
RPE1	symmetric KL	training mean > regression, regression > training mean
Replogle-8	energy distance	linear > regression, regression > linear
Tian act.	Pearson distance	training mean > linear, linear > training mean
Tian act.	symmetric KL	regression > linear, linear > regression
Tian inh.	Pearson distance	linear > training mean, training mean > linear

Table I.4: **Equalising the completion reverses the Wasserstein comparison in all three seeds and leaves the sign of the other four metrics unchanged.** The cross-path comparison, per seed. Cell-emitting entrant minus mean-emitting baseline on the largest screen, paired over 323 held-out perturbations, with 95% bootstrap intervals over perturbations. Negative favours the cell-emitting entrant. Equalised completes the entrant's mean with the harness's Gaussian, the rule the baseline receives, which is not mean-exact.

metric	seed	as routed	equalised
mean squared error	1	+0.0090 [+0.0070, +0.0113]	+0.0106 [+0.0086, +0.0128]
	2	+0.0069 [+0.0051, +0.0088]	+0.0084 [+0.0067, +0.0102]
	3	+0.0097 [+0.0074, +0.0120]	+0.0110 [+0.0088, +0.0134]
Pearson distance	1	+0.1586 [+0.1211, +0.1963]	+0.2428 [+0.2140, +0.2726]
	2	+0.2033 [+0.1674, +0.2389]	+0.2424 [+0.2144, +0.2696]
	3	+0.2152 [+0.1746, +0.2549]	+0.2630 [+0.2318, +0.2937]
energy distance	1	+1.3102 [+1.2637, +1.3587]	+0.1259 [+0.1014, +0.1535]
	2	+1.2772 [+1.2353, +1.3196]	+0.1028 [+0.0825, +0.1237]
	3	+1.3244 [+1.2742, +1.3749]	+0.1361 [+0.1083, +0.1661]
symmetric KL divergence	1	+2.3121 [+1.9560, +2.6464]	+0.0277 [+0.0186, +0.0378]
	2	+1.9898 [+1.5983, +2.3635]	+0.0223 [+0.0138, +0.0307]
	3	+2.1299 [+1.7891, +2.4642]	+0.0267 [+0.0161, +0.0383]
Wasserstein distance	1	-10.1054 [-10.3786, -9.8235]	+1.1860 [+0.9632, +1.4293]
	2	-10.2698 [-10.4987, -10.0372]	+0.9342 [+0.7456, +1.1295]
	3	-9.9142 [-10.2029, -9.6117]	+1.2208 [+0.9782, +1.4781]

The cause of the three Pearson reversals. All three occur at the step from the Gaussian rule to its recentred version. Those two rules use the same random draws, and the only difference between them is whether the emitted sample is shifted so its mean equals the prediction. The Pearson distance is a function of the group means alone, verified above with no disagreement in 60 of 60 combinations. So on this metric, which of the linear model and the training mean ranks higher is decided by a Monte Carlo error of order $\boldsymbol{\sigma}/\sqrt{n_t}$ that the harness injects and then scores. The division of published Wasserstein ranks by output representation is in Appendix F, as context, since the two groups differ in more than their output format.

I.4 Every reordering, and the denominator behind the count

Five screens, five metrics, six pairs of the four baselines and six pairs of the four rules give 900 opportunities for a pair to reverse, each on a gap averaged over the three seeds. Eighteen do, 2.0%, and they fall in the six combinations of Appendix I.3.

I.5 The cross-path comparison, per seed

Figure I.1(a) draws the per-seed comparisons of Table I.4.

The known-answer check on this experiment. The mean routed through the completion is extracted from the cell-emitting entrant's own output. A rule that preserves the mean cannot change a metric that sees only the mean, so that entrant's native cells and its mean re-completed by the three mean-exact rules must score identically on the mean squared error and the Pearson distance. They do, with no disagreement in all six metric-and-seed combinations. The equalised column of Tables I.1 and I.4 completes both means with the harness's own Gaussian, which is not mean-exact, so its mean-only metrics move by that rule's Monte Carlo error, from $+0.0090$ to $+0.0106$ on the mean squared error in the first seed. With both means completed by the recentred Gaussian instead, on the same draws, the Wasserstein comparison still reverses, to $+1.19$ $[+0.96, +1.43]$, $+0.94$ $[+0.75, +1.13]$ and $+1.22$ $[+0.98, +1.48]$ in the three seeds.

