SHRINKAGELENS: INFORMATION SOURCES IN PERTURBATION DISCRIMINATION SCORES

Anonymous authors
Paper under double-blind review

ABSTRACT

Perturbation discrimination ranks held-out responses against a prediction, often after magnitude correction. A corrected score can, however, obtain perturbation-specific information from scoring inputs instead of the prediction. ShrinkageLens separates these sources by specifying each rule's inputs and comparing scores using identical predictions, matched permutations and the inverse of a known shrinkage map. Identical predictions score exactly $(N+1)/(2N)$ among N candidates under official ties for nine rules without row-specific inputs. Scaling each prediction against its own truth raises the shared response from near 0.51 to 0.80–0.84, whereas own-truth gene weights lower it. On five Perturb-seq panels, magnitude-stratified permutations raise most null means but move no own-truth ℓ_1 score across its range. Of 78 published model–context pairs, 32 beat the mean and 28 of these exceed their global matched null (25 under strata). GEARS on one RPE1 split and scBERT, whose predictions are effectively constant, do not. Inverting an unclipped empirical-Bayes map on positive-weight coordinates restores the unshrunk ℓ_2 score within 10^{-3} in all 90 settings. Scalar recovery varies with panel, center and weights. These controls locate a score's information source without ranking corrections by biological validity.

1 INTRODUCTION

Virtual cell models predict how a cell's transcriptome changes under a genetic perturbation (Roohani et al., 2024; Lotfollahi et al., 2023; Adduri et al., 2026), including models built on pretrained transcriptome representations (Cui et al., 2024; Theodoris et al., 2023). Their evaluation must separate response accuracy from retention of perturbation identity. The perturbation discrimination score (PDS) ranks held-out true responses by distance to a prediction and locates the matching perturbation (Roohani et al., 2025; Viñas Torné et al., 2026). Benchmarks report it alongside response error (Arc Institute, 2025; Wu et al., 2025). The distinction matters because naive controls remain competitive with trained models in several comparisons (Arc Institute, 2025; Ahlmann-Eltze et al., 2025; Wong et al., 2025). Recent proposals also rescale or reweight predictions before ranking (Liu et al., 2025; Vollenweider & Bühlmann, 2026).

However, a corrected discrimination score does not reveal whether its perturbation-specific information came from the prediction or from inputs supplied during scoring. Liu et al. (2025) established that discrimination is sensitive to prediction magnitude and recommend a directional distance or a fixed prediction norm before an ℓ_p distance. Their fixed-norm implementation matches each prediction to the norm of its own true response, which a deployed model does not have. Truth used to score a prediction differs from truth used to transform it before ranking. Rescaling alone also cannot tell whether a lost score reflects changed magnitude, changed direction or discarded information.

Empirical-Bayes shrinkage makes these questions concrete. Shrinking a noisy estimate toward a pooled mean is classical (James & Stein, 1961; Stephens, 2017) and common in differential-expression analysis (Ritchie et al., 2015; Love et al., 2014). Here it maps $\hat{\delta}$ to $\bar{\delta} + B(\hat{\delta} - \bar{\delta})$, with a fitted shared response $\bar{\delta}$ and known nonnegative diagonal weights B. Unequal weights and a nonzero center can each rotate a prediction. One scalar cannot generally undo this rotation, but the known map is invertible wherever $B_g > 0$. Shrinkage therefore produces predictions whose loss of score has an explicit cause and whose recovery can be tested without test labels.

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

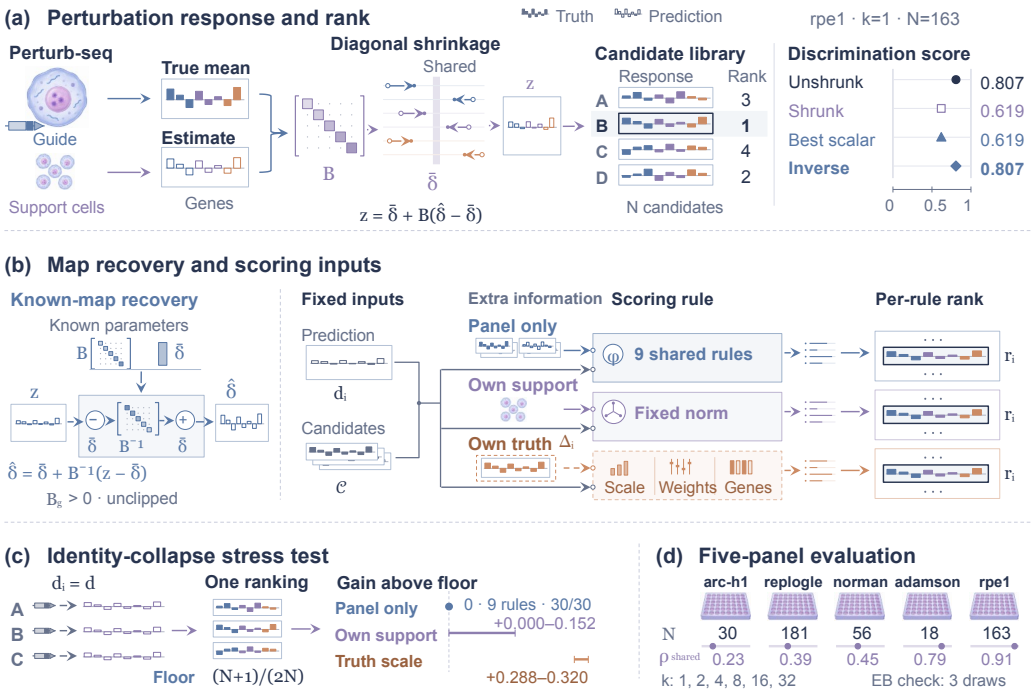


Figure 1: **ShrinkageLens distinguishes recoverable changes in predictions from information introduced during scoring. (a) Perturbation response and rank.** Filled and hollow profiles denote truth and prediction, and dark outlines mark matching candidates. On rpe1 with $k = 1$ support cell under ℓ_2, shrinkage lowers the score from 0.807 to 0.619. The best shared scalar (a tie-optimistic bound) leaves it at 0.619, and the known-map inverse restores 0.807. **(b) Map recovery and scoring inputs.** Known-map recovery is an algebraic control on unclipped genes with shrinkage weight $B_g > 0$ and leaves zero-weight genes at the shared response. Scoring rules share the prediction and candidates (neutral lines) and differ in panel-wide, own-support or own-truth access. The dashed orange branch reads the matched true response Δ_i before distances are formed. **(c) Identity-collapse stress test.** Identical predictions yield the exact floor $(N + 1)/(2N)$ under official ties, and nine panel-only rules keep it in 30 settings. Bars are ranges across settings. The own-truth gain is for scale selection on the fitted shared response (Table 1). **(d) Five-panel evaluation.** Five estimators on five panels and six support sizes, with the unshrunk versus empirical-Bayes comparison on three draws. N is the number of candidates. Icons, profiles, matrices, distances and ranks are schematic.

To separate the prediction's information from information introduced during scoring, we introduce ShrinkageLens, an information-source audit organized by what each scoring rule reads (Figure 1). Giving every perturbation the same prediction yields the exact official chance level $(N + 1)/(2N)$ for any rule that reads nothing specific to one perturbation. Choosing each prediction's scale against its own truth instead raises the fitted shared response from a floor near 0.51 to 0.80–0.84, whereas own-truth gene weights and gene selection lower it. Matched permutations of 270 internal prediction sets and 78 published model–context pairs then test directions while retaining either the bundled or the receiving perturbation's magnitude, under global and magnitude-stratified references. Of the 32 published pairs that beat the mean on the same targets, 28 also exceed their global matched null range and 25 their stratified range. GEARS on RPE1's second split, whose null reaches 95% of its gain, and scBERT, whose predictions are effectively constant, do not. Finally, inverting the known, unclipped empirical-Bayes map on its positive-weight coordinates requires no test labels and restores the unshrunk ℓ_2 score within 10^{-3} in all 90 settings. This is an algebraic positive control, not a general inverse for unknown maps. Scalar remedies instead recover a share of the gap that falls from 95.3% to 1.3% as the shared response comes to dominate, across five panels. The ordering holds under ℓ_2 on these panels and is not a general law.

Our contributions are as follows:

- We introduce controls that audit each scoring rule's information sources. Identity-free predictions give an exact official chance reference, and rules that read the receiving perturbation's support or truth depart from it in both directions (Sections 3 and 4).
- We construct matched bundle and direction permutations, global and magnitude-stratified, and apply them to 270 internal prediction sets and to every usable method in a published benchmark archive. Of the 32 published gains over the mean, 28 exceed the global matched null and 25 the stratified one (Section 5).
- We show that a label-free inverse of the known empirical-Bayes map restores the ℓ_2 score within 10^{-3} in all 90 settings and the exact rank in 7,980 of 8,064 perturbation-settings. Scalar recovery, by contrast, varies with panel, center and weights (Sections 6–7).

2 Related work

Discrimination and shared responses. The shared response is also an established object of analysis. Systema measures the cosine between each perturbation's shift and the average effect (Viñas Torné et al., 2026), and Wang et al. (2026) use squared cosine to that response. Our norm ratio ρ_{shared}, the norm of the shared response over the mean perturbation-specific spread (Equation (5)), measures the same component to describe scalar recovery after shrinkage. A Systema-style cosine orders these five panels differently (Appendix T). Neither statistic makes five panels a test of a general law.

Magnitude sensitivity and its remedies. Liu et al. (2025) is the closest work. They show that ℓ_1 and ℓ_2 discrimination differ from cosine after norm matching, explain why, and recommend directional distance or a fixed prediction norm. We take this sensitivity as established. Their forward scaling family does not invert an estimator's map, and their implementation uses an own-truth norm. We add controls that identify the information each remedy introduces and matched references that condition on it, not a new account of scale.

Related evaluation diagnostics. Marrakchi et al. (2026) propose a classifier discrimination score with a different order of averaging, whereas we retain the score and examine corrections to its inputs. A protocol taxonomy catalogs signal-recovery controls, including the shared training mean (Schäfer et al., 2026), for which Proposition 2 supplies an exact reference. Vollenweider & Bühlmann (2026) calibrate scores between replicate and control bounds and weight them by each perturbation's responsive genes. Nicol et al. (2026) show that shared controls inflate correlation for uninformative predictions.

Known maps and nuisance transformations. The empirical-Bayes map makes the estimator's action explicit, and published-model predictions provide a separate application that assumes no map (Appendix L). Neither covers all large perturbation models (Miladinović et al., 2025), conditional generators (Lotfollahi et al., 2019; Piran et al., 2024) or transport methods (Bunne et al., 2023b;a; Chi et al., 2026). Quotienting a nuisance group is standard in representational analysis, where excessive invariance can make a similarity measure degenerate (Kornblith et al., 2019), and maximization can inflate reported performance (Smith & Winkler, 2006; Gröger et al., 2026). We therefore restrict optimization to a scalar and report it as a bound. Proposition Q.1 shows why unrestricted positive diagonal invariance discards within-sign magnitude information.

3 Scores and the information they read

Official and tie-optimistic scores. Let $\mathcal{C}$ be the held-out perturbation panel, $|\mathcal{C}| = N$, with true pseudobulk responses $\Delta_i \in \mathbb{R}^G$ measured relative to control. A model predicts d_i for perturbation i. For a distance $\rho(\cdot,\cdot)$, the official scorer deterministically sorts the N values $\rho(d_i, \Delta_j)$ over alphabetically listed candidates. Identical rows therefore have identical orderings. The scorer records the true candidate's zero-based position r_i. The score is

$$\mathrm{PDS}_i \;=\; 1 - \frac{r_i}{N}, \qquad \mathrm{PDS} = \frac{1}{N}\sum_i \mathrm{PDS}_i, \tag{1}$$

Table 1: **A scoring rule's inputs determine whether an identity-free prediction stays on the chance floor.** Excess is the score of an identity-free prediction minus $(N+1)/(2N)$, over 30 settings (five panels, $k \in \{1, 32\}$, three constant vectors) or the count in parentheses, computed with Equation (2). Under the official Equation (1) the shared-scalar excess is zero. Ranges span settings and are not uncertainty intervals. Own-truth rows are evaluation-side diagnostics or bounds, not deployable corrections (Appendix I).

	additional input read			
Scoring rule	panel	own support	own truth	excess over floor
Raw ℓ_1 or ℓ_2, cosine	no	no	no	0
Common norm, panel centering, gene standardization	yes	no	no	0
Hubness correction, per-candidate scale	yes	no	no	0
One least-squares scalar fitted to all truths	yes	no	no	0
Best shared scalar, $\overline{\mathrm{PDS}}^{\star}_{\mathrm{glob}}$	yes	no	no	$1/N^2$
Norm of own unshrunk support estimate	no	yes	no	$+0.000$ to $+0.152$
Best scale per prediction, $\overline{\mathrm{PDS}}^{\star}_{\mathrm{pred}}$	yes	no	yes	$+0.138$ to $+0.325$
Own-truth norm (published fixed norm)	no	no	yes	$+0.000$ to $+0.165$
Own-response gene weights (45)	yes	no	yes	-0.369 to -0.111
Own top-k differentially expressed genes (270)	yes	no	yes	-0.439 to -0.074

and official evaluations and matched nulls use it. Counting only strictly closer candidates instead gives the tie-optimistic score

$$\overline{\mathrm{PDS}}_i \;=\; 1 - \frac{\#\{j : \rho(d_i, \Delta_j) < \rho(d_i, \Delta_i)\}}{N} \;\geq\; \mathrm{PDS}_i, \tag{2}$$

which credits every tie to the prediction. The two agree whenever no candidate ties the true one. On 163 perturbations under three distances, Equation (2) matches the official package to 1.1×10^{-16} on such inputs, and constructed ties show that the package credits a tie only when its sort places the true candidate first (Appendix F). The bounds, the identity-free controls and the inverse and scalar-recovery analyses use Equation (2) because an optimized scale can lie exactly on a tie. Elsewhere the two agree to within ties that real predictions do not produce at their own scale.

Two gauge groups. The remedies under test remove magnitude in one of two ways. A single positive scalar acts on the whole prediction set, or each prediction carries its own, giving

$$\overline{\mathrm{PDS}}^{\star}_{\mathrm{glob}} = \max_{c>0} \frac{1}{N} \sum_i \overline{\mathrm{PDS}}_i(c\, d_i), \qquad \overline{\mathrm{PDS}}^{\star}_{\mathrm{pred}} = \frac{1}{N} \sum_i \max_{c_i>0} \overline{\mathrm{PDS}}_i(c_i d_i). \tag{3}$$

These answer different questions and $\overline{\mathrm{PDS}}^{\star}_{\mathrm{pred}} \geq \overline{\mathrm{PDS}}^{\star}_{\mathrm{glob}}$ always. Because $\overline{\mathrm{PDS}}_i \geq \mathrm{PDS}_i$ pointwise, each also bounds the corresponding maximum of the official score. We report the uncorrected score, the train-only scale and these bounds separately. Maximization over the candidate library is not credited to a method. Cosine and fixed-norm rescaling act per prediction, whereas the global group asks whether one factor explains the whole deficit. On rpe1 at $k = 8$, the reported score is 0.8705, the global bound 0.8721 and the per-prediction bound 0.9497.