I.6 Full attribution tables

Table I.5: **Which factor dominates depends on the screen and the metric, while the interaction stays small throughout.** Per-seed two-way attribution, five predicting arms. The predictor's, the generation rule's and the interaction's shares of the between-arm variance, as percentages. A rule share is bold where it exceeds the predictor's. Held-out sizes in parentheses.

		seed 1			seed 2			seed 3		
screen	metric	pred	rule	int	pred	rule	int	pred	rule	int
K562 (323)	mean squared error	99.3	0.7	0.0	99.3	0.7	0.0	99.6	0.4	0.0
	Pearson distance	96.2	2.7	1.2	96.2	2.5	1.3	97.1	1.8	1.1
	energy distance	97.0	3.0	0.0	96.8	3.2	0.0	97.4	2.5	0.0
	symmetric KL divergence	43.8	**56.1**	0.1	22.1	**77.8**	0.1	46.1	**53.8**	0.1
	Wasserstein distance	86.3	13.7	0.0	85.3	14.7	0.0	86.9	13.0	0.0
RPE1 (266)	mean squared error	99.9	0.1	0.0	99.8	0.2	0.0	99.9	0.1	0.0
	Pearson distance	99.0	0.5	0.4	99.1	0.7	0.2	99.1	0.6	0.3
	energy distance	98.0	1.9	0.0	97.5	2.5	0.0	98.0	1.9	0.0
	symmetric KL divergence	19.8	**77.8**	2.4	20.3	**78.1**	1.7	24.9	**73.7**	1.4
	Wasserstein distance	69.3	30.7	0.0	63.6	36.4	0.0	68.4	31.5	0.0
Replogle-8 (12)	mean squared error	99.9	0.0	0.0	99.9	0.1	0.0	100.0	0.0	0.0
	Pearson distance	99.2	0.2	0.6	99.6	0.1	0.3	99.3	0.1	0.6
	energy distance	93.0	5.5	1.5	93.5	5.2	1.3	89.0	9.7	1.3
	symmetric KL divergence	0.2	**99.6**	0.2	0.0	**99.9**	0.0	0.0	**100.0**	0.0
	Wasserstein distance	3.1	**96.9**	0.0	3.2	**96.8**	0.0	3.1	**96.9**	0.0
Tian act. (19)	mean squared error	99.3	0.7	0.0	99.6	0.4	0.0	98.5	1.5	0.0
	Pearson distance	98.2	1.1	0.6	97.7	2.1	0.2	98.1	1.7	0.2
	energy distance	97.1	2.6	0.3	96.5	3.4	0.1	91.9	7.7	0.4
	symmetric KL divergence	0.0	**100.0**	0.0	0.0	**100.0**	0.0	0.0	**100.0**	0.0
	Wasserstein distance	25.9	**74.1**	0.0	43.1	**56.8**	0.0	30.9	**69.0**	0.1
Tian inh. (36)	mean squared error	93.0	7.0	0.0	93.7	6.3	0.0	91.2	8.8	0.0
	Pearson distance	91.5	6.6	1.9	91.7	6.1	2.2	89.4	8.5	2.1
	energy distance	78.1	21.9	0.0	80.4	19.5	0.0	78.0	21.9	0.0
	symmetric KL divergence	0.1	**99.9**	0.0	0.2	**99.8**	0.0	0.3	**99.7**	0.0
	Wasserstein distance	19.9	**80.1**	0.0	15.8	**84.2**	0.0	14.2	**85.8**	0.0

The interaction term, per seed in Table I.5, is small throughout, at most 2.4% of the total in any screen, metric and seed and at most 2.1% once the three seeds are averaged. The two main effects can therefore be read separately for this arm set.

J What we could not reproduce

This appendix states what we could not reproduce, which bounds the comparisons of Sections 4.1 to 4.3 and recommendation P3.

The score transform. The benchmark publishes, for each method and perturbation, metric scores and ranks derived from them. The transform from a raw distance to a published score is absent from the released implementation, which reads the finished results without exposing the preceding transform. We therefore cannot recompute a published score or rank from a raw distance (recommendation P3). Every rank this paper reports is either the benchmark's own published rank, aggregated as its published analysis does (Appendix F), or a rank we compute per perturbation from

raw distances on the same cells (Table 1). We do reproduce the raw distance, through the harness's own scoring call, for every arm we run.

Four screens whose split we cannot reproduce. The harness draws its held-out perturbations with a seeded routine we reimplement, and the benchmark's released results name the perturbations it scored. On five of the nine genetic-track screens our test sets are exactly those, on every seed, and those five support our matched entrant comparisons. On the other four, Adamson, Frangieh, Papalexi and Replogle-7, twelve to thirteen of the perturbations the benchmark scored are missing from ours on every seed, so nothing computed on them can be set beside a published number. Appendix K uses Frangieh and Papalexi only for a separate measured-population diagnostic. It does not repair these splits.

A duplicated metric list. The benchmark's released implementation assigns its metric list twice, and the second assignment silently overwrites the first. The published score columns and the repository's own documentation both correspond to the first list, which is the one we reproduce, and the published numbers are consistent with it.

K ADDITIONAL CALIBRATION, SENSITIVITY AND DISJOINT-CELL VALIDATION

This appendix supports the calibration statement of Section 3 and the transfer and out-of-sample results of Section 4.4.

These analyses were specified before their outcomes, after the main experiments. They test whether the complete selection and averaging procedure preserves the simulated null and whether the recommended estimator penalises prescribed mean and dispersion errors. They also test whether a calibration margin predicts a transport preference on separate cells, and whether it does so better than the number of genes and cells on two screens no earlier analysis used. All conditions are retained. Intervals are pointwise 95% percentile bootstrap intervals from 10,000 resamples, with technical repeats averaged before resampling the independent units. They are descriptive intervals, without multiplicity correction or an equivalence margin.