Proposition 1 (Finite-scale closed form under ℓ_2)**.** *Let ρ be the Euclidean distance. Along an affine path $d_i(\beta) = a_i + \beta v_i$, each squared-distance difference to the true candidate is affine in β. Each nonconstant difference has one root, giving at most $N-1$ breakpoints per prediction. Both maxima in Equation* (3) *are attained by checking the positive breakpoints and one point in each intervening interval. The global maximum uses the union of all predictions' breakpoints.*

The quadratic terms cancel, leaving a constant and a linear term. The constant fixes finite-scale crossings and disappears in the $c \to \infty$ limit analyzed by Liu et al. (2025). Under ℓ_1 the differences are only piecewise affine, and we report no ℓ_1 maxima. Unless crossings in opposite directions coincide, a breakpoint's value is also attained on an adjacent open interval. No candidate ties the true one there, so the two scores agree. A search of all 270 prediction sets finds no such coincidence behind any reported maximum, so each reported bound is also an official maximum (Appendix G).

Task, estimators and data. Each held-out perturbation supplies k perturbed support cells, from which an estimator predicts its response, and 32 disjoint query cells whose pseudobulk is its true response Δ_i. We compare the unshrunk k-shot mean, empirical-Bayes affine map, affine optimal transport (Makkuva et al., 2020; Bunne et al., 2023b), rank-limited projection and a trained neural pooler (Appendix A). Five Perturb-seq panels (Adamson et al., 2016; Norman et al., 2019; Replogle et al., 2022; Roohani et al., 2025) and $k \in \{1, 2, 4, 8, 16, 32\}$ give 150 first-draw estimator settings. The unshrunk and empirical-Bayes comparison is repeated on three draws, giving 90 paired settings (Appendix V). These settings reuse held-out perturbations. The cross-panel evidence comprises five panels, not 90 biological replicates. Within-panel intervals resample query perturbations against the fixed candidate library, preserving method pairing. They do not cover a new library, refitted estimators or new screens. Data, scoring requirements, per-setting values, support-limit and batch sensitivities, reproduction and configurations are in Appendices C–E, H, X and Y.

4 THE IDENTITY-FREE INPUT AND WHAT EACH RULE READS

Giving every held-out perturbation the same vector tests whether scoring adds perturbation-specific information. It carries no identity, so any score above chance comes from the rule.

Proposition 2 (Evaluations that read nothing specific to one row sit on the floor)**.** *Score perturbation i against candidate j through $\phi(d_i, \Delta_j; \mathcal{C}, \mathcal{D})$, where ϕ may read the whole candidate panel $\mathcal{C}$ and the whole multiset of predictions $\mathcal{D}$ but not the index i. Let i score $1 - r_i/N$ as in Equation* (1)*, with ties broken by a deterministic rule that does not depend on i. If $d_i = d$ for every i, the mean score is exactly $(N{+}1)/(2N)$. The tie-optimistic score of Equation* (2) *returns $(N{+}1)/(2N) + T/N^2$ instead, with T the number of tied pairs among the N values.*

Every row is then the same, so one sorted order serves all N rows and the true candidates occupy its positions $0, \ldots, N{-}1$ once each. Equation (2) counts only strictly closer candidates, and summed over i these count each of the $N(N{-}1)/2 - T$ pairs of unequal values once. As the positions form a permutation, recall at m is m/N and mean reciprocal rank is H_N/N, with H_N the Nth harmonic number. The statement thus covers any score of the true candidate's position, ties included.

Nine of the 15 rules in Table 1 read nothing specific to the receiving perturbation and sit at the exact floor in all 30 identity-free settings. A shared scalar chosen to maximize the score belongs to the same class once chosen, and its $1/N^2$ under Equation (2) is one credited tie (Sun et al., 2020).

A rule leaves this class when it reads something specific to the receiving perturbation. Support cells can legitimately supply such information, and support-norm correction raises the constant-prediction score by up to 0.152. Own truth instead supplies evaluation-side information, and per-prediction scale selection adds 0.288–0.320 on the fitted shared response. Such access can also lower the score. Own-response gene weights, as in a recent weighted score (Vollenweider & Bühlmann, 2026), lower it in all 45 settings, and own top-k differentially expressed genes do so in all 270. Sharing the weights or selected genes across the panel restores the floor. Reading the truth is not itself a violation, since every evaluation reads it to score. The distinction is whether the receiving perturbation's truth shapes its prediction or comparison before candidates are ranked. Such a result is an evaluation-side diagnostic, not a deployable score, and is a bound only where the truth is optimized over. These controls identify the source of scoring information and do not measure which correction best captures biology. The per-prediction bound changes the top estimator in 11 of 30 settings, all on the two smallest panels, without showing which ranking is correct (Appendix J).

5 MATCHED PERMUTATIONS OF REAL PREDICTIONS

The identity-free input calibrates a rule, not a prediction set. For real predictions we compare the observed score with 2,000 paired permutations that reassign predictions across receiving perturbations. Let u_i denote the direction of prediction i and s_i its norm, or its support norm under support-norm correction. A permutation π gives receiving perturbation i the prediction $s_{\pi(i)}u_{\pi(i)}$ in the bundle null and $s_i u_{\pi(i)}$ in the direction null. The latter retains the receiving magnitude and any support information it carries. Own-truth norms always read the receiving answer, so the two nulls coincide for them (Appendix K). We apply both to 270 internal prediction sets, five estimators at six support sizes on the first draw and two estimators on two further draws. Figure 2 shows one prespecified

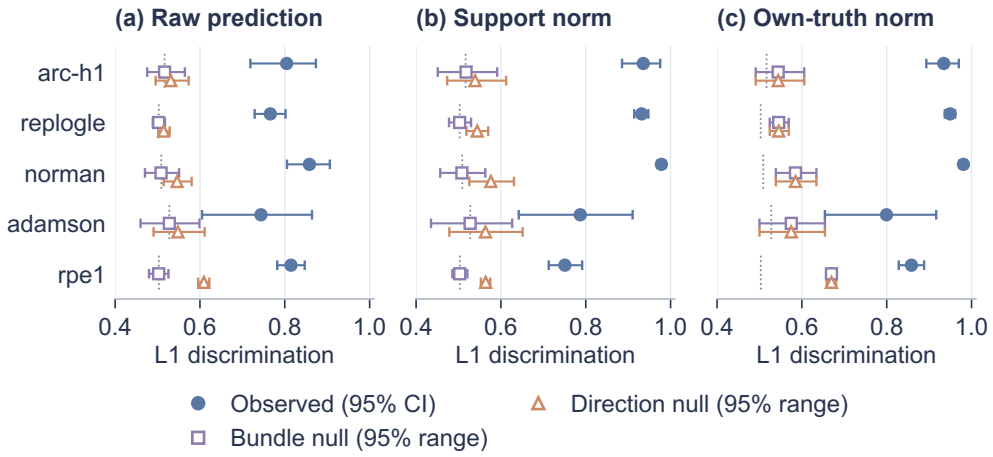


Figure 2: **Matched permutations distinguish directional information from retained magnitudes.** Empirical Bayes at $k = 8$, first support draw, on all five panels. Filled points show observed ℓ_1 scores with 95% intervals from resampling query perturbations against the fixed candidate library. Open squares and triangles show means and central 95% ranges of 2,000 paired permutations, not confidence intervals. Dotted ticks mark $(N + 1)/(2N)$. Bundle permutations move predictions and support norms together. Direction permutations retain the receiving magnitude. Own-truth norms read the receiving answer under both nulls, so those null distributions coincide. Appendix K gives all estimators and support sizes, and Appendix K.1 the magnitude-stratified references.

Table 2: **Most published gains over the mean also exceed the matched null.** Own-truth ℓ_1 classes of 78 model–context pairs, each on its own library against the mean on the same targets. A: beats the mean (paired 95% interval above zero) inside the central 95% range of 2,000 global direction permutations. B: beats it above that range. C: does not beat it. $*$: inside the stratified range. Pert.: perturbation embedding. Coverage and other rules are in Appendix M.

	Adamson		K562		RPE1			Adamson		K562		RPE1	
Model	1	2	1	2	1	2	Linear model	1	2	1	2	1	2
GEARS	B	B	B	B	B	A	Self-trained	C	C	B	B	B	B
scGPT	B	C	C	B	C	C	GEARS pert.	C	C	B	B	B	B
CPA	C	C	C	C	C	C	K562 pert.	B*	B	B	B	B	B
Geneformer	C	C	C	C	C	B	RPE1 pert.	B*	B*	B	B	B	B
scBERT	C	C	A	A	C	C	Random pert.	C	C	C	C	A	C
UCE, 4 or 33 layers	C	C	C	C	C	C	Random gene	C	C	C	C	C	C

support size and draw on all five panels. For empirical Bayes on rpe1, support-norm correction scores 0.751, with null means 0.503 when the bundle moves and 0.564 when the receiving support norm remains. Own-truth norm correction scores 0.858 while its null mean is already 0.670. Conversely, on adamson at $k = 1$ the support-norm score of 0.608 lies inside the direction-null range 0.503–0.611. Across settings, own-truth ℓ_1 null means range from 0.508 to 0.671, above the chance floor throughout. A score above the constant floor therefore cannot, by itself, identify the input that supplied discrimination. We define no null-subtracted metric.

Published-model predictions. Published predictions for unseen perturbations (Ahlmann-Eltze et al., 2025), made without support cells, keep their source partitions and are not retrained or inverted. We score the 1,000 genes with the highest control expression for the mean and for all 13 methods in that archive whose predictions cover these genes, which excludes two linear models. Scoring each method on its own finite test library against the mean on the same targets gives 78 model–context pairs from three biological screens (Appendices L and M). Under own-truth ℓ_1, 32 pairs beat the mean with a paired 95% interval above zero, and 28 of them also lie above their global

direction-null range (Table 2). Under raw and training-fixed ℓ_1, 30 of 31 and 31 of 32 gains lie above their global range, and the remaining raw gain lies below it. Most published gains over the mean therefore come with resolved directional identity. The two comparisons still answer different questions, because 7 of the 46 pairs that do not beat the mean lie above their null range. Training-fixed normalization, which reads no held-out response, changes the original four methods' raw ℓ_1 scores by -0.0671 to $+0.0444$ on their common libraries and is not uniformly better.

Magnitude-stratified references. A global permutation need not preserve structure shared by similar perturbations. If strong responders share a direction, a prediction that knows only whether its perturbation responds strongly could clear a global null. As a sensitivity analysis, specified after the global results and before any stratified draw, we repeat both nulls with permutations restricted to three strata of similar true-response norm (Appendix K.1). Stratification raises most null means, by a median of 0.050 for support-norm and 0.035 for own-truth ℓ_1 across the internal sets. Under ℓ_1, however, no internal observation changes side. Support-norm and own-truth scores still exceed the upper direction-null quantile in 269 of 270 internal sets. The only internal changes, two $k = 1$ settings under ℓ_2 or cosine, move inside the stratified range. The three published own-truth ℓ_1 pairs marked in Table 2, whose libraries hold 8 to 13 targets, also move inside. This leaves 7 of the 32 published gains inside, compared with 4 globally. Separation of the internal sets above the global ℓ_1 nulls is therefore not explained by the three magnitude strata tested.

Using the controls. Two published-model exceptions separate mean-beating gains from directional evidence. On RPE1's second split, GEARS exceeds the mean on all 384 targets by 0.0223 under own-truth ℓ_1 (95% paired interval $[0.0145, 0.0302]$), so a comparison with the mean alone would credit it with identifying perturbations. Its direction null, which reassigns its own directions at the receiving norms, already reaches 95% of that gain, with mean 0.7846 against an observed 0.7856. scBERT is the limiting case of mean-beating without resolved directional identity. Its predictions are effectively constant, with pairwise cosines of 1 to nine decimals and a norm coefficient of variation below 4×10^{-8}. Its raw ℓ_1 score equals $(N + 1)/(2N)$ to machine precision, and its null equals its score. Under own-truth normalization, which supplies each target's measured norm, that single direction still beats the mean on both K562 splits, by 0.0012 and 0.0040. Null overlap leaves directional identity unresolved. It does not establish that biological effects are absent.

6 THE LABEL-FREE INVERSE OF A KNOWN MAP

For the empirical-Bayes map the action is $z = \bar{\delta} + B(\hat{\delta} - \bar{\delta})$ with B a known nonnegative diagonal matrix derived in Appendix B. Depending on the dataset, 0.00% to 13.10% of its diagonal entries are exactly zero. On the coordinates where $B_g > 0$ the map is a bijection and

$$\hat{\delta}_g \;=\; \bar{\delta}_g + (z_g - \bar{\delta}_g)/B_g, \qquad B_g > 0, \tag{4}$$

is exact. On coordinates with $B_g = 0$ the shrunk prediction is already $\bar{\delta}_g$, and the inverse leaves them at the shared response. The center $\bar{\delta}$ and the variances in B are fitted on training perturbations and control cells, and B depends only on the support size k, so Equation (4) uses no test labels. We apply it to the unclipped prediction, since the clip is not invertible, and report the clip's cost separately (Appendix O). The inverse is a recoverability control. That a known map can be undone does not imply that a fixed-coordinate distance should be invariant to it, and a score may legitimately penalize shrinkage that a deployed model would output.

Figure 3d and Tables N.1 and N.3 give the inverse results by setting. Under ℓ_1, used by the Virtual Cell Challenge, and cosine, all 30 first-draw settings recover within 6.4×10^{-4} and 3.2×10^{-4}, respectively. Under ℓ_2, all 90 settings across three draws remain below 10^{-3}. Each score averages discrete ranks, so a residual moves in steps of $1/N^2$, and across 8,064 perturbation-settings, 7,980 ranks recover exactly and none moves by more than two (Appendix N). Recovery persists across nearly four orders of magnitude in conditioning and up to 13.10% zero-weight coordinates. Their small effect on these ranks is a measurement on five panels, not evidence that they carry no signal (Table N.2). Failure to invert likewise does not imply information loss. On official ℓ_2 scores from clipped cells, a rank-limited projector outranks the empirical-Bayes estimator in 60 of the 90 settings. At adamson with $k = 1$ the shrunk prediction scores 0.5370, the inverse restores the unshrunk 0.7747, and a rank-15 projector still outranks the shrunk prediction. That comparison concerns the directions each estimator retains, not invertibility (Appendices G and V).

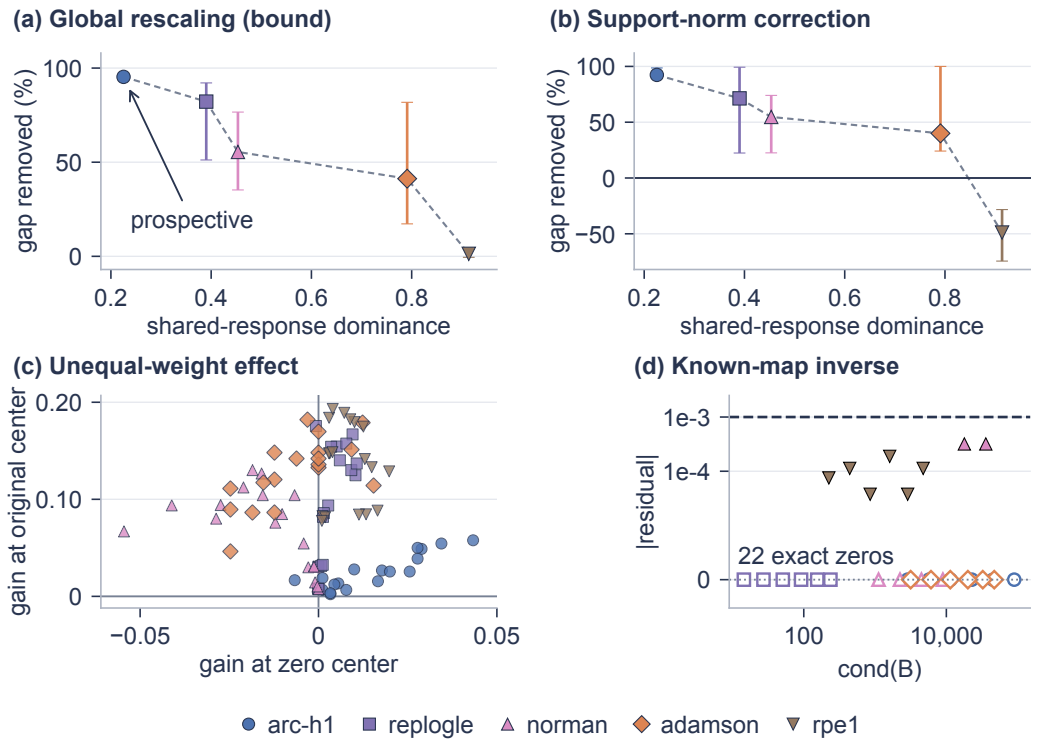


Figure 3: **Scalar recovery depends on the panel and fitted map. The known-map inverse restores the score.** First support draw in (a), (b) and (d), ℓ_2 throughout. **(a)** The global bound $\overline{\mathrm{PDS}}^{\star}_{\mathrm{glob}}$ removes 95.3% of the gap at the smallest ρ_{shared} and 1.3% at the largest. The arrowed point was predicted before arc-h1 was scored (Appendix V). Only the extremes keep their order under query resampling (Appendix W). **(b)** Support-norm correction follows the same ordering and turns negative on rpe1. Marks in (a) and (b) are medians over settings with $|\mathrm{gap@1}| > 0.02$. Bars span those settings. **(c)** Replacing scalar weights $\bar{B}I$ by the diagonal B raises the tie-optimistic bound at the original center in all 90 settings but has both signs at zero center. **(d)** Inverse residuals against $\mathrm{cond}(B)$, the ratio of the largest to the smallest positive weight, all below the dashed 10^{-3} line, with exact zeros on the floor. Panels, not support sizes, are the units of the ordering.