K.1 THE NULL AFTER FEATURE SELECTION AND AVERAGING

We simulate 1,000 independent Gaussian populations on 5,822 genes with standard deviation 0.4. Each population provides independent pools of 150 measured cells, 150 prediction cells and 150 controls. Measured and predicted populations have mean 0.2 on the first 100 coordinates and zero elsewhere. Controls have mean zero. Three permutations each divide the measured pool into discovery, floor and reference thirds of 50 cells. Each repeat selects 50 cells from the prediction pool. The original pools are reused across repeats. No iid sample is recentred. Discovery Welch tests against controls select 100 genes under two policies. One ranks the full gene axis, and the other re-ranks within 5,000 candidates first chosen by a Welch test on the entire measured pool. The latter exercises the candidate-filter dependence discussed in Appendix D.

For each repeat, we compare the independently sampled correct candidate with the floor using both exact squared Wasserstein and the benchmark's rounded entropic call. We also contract the same prediction sample around its own mean by a factor of 0.29. Table K.1 reports differences after averaging repeats within population. The perfect-sampler mean intervals include zero, but this is not evidence of equivalence. Its below-floor frequency is 46.3–46.9%, and three of the four intervals exclude 50%. Thus the three-repeat aggregate does not justify a half-below-floor null. Single-repeat summaries are retained separately in the released results and vary across repeats. Some pointwise mean intervals exclude zero. The contracted candidate falls below the floor in every simulated population for both policies and estimators. These findings support sensitivity to contraction in this regime while limiting the calibration claim to the exchangeability conditions and single comparison stated in Section 3.

Table K.1: **After gene selection and averaging over three splits, a correctly distributed candidate falls below the floor in 46.3 to 46.9% of simulated populations.** An independent correctly distributed candidate is compared with the split-half floor in 1,000 simulated populations. Each policy selects 100 genes, either from the full gene axis or within the 5,000 genes ranked by Welch tests on the measured pool versus controls. The final 100 genes are ranked on discovery cells versus controls. Candidate-minus-floor differences are averaged over three technical splits before calculating population means and below-floor percentages. Brackets are pointwise 95% percentile bootstrap intervals over independent simulated populations (10,000 resamples). The below-floor probability is measured, not fixed at one half. The mean-preserving 0.29-scale candidate falls below the floor in all 1,000 populations under both policies and both estimators.

Gene-selection scope	Estimator	Candidate − floor	Below floor (%)
Full gene axis	Exact W_2^2	$0.015\ [-0.003, 0.032]$	$46.9\ [43.8, 50.0]$
Full gene axis	Benchmark call	$0.015\ [-0.003, 0.033]$	$46.5\ [43.4, 49.6]$
Measured top 5,000	Exact W_2^2	$0.014\ [-0.003, 0.031]$	$46.3\ [43.2, 49.4]$
Measured top 5,000	Benchmark call	$0.014\ [-0.003, 0.032]$	$46.8\ [43.7, 49.9]$

K.2 Sensitivity to prescribed mean and dispersion errors

We draw independent truth and prediction samples with 50 cells and 100 genes from a zero-mean Gaussian with standard deviation 0.4, for each of 200 independent populations. A variant multiplies the prediction by a scale in $\{0.5, 0.8, 1, 1.2\}$ and shifts every coordinate by $\{0, 0.1, 0.2\}$ true standard deviations. The shift direction is fixed. Scaling is around the known population mean, not the sample mean. We score all twelve variants using exact and benchmark Wasserstein and both energy estimators. Energy uses unsquared Euclidean distances. The U-statistic excludes self-pairs and retains negative estimates. Each variant is paired with the same unmodified prediction sample.

Table K.2 contains the full grid. Both energy estimators penalise every non-reference variant on average, with all corresponding pointwise intervals above zero. Both transport estimators favour the contracted variants even when a mean shift is added. At unit scale, all four estimators penalise the prescribed mean shifts on average. The energy scores therefore retain the transport scores' sensitivity to mean shifts. The controls establish neither sensitivity to untested error types nor superiority for choosing biological models. Raw differences cannot be compared between metrics on different scales.

K.3 Transfer to separate cells in additional biological contexts

We use matrices with the released preprocessing from the Frangieh melanoma experiment (Frangieh et al., 2021) and the Papalexi THP-1 experiment (Papalexi et al., 2021). Their prior exclusion from matched benchmark reproduction remains in force (Appendix J). These data had already been inspected and their published ranks audited. This is a new measured-cell analysis, not an untouched-dataset test or a comparison with trained entrants. Source metadata supplies the three Frangieh conditions and three Papalexi replicates, with exact perturbation-label row alignment. We require at least 100 cells per target within its context and use matching controls, retaining 175, 167 and 183 targets for Frangieh and 23, 20 and 22 for Papalexi. The corresponding numbers of excluded target–context pairs are 108 and 10. Conditions and replicates share targets and are never pooled as independent biological systems.