Scope of map recovery. Only the empirical-Bayes map is inverted here. A pre-registered fit on training pairs, which regresses the estimator's output on a noisy held-out response and not on the map's own input, had to reach rank correlation ≥ 0.9 with the known map in all 30 settings. It reached it in 23, missing adamson $k = 1$ and every norman support size, and the cause remains unresolved (Appendix P.1). As tested, corrections that ignore the map do not restore the score, whether per-gene least-squares rescaling or whitening on rpe1 at $k = 8$ or denoising inside the response subspace (Appendix P). Invariance to positive gene-wise rescaling is no substitute. It would retain only signs (Proposition Q.1), equating predictions whose cosines to truth are 0.998 and 0.756 (Appendix Q.1), whereas cosine removes only a scalar magnitude (Appendix Q).

7 Scalar recovery after shrinkage

One global factor suffices to undo the map when $\bar{\delta} = 0$ and $B = bI$ with $b > 0$. Unequal weights or a large shared response rotate the prediction, which a scalar generally cannot undo (Figure 1a and Appendix T). The quantity that interpolates between these cases is shared-response dominance,

$$\rho_{\mathrm{shared}} = \frac{\|\bar{\delta}\|}{\mathrm{mean}_{p\in\mathcal{C}}\|\Delta_p - \bar{\delta}\|}, \tag{5}$$

with the numerator fitted on training perturbations and the denominator measured over the held-out panel, so it needs the published candidate library but no prediction (Appendix R).

Recovery. With g_1 the shrunk-minus-unshrunk ℓ_2 score gap and g_c the same gap after correcting each arm separately, we measure recovery as the absorbed fraction $1 - |g_c|/|g_1|$, taking the median over settings with $|g_1| > 0.02$ (Appendix S). The global bound $\overline{\mathrm{PDS}}^{\star}_{\mathrm{glob}}$ absorbs 95.3% of the gap at $\rho_{\mathrm{shared}} = 0.23$ and 1.3% at $\rho_{\mathrm{shared}} = 0.91$. The point estimates are strictly ordered across the five panels on each of three support draws (Figure 3a, Table R.1 and Appendix V). Only the extremes are stable. Jointly resampling the query perturbations, with each arm's bound reoptimized, keeps the strict order in 41–51% of replicates per draw and 20% on all three. The rpe1 panel stays lowest in at least 97.9% (Appendix W). Matching each prediction to the norm of its unshrunk support estimate reads no test labels. On the first draw it ranges from closing 92.4% of the gap to widening it by 48.9% in the same order. Only rpe1 worsens at all six support sizes (Figure 3b). The train-only scalar $1/\bar{B}(k)$, with $\bar{B}(k)$ the mean fitted weight at support size k, worsens the score in 14 of 30 settings. In a 2×2 design replacing the center by zero and the weights by their scalar mean, unequal weights raise the global bound in all 90 settings at the original center but have both signs at zero center. The deficit therefore has no additive center and weight shares (Figure 3c and Appendix U).

Scope. Before arc-h1 was scored, its $\rho_{\mathrm{shared}} = 0.225$ predicted the largest global recovery, and 95.3% was observed, although the registration was not a fully specified test (Appendix V). Support-norm recovery is ordered on two of the three draws. A train-only denominator breaks the relation on all three (Appendix R.1). The within-panel bootstrap interval for the ρ_{shared}–cosine-gap association is $[0.00, 1.00]$, or $[0.30, 1.00]$ with norman blocked by target (Appendices W and W.1). The ordering therefore describes five panels under ℓ_2 and is not a general law. Its ℓ_1 counterpart remains unresolved (Appendix W.2).

8 Limitations

The ordering rests on five panels, and only its extremes are stable to query resampling. The inverse requires a known map. Support and query cells come from the same experiment, so batch structure removed by shrinkage would make the deficit overstate the biological signal lost. On arc-h1, the only panel with batches, a batch-disjoint draw changes the deficit by less than the resolution of three draws at four of five sizes. The change exceeds that resolution at the fifth (Appendix X). Except for official scores, we score unclipped predictions. The pipeline's cellwise clipping is not invertible, costs estimators differently and changes ℓ_1 ranks that a common pseudobulk floor preserves in all 240 eligible prediction sets. Affine OT cannot enter this control because it clips internally (Appendices O and O.1). The data are five datasets from one species and three cell lines, K562 in three of them. The external evaluation reuses three screens, six of its methods share the linear architecture, and missing genes or predictions exclude two methods and restrict several libraries (Appendix M). Global and stratified permutation ranges are conditional references, not a test of population-level exchangeability. No perturbation-level batch or class strata were available (Appendix K.1). The support grid is discrete. The center–weight intervention tests one same-mean scalar replacement, not a family of gene weights (Appendix U). Conclusions about the rank-limited projector rest on one architecture at one rank, and Proposition 2 says little about partial collapse (Appendix I). We neither claim the score is flawed nor propose replacing it, and its scale sensitivity was established by Liu et al. (2025).

9 Conclusion

ShrinkageLens audits whether corrected discrimination reflects perturbation-specific information from predictions or scoring inputs. Identical predictions attain exactly $(N + 1)/(2N)$ under official ties when scoring reads nothing specific to one perturbation. A score far above chance, or above the mean, can still lie inside its matched null. Known-map inversion restores the unshrunk ℓ_2 score within 10^{-3} in all 90 settings. Scalar recovery depends on the panel, center and weights. We recommend declaring what each correction reads, reporting uncorrected and deployable scores with matched nulls, reporting inverse residuals and non-invertible coordinates for known maps, and treating truth-dependent optimization as an upper bound on the evaluated candidate library.

ETHICS STATEMENT

This study analyzes publicly released Perturb-seq cell-line data and published model predictions. It involves neither participant recruitment nor collection of new human or animal samples. The results concern prediction evaluation and do not establish clinical validity or the safety of biological interventions. We report exclusions, unmet criteria and adverse results to limit overinterpretation. Reuse should follow the source datasets' access conditions and applicable licenses. Biological icons are schematic illustrations, not experimental evidence.

REPRODUCIBILITY STATEMENT

The code supplement provides estimator implementations, scoring and analysis routines, figure scripts, tests, pinned dependencies and public-data reproduction commands. It excludes datasets and prediction arrays. Data sources, preprocessing, partitions, leakage checks and configurations are specified in Appendices C, D and Y.1. Assumptions and derivations accompany the theoretical claims in Sections 3, 6 and 4 and Appendices B, F and Q. A fresh public-data run in an isolated environment reproduced all 90 primary settings and 450 official evaluations. All 90 settings met the leakage checks. Primary comparisons agreed within 3.1×10^{-16}. Appendix Y reports cross-environment agreement and differences in official metrics. The published-model protocol and coverage analyses are in Appendices L and M. Appendix analyses outside the 90 primary settings were not repeated in that run. Their negative results and unmet criteria remain reported.

AI USE STATEMENT

We used large language models solely for code implementation assistance, manuscript copy-editing and polishing, and generating conceptual schematic icons. All experimental workflows, mathematical formulations, data analyses and scientific claims were designed, conducted and verified by the authors.

REFERENCES

Britt Adamson, Thomas M. Norman, Marco Jost, Min Y. Cho, James K. Nuñez, Yuwen Chen, Jacqueline E. Villalta, Luke A. Gilbert, Max A. Horlbeck, Marco Y. Hein, Ryan A. Pak, Andrew N. Gray, Carol A. Gross, Atray Dixit, Oren Parnas, Aviv Regev, and Jonathan 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.

Abhinav K. Adduri, Dhruv Gautam, Beatrice Bevilacqua, et al. Predicting cellular responses to perturbation across diverse contexts with State. *Cell*, 189(19):5914–5931.e20, 2026. doi: 10.1016/j.cell.2026.07.052.

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(8): 1657–1661, 2025. doi: 10.1038/s41592-025-02772-6.

Arc Institute. Virtual Cell Challenge 2025 wrap-up: Winners and reflections. `https://arcinstitute.org/news/virtual-cell-challenge-2025-wrap-up`, 2025.

Charlotte Bunne, Ya-Ping Hsieh, Marco Cuturi, and Andreas Krause. The Schrödinger bridge between Gaussian measures has a closed form. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 5802–5833, 2023a.

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(11):1759–1768, 2023b. doi: 10.1038/s41592-023-01969-x.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(25):20544–20552, 2026. doi: 10.1609/aaai.v40i25.39190.

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

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(8):1470–1480, 2024. doi: 10.1038/s41592-024-02201-0.

Fabian Gröger, Shuo Wen, and Maria Brbić. Revisiting the Platonic representation hypothesis: An Aristotelian view. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

W. James and C. Stein. Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pp. 361–379, 1961.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529, 2019.

Qiyuan Liu, Qirui Zhang, Jinhong Du, Siming Zhao, and Jingshu Wang. Effects of distance metrics and scaling on the perturbation discrimination score. *arXiv preprint arXiv:2511.16954*, 2025.

Mohammad Lotfollahi, F. Alexander Wolf, and Fabian J. Theis. scGen predicts single-cell perturbation responses. *Nature Methods*, 16(8):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(6):e11517, 2023. doi: 10.15252/msb.202211517.

Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, 2014. doi: 10.1186/s13059-014-0550-8.

Ashok Makkuva, Amirhossein Taghvaei, Sewoong Oh, and Jason D. Lee. Optimal transport mapping via input convex neural networks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6672–6681, 2020.

Youssef Marrakchi, Davide D'Ascenzo, and Sebastiano Cultrera di Montesano. Score distributions, not cells: Evaluating single-cell perturbations under class overlap. *arXiv preprint arXiv:2607.04595*, 2026.

Djordje Miladinović, Tobias Höppe, Mathieu Chevalley, Andreas Georgiou, Lachlan Stuart, Arash Mehrjou, Marcus Bantscheff, Bernhard Schölkopf, and Patrick Schwab. In silico biological discovery with large perturbation models. *Nature Computational Science*, 5(11):1029–1040, 2025. doi: 10.1038/s43588-025-00870-1.

Phillip B. Nicol, Shriya Shivakumar, and Rafael A. Irizarry. Spurious correlation inflates performance in single-cell perturbation prediction. *bioRxiv*, 2026. doi: 10.64898/2026.05.07.723486.

Thomas M. Norman, Max A. Horlbeck, Joseph M. Replogle, Alex Y. Ge, Albert Xu, Marco Jost, Luke A. Gilbert, and Jonathan S. Weissman. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science*, 365(6455):786–793, 2019. doi: 10.1126/science.aax4438.

Zoe Piran, Niv Cohen, Yedid Hoshen, and Mor Nitzan. Disentanglement of single-cell data with biolord. *Nature Biotechnology*, 42(11):1678–1683, 2024. doi: 10.1038/s41587-023-02079-x.

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

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

Matthew E. Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47, 2015. doi: 10.1093/nar/gkv007.

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

Yusuf H. Roohani, Tony J. Hua, Po-Yuan Tung, et al. Virtual Cell Challenge: Toward a Turing test for the virtual cell. *Cell*, 188(13):3370–3374, 2025. doi: 10.1016/j.cell.2025.06.008.

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.

James E. Smith and Robert L. Winkler. The optimizer's curse: Skepticism and postdecision surprise in decision analysis. *Management Science*, 52(3):311–322, 2006. doi: 10.1287/mnsc.1050.0451.

Matthew Stephens. False discovery rates: a new deal. *Biostatistics*, 18(2):275–294, 2017. doi: 10.1093/biostatistics/kxw041.

Zhiqing Sun, Shikhar Vashishth, Soumya Sanyal, Partha Talukdar, and Yiming Yang. A re-evaluation of knowledge graph completion methods. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5516–5522, 2020. doi: 10.18653/v1/2020.acl-main.489.

Christina V. Theodoris, Ling Xiao, Anant Chopra, Mark D. Chaffin, Zeina R. Al Sayed, Matthew C. Hill, Helene Mantineo, Elizabeth M. Brydon, Zexian Zeng, X. Shirley Liu, and Patrick T. Ellinor. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, 2023. doi: 10.1038/s41586-023-06139-9.

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(6):1050–1059, 2026. doi: 10.1038/s41587-025-02777-8.

Michael Vollenweider and Peter Bühlmann. Signal, bounds, and baselines: Principles for evaluating virtual cell perturbation models. *bioRxiv*, 2026. doi: 10.64898/2026.04.20.719650.

Xi Wang, Jack Kuipers, Florian Hugi, Randall J. Platt, and Niko Beerenwinkel. Reliable single-cell perturbations explain and improve model performance. *bioRxiv*, 2026. doi: 10.64898/2026.08.11.744177.

Daniel R. Wong, Abby S. Hill, and Rob Moccia. Simple controls exceed best deep learning algorithms and reveal foundation model effectiveness for predicting genetic perturbations. *Bioinformatics*, 41(6):btaf317, 2025. doi: 10.1093/bioinformatics/btaf317.

Yan Wu, Esther Wershof, Sebastian M. 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, pp. 106937–106977, 2025. doi: 10.52202/085713-3225.

APPENDIX

Following the main-text order, Appendices A–H define the task, estimators, data, scoring requirements and official metrics, and verify the scoring convention and quotient bounds (Section 3). Appendices I and J report the identity-free controls (Section 4). Appendices K–M cover matched permutations and the published-model application (Section 5), and Appendices N–Q the known-map inverse and its limits (Section 6). Appendices R–W address scalar recovery and shared-response dominance (Section 7). Appendix X reports the batch sensitivity and Appendix Y reproduction and compute.

CONTENTS

A Task and estimator definitions

B Derivation of the empirical-Bayes weight

C Datasets and protocol

D Scoring requirements and leakage checks

E The official metric suite

F Verification of the scoring convention

G Criteria for the scalar-quotient bound

H Exact values by support size

I Truth-dependent corrections of identity-free predictions

J Estimator ranking changes by dataset

K Matched permutations of real predictions

L Application to published external-model predictions

M Matched references for every published method

N Per-point inverse recovery

O Effect of cellwise clipping

P Corrections that do not use the known map

Q Invariance of candidate rankings

R Shared-response dominance in full

S Deployable corrections and oracle bounds by setting

T Directional changes from unequal gene weights

U Separating the shared center and unequal weights

V Support-draw robustness

W Resampling stability of the scalar-recovery ordering

X Cross-batch holdout

Y Reproduction across execution environments

A TASK AND ESTIMATOR DEFINITIONS

The task and five estimators are defined below, including the inputs available to each estimator.