For a target with n cells, set $h = \min(\lfloor n/5 \rfloor, 100)$ and permute once to reserve h discovery cells, $2h$ calibration cells and $2h$ validation cells. These pools remain disjoint throughout the experiment. Discovery cells select 10, 100 and 1,000 genes by absolute Welch statistics against the matching controls on the full released gene axis. Three technical repeats repartition the fixed calibration and validation pools into equal halves, without changing discovery cells or genes. In calibration halves A, B, compute $q = (W(A, B) - b - \Delta)/b$, where b is the empirical variance of B and Δ is the squared sample-mean gap. A positive q predicts a collapse preference. On separate validation halves C, D, compare $W(C, D)$ with $W(\bar{C}, D)$, where $\bar{C}$ denotes copies of C's sample mean. Divide their difference by calibration b. This gives a validation margin whose positive sign favours collapse. Ratios and signs used for inference are formed by first averaging repeat-level ratios within target. A

Table K.2: **Energy V and U penalise every prescribed error on average, while both transport estimators favour contraction even with a mean shift.** All 12 combinations of coordinate-wise mean shift (in true standard deviations) and dispersion scale are shown for each estimator. Entries are the mean score difference, variant minus the correct candidate, with pointwise 95% percentile bootstrap intervals over 200 independent simulated populations (10,000 resamples). Positive values favour the correct candidate. Each population has 50 cells per sample and 100 genes. The zero-shift, unit-scale row is the identical reference candidate and ties in every population under all four estimators. Metric magnitudes are on different scales and should not be compared across estimators.

Variant − correct candidate: mean [95% interval]			
Transport			
Shift	Scale	Exact W_2^2	Benchmark call
0.0	0.5	-8.722 $[-8.753, -8.692]$	-9.031 $[-9.062, -8.999]$
0.0	0.8	-4.449 $[-4.464, -4.434]$	-4.594 $[-4.609, -4.579]$
0.0	1.0	0.000 $[0.000, 0.000]$	0.000 $[0.000, 0.000]$
0.0	1.2	5.729 $[5.711, 5.747]$	5.903 $[5.884, 5.921]$
0.1	0.5	-8.566 $[-8.596, -8.535]$	-8.871 $[-8.902, -8.840]$
0.1	0.8	-4.293 $[-4.309, -4.277]$	-4.435 $[-4.451, -4.418]$
0.1	1.0	0.155 $[0.147, 0.164]$	0.159 $[0.150, 0.167]$
0.1	1.2	5.884 $[5.863, 5.905]$	6.061 $[6.039, 6.082]$
0.2	0.5	-8.089 $[-8.121, -8.057]$	-8.384 $[-8.416, -8.351]$
0.2	0.8	-3.817 $[-3.837, -3.797]$	-3.948 $[-3.969, -3.928]$
0.2	1.0	0.631 $[0.614, 0.647]$	0.644 $[0.628, 0.662]$
0.2	1.2	6.359 $[6.332, 6.386]$	6.546 $[6.519, 6.573]$
Energy distance			
Shift	Scale	V-statistic	U-statistic
0.0	0.5	0.402 $[0.399, 0.404]$	0.458 $[0.456, 0.461]$
0.0	0.8	0.040 $[0.039, 0.041]$	0.063 $[0.061, 0.064]$
0.0	1.0	0.000 $[0.000, 0.000]$	0.000 $[0.000, 0.000]$
0.0	1.2	0.074 $[0.073, 0.075]$	0.051 $[0.050, 0.052]$
0.1	0.5	0.437 $[0.434, 0.440]$	0.493 $[0.490, 0.496]$
0.1	0.8	0.070 $[0.069, 0.072]$	0.093 $[0.091, 0.095]$
0.1	1.0	0.027 $[0.026, 0.029]$	0.027 $[0.026, 0.029]$
0.1	1.2	0.098 $[0.097, 0.100]$	0.076 $[0.074, 0.078]$
0.2	0.5	0.542 $[0.538, 0.546]$	0.598 $[0.595, 0.602]$
0.2	0.8	0.162 $[0.159, 0.166]$	0.185 $[0.182, 0.188]$
0.2	1.0	0.111 $[0.108, 0.114]$	0.111 $[0.108, 0.114]$
0.2	1.2	0.174 $[0.171, 0.177]$	0.151 $[0.148, 0.154]$

zero calibration variance would exclude the target–space combination. None occurred. Perturbation targets within each context are the bootstrap units.

Table K.3 reports every context and space. In all six contexts at 100 and 1,000 genes, the calibration-positive subset has at least ten targets and a positive validation-mean interval. At ten genes, only zero to three targets are flagged per context. The three estimable flagged-subset intervals are negative, contradicting the predicted direction there, and the small subsets give imprecise evidence about transfer. The diagnostic therefore does not work uniformly across scoring spaces.

We compare direction accuracy with two fixed rules, one that always predicts collapse and one that predicts it only at 100 and 1,000 genes. The dimension-only rule matches the diagnostic's perfect accuracy throughout Frangieh's larger spaces and every 1,000-gene space. Only Papalexi replicate 1 at 100 genes has a paired accuracy-gain interval strictly above zero, 91.3% against 60.9%, a gain of 30.4 percentage points $[13.0, 47.8]$. This is one of eighteen pointwise comparisons and does not establish general incremental prediction. Energy V and U both penalise collapse on average in every context and space, with pointwise intervals above zero. Their collapsed candidate depends on the validation sample mean, so this is a mechanism control, not an iid-estimator proof of propriety. These results support transfer of the high-dimensional transport preference beyond K562 and RPE1. On these two experiments they show no broad gain over dimension alone, which Appendix K.4 tests on two further screens, and they leave biological model selection unestablished.

Table K.3: **The calibration margin transfers to disjoint cells at 100 and 1,000 genes, while the estimable ten-gene flagged subsets point the wrong way.** All dataset, context, and gene-count combinations are reported separately. Discovery cells select genes. Discovery, calibration and validation pools remain disjoint across all technical repeats. Flagged targets have a positive calibration margin. Direction accuracy compares the sign of that margin with the validation sign across all targets. The fixed dimension-only rule predicts a collapse preference at $k \geq 100$ and a spread preference at $k = 10$. The always-collapse rule predicts collapse at every dimension. All three accuracy columns are percentages. The last column reports the flagged targets' validation margin, truthful-minus-collapsed exact W_2^2, divided by calibration within-population dispersion. Positive values mean that collapse scores better. Ratios are calculated within each repeat and averaged over three technical repeats before summarising or bootstrapping perturbation targets within each context. Brackets are pointwise 95% percentile bootstrap intervals (10,000 resamples). A dash denotes fewer than two flagged targets. Intervals from two or three targets have limited resolution. Contexts are not pooled as independent biological systems.

Context	k	n	Flagged	Diagnostic	Dimension	Always	Flagged margin
Frangieh							
Co-culture	10	175	1	99.4	100.0	0.0	—
Co-culture	100	175	175	100.0	100.0	100.0	$0.437\ [0.425, 0.450]$
Co-culture	1000	175	175	100.0	100.0	100.0	$0.808\ [0.802, 0.813]$
Control	10	167	3	97.0	98.8	1.2	$-0.203\ [-0.224, -0.170]$
Control	100	167	167	100.0	100.0	100.0	$0.519\ [0.509, 0.528]$
Control	1000	167	167	100.0	100.0	100.0	$0.818\ [0.811, 0.824]$
IFNγ	10	183	1	98.9	99.5	0.5	—
IFNγ	100	183	183	100.0	100.0	100.0	$0.410\ [0.397, 0.422]$
IFNγ	1000	183	183	100.0	100.0	100.0	$0.799\ [0.795, 0.804]$
Papalexi							
Replicate 1	10	23	2	91.3	100.0	0.0	$-0.381\ [-0.461, -0.300]$
Replicate 1	100	23	16	91.3	60.9	60.9	$0.196\ [0.121, 0.275]$
Replicate 1	1000	23	23	100.0	100.0	100.0	$0.551\ [0.524, 0.578]$
Replicate 2	10	20	2	85.0	95.0	5.0	$-0.642\ [-0.646, -0.638]$
Replicate 2	100	20	19	90.0	95.0	95.0	$0.219\ [0.148, 0.298]$
Replicate 2	1000	20	20	100.0	100.0	100.0	$0.601\ [0.577, 0.627]$
Replicate 3	10	22	0	100.0	100.0	0.0	—
Replicate 3	100	22	21	86.4	81.8	81.8	$0.211\ [0.144, 0.275]$
Replicate 3	1000	22	22	100.0	100.0	100.0	$0.580\ [0.553, 0.606]$

K.4 Prediction beyond gene count and cell number on two untouched screens

The comparison above cannot separate the measured margin from the number of scoring genes, because on Frangieh and Papalexi the gene count alone predicts its sign. Pre-registration P therefore scored two screens whose cells no earlier analysis had used, Adamson and Replogle-7, the two scPerturBench screens excluded from the entrant comparison because their benchmark split does not reproduce (Appendix J). Each is one context, with 72 and 104 targets of at least 100 cells. The protocol is that of the previous subsection with a new seed and all control cells of the screen. The validation margin $m = (W(C, D) - W(\bar{C}, D))/b_D$ is normalised by its own reference variance. By Corollary 1, it is the calibration statistic $\rho - 1 - \Delta/b$ computed on the validation halves. Before either screen was scored, three predictors of m were fitted by least squares on the Frangieh and Papalexi margins, one per gene count, and frozen. B uses the logarithm of the half size, C adds the calibration margin, and D is the gene-count mean. A is the unfitted calibration margin. The primary endpoint, one per screen, is the paired mean over targets of $|m - \mathrm{B}| - |m - \mathrm{C}|$ averaged over the three gene counts. It counts as support if its interval lies above zero, and the pre-specified criterion requires support on both screens.