Estimators. The five estimators use the same held-out support cells but pool information differently. The unshrunk k-shot mean, labeled naive k-shot mean in the tables, applies a mean shift, whereas affine OT also estimates a support covariance. EB-affine shrinks the mean shift toward a training-perturbation prior. Projection uses a fixed train-only response subspace, and the trained pooler learns a nonlinear map from the same support statistic. The first three define the closed-form path below. The latter two are information-matched comparators configured as described in Appendix Y.1. Projection takes the top $r = \min(20, n_{\text{train}} - 1)$ right singular vectors V of the training perturbations' mean shifts centered at their mean $\bar{\delta}$ and predicts $\bar{\delta} + V^\top V(\hat{\delta} - \bar{\delta})$. The trained pooler predicts $\bar{\delta} + V^\top f(V(\hat{\delta} - \bar{\delta}))$, with f a network trained on four k-cell draws from each training perturbation that has at least $k + 2$ cells, each draw's mean paired with the mean of that perturbation's remaining cells. Its early-stopping split holds out 10% of those pairs, sampled by pair, not by perturbation. Its seed does not depend on the support draw.

A.1 FEW-SHOT PERTURBATION TASK

Let μ_0 be the control-cell distribution and μ_p that under perturbation p. For a held-out perturbation, a method observes a small support set $\mathcal{S}_p$ of k perturbed cells and predicts the response of query control cells under the same perturbation. The held-out perturbation is absent from training, and support cells are disjoint from query cells. All preprocessing, including highly variable gene selection and latent projections, is fitted only on training cells.

A.2 AFFINE MAP FAMILY

Optimal transport between two Gaussians is affine. Estimating the support mean $\hat{\boldsymbol{\mu}}_p$ and covariance $\hat{\Sigma}_p$ from the k support cells, and the control statistics $\boldsymbol{\mu}_c, \Sigma_c$ from training controls, the closed-form Gaussian (Bures) OT map (Bunne et al., 2023a) is

$$T(\mathbf{x}) = \hat{\boldsymbol{\mu}}_p + A\,(\mathbf{x} - \boldsymbol{\mu}_c), \quad A = \Sigma_c^{-\frac{1}{2}} \left(\Sigma_c^{\frac{1}{2}} \hat{\Sigma}_p \Sigma_c^{\frac{1}{2}}\right)^{\frac{1}{2}} \Sigma_c^{-\frac{1}{2}}, \tag{A.1}$$

applied in a 64-component principal component analysis (PCA) space fitted on the control pool, then reconstructed to gene space. The map clips its predictions between zero and the largest value in the control pool and support cells. Both covariances are sample covariances. Eigenvalues inside the matrix square roots are floored at 10^{-4} to regularize the rank-deficient support covariance at small k. At k=1, no covariance is estimable from one cell, so the support covariance is replaced by 10^{-4} times the identity. The two endpoints of this construction give two closed-form estimators. Setting $A = I$ in gene space yields the unshrunk k-shot mean-shift map, with no contraction. The full coupling is affine OT. Both are closed-form and require no gradient training. Their held-out perturbation inputs are restricted to the support cells.

A.3 CROSS-PERTURBATION EMPIRICAL-BAYES AFFINE MAP

The maps above estimate each held-out perturbation's mean shift $\hat{\boldsymbol{\delta}}_p = \hat{\boldsymbol{\mu}}_p - \boldsymbol{\mu}_c$ from only its k support cells. At small k, this estimate is dominated by sampling noise. Perturbation responses share a common component and are therefore not independent. The empirical-Bayes map uses a cross-perturbation prior fitted only on training perturbations to shrink each held-out estimate toward that component.

Model the true per-perturbation shift as $\boldsymbol{\delta}_p \sim \mathcal{N}(\bar{\boldsymbol{\delta}}, \mathrm{diag}(\boldsymbol{\tau}^2))$, with $\bar{\boldsymbol{\delta}}$ the shared response and τ_g^2 the cross-perturbation variance of gene g. The k-shot estimate is $\hat{\boldsymbol{\delta}}_p = \boldsymbol{\delta}_p + \boldsymbol{\eta}$, $\boldsymbol{\eta} \sim \mathcal{N}(0, \mathrm{diag}(\boldsymbol{\sigma}^2)/k)$ where σ_g^2 is the within-population per-gene variance. The posterior mean, the minimum-MSE estimator of $\boldsymbol{\delta}_p$ under this model, is the per-gene empirical-Bayes (conjugate-Gaussian) shrinkage

$$\delta_{p,g}^{\mathrm{EB}} = \bar{\delta}_g + B_g\,(\hat{\delta}_{p,g} - \bar{\delta}_g), \qquad B_g = \frac{\tau_g^2}{\tau_g^2 + \sigma_g^2/k}, \tag{A.2}$$

and the predictor is $T^{\mathrm{EB}}(\mathbf{x}) = \mathbf{x} + \boldsymbol{\delta}_p^{\mathrm{EB}}$. We estimate $\bar{\boldsymbol{\delta}}$ and $\boldsymbol{\tau}^2$ from the held-in perturbations (τ_g^2 debiased for each training perturbation's own estimation noise) and $\boldsymbol{\sigma}^2$ from the control pool. Neither query cells nor held-out labels enter these estimates. Let q denote a training perturbation and Δ_q its mean shift over its n_q cells relative to the control mean. Every training perturbation with at least 8 cells receives equal weight. Here $\bar{\delta}_g$ is the mean of Δ_{qg}, σ_g^2 is the control pool's variance with one degree of freedom, floored at 10^{-6}, and $\tau_g^2 = \max\{0,\ \mathrm{Var}_q(\Delta_{qg}) - \mathrm{mean}_q(\sigma_g^2/n_q)\}$, again with one degree of freedom. There is no positive floor on τ_g^2, so a gene whose spread across perturbations does not exceed its estimation noise gets $B_g = 0$ exactly, accounting for the zero weights in Section 6. The weight $B_g \in [0, 1]$ is data-adaptive. At small k the noise σ_g^2/k dominates and $\boldsymbol{\delta}_p^{\mathrm{EB}} \to \bar{\boldsymbol{\delta}}$ (the shared response). For genes with $\tau_g^2 > 0$, $B_g \to 1$ as k grows and the estimator approaches the unshrunk k-shot mean. Genes with $\tau_g^2 = 0$ keep $B_g = 0$ and stay at the shared response for every k. The map is closed-form and requires no gradient training. Under this model, it is the conjugate-Gaussian posterior mean, the James–Stein-style shrinkage estimator for this regime (James & Stein, 1961), not a tuned heuristic.

Proposition A.1 (Posterior-mean optimality of the EB shift)**.** *Under the cross-perturbation Gaussian model $\boldsymbol{\delta}_p \sim \mathcal{N}(\bar{\boldsymbol{\delta}}, \mathrm{diag}(\boldsymbol{\tau}^2))$ with k-shot likelihood $\hat{\boldsymbol{\delta}}_p \mid \boldsymbol{\delta}_p \sim \mathcal{N}(\boldsymbol{\delta}_p, \mathrm{diag}(\boldsymbol{\sigma}^2)/k)$, the empirical-Bayes shift of Equation* (A.2) *is the Bayes estimator of $\boldsymbol{\delta}_p$ under squared-error loss, minimizing the posterior risk $\mathbb{E}[\,\|\boldsymbol{\delta}_p - \mathbf{d}\|^2 \mid \hat{\boldsymbol{\delta}}_p\,]$ over all estimators $\mathbf{d}(\hat{\boldsymbol{\delta}}_p)$, affine or not, and hence attains the smallest mean-squared error of any estimator built from the support statistic under this model. On coordinates with $\tau_g^2 > 0$, it interpolates between the prior mean $\bar{\boldsymbol{\delta}}$ ($k \to 0$) and the unshrunk support estimate $\hat{\boldsymbol{\delta}}_p$ ($k \to \infty$). On coordinates with $\tau_g^2 = 0$, it equals the prior mean.*

Appendix B gives the full proof. The posterior mean minimizes squared-error risk over all estimators of the same statistic without a Gaussianity assumption, and the diagonal conjugate-Gaussian posterior mean is Equation (A.2). The optimality is conditional on the model, the statistic and the loss. It

does not extend to a map that reads the support beyond its mean, to another loss or to a misspecified prior. In practice $\bar{\boldsymbol{\delta}}, \boldsymbol{\tau}^2, \boldsymbol{\sigma}^2$ are plug-in estimates, so the guarantee is approximate (empirical Bayes).

B Derivation of the empirical-Bayes weight

Model the per-perturbation shift as $\boldsymbol{\delta}_p \sim \mathcal{N}(\bar{\boldsymbol{\delta}}, \mathrm{diag}(\boldsymbol{\tau}^2))$ with the k-shot likelihood $\hat{\boldsymbol{\delta}}_p \mid \boldsymbol{\delta}_p \sim \mathcal{N}(\boldsymbol{\delta}_p, \mathrm{diag}(\boldsymbol{\sigma}^2)/k)$, both covariances diagonal. *Claim.* The empirical-Bayes shift

$$\delta^{\mathrm{EB}}_{p,g} = \bar{\delta}_g + B_g(\hat{\delta}_{p,g} - \bar{\delta}_g), \qquad B_g = \frac{\tau_g^2}{\tau_g^2 + \sigma_g^2/k}, \tag{B.1}$$

minimizes the posterior risk $R(\mathbf{d}) = \mathbb{E}\Big[\|\boldsymbol{\delta}_p - \mathbf{d}\|^2 \mid \hat{\boldsymbol{\delta}}_p\Big]$ over all estimators $\mathbf{d} = \mathbf{d}(\hat{\boldsymbol{\delta}}_p)$, affine or not, and so attains the smallest mean-squared error of any function of the support statistic under this model.

Step 1. The posterior mean uniquely minimizes posterior risk. This step uses no Gaussianity. Write $\mathbf{m} = \mathbb{E}[\boldsymbol{\delta}_p \mid \hat{\boldsymbol{\delta}}_p]$, which exists because $\boldsymbol{\delta}_p$ has finite second moment. For any estimator $\mathbf{d}$ that is a function of $\hat{\boldsymbol{\delta}}_p$,

$$\begin{aligned} R(\mathbf{d}) &= \mathbb{E}\big[\|(\boldsymbol{\delta}_p - \mathbf{m}) + (\mathbf{m} - \mathbf{d})\|^2 \mid \hat{\boldsymbol{\delta}}_p\big] \\ &= \mathbb{E}\big[\|\boldsymbol{\delta}_p - \mathbf{m}\|^2 \mid \hat{\boldsymbol{\delta}}_p\big] + \|\mathbf{m} - \mathbf{d}\|^2, \end{aligned}$$

where the cross term $2(\mathbf{m} - \mathbf{d})^\top \mathbb{E}[\boldsymbol{\delta}_p - \mathbf{m} \mid \hat{\boldsymbol{\delta}}_p]$ vanishes because $\mathbf{m}$ and $\mathbf{d}$ are fixed given $\hat{\boldsymbol{\delta}}_p$ and can be taken outside the conditional expectation, while $\mathbb{E}[\boldsymbol{\delta}_p - \mathbf{m} \mid \hat{\boldsymbol{\delta}}_p] = \mathbf{m} - \mathbf{m} = \mathbf{0}$. Hence $R(\mathbf{d}) = R(\mathbf{m}) + \|\mathbf{m} - \mathbf{d}\|^2 \geq R(\mathbf{m})$, with equality iff $\mathbf{d} = \mathbf{m}$ almost surely. The Bayes estimator under squared-error loss is therefore the posterior mean, uniquely among all measurable $\mathbf{d}$.

Step 2. Posterior mean under the diagonal Gaussian model. Both covariances are diagonal, so the prior, likelihood and posterior factorize across genes, and it suffices to treat one coordinate g. There the prior is $\delta_g \sim \mathcal{N}(\bar{\delta}_g, \tau_g^2)$ and the likelihood $\hat{\delta}_g \mid \delta_g \sim \mathcal{N}(\delta_g, \sigma_g^2/k)$. If $\tau_g^2 = 0$ the prior is a point mass at $\bar{\delta}_g$, the posterior mean is $\bar{\delta}_g$, and $B_g = 0$ gives the same value. For $\tau_g^2 > 0$, writing the prior precision $a = 1/\tau_g^2$ and the likelihood precision $b = k/\sigma_g^2$, conjugacy of the Gaussian family gives a Gaussian posterior with mean

$$m_g = \frac{a\,\bar{\delta}_g + b\,\hat{\delta}_g}{a + b} = \frac{\bar{\delta}_g/\tau_g^2 + \hat{\delta}_g\,k/\sigma_g^2}{1/\tau_g^2 + k/\sigma_g^2}.$$

Multiplying numerator and denominator by $\tau_g^2\sigma_g^2/k$ and collecting terms,

$$m_g = \frac{(\sigma_g^2/k)\,\bar{\delta}_g + \tau_g^2\,\hat{\delta}_g}{\tau_g^2 + \sigma_g^2/k} = \bar{\delta}_g + B_g(\hat{\delta}_g - \bar{\delta}_g),$$

the last step by the definition $B_g = \tau_g^2/(\tau_g^2 + \sigma_g^2/k)$, which is Equation (B.1) coordinatewise. Stacking the G coordinates gives $\mathbf{m} = \boldsymbol{\delta}_p^{\mathrm{EB}}$. With Step 1 this proves the claim. As $k \to 0$, $\sigma_g^2/k \to \infty$ and $B_g \to 0$, so $\boldsymbol{\delta}_p^{\mathrm{EB}} \to \bar{\boldsymbol{\delta}}$. As $k \to \infty$, $B_g \to 1$ on every coordinate with $\tau_g^2 > 0$, where $\boldsymbol{\delta}_p^{\mathrm{EB}}$ approaches $\hat{\boldsymbol{\delta}}_p$. ■

Scope. The optimality is conditional on the model and on the statistic. Step 1 applies only to the mean shift. An estimator using support information beyond its mean $\hat{\boldsymbol{\delta}}_p$ is not a function of that statistic and lies outside the argument. A different loss changes the minimizer. The parameters $\bar{\boldsymbol{\delta}}, \boldsymbol{\tau}^2, \boldsymbol{\sigma}^2$ are plug-in estimates, so the practical guarantee is approximate (empirical Bayes). The affine map can therefore overtake Equation (B.1) once the support supplies a usable covariance (Appendix H). Perturbation discrimination can worsen because the posterior mean does not optimize it.

Table C.1: **The five panels differ in library size and shared-response dominance.** n_{test} is the number of held-out perturbations, which is also the size of the candidate library that discrimination retrieves against.

dataset	cells	genes	n_{test}	ρ_{shared}
adamson	68603	5060	18	0.791
arc-h1	221193	18080	30	0.225
norman	91205	5045	56	0.453
rpe1	162733	5000	163	0.914
replogle	162751	5000	181	0.390

C DATASETS AND PROTOCOL

Table C.1 summarizes the cell, gene and candidate counts and the shared-response dominance of the five panels.

Among the five panels, the labels adamson, norman, replogle and rpe1 denote the Adamson screen (Adamson et al., 2016), Norman screen (Norman et al., 2019), and K562 and RPE1 essential-gene screens of Replogle et al. (2022), respectively, in the processed form distributed with GEARS (Roohani et al., 2024). The label arc-h1 denotes the H1 human embryonic stem cell training panel released for the Virtual Cell Challenge (Roohani et al., 2025).