Table K.4 reports every setting. The increment is supported on both screens. The secondary comparison refits B and C within each screen over target folds, so it excludes shifts between development and test screens. It is positive on both screens at every gene count. It shows that the margin carries information on these screens, not that the frozen fit transfers. The new screens differ from the development contexts in a way no dimension rule anticipates. On Replogle-7, the median ratio at

Table K.4: **The calibration margin predicts the margin on disjoint cells better than gene count and cell number on both untouched screens.** Pre-registration P, exact unregularised W_2^2, targets as units, three technical repeats averaged within target. The margin is $\rho - 1 - \Delta/b$, positive when a sample-mean copy beats a sample of the truth. Favoured is the share of targets whose validation margin is positive. Mean absolute errors are given for four predictors of the validation margin, A, the calibration margin itself; B, gene count and log half size; C, B plus the calibration margin; D, gene count only. B, C and D were fitted on Frangieh and Papalexi and frozen before scoring. B − C is paired over targets with 95% percentile bootstrap intervals (10,000 resamples), and positive values favour C. Sign accuracy compares the calibration margin's sign and the dimension-only rule of Table K.3 with the validation sign. The cross-fitted comparison refits B and C within each screen over five target folds, out-of-fold errors only.

		Favoured	Mean absolute error					Sign accuracy (%)	
Screen	k	(%)	A	B	C	D	B − C	Margin	Dimension
Adamson	10	0.0	0.113	0.150	0.124	0.222	0.026 [0.017, 0.035]	98.6	100.0
Adamson	100	44.4	0.079	0.220	0.104	0.475	0.116 [0.080, 0.151]	93.1	44.4
Adamson	1,000	100.0	0.038	0.146	0.044	0.241	0.102 [0.089, 0.116]	100.0	100.0
Adamson, primary endpoint, 72 targets							0.082 [0.068, 0.095]		
Adamson, cross-fitted over five target folds							0.052 [0.039, 0.064]		
Replogle-7	10	1.0	0.067	0.334	0.253	0.528	0.081 [0.074, 0.087]	99.0	99.0
Replogle-7	100	1.9	0.046	0.843	0.280	1.136	0.563 [0.547, 0.577]	100.0	1.9
Replogle-7	1,000	16.3	0.044	0.863	0.148	0.960	0.715 [0.688, 0.742]	94.2	16.3
Replogle-7, primary endpoint, 104 targets							0.453 [0.439, 0.466]		
Replogle-7, cross-fitted over five target folds							0.059 [0.047, 0.071]		

100 genes is 0.27 and collapse is favoured for 1.9% of targets. The rule that predicts collapse from 100 genes upward is right on 1.9% of them, while the calibration margin is right on all 104. The unfitted margin A has the lowest error of the four predictors in every setting, and its validation-minus-calibration bias interval includes zero in all six. The frozen fit C is not calibrated on the new screens, with slopes of 0.84 to 2.26. Energy V and U penalise the collapsed copy with every interval above zero in all six settings. These results concern the transport preference on disjoint cells of the same perturbation, not the rank of a trained model, and the two screens are reported separately.

K.5 The margin against a trained entrant's advantage

The previous subsection predicts the margin of a sample-mean copy. Pre-registration Q asks the same question of the reproduced entrants, using their saved predictions and no retraining. The twelve runs comprise CPA on K562 and on RPE1, and scGPT and scouter on K562, each at benchmark seeds 1, 2 and 3. Every held-out perturbation with at least 100 cells is included. Each population is partitioned as in K.3 with a new seed. The margin q comes from calibration halves disjoint from the entrant's evaluation halves. On the evaluation halves, the entrant is represented by as many of its own cells as a half holds and scored against one half. Its margin is its score minus the split-half floor of the two evaluation halves, divided by the reference variance. The completed linear baseline is scored the same way. Within each run and gene count, five target folds give out-of-fold errors for two least-squares models of the entrant's margin. One uses log half size, and the other adds q. The primary endpoint is the gain from adding q, averaged over the three gene counts, and an entrant and screen pass only if all three seeds show a gain interval above zero. These screens were analysed before, so Q is a pre-registered analysis of existing predictions on newly partitioned cells rather than a test on untouched data.

Table K.5 reports every run. Eleven of the twelve show the gain, and none contradicts it. CPA on RPE1, scGPT and scouter pass the rule, while CPA on K562 does not because its third seed's interval includes zero. The third scouter seed passes with a lower bound of 0.0002. The gain is small, 0.014 to 0.056 in units of the reference variance, carried by 100 and 1,000 genes and absent at ten. The Spearman correlation of q with the entrant's margin at 100 genes is -0.57 to -0.77. At ten genes, where the entrants beat the completed linear baseline on 64 to 96% of targets, q orders the targets on which they do with an AUC of 0.66 to 0.94. A post hoc comparison shows that a smaller population alone reaches 0.68 to 0.87. That ordering is therefore not evidence of information beyond

Table K.5: **Adding the cell-only margin to gene count and cell number predicts a reproduced entrant's margin below an independent floor in eleven of twelve runs.** Pre-registration Q, exact unregularised W_2^2. The entrant's margin is its score minus the split-half floor on evaluation cells, divided by the reference variance. The margin q is measured on separate calibration cells disjoint from the evaluation cells. Gain is the out-of-fold drop in absolute error when q is added to a per-gene-count model on log half size, using five target folds within the run. It is reported averaged over 10, 100 and 1,000 genes, with a 95% target-bootstrap interval (10,000 resamples), and then per gene count. r_S is the Spearman correlation of q with the entrant's margin at 100 genes. The last two columns give, at ten genes, the AUC for the entrant beating the completed linear baseline, of q and, post hoc, of a smaller half size. At 100 and 1,000 genes the entrant beats that baseline on 99 to 100% of targets, so the AUC is not informative there. Targets are the independent units, and seeds change the held-out set. The lower bound printed as 0.000 for scouter's third seed is 0.0002.

				Gain by gene count				AUC, ten genes	
Entrant, screen	Seed	Targets	Gain	10	100	1,000	r_S	q	half size
CPA, K562	1	248	0.027 [0.010, 0.044]	-0.003	0.060	0.023	-0.69	0.87	0.82
	2	225	0.019 [0.008, 0.031]	0.008	0.038	0.012	-0.61	0.91	0.87
	3	231	0.016 [−0.001, 0.033]	-0.002	0.039	0.011	-0.68	0.88	0.85
CPA, RPE1	1	121	0.055 [0.033, 0.076]	-0.026	0.111	0.080	-0.64	0.84	0.68
	2	131	0.056 [0.031, 0.080]	-0.011	0.127	0.051	-0.57	0.74	0.70
	3	126	0.055 [0.033, 0.077]	-0.003	0.102	0.065	-0.69	0.94	0.87
scGPT, K562	1	248	0.031 [0.022, 0.041]	0.007	0.063	0.024	-0.77	0.66	0.68
	2	225	0.016 [0.007, 0.024]	0.003	0.033	0.011	-0.69	0.73	0.74
	3	231	0.018 [0.009, 0.027]	0.002	0.040	0.011	-0.76	0.74	0.77
scouter, K562	1	248	0.025 [0.011, 0.040]	0.001	0.053	0.020	-0.72	0.74	0.72
	2	225	0.018 [0.007, 0.030]	0.013	0.032	0.010	-0.63	0.80	0.78
	3	231	0.014 [0.000, 0.028]	-0.006	0.037	0.011	-0.72	0.77	0.79

cell number. After Q's outcome, pre-registration R replaced the log half size with a quadratic in it and pooled the runs with equal weight on gene counts, seeds and entrant-screen pairs. Adding q then lowers the out-of-fold error by 10.4% $[6.8, 12.4]$. The target-bootstrap interval, computed by refitting both models in every resample, exceeds the registered 5% threshold. The error falls in every run, by 4.5 to 20.1%. The tests concern the entrants' advantage over the floor and over one baseline, not a biologically correct choice between models.

L SAVED-PREDICTION RESCORING AND MEASURED-CELL CONTROLS

This appendix supports recommendation P2, the nine-arm extension of the seven-arm ordering in Section 4.5 and the full-axis result of Section 4.3 with saved-prediction rescoring and measured-cell error controls.

These analyses were specified after the earlier outcomes. The continuous margin analysis was explicitly post hoc. The saved-prediction rescoring used the existing model outputs without retraining. Twenty-four scPerturBench runs and four Norman19 spaces reproduce their stored energy values exactly. Both V and U energy estimates put every reproduced entrant above the floor on every eligible target in those runs. In all four Norman19 spaces, the truth sample ranks first on all 45 held-out combinations under both estimates, and each model remains above the floor on every combination. The train-fitted PCA rescore reproduces the stored values only under the original eight-thread numerical setting. In that setting all 405 rounded-V and 405 native-energy rows match exactly, with the scorer, data and tolerance unchanged. The default setting does not reproduce them, and the numerical cause has not been isolated. These four spaces share combinations and do not provide four independent datasets.

In the original seven-arm K562 comparison, CPA, scGPT and scouter retain the bottom three mean energy ranks under both V and U. Adding two raw mean point masses places those raw means at the bottom two positions among nine arms. The seven-arm result therefore depends on its stated comparator set. In all 24 scPerturBench runs, raw control and training means score worse than

Table L.1: **Full-axis gene selection keeps every K562 CPA-minus-floor interval below zero, while every RPE1 interval includes zero.** Exact unregularised squared Wasserstein CPA-minus-floor means and pointwise 95% target-bootstrap intervals are shown for both fixed gene-selection policies. Negative values favour CPA. Three technical repeats are averaged within target. Model seed 1117 duplicates the original K562 seed-1 prediction and adds no independent evidence. Energy V and U intervals are positive for every row under both policies.

Run	Targets	Candidate-list genes	Full-axis genes
K562, seed 1	310	$-3.66\,[-4.31, -2.93]$	$-3.59\,[-4.24, -2.87]$
K562, seed 2	312	$-4.25\,[-4.78, -3.65]$	$-4.21\,[-4.73, -3.60]$
K562, seed 3	311	$-3.79\,[-4.37, -3.15]$	$-3.73\,[-4.31, -3.10]$
RPE1, seed 1	232	$-0.15\,[-1.16, 0.99]$	$-0.21\,[-1.21, 0.92]$
RPE1, seed 2	232	$-0.68\,[-1.69, 0.45]$	$-0.71\,[-1.71, 0.41]$
RPE1, seed 3	238	$-0.68\,[-1.58, 0.32]$	$-0.75\,[-1.63, 0.23]$
K562, model 1117	310	$-3.66\,[-4.31, -2.93]$	$-3.59\,[-4.24, -2.87]$
K562, model 1118	310	$-3.72\,[-4.37, -3.00]$	$-3.66\,[-4.30, -2.94]$
K562, model 1119	310	$-3.69\,[-4.34, -2.97]$	$-3.62\,[-4.27, -2.91]$

their completed versions under both energy estimates. Completion still matters after changing the estimator. The first K562 model-seed repeat duplicates the original saved predictions and does not increase the independent sample count.