The primary protocol specifies the eligibility rule, support and query partition, and candidate library for every scored run, with one exception, the response-side cross-batch control in Appendix X, which used an earlier protocol. The batch-disjoint discrimination control in that appendix uses partitions derived from the primary protocol, changing only the source of each support draw. Query cells are held at a fixed count across support sizes so that the query-side noise does not shrink with k. Support and query cell sets are disjoint by construction. Every scored setting met all five leakage checks.

Deterministic split construction. The protocol uses seeded deterministic orderings specific to the dataset and sampling purpose, with protocol seed 2027. The resulting partition is a deterministic function of the data. A perturbation label is eligible if it has at least 64 cells, 32 for the largest support and 32 for the query. The first 20% of eligible labels in seeded order, rounded up, are held out, which gives 18 of 86 on adamson, 30 of 147 on arc-h1, 56 of 279 on norman, 181 of 901 on replogle and 163 of 815 on rpe1. For each held-out label, the first 32 cells of its seeded order form the query, fixed across support sizes and draws. The support for draw d consists of the first k remaining cells in a second ordering, seeded $2027 + d$. A smaller support is therefore a prefix of a larger one. Training cells have labels that are not held out and share no atomic target with a held-out label. A combined label such as BCL2L11+BAK1 counts as both of its targets. The estimators' priors use the training perturbations with at least 8 cells, and predictions transport the first 2000 control cells in seeded order. Expression values are the processed matrices distributed with GEARS, and the arc-h1 panel, rebuilt from its source, was normalized to 10^4 counts per cell and log-transformed. The 2000 highly variable genes are selected by scanpy's Seurat-flavor dispersion ranking on the training cells.

D SCORING REQUIREMENTS AND LEAKAGE CHECKS

Appendix C specifies the protocol. Reported official scores come only from the official scorer and require all leakage checks to be satisfied, along with a recorded split, configuration and software environment. The released code computes derived quantities from saved results. Observed scores and matched nulls use the official sort of Equation (1). The gauge maxima and inverse residuals use Equation (2), which Appendix F shows equals the official score whenever no candidate ties the true one. For each scored setting, we record the pre-scoring clipping bounds and the number of predicted values outside them.

Across the k=32 runs, the fraction of predicted values clipped to zero is 0.285 to 0.598 for the unshrunk mean and 0.177 to 0.338 for EB-affine, a ratio of 1.03 to 3.30 depending on the dataset. The unshrunk mean clips more than every shrinking method on all five datasets. The ordering among

shrinking methods varies. Projection and the trained pooler have marginally lower clipping fractions than EB-affine on arc-h1 and rpe1. Upper clipping affects at most 1.6×10^{-6} of values on any row, so the clip is a floor in practice. The affine OT map records zero because it clips its own output, to between zero and the largest value in the control pool and its support cells, before the scoring pipeline sees it. Its output before the scoring pipeline's clip is therefore already floored at zero, which Appendices O and J take into account.

Clipping frequencies alone do not determine its effect on the score. Although every measured value lies in $[0, \max]$ and projection onto a box containing the target is non-expansive, this argument does not apply directly here. The clip acts on predicted cells, and the metric is computed on the pseudobulk mean over them. Clipping cellwise and then averaging is not the projection of the average, and it can move the mean away from the truth even when it moves every cell toward the feasible range.

Clipping frequencies alone do not separate estimator and transformation effects. We therefore apply the same lower floor to each method's predicted pseudobulk and compare it with both the raw prediction and the stored cellwise-clipped result (Appendix O.1). This comparison covers the four estimators with genuinely preclip outputs. It excludes affine OT, whose internal clipping prevents the corresponding control.

A scored row is excluded unless it meets all five leakage checks. Two compare identity sets, requiring disjoint support and query cells and no held-out perturbation labels among the training labels. Atomic targets are not re-checked there, because split construction excludes every training label that shares a target with a held-out one when it is built (Appendix C). The check that preprocessing was fitted on training cells compares a recorded fit set that the split builder fills with the training cells, so it holds by construction, while the highly variable genes are in fact selected on the training cells. The two provenance checks, that no embedding was fitted on expression and that the mean shift reads support cells only, read fields the scoring script writes about its own configuration. Split construction additionally enforces two invariants. The held-out set must match the $k_{\max}$ set, and the training set must not move across k. The leakage summary records one row per scored run, defined by one dataset, support size and draw across all five estimators. All 90 runs met all five checks. Set sizes accompany overlap counts to distinguish disjoint nonempty sets from vacuous checks.

The confirmatory family. The protocol registered 15 tests before the primary runs, one for each of five datasets and three contrasts of EB-affine against the unshrunk mean, affine OT and the trained pooler. Each compares the normalized area under the Pearson δ curve over $\log_2 k$, support draws averaged within a perturbation, by a two-sided paired Wilcoxon test, and Holm's correction runs once over all 15. Table D.1 gives every row. It concerns response accuracy, not discrimination, and no claim of the main text rests on it.

The confirmatory family pairs by perturbation label, which requires care when labels share targets. On the four single-target screens the labels are target-disjoint and the paired differences are plausibly independent. The norman panel is not, because its 56 held-out labels draw on 54 atomic targets, 19 of them appearing in more than one label, and joining labels that share a target leaves only 25 target-disjoint components, the largest holding 27 labels. A signed-rank test over 56 labels therefore treats dependent differences as independent, and Holm corrects for multiplicity, not for that.

Norman's three contrasts were repeated with only the unit of analysis changed. Label-level differences were averaged within each component and tested over the 25 components, then over the 20 singleton components alone, excluding every label that shares a target. The effects move little. Against the unshrunk mean, the label-level $+0.082$ becomes $+0.065$ and $+0.073$. Against affine OT, $+0.089$ becomes $+0.087$ and $+0.085$. Against the trained pooler, $+0.048$ becomes $+0.051$ and $+0.041$. Every interval still excludes zero. Substituting either set of p-values into the 15-test Holm family leaves 12 of 15 significant, with no changed classification. The closest case, the singleton-unit pooler contrast at $p_{\text{Holm}} = 0.048$, is only just significant. The original confirmatory analysis is retained. These component-level analyses assess sensitivity to the unit of analysis and do not replace it.

Table D.1: **Twelve of the 15 confirmatory tests are significant after Holm's correction, nine with EB-affine higher and three lower.** Effect is the mean paired difference in normalized area under the Pearson δ curve, positive when EB-affine is higher, with a percentile interval from 10,000 bootstrap resamples of perturbations.

dataset	EB-affine against	n	effect	95% interval	p	p_{Holm}
adamson	unshrunk mean	18	$+0.115$	$[+0.017, +0.193]$	0.027	0.081
	affine OT	18	-0.002	$[-0.115, +0.085]$	0.074	0.15
	trained pooler	18	-0.106	$[-0.207, -0.033]$	5.3×10^{-5}	3.2×10^{-4}
norman	unshrunk mean	56	$+0.082$	$[+0.061, +0.102]$	4.6×10^{-8}	4.2×10^{-7}
	affine OT	56	$+0.089$	$[+0.069, +0.109]$	1.7×10^{-8}	1.7×10^{-7}
	trained pooler	56	$+0.048$	$[+0.029, +0.070]$	4.1×10^{-5}	2.9×10^{-4}
replogle	unshrunk mean	181	$+0.114$	$[+0.105, +0.122]$	1.0×10^{-30}	1.5×10^{-29}
	affine OT	181	$+0.059$	$[+0.048, +0.070]$	8.8×10^{-19}	1.2×10^{-17}
	trained pooler	181	-0.063	$[-0.075, -0.050]$	2.1×10^{-17}	2.7×10^{-16}
rpe1	unshrunk mean	163	$+0.110$	$[+0.089, +0.130]$	5.8×10^{-16}	7.0×10^{-15}
	affine OT	163	$+0.088$	$[+0.065, +0.109]$	4.0×10^{-15}	4.4×10^{-14}
	trained pooler	163	-0.012	$[-0.023, -0.002]$	2.4×10^{-3}	0.012
arc-h1	unshrunk mean	30	$+0.050$	$[+0.032, +0.067]$	3.0×10^{-5}	2.4×10^{-4}
	affine OT	30	$+0.052$	$[+0.018, +0.092]$	4.0×10^{-3}	0.016
	trained pooler	30	$+0.020$	$[-0.024, +0.071]$	0.66	0.66

D.1 Eligibility-floor sensitivity

The primary protocol fixes eligibility at $k_{\max}$=32. This 64-cell floor removes almost half of rpe1 and, in a proliferation screen, plausibly the perturbations that killed or arrested their cells. To test whether selection drives the pattern, we rebuilt the partitions at $k_{\max}$=16 (a 48-cell floor, support grid $\{1, 2, 4, 8, 16\}$, the same result-independent draw), admitting smaller-count perturbations excluded by the main floor. The largest increases are on rpe1, from 815 to 1069 eligible perturbations, and replogle, from 901 to 990, enlarging the held-out sets to 214 and 198. All estimators were rerun with the same official scorer and leakage checks.

The pattern retains its sign and its magnitude increases on some panels (Table D.2). At $k_{\max}$=16, EB-affine still improves response accuracy over the unshrunk mean on all five datasets ($+0.06$ to $+0.14$) and lowers perturbation discrimination on all five (-0.19 to -0.27). The unshrunk mean remains the best discriminator among the estimators on all five. On the two datasets with the largest increases in eligible perturbations, both effects grow. Rpe1's response gain rises from $+0.11$ to $+0.14$ and its discrimination deficit from -0.17 to -0.20. This is consistent with the shrinkage mechanism, because the newly admitted low-count perturbations are noisier at small k, so shrinkage gains more accuracy and loses more identity. The pattern holds at both floors tested. Because lowering the floor also changes perturbation eligibility and shortens the support grid, the comparison in Table D.2 shows that the pattern is not an artifact of one floor, not that nothing else contributes.

E The official metric suite

Table E.1 reports both members of all three official metric families at k=32, for every method, to test whether the observed pattern depends on choosing one metric per axis.

The PDS ℓ_1 implementation defines discrimination through rank-based retrieval. For each of the N held-out perturbations, the official scorer measures the L_1 distance from that perturbation's predicted pseudobulk profile to the measured profile of every held-out perturbation, and scores the prediction by the normalized rank r of the correct match, $1-r/N$ with r counted from zero. The score equals 1 when the nearest measured profile matches the predicted perturbation and falls toward 0 as unrelated perturbations move closer. Scores are averaged over perturbations, with higher values better. The ℓ_2 and cosine variants replace the distance. The score therefore rewards predictions that stay mutually

Table D.2: **The response gain and discrimination loss persist at both eligibility floors.** Δ is EB-affine minus the unshrunk mean on the area under the response curve and under the perturbation-discrimination curve, each taken over $\log_2 k$ and normalized by its width. The k_{16} columns use the $k_{\max}$=16 eligibility floor and the k_{32} columns the $k_{\max}$=32 floor every other table uses, and n_{16} is the number of held-out perturbations at $k_{\max}$=16. The lower floor admits the smaller-count perturbations. The response gain and the discrimination loss both persist on all five datasets, and grow where the eligible set grows most.

		response Δ		discrimination Δ	
dataset	n_{16}	k_{16}	k_{32}	k_{16}	k_{32}
adamson	18	$+0.135$	$+0.115$	-0.265	-0.225
norman	57	$+0.099$	$+0.082$	-0.192	-0.164
replogle	198	$+0.128$	$+0.114$	-0.237	-0.208
rpe1	214	$+0.142$	$+0.110$	-0.195	-0.165
arc-h1	30	$+0.056$	$+0.050$	-0.194	-0.160

distinguishable and penalizes ones that collapse toward a shared response, the quantity shrinkage moves.

The chance level is $(N{+}1)/(2N)$, slightly above $1/2$ and dependent on N. It is attained exactly, not just in expectation. A prediction that is identical for every perturbation induces one fixed distance ordering, so across the N perturbations the correct match occupies each rank $0, \ldots, N-1$ once, and the average of $1 - r/N$ is $1 - (N-1)/(2N)$. The constant prediction of Table I.1 is such a prediction, and matches this identity to within 1.1×10^{-16} on all five datasets, reaching 0.527778 on adamson at N=18 and 0.502762 on replogle at N=181, against $19/36$ and $91/181$. The floor varies with N, ranging from 0.5028 to 0.5278 on these five screens. Approximating it as one half obscures this dependence on N.

Preprocessing keeps the target gene out of the scored space, with three exceptions counted below. The metric's own exclusion rule removes nothing. The official scorer's discrimination call enables target-gene exclusion and implements it as a lookup of the perturbed target's gene index, which drops a gene only if its name equals the full perturbation label. Labels in these screens take the form AARS+ctrl, and on norman BCL2L11+BAK1, so no gene name ever matches and the mask removes nothing. The flag is inert.

A knocked-down target gene provides the most direct identity signal, and shrinkage attenuates it along with everything else. A discrimination deficit confined to that one coordinate would say little about perturbation identity. The deficit does not rest on that coordinate, because the scored space excludes those genes apart from the three exceptions below. HVG selection is fit on training cells only (Appendix D), where no held-out target is directly perturbed, and the measured membership shows the outcome. Of the 446 distinct held-out atomic targets summed over the five datasets, 3 survive into the scored space, all three on arc-h1, where they carry 0.29% of that dataset's between-perturbation L_1 distance mass, and on the other four the count is 0. Recomputing the retrieval with each perturbation's own atomic targets removed leaves the unshrunk mean's discrimination unchanged on all five. For that estimator, preprocessing supplies the intended exclusion, not the metric.

The target-gene exclusion analysis produced both counts and recomputed scores. This raw-expression reanalysis is outside the supplied post-hoc analysis package. It reimplements the metric without calling the official pipeline, so its absolute scores are close to, but not identical with, the official pipeline's. The claims concern membership counts and score differences, both invariant to that gap.

To assess metric dependence, we counted the best method in each (family, metric, dataset) setting, with ten settings per family. The response family is split four ways (EB-affine 4, the trained pooler 3, projection 2, affine OT 1) and the differential-expression family two (affine OT 6, EB-affine 4). The discrimination family is not split at all. The unshrunk k-shot mean takes all ten, on PDS ℓ_1 and PDS ℓ_2 alike, on every dataset. No shrinking, projecting or trained estimator wins at any setting under either discrimination metric, even on datasets where the other metric families have no consistent winner.

The response axis requires a separate distinction. Proposition A.1 establishes optimality under squared error on the mean shift, while the pre-specified primary endpoint, Pearson δ, is scale-insensitive. The theory and primary endpoint therefore measure different quantities. The official scorer also reports response squared error for every setting. Under this risk, which Proposition A.1 minimizes, EB-affine improves on the unshrunk mean in all 30 dataset-by-k settings, with large gains where shrinkage is expected to matter. At k=1, the unshrunk mean's squared error is 5.7 to 12.5 times EB-affine's, on replogle 0.140 against 0.011. The ratio then closes monotonically with k, to between 1.00 and 1.26 at k=32, where norman is a near-tie (0.002452 against 0.002448). Mean absolute error (MAE) of δ agrees in all 30 settings as well.

These error metrics support the variance-reduction interpretation. Squared error directly penalizes the unshrunk estimator's variance, which shrinkage trades against bias. The agreement therefore reflects the mechanism, not a coincidence. Both the error advantage and the identity cost decrease toward k=32, without reaching zero in every setting. Pearson δ remains the primary endpoint, as registered in advance. The error metrics are more favorable on the same runs, but selecting them after observing the results would change the endpoint post hoc. The analysis code computes the per-support-size error values from the prediction inputs.