We keep CPA's predictions, cells and top-100 gene count fixed while changing the discovery ranking from the released candidate list to the full released gene axis. This full-axis policy cannot undo selection in upstream preprocessing. Table L.1 reports exact unregularised W_2^2, which is a new score on these cells rather than a replay of the benchmark's regularised call. Both policies retain a negative CPA-minus-floor mean interval in every K562 run. All RPE1 intervals include zero. Energy V and U intervals remain positive in every run and policy. The K562 direction is not explained by the candidate-list filter alone, while this exact-Wasserstein comparison does not establish the same mean advantage on RPE1.

L.1 Prescribed errors on separate measured cells

The real-cell controls reuse the six contexts and three gene counts of the disjoint-cell validation without changing its eligible targets or discovery, calibration and validation pools. The earlier per-repeat observations and exclusions are reproduced exactly. A correct arm ties exactly by construction. Within each context and gene space, we compare independent gene-wise row permutations, collapse to the candidate sample mean and a prediction from another target against the unchanged candidate. The wrong-target arm pairs eligible targets without replacement and averages both directions within pair. One odd target is excluded from this arm in each of Frangieh's three conditions and Papalexi replicate 1. Its independent unit is the pair. Other controls and the margin analysis use the target.

Table L.2 reports every context, gene count and metric, including intervals that cross zero or favour an error. For collapse, energy V and U have positive error-minus-correct intervals in all 18 settings. Exact W_2^2 instead has six positive, eleven negative and one inconclusive interval. Gene permutation has positive intervals in 16 settings under energy V and 15 under energy U. The remaining two and three settings are inconclusive. Both energy estimates have positive wrong-target intervals in 15 settings. All three inconclusive settings are Frangieh Control, where each comparison contains 83 independent target pairs. These are tests of specified errors on measured cells. A wrong target can have a similar response, and none of the controls proves general biological model-selection superiority.

The post hoc continuous analysis compares calibration and validation margins without fitting a threshold. With each margin divided by its own reference variance, the target-level Spearman correlation ranges from -0.005 to 0.897 across the 18 settings. Four pointwise intervals include zero, including Frangieh Control at 1,000 genes. Papalexi replicate 1 at ten genes has a negative evaluation-minus-calibration bias interval despite a positive correlation. Associations between contexts that share targets are not pooled. A correlation does not validate predictive calibration.

Table L.2: **Energy V and U penalise collapse in all 18 measured-cell settings, where exact W_2^2 does so in six.** Each sign triplet orders exact W_2^2, energy V and energy U. A plus or minus means the pointwise 95% bootstrap interval for error-minus-correct mean lies strictly above or below zero. A question mark means it includes zero. The correct arm ties exactly in all settings. Target counts apply to permutation and collapse. Pair counts apply only to wrong-target controls. r is the post hoc Spearman correlation between calibration and independent validation margins, each divided by its own reference variance. Brackets give the pointwise target-bootstrap interval for r. Contexts are separate.

Context	Genes	Targets	Pairs	Permute	Collapse	Wrong target	r [95% interval]
Co-culture	10	175	87	???	+++	+++	0.358 [0.21,0.50]
Co-culture	100	175	87	?++	-++	+++	0.747 [0.66,0.81]
Co-culture	1000	175	87	+++	-++	+++	0.217 [0.07,0.35]
Control	10	167	83	???	+++	???	0.178 [0.03,0.32]
Control	100	167	83	?+?	-++	???	0.499 [0.38,0.60]
Control	1000	167	83	+++	-++	???	-0.005 [-0.16,0.16]
IFNγ	10	183	91	?++	+++	+++	0.321 [0.17,0.46]
IFNγ	100	183	91	+++	-++	+++	0.733 [0.65,0.80]
IFNγ	1000	183	91	+++	-++	+++	0.108 [-0.04,0.26]
rep1	10	23	11	+++	+++	+++	0.792 [0.50,0.93]
rep1	100	23	11	+++	?++	+++	0.897 [0.72,0.97]
rep1	1000	23	11	+++	-++	+++	0.659 [0.29,0.87]
rep2	10	20	10	+++	+++	+++	0.344 [-0.23,0.75]
rep2	100	20	10	+++	-++	+++	0.786 [0.52,0.92]
rep2	1000	20	10	+++	-++	+++	0.641 [0.26,0.86]
rep3	10	22	11	+++	+++	+++	0.455 [-0.03,0.78]
rep3	100	22	11	+++	-++	+++	0.661 [0.25,0.91]
rep3	1000	22	11	+++	-++	+++	0.702 [0.41,0.85]