Differential-expression (DE) ρ_s is the one metric on which the trained pooler collapses (to 0.052 on adamson). The official implementation is not a per-gene rank correlation. It computes the Spearman correlation across held-out perturbations between the numbers of significant genes in truth and prediction. It therefore measures calibration of the differential-expression set's size, not which genes change. The naive mean, affine OT, and EB-affine score 0.48 to 0.84 on it across all five datasets (Table E.1). The only low values are the pooling methods, projection and the trained pooler, on the two smallest screens, and even those recover above 0.57 on the three larger datasets. Thus this metric distinguishes the methods without assigning uniformly low scores.

F VERIFICATION OF THE SCORING CONVENTION

We compared Equation (2) with the official scorer's per-perturbation output on rpe1 at $k = 8$, with 163 perturbations, under ℓ_1, ℓ_2 and cosine. The form $1 - \mathrm{closer}/N$ with the true candidate counted in N agrees to 1.1×10^{-16} under all three distances. Four tested alternatives all differ by at least $6 \times 10^{-3} \approx 1/N$. They are $1 - \mathrm{closer}/(N-1)$, $\mathrm{further}/(N-1)$, $\mathrm{further}/N$, and a variant halving ties over $N - 1$. These inputs identify the denominator unambiguously but do not identify the tie rule. No candidate ties the true one, so every tie rule agrees, including mid-ranks over N, which the four alternatives did not include.

The tie rule, measured by forcing ties. We called the package's discrimination function, version 0.7.2 in the environment that produced the reported scores, on constructed panels. Each perturbation is one cell with a small integer effect against an all-zero control, making every tie exact under both ℓ_1 and ℓ_2. Table F.1 gives the outcome. The package does not count strictly closer candidates. It scores the true label's position after sorting the row's distances over alphabetically listed labels. A tie favors the prediction only if the sort places the true label first. Under ℓ_1 it kept the alphabetical order at 8 and 12 labels and not at 40, consistent with a sort that is stable only on short arrays. Under ℓ_2 it kept it at 8 and not at 12, consistent with the Euclidean computation turning an exact tie into a rounding-level difference. Equation (2) is therefore the official score on inputs without ties and a different score on inputs with them.

Tie-dependent and tie-independent quantities. Without a tie involving the true candidate, the two scores agree. With continuous responses, such a tie at an estimator's own scale has probability zero, leaving the reported scores at that scale unaffected. Three statements are properties of Equation (2), and the main text states them in that scope. The first is the T/N^2 term of Proposition 2, which the package does not have. A shared prediction fills the positions of one sorted order once each whatever its ties, and mid-ranks return the same floor by the counting argument of Appendix I. The second is the $1/N^2$ gain from a maximized shared scalar on an identity-free prediction. The official package gives no such gain. The third is the value of every maximum of Proposition 1, since those are evaluated on breakpoints, where a tie is exact. A breakpoint's value is also attained on a neighboring open interval, where no candidate is tied, unless crossings in opposite directions coincide at the breakpoint. Identical predictions under one shared scalar make them coincide by

Table E.1: **The discrimination family has one winner everywhere, unlike the other two families.** Two official metrics per family at k=32, with support draws $\{1, 2, 7\}$ averaged within a perturbation. Arrows indicate the direction of better performance, and the best value per column within each dataset is in bold.

	response		differential expression		discrimination	
Method	Pearson δ ↑	MAE δ ↓	DE sign ↑	DE ρ_s ↑	PDS ℓ_1 ↑	PDS ℓ_2 ↑
adamson						
naive k-shot mean	0.799	0.043	0.802	0.500	**0.979**	**0.984**
affine OT	0.836	0.037	**0.974**	0.485	0.967	0.974
EB-affine	0.823	0.040	0.951	**0.556**	0.958	0.972
projection	0.847	**0.037**	0.935	0.375	0.963	0.971
trained pooler	**0.848**	0.037	0.951	0.052	0.945	0.950
norman						
naive k-shot mean	0.751	0.019	0.540	0.621	**0.993**	**0.996**
affine OT	0.699	0.020	**0.929**	**0.655**	0.928	0.954
EB-affine	**0.763**	**0.018**	0.856	0.529	0.979	0.992
projection	0.682	0.020	0.798	0.144	0.870	0.899
trained pooler	0.663	0.021	0.803	0.071	0.781	0.825
replogle						
naive k-shot mean	0.485	0.074	0.924	0.681	**0.987**	**0.992**
affine OT	0.491	0.067	0.934	**0.801**	0.923	0.942
EB-affine	0.526	0.066	**0.949**	0.586	0.945	0.979
projection	0.534	0.064	0.942	0.579	0.935	0.949
trained pooler	**0.539**	**0.063**	0.948	0.610	0.918	0.931
rpe1						
naive k-shot mean	0.620	0.076	0.943	0.682	**0.977**	**0.988**
affine OT	0.581	**0.070**	0.923	**0.808**	0.922	0.927
EB-affine	**0.636**	0.071	**0.950**	0.741	0.946	0.977
projection	0.605	0.070	0.925	0.576	0.893	0.917
trained pooler	0.604	0.070	0.920	0.668	0.862	0.889
arc-h1						
naive k-shot mean	0.446	0.035	0.593	0.820	**0.966**	**0.970**
affine OT	0.416	0.033	**0.882**	0.806	0.895	0.883
EB-affine	**0.469**	0.031	0.792	**0.838**	0.957	0.958
projection	0.406	**0.031**	0.774	0.797	0.861	0.859
trained pooler	0.392	0.031	0.775	0.813	0.807	0.805

Table F.1: **The package credits a tie to the prediction only when the true label sorts first, and returns the floor exactly for a prediction shared by every row.** The package column calls its own discrimination function. Rows with one tie report the true perturbation's score and are the same under ℓ_1 and ℓ_2. Rows with a shared prediction report the mean, whose floor $(N+1)/(2N)$ is 0.5417, 0.5125 and 0.55 at the three sizes. Where two tie counts or two values are given they are ℓ_1 then ℓ_2. The last row scans the shared prediction's scale over 81 values including ones that create ties, and every value the package returns is the floor to 1.1×10^{-16}.

construction	N	package	Equation (2)	mid-ranks
no tie	12	identical, largest difference 0		
one tie, true label sorts first	8	1.000	1.000	0.9375
one tie, true label sorts second	8	0.875	1.000	0.9375
same prediction for every row, 18 tied pairs	12	0.5417	0.6667	
same prediction for every row, 188 or 252 tied pairs	40	0.5125	0.63 or 0.67	
the same, times the best of 81 shared scalars	10	0.55	0.71	

construction, and that is the second statement. For distinct continuous predictions a coincidence has probability zero, and a search of the real predictions finds none that moves a reported maximum (Appendix G). For the per-prediction best case on an identity-free prediction we recomputed the

maximum over open intervals alone. On every open interval, each candidate's side, closer or farther than the true candidate, was determined from the sign and order of its crossing, without evaluating a sampled scale. The maximum equals the breakpoint value in all ten settings, five panels at $k \in \{1, 32\}$. No perturbation reaches its best only on a breakpoint, so the abstract's 0.80 to 0.84 does not depend on ties.

The closed-form maxima are checked against brute-force grids (Appendix G).

G CRITERIA FOR THE SCALAR-QUOTIENT BOUND

$\overline{\mathrm{PDS}}^{\star}$ is constant on scalar orbits, is the tightest scalar-invariant upper bound on the tie-optimistic score $\overline{\mathrm{PDS}}$ and therefore an upper bound, not always the tightest, on the official PDS. It is at least the score at the reported scale, with equality only when that scale is already optimal, and is exactly computable under ℓ_2 by Proposition 1. We checked the closed form against two differently constructed brute-force grids. Neither exceeded it (0 of 8001 scale values at four settings, 0 of 163 perturbations in the per-prediction check). The grids and the closed form agree at $\beta = 1$ to $0.000\mathrm{e}{+00}$. The closed form also finds optima missed by the grid (0.9414 against 0.9383 on adamson at $k = 8$).

Finite-scale derivation. Let $d_i(\beta) = a_i + \beta v_i$ be an affine path through the prediction. Under squared Euclidean distance, for any candidate j,

$$\|a_i + \beta v_i - \Delta_i\|^2 - \|a_i + \beta v_i - \Delta_j\|^2 = \underbrace{\left(\|a_i - \Delta_i\|^2 - \|a_i - \Delta_j\|^2\right)}_{A_j} + 2\beta \underbrace{\langle v_i,\, \Delta_j - \Delta_i \rangle}_{B_j}. \quad \text{(G.1)}$$

The $\beta^2\|v_i\|^2$ term is identical on both sides and cancels. Each margin is affine in β and contributes the breakpoint $\beta_j = -A_j/(2B_j)$ if $B_j \neq 0$ and none if $B_j = 0$, when candidate j is closer, farther or tied at every β. The true candidate's own column is identically zero, so there are at most $N - 1$ breakpoints and $\mathrm{PDS}_i(\beta)$ is a step function with at most N pieces. For the groups of Equation (3), $a_i = 0$, $v_i = d_i$ and $\beta = c > 0$, and both maxima are attained by evaluating Equation (2) at the positive breakpoints, where a tie counts for the prediction, and at one point inside each interval between them, per perturbation for the per-prediction group and over the union of breakpoints for the global one.

Breakpoint enumeration. For perturbation i on the path $a_i + \beta v_i$ the released code computes A_j and B_j of Equation (G.1) for every candidate, keeps the roots $\beta_j = -A_j/(2B_j)$ in the group's range, $\beta > 0$ for both groups, and evaluates Equation (2) at three kinds of scale values, comprising every root, the midpoint of each interval between consecutive roots with one point beyond the last, and $\beta = 1$. At each value, candidate j counts as strictly closer when $A_j + 2\beta B_j > 0$, unless it is tied, and it is tied when the evaluated scale is its own root or when $A_j = B_j = 0$, which is the true candidate's column. The per-prediction maximum is the best value for each perturbation, requiring $O(N)$ evaluations at $O(N)$ each. The global maximum uses the union of all perturbations' roots, up to $N(N-1)$ values, and evaluates all N perturbations at each, for $O(N^4)$ overall. The tie rule has one numerical limitation. Two candidates' exactly coincident crossings, computed by different arithmetic, can yield two floating-point roots differing by one unit in the last place. Each evaluation then marks only its own candidate as tied, so a joint tie is credited only when rounding favors it. Constructed inputs with two opposite crossings at one scale received this credit in some trials but not others. The error can only lower a maximum, since it withholds a tie, and it needs two crossings that coincide exactly. Appendix F rules that out for the identity-free inputs by recomputing on open intervals. For real predictions, we searched all 270 sets underlying the reported maxima. These cover five estimators on every panel and support size on the first draw, and the two estimators in the draw analysis on the other two. The search looks for a root where one candidate starts to beat the true one within a relative 10^{-9} of a root where another stops, seven orders of magnitude above rounding. Within one prediction the closest such pair is 5.2×10^{-7} apart. Pooled across predictions, which is what the global maximum reads, three sets on rpe1 hold a pair within 10^{-9}, the closest 6.2×10^{-11} apart. Crediting every such pair as a joint tie gives 0.749, 0.691 and 0.901 at those points, below the released global maxima of 0.803, 0.697 and 0.939, so no reported maximum depends on the gap. Both constructed cases, an exactly representable joint tie that is credited and a coincidence that

is bounded between the open-interval value and the joint-tie value, are included in the numerical validation, along with the search.

A semi-synthetic comparison distinguishes it from per-gene alternatives on two pre-specified criteria for measuring retained identity. Bijections of the estimate carry the same information, so such a score should give them the same value. The spread across bijections is 0.1311 for PDS@1 and 0.0026 for $\overline{\mathrm{PDS}}^{\star}$. Such a score should also be monotone in surviving rank without inflating the chance floor. At rank 2, the lowest rank tested, a per-gene coordinate-ascent version returns 0.6454 against a chance floor of 0.5 while $\overline{\mathrm{PDS}}^{\star}$ returns 0.5237. These two tests on one semi-synthetic panel concern the global quotient, approximated here by its maximum over 241 scales between 0.05 and 6. They motivate reporting a scalar quotient as the bound, but they do not make a quotient a measure of retained identity. In particular, the per-prediction quotient lifts an identity-free prediction to about 0.8 (Appendix I). The same experiment also limits the interpretation of non-invertibility. Zeroing 30% of genes is not invertible, yet it scores 0.8834, indistinguishable from the bijections at 0.885. The rank-limited projector's information restriction is therefore supported by its retention of 20 of 163 directions, not by non-invertibility alone.

H Exact values by support size

Tables H.1 and H.2 give every estimator's response and discrimination score at every $(\text{dataset}, k)$ setting. They support the inverse analysis in Section 6 and the ranking comparison in Appendix J.

Across the 30 settings, the trained pooler leads in response accuracy in 19 and EB-affine in 11. Neither dominates all settings. On discrimination, the unshrunk k-shot mean leads across all 30 settings, spanning every dataset and support size. That is stronger than the same count aggregated over k and than the 10-of-10 across both discrimination metrics in Table E.1. None of the tested shrinking, projecting or trained estimators exceeds the unshrunk estimate's discrimination score. The paper asks not why this happens but whether the score change is recoverable, and Section 6 shows that it is for the empirical-Bayes map.

On adamson, EB-affine's discrimination rises from 0.562 to 0.958 across support sizes, while the unshrunk mean rises from 0.843 to 0.979. The map starts far below and closes most of the gap as its shrinkage weight relaxes with growing support size. Projection and the trained pooler instead use a fixed response subspace, so their changes with k need not follow the empirical-Bayes weight.

I Truth-dependent corrections of identity-free predictions

We test both quotient bounds on identity-free predictions. A high score produced by group maximization alone would not establish retained perturbation identity.

We assign every held-out perturbation the same vector, the fitted shared response $\bar{\delta}$. This constant prediction cannot distinguish perturbations by direction, the only feature a scalar group leaves unchanged. Any excess over the finite-library chance floor $(N{+}1)/(2N)$ therefore comes entirely from the gauge.

Table I.1: **Per-prediction maximization raises a constant prediction above chance.**

dataset	N	chance	PDS@1	cosine	common norm	$\overline{\mathrm{PDS}}^{\star}_{\mathrm{glob}}$	excess	$\overline{\mathrm{PDS}}^{\star}_{\mathrm{pred}}$
arc-h1	30	0.516667	$+0$	$+0$	$+0$	0.517778	$1/N^2$	0.8044
replogle	181	0.502762	$+0$	$+0$	$+0$	0.502793	$1/N^2$	0.7953
norman	56	0.508929	$+0$	$+0$	$+0$	0.509247	$1/N^2$	0.8093
adamson	18	0.527778	$+0$	$+0$	$+0$	0.530864	$1/N^2$	0.8364
rpe1	163	0.503067	$+0$	$+0$	$+0$	0.503105	$1/N^2$	0.8230

The values in Table I.1 do not depend on k and repeat at all six support sizes. The three columns marked $+0$ have no measurable excess. The reported score, cosine distance and rescaling to one common norm agree with the floor in all 30 settings, exactly in 24 and to one unit in the last place, 1.1×10^{-16}, in adamson's six. The independent recomputation behind Table I.2 agrees exactly in

Table H.1: **The response ranking depends on dataset and support size.** Pearson δ (higher is better), with support draws $\{1, 2, 7\}$ averaged within a perturbation. The best value per column within each dataset is in bold.

Method	k=1	k=2	k=4	k=8	k=16	k=32
adamson						
naive k-shot mean	0.342	0.455	0.570	0.660	0.741	0.799
affine OT	0.496	0.624	0.716	0.767	0.808	0.836
EB-affine	0.605	0.643	0.690	0.738	0.786	0.823
projection	0.619	0.722	0.787	0.812	0.834	0.847
trained pooler	**0.768**	**0.794**	**0.822**	**0.833**	**0.843**	**0.848**
norman						
naive k-shot mean	0.328	0.421	0.526	0.615	0.694	0.751
affine OT	0.352	0.442	0.530	0.601	0.661	0.699
EB-affine	**0.503**	**0.559**	**0.619**	**0.671**	**0.724**	**0.763**
projection	0.439	0.526	0.592	0.634	0.666	0.682
trained pooler	0.463	0.544	0.589	0.621	0.646	0.663
replogle						
naive k-shot mean	0.132	0.189	0.254	0.327	0.409	0.485
affine OT	0.189	0.256	0.327	0.391	0.448	0.491
EB-affine	0.315	0.345	0.384	0.428	0.478	0.526
projection	0.259	0.340	0.413	0.465	0.506	0.534
trained pooler	**0.378**	**0.426**	**0.466**	**0.497**	**0.521**	**0.539**
rpe1						
naive k-shot mean	0.217	0.296	0.386	0.477	0.555	0.620
affine OT	0.269	0.345	0.429	0.498	0.546	0.581
EB-affine	0.459	0.479	0.511	0.550	**0.594**	**0.636**
projection	0.313	0.405	0.485	0.545	0.581	0.605
trained pooler	**0.488**	**0.509**	**0.540**	**0.563**	0.586	0.604
arc-h1						
naive k-shot mean	0.130	0.204	0.268	0.326	0.391	0.446
affine OT	0.139	0.219	0.275	0.324	0.376	0.416
EB-affine	0.215	0.270	0.325	**0.368**	**0.424**	**0.469**
projection	0.213	0.281	0.321	0.346	0.383	0.406
trained pooler	**0.234**	**0.282**	**0.327**	0.341	0.365	0.392

all 30. The common norm is not the fixed-norm remedy as it is implemented, which matches each prediction to its own true norm, and that version is not on the floor (Table I.2). Under Equation (2), which the excess column uses, the one-scalar maximum exceeds it by exactly $1/N^2$ on all five panels, with agreement to 5.1×10^{-12} here and exactly in the independent recomputation. Under the official package it is the floor itself, because a shared scalar leaves every row identical (Appendix F). The per-prediction maximum exceeds the floor by 0.288 to 0.320, mean 0.302, under both. Six decimal places are needed because four would round the 3×10^{-5} excesses on replogle and rpe1 to zero.

Panel-wide, support-dependent and truth-dependent evaluation. Consulting the candidate panel does not by itself violate the floor condition. Five constructions in Table I.2 consult the candidates while remaining on the floor. Proposition 2 asks only that the dissimilarity used to rank perturbation i not take i as an argument, and it may read the whole panel. If every prediction is the same vector then every row of the distance matrix is the same function f of the candidate index, the count of candidates strictly beating the true one for perturbation i is the strict rank of $f(i)$, and the strict ranks sum over the N perturbations to $N(N-1)/2-T$, with T the number of tied pairs among the N values, giving $(N+1)/(2N) + T/N^2$ under Equation (2). Counting a tie as half a rank adds the T back and returns the floor whether or not there are ties. The official package does neither. It takes the true candidate's position in a sort of the row, and identical rows share one sorted order, so the true candidates fill its N positions once each and the mean is the floor exactly, again whatever the ties. Ranks that form a permutation of $1, \ldots, N$ put recall at k at k/N and mean reciprocal rank at H_N/N, exactly, and under Equation (2) the strict ranks form one only when $T = 0$.

Table H.2: **The unshrunk k-shot mean has the best discrimination in all 30 settings.** PDS ℓ_1 (higher is better) on the same runs and settings, with the best value per column within each dataset in bold.

Method	k=1	k=2	k=4	k=8	k=16	k=32
adamson						
naive k-shot mean	**0.843**	**0.913**	**0.951**	**0.958**	**0.966**	**0.979**
affine OT	0.809	0.886	0.930	0.938	0.950	0.967
EB-affine	0.562	0.572	0.629	0.752	0.861	0.958
projection	0.808	0.887	0.938	0.945	0.958	0.963
trained pooler	0.779	0.837	0.882	0.917	0.926	0.945
norman						
naive k-shot mean	**0.845**	**0.904**	**0.952**	**0.973**	**0.987**	**0.993**
affine OT	0.801	0.839	0.861	0.889	0.904	0.928
EB-affine	0.590	0.641	0.730	0.839	0.921	0.979
projection	0.784	0.810	0.830	0.847	0.850	0.870
trained pooler	0.599	0.662	0.697	0.725	0.739	0.781
replogle						
naive k-shot mean	**0.794**	**0.876**	**0.928**	**0.960**	**0.979**	**0.987**
affine OT	0.760	0.803	0.852	0.883	0.910	0.923
EB-affine	0.551	0.592	0.653	0.741	0.863	0.945
projection	0.775	0.823	0.870	0.902	0.923	0.935
trained pooler	0.691	0.747	0.811	0.854	0.893	0.918
rpe1						
naive k-shot mean	**0.794**	**0.852**	**0.901**	**0.941**	**0.963**	**0.977**
affine OT	0.723	0.781	0.832	0.873	0.903	0.922
EB-affine	0.570	0.616	0.687	0.780	0.879	0.946
projection	0.738	0.782	0.825	0.860	0.880	0.893
trained pooler	0.663	0.714	0.766	0.809	0.844	0.862
arc-h1						
naive k-shot mean	**0.773**	**0.864**	**0.900**	**0.939**	**0.960**	**0.966**
affine OT	0.709	0.787	0.825	0.846	0.878	0.895
EB-affine	0.550	0.601	0.682	0.781	0.914	0.957
projection	0.716	0.789	0.825	0.833	0.861	0.861
trained pooler	0.658	0.699	0.753	0.756	0.788	0.807

The tie term belongs to Equation (2), not to the official score. Under Equation (2), where two candidates are exactly equidistant from the prediction the strict rank is lower and the score is higher than the floor, by $1/N^2$ per tie. For example, a four-candidate instance with one tie scores 0.6875 against a floor of 0.6250, and with all four tied it scores 1. That crediting ties to the true candidate lets a constant model score perfectly is known from knowledge-graph evaluation (Sun et al., 2020), and we claim no novelty for it. The official package breaks a tie by a sort instead, and on a constant prediction it returns the floor whatever the ties (Table F.1). On the five panels, candidate distances are distinct, with the closest pair separated by 8.4×10^{-6} under cosine and 6.0×10^{-5} under ℓ_2. The two scores therefore agree, putting the cosine and common-norm scores on the floor, not merely near it. This is an observation on these panels, not a general theorem.

Maximizing is a different operation, and the excess measures how much freedom it is given. Equation (2) counts candidates strictly closer, so where a prediction is exactly equidistant from its own candidate and a competitor the tie counts in its favor, worth $1/N$ to that perturbation and $1/N^2$ to the mean. A shared scalar leaves predictions identical, so ties are its only source of score increases. The measured excess is exactly $1/N^2$ on all five panels, one rank for one perturbation and the smallest possible score change. Under the official package, identical rows keep the score on the floor at every scale, as the forced-tie scan in Table F.1 confirms on a toy panel. Per-prediction scalars allow much more freedom. Candidate norms differ, giving each prediction a different step function. Maximizing over at most N pieces of a rank statistic exceeds any single piece regardless of prediction information, accounting for the observed 0.30 excess. The closed form agrees with an 8002-point brute-force sweep to $0.00\text{e}{+00}$ on this input, and the same maximum is reached on open intervals, where no candidate is tied, so the official package reaches it too (Appendix F).

Table I.2: **Only constructions that read something specific to one perturbation move an identity-free prediction off the floor.** Each row covers 30 settings spanning five panels, $k \in \{1, 32\}$, and three identity-free vectors, the fitted shared response $\bar{\delta}$, the panel mean and a seeded Gaussian direction scaled to $\|\bar{\delta}\|$ that encodes nothing about any perturbation. Every entry is computed under Equation (2). Above the first rule, every entry equals $(N+1)/(2N) + T/N^2$ to 10^{-12}, with $T = 0$ for the nine fixed constructions and $T = 1$ for the maximized shared scalar. Its $1/N^2$ excess comes from one tied pair, which the official package does not credit, returning the floor (Appendix F). Recall at 1 and 5 and mean reciprocal rank equal their permutation means. The hubness correction is cross-domain similarity local scaling (CSLS) (Conneau et al., 2018), $2\cos - r_x - r_y$ with each r averaged over the ten nearest neighbors in the candidate panel, a standard remedy for the hubness of high-dimensional nearest-neighbor retrieval (Radovanović et al., 2010), and the scale quotient per candidate ranks by $\inf_{c>0} \|c\,d - \Delta_j\|$, choosing a scale separately against each candidate. The infimum need not be attained. For $\langle d, \Delta_j \rangle \leq 0$, the value decreases as $c \downarrow 0$ toward $\|\Delta_j\|$, the boundary value used by the implementation. For $d = 0$ every candidate gets $\|\Delta_j\|$, which is a constant row and therefore on the floor. The one scalar fitted to every truth reads all of them but applies one number to every perturbation, which is why it stays above the first rule. Above that rule a construction reads neither perturbation i's support cells nor its truth, which is the condition Proposition 2 covers. Between the rules, the deployable fixed norm reads i's support cells, which are available at deployment and supply the gained information. Below the second rule each construction chooses perturbation i's representative, scale, weights or genes from i's own true response. The criterion concerns inputs, not the direction of a score change. Using perturbation i's ground truth to shape the comparison for i supplies evaluation-side information, whether the score rises or falls. The result is a diagnostic or bound, not a deployable score. Reading i's support is legitimate although it raises the floor. The last two rows cover their own settings, 45 and 270, in the tables they cite. The per-prediction best case is recomputed from the one-dimensional geometry of an identity-free prediction, independently of Proposition 1, and the two agree to $0.00\text{e}{+00}$.

	reads			
construction	panel	i's support	i's truth	excess over floor
reported score, ℓ_2 and ℓ_1	no	no	no	0
cosine	no	no	no	0
common norm (the mean candidate norm)	yes	no	no	0
centering both sides on the panel mean	yes	no	no	0
per-gene standardization over the panel	yes	no	no	0
hubness correction (CSLS, 10 neighbors)	yes	no	no	0
scale quotient per candidate	yes	no	no	0
one scalar fitted to every truth by least squares	yes	no	no	0
best shared scalar, $\overline{\text{PDS}}^{\star}_{\text{glob}}$	yes	no	no	exactly $1/N^2$
norm matched to own unshrunk estimate	no	yes	no	$+0.000$ to $+0.152$
best scalar per prediction, $\overline{\text{PDS}}^{\star}_{\text{pred}}$	yes	no	yes	$+0.138$ to $+0.325$
norm matched to own truth (oracle fixed norm)	no	no	yes	$+0.000$ to $+0.165$
best of $\{d, -d\}$ per prediction	yes	no	yes	$+0.004$ to $+0.193$
best of 10 random directions per prediction	yes	no	yes	$+0.010$ to $+0.037$
own-response weights (Table I.4)	yes	no	yes	-0.369 to -0.111
own top-k differentially expressed genes (Table I.5)	yes	no	yes	-0.439 to -0.074

The fixed-norm remedy, three ways. Fixing the norm can mean three different maps, and Proposition 2 separates them. Rescaling every prediction to one common norm treats the perturbations alike and sits on the floor. Matching each prediction to its own true norm, which is how the remedy is implemented (Liu et al., 2025), reads each perturbation's answer, and on the identity-free input it lifts the floor by up to 0.165. Matching each prediction to the norm of its own unshrunk estimate, the deployable form of Section 7, reads perturbation i's support cells, not its truth. Its gain therefore reflects information in the support. It is near zero at $k = 1$, where a one-cell estimate says little about magnitude, and reaches 0.144 at $k = 32$ for the fitted shared response, comparable to what the oracle norm reaches on the same panels.

Table I.3: **Own-support and own-truth corrections have distinct null effects.**

		best case per prediction			oracle norm		deployable norm, $\bar{\delta}$	
dataset	N	$\bar{\delta}$	panel mean	Gaussian	$\bar{\delta}$	panel mean	$k = 1$	$k = 32$
arc-h1	30	0.288	0.294	0.216 to 0.219	0.031	0.084	0.000	0.048
replogle	181	0.293	0.293	0.138 to 0.162	0.107	0.103	0.001	0.096
norman	56	0.300	0.325	0.201 to 0.239	0.149	0.159	0.008	0.140
adamson	18	0.309	0.309	0.219 to 0.256	0.003	0.136	0.025	0.000
rpe1	163	0.320	0.323	0.147 to 0.188	0.157	0.165	0.002	0.144

Table I.3 gives excess over the floor on identity-free inputs. The fitted shared response and the panel mean do not depend on k, so their columns hold one value each. The Gaussian direction is redrawn at each k and its column is the range. A random direction gains at most 0.003 from the oracle norm, because it is nearly orthogonal to every candidate, causing its distances to rank the candidates by their own norms, which is the same ranking for every perturbation.

Perturbation-specific response weights. Vollenweider & Bühlmann (2026) weight the distances used to rank perturbation p by weights of the query perturbation p, and say so in those terms: "when the dissimilarity d is a weighted metric … the weights used are always those of the query perturbation p, ensuring that each perturbation is evaluated on the genes where it induces the strongest signal". They apply it to the discrimination score by name, as one of three weighted metrics they report, and they state that carrying the weighting onto that score "has not been done previously". The weights are derived from p's differential-expression statistics computed over the whole dataset, test partition included, which they also state and defend, as the weights serve only for evaluation and a test-set-dependent metric is one they "consider appropriate". This deliberate use of own-truth information places the rule outside Proposition 2. Table I.4 measures what an identity-free prediction scores under it, with their weights computed from cells exactly as they define them. For each panel we draw 100 synthetic reference profiles, each the mean of 100 cells sampled without replacement from the non-control cells, run a Welch t-test per gene between a perturbation's cells and those profiles, and take $|t|$ min-max normalized across genes and squared. The whole computation is repeated at three seeds. All 45 values, five panels by three seeds by three identity-free vectors, fall below the floor, by 0.111 to 0.369, and the three seeds move a panel by at most 0.0045. The mechanism is that weighting on p's own responsive genes places p's true response far from a vector near the center of the panel, so the truth ranks near last. It is a statement about the three inputs tested here, all of which are central, and not about every constant vector. A constant vector placed beyond the responses is nearest to its own truth in each selected coordinate and scores above chance instead. With three truths e_1, e_2, e_3 scored on their own coordinates, the constant $(1, 1, 1)$ scores 1 against a floor of $2/3$ while the panel mean scores $1/3$. Which way the floor moves is therefore empirical, and what the proposition settles is only that it is free to move. Averaging the same weights into one vector shared by every perturbation returns all 45 to the floor to 10^{-16}. The weighting itself is therefore harmless once it stops reading each perturbation's answer. A pseudobulk proxy of the same weighting agrees in sign on every panel, with its implementation included in the code. This does not imply that their model rankings are flawed, as their calibration relies on an empirical control bound. The analysis identifies a directional effect of own-response weighting, placing an uninformative prediction below the rank assigned by a uniform ordering. Consequently, the zero of their rescaled axis does not correspond to where such a prediction sits.

Perturbation-specific gene selection. Weighting is one way to read a perturbation's own response. Restricting the score to its own top-k differentially expressed genes is another. Restriction to differentially expressed genes is ordinary practice for point and distributional metrics, cataloged as such by Schäfer et al. (2026), and Vollenweider & Bühlmann (2026) carry it to the discrimination score: "an alternative approach is to restrict evaluation to the top n differentially expressed genes per perturbation", compared in their Figure 5a against "the same selection of genes for all perturbations". Both arms of that published comparison are reproduced below using their control, not ours. The gene ranking is the one behind Table I.4, the squared min-max normalized $|t|$ against synthetic non-control references, which is strictly increasing in $|t|$. Ranking by those weights or by $|t|$ therefore gives exactly the same top-k sets. Table I.5 scores the same three identity-free inputs when perturbation p is scored only on its own top-k genes, at $k \in \{50, 100, 200\}$, under both distances,

Table I.4: **Own-response weighting puts identity-free predictions below the floor.** The weighting of Vollenweider & Bühlmann (2026) is computed from cells for each perturbation. Nine values are evaluated per panel, spanning three seeds for synthetic references across three identity-free inputs, the control mean, the fitted shared response and the panel mean. The shared-weight column scores the same predictions with the per-perturbation weights averaged into one vector. The last column is the median number of cells a held-out perturbation contributes to its own weights.

dataset	N	floor	excess over floor own-response weights	shared weights |excess|	cells per pert
arc-h1	30	0.5167	-0.276 to -0.249	1×10^{-16}	769
replogle	181	0.5028	-0.369 to -0.277	0	133
norman	56	0.5089	-0.188 to -0.111	0	259
adamson	18	0.5278	-0.170 to -0.117	0	536
rpe1	163	0.5031	-0.355 to -0.235	1×10^{-16}	100

Table I.5: **Own top-k genes put an identity-free prediction below the floor.** Each entry spans three seeds for the reference sampling by three identity-free inputs, the control mean, the fitted shared response and the panel mean. The seeds move an entry by at most 0.016. Under ℓ_1 the same 135 settings run from -0.439 to -0.074. One gene set chosen once for the panel, at the same k, returns the floor to 1.1×10^{-16} in every entry, as required by the criterion fixed in advance.

			excess over the floor, own top-k genes, ℓ_2		
dataset	N	floor	$k = 50$	$k = 100$	$k = 200$
arc-h1	30	0.5167	-0.279 to -0.232	-0.252 to -0.210	-0.201 to -0.168
replogle	181	0.5028	-0.378 to -0.277	-0.363 to -0.254	-0.340 to -0.221
norman	56	0.5089	-0.195 to -0.118	-0.172 to -0.111	-0.167 to -0.091
adamson	18	0.5278	-0.157 to -0.123	-0.139 to -0.108	-0.127 to -0.099
rpe1	163	0.5031	-0.371 to -0.220	-0.341 to -0.189	-0.300 to -0.151

at three seeds. Every one of the 270 settings lands below the floor, by 0.074 to 0.439, and the effect shrinks as the gene set grows, with ℓ_2 medians -0.244, -0.210 and -0.169. The fewer genes retained, the more completely the subspace isolates directions where that perturbation's response is extreme, and at the full panel the effect is zero by Proposition 2. The same count of genes chosen once for the whole panel, which treats perturbations alike, returns the floor to 1.1×10^{-16} in all 270, providing a panel-wide selection control. The Virtual Cell Challenge scores its discrimination metric on the full panel, which this paper's reproduction matched on all 2000 genes, so none of this describes the challenge.

The floor is that evaluation's own chance level. $(N{+}1)/(2N)$ is the chance level of an evaluation that compares every perturbation in the same way, and the top-k restriction is by design not one, so a score below $(N{+}1)/(2N)$ could mean only that its chance level lies elsewhere. We computed the chance level under the evaluation itself using a null specified before computation. The null still scores every row on the gene set of some perturbation, so the evaluation stays exactly as uneven across rows. Uniformly assigning these sets to rows breaks only the coupling between a row and its own truth. Its mean has a closed form, which 2000 permutations per setting matched to within 2.76 standard errors against the specified limit of 4. In all 270 settings behind Table I.5 that mean equals $(N{+}1)/(2N)$ to 1.1×10^{-16}, with no tie term, and the observed deficit of 0.074 to 0.439 lies 3.5 to 53.3 null standard deviations below it, with at most 2 of the 2000 permutations reaching the observed score in any setting. The floor is therefore the evaluation's own chance level, and what drives the score below it is the coupling between a row's gene set and its own truth.

Figure I.1 compares the ranges produced by panel-only and own-truth scoring rules on identity-free predictions.

The per-prediction quotient remains a bound. $\overline{\mathrm{PDS}}^{\star}_{\mathrm{pred}}$ is still the tightest scalar-invariant upper bound on $\overline{\mathrm{PDS}}$, and Proposition 1 still computes it exactly. It is not evidence of retained identity, because a value of 0.80 under that quotient is reachable without biological signal. In Appendix N,

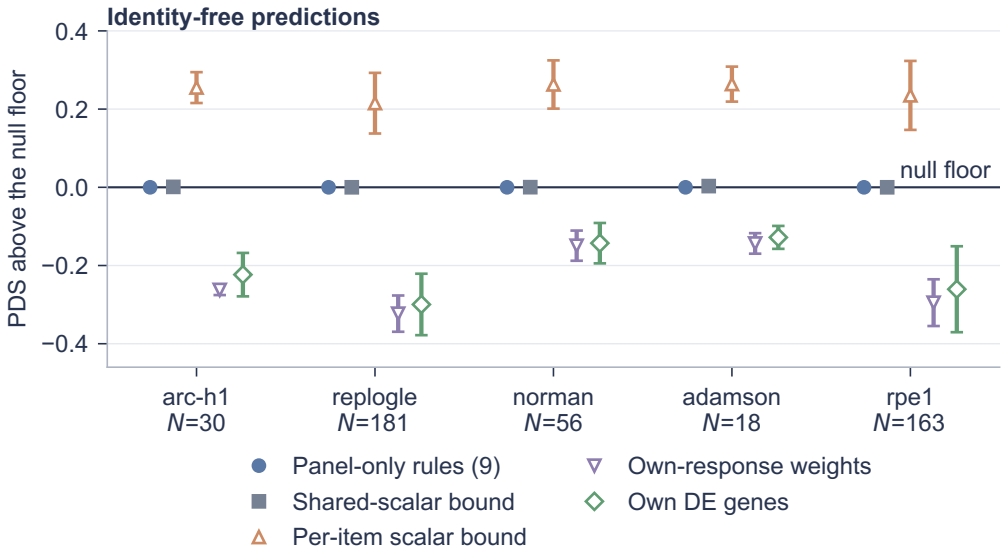


Figure I.1: **An evaluation that reads nothing specific to one perturbation returns the chance floor when every perturbation is given the same vector.** None is then recognizable. Filled marks denote evaluations with no row-specific input. The nine fixed rules of Table I.2 sit on the floor exactly. The best case over one shared scalar clears it by $1/N^2$ under Equation (2), one rank for one perturbation and invisible at this scale, and not at all under the official package. Hollow marks let each perturbation's own truth choose how it is compared. Specifically, a best case chosen per prediction lifts the floor, weighting by each perturbation's own response and restricting to its own top-k differentially expressed genes push it below. The deployable fixed norm reads each perturbation's support cells, legitimately falling outside the condition, and is not shown (Table I.2). The figure shows the condition is sufficient, not that it is necessary. Bars are min–max ranges over the settings behind each construction, three identity-free vectors for the first three, and additionally three seeds for the last two and three gene counts for the last. The central mark is the midpoint of the range, not a mean or a median, and the vertical axis is each panel's score minus its own floor. ℓ_2 throughout, so the last mark covers 135 of the 270 settings and its ℓ_1 half, which carries both extremes of the range quoted in the text, appears only in Table I.5. The plotted ranges are derived from the corresponding three tables.

the reported score penalizes an information-preserving map 18.6 times as much as the per-prediction quotient. This measures the penalty removed by one scalar per prediction, not the quotient's accuracy as a measure of identity. The same caution applies to the rescoring of Appendix J, which ranks estimators under this quotient. The ρ ordering does not inherit the problem, because it is evaluated under the global group, which the first table of this section shows is inert on an identity-free prediction, and the exact inverse of Section 6 is measured at $\beta = 1$ under no quotient at all.

These conclusions apply only to the evaluation constructions tested. Cosine and a common norm do not have this defect, and what the deployable norm gains is information its support cells carry. The oracle norm with which the fixed-norm remedy is implemented does have it, as does the best-case quantity one would naturally compute in support of the remedies. Interpreting the latter as a bound therefore requires an accompanying identity-free control.

Near-collapse bounds are too loose for diagnosis. Proposition 2 assumes every prediction is the same vector, which is the endpoint $\beta = 0$ of the shrinkage path and not the partial shrinkage the rest of this paper measures. It has a continuous version under Equation (2). For a dissimilarity φ that does not take the row index and is L-Lipschitz in the prediction, fix a reference $\bar{d}$ and write $v_j = \varphi(\bar{d}, \Delta_j)$ and $\eta_i = L\|d_i - \bar{d}\|$. Row i's values then lie within η_i of the reference row, so only

Table J.1: **The top-ranked estimator changes under rescoring only on the two panels with the smallest libraries, under either clipping convention.** The text uses predictions before the scoring pipeline's clip. Official scores use predictions after it. All changes, 11 before the clip and 12 after, fall on the two datasets with at most 30 held-out perturbations, and none on the three with at least 56. The absorbed fraction does not order the same way, and library size cannot be separated from dataset here.

		before the clip		after the clip		
dataset	n_{test}	top-1	pairs	top-1	pairs	global absorption
adamson	18	5/6	13/60	6/6	12/60	41.2%
arc-h1	30	6/6	27/60	6/6	26/60	95.3%
norman	56	0/6	2/60	0/6	2/60	55.4%
rpe1	163	0/6	5/60	0/6	7/60	1.3%
replogle	181	0/6	11/60	0/6	11/60	82.2%

a candidate whose v_j lies within $2\eta_i$ of v_i can change sides, and

$$\left|\overline{\mathrm{PDS}} - \frac{N+1}{2N}\right| \;\le\; \frac{T}{N^2} + \frac{1}{N^2}\sum_i \#\{j \neq i : 0 < |v_i - v_j| \le 2\eta_i\},$$

with T the number of tied pairs among the v_j. At $\eta_i = 0$ this is Proposition 2. Along the path $\bar{\delta} + \beta(z - \bar{\delta})$ from the fitted shared response to the unclipped empirical-Bayes prediction z, on the five panels at $k \in \{1, 32\}$ under ℓ_1 and ℓ_2, the bound held at all 580 points evaluated, to 1.1×10^{-16}. It is too loose for diagnosis. The pre-specified test required a half-width at or below 0.1 beyond negligible shrinkage. In each of the 20 panel, support-size and distance combinations, this holds only up to β between 0.02 and 0.10. Under ℓ_2, the median across the ten combinations reaches 0.91 at $\beta = 1$. So Proposition 2 describes prediction sets within a few percent of one vector and says little about partial collapse.

J Estimator ranking changes by dataset

Before inspecting results, we specified comparisons of the top-ranked estimator under PDS@1 and $\overline{\mathrm{PDS}}^{\star}_{\mathrm{pred}}$, Kendall τ between the orderings, and pairwise inversions. Across the 30 settings the top-ranked estimator changes in 11, the median τ is 0.80, and 58 of the 300 pairs invert. Eleven of the twelve settings with $n_{\text{test}} \le 30$ change and none of the eighteen with $n_{\text{test}} \ge 56$ do. Absorption does not separate these groups. On replogle 82.2% is absorbed with no changes, and on adamson 41.2% with five (Table J.1).

Two limitations apply. $\overline{\mathrm{PDS}}^{\star}_{\mathrm{pred}}$ lifts a prediction with no identity to about 0.8 (Appendix I), so a change of order under it measures how far the quotient can reorder estimators and is not a corrected leaderboard. Library size is confounded with dataset, since the twelve small-library settings come from two panels.

The pattern persists under both clipping conventions. The counts above read every estimator before the scoring pipeline's clip. For the linear transport map this is not an unclipped prediction, because the estimator clips its own output to the observed expression range before returning it (Appendix O). After clipping every estimator, as in official scoring, the top-ranked estimator changes in 12 of the 30 settings, covering all twelve small-library settings and none of the eighteen large ones. There are 58 inversions among the 300 pairs, and median τ remains 0.80. The one added change is adamson at $k = 4$. Both conventions are reported in Table J.1.

K Matched permutations of real predictions

The constant-prediction control isolates a scoring rule's access to perturbation-specific information. Matched permutations extend the check to the actual prediction distribution. We analyze all 270 saved prediction arrays from five estimators at six support sizes on the first draw and the unshrunk and empirical-Bayes estimates on two further draws, across all five panels. No setting is excluded.

This post hoc analysis neither retrains nor selects models. Each array has 2,000 uniformly sampled permutations, including fixed points. A panel, support size and draw share the same permutations across estimators and corrections. Perturbation names determine the common alphabetical candidate order.

Let $u_i = d_i/\|d_i\|$ and $s_i = \|\hat{\delta}_i\|$, where s_i is computed from the unshrunk support estimate. Zero predictions remain zero. For a permutation π, the whole-bundle null moves (d_i, s_i) together to $(d_{\pi(i)}, s_{\pi(i)})$. Under support-norm correction its prediction is $s_{\pi(i)}u_{\pi(i)}$. The direction-only null instead uses $s_i u_{\pi(i)}$, retaining the receiving perturbation's support information. For uncorrected distances it uses $\|d_i\|u_{\pi(i)}$, retaining the receiving prediction's original norm. Cosine gives the same null in both cases. The own-truth rule always uses $\|\Delta_i\|u_{\pi(i)}$, so it reads the receiving answer under either permutation and the two null distributions coincide. A direction-only null is therefore not an identity-free input when its retained norm contains perturbation information.

We score raw predictions by ℓ_1, ℓ_2 and cosine, and support- and own-truth-normalized predictions by both ℓ_p distances. All use the same genes, candidate set and official sorting/tie convention. The implementation matches valid real-data scores and rejects a misaligned pairing. Constant predictions give the exact reference floor, truth predictions give a score of one, and cyclically mismatched truths score lower. Zero-vector counts are recorded, with zero cosine similarity whenever either vector is zero. Affine OT retains its internal clip, as in the original estimator comparison.

Information retained by matched nulls. Whole-bundle null means for raw ℓ_1, raw ℓ_2, cosine and support-norm corrections are within 0.00414 of the relevant panel's exact chance reference. This finite Monte Carlo variation is largest on the smallest panel. Retaining a receiving support norm raises the ℓ_1 direction-null mean as high as 0.643. Own-truth ℓ_1 null means range from 0.508 to 0.671 and exceed the corresponding reference floor at every setting. The ℓ_2 range is 0.509–0.670. Matched nulls thus retain information supplied by a correction. This does not invalidate support information or justify replacing an observed score with its difference from a null mean.

Observed scores do not exceed the matched null range everywhere. For empirical Bayes on adamson at $k = 1$ on the first draw, support-norm ℓ_1 is 0.608025, within the direction-null range [0.503086, 0.611111]. Support-norm ℓ_2 is 0.614198 and lies inside both null ranges. Own-truth ℓ_1 scores 0.537037, below its null mean of 0.538866. These exceptions prevent a claim that every corrected prediction retains distinguishable directional identity. We do not use overlap counts as multiplicity-adjusted significance tests.

Table K.1 expands the main figure to all five estimators at the prespecified $k = 8$, first-draw operating point. Results for all six sizes and three draws include every observed perturbation score, null mean and permutation quantile, without selection by performance. Figure 2 distinguishes observed-score uncertainty from randomization variation. Its observed intervals use 2,000 paired perturbation bootstrap samples, conditional on the fixed candidate library. Null bars are the 2.5th and 97.5th percentiles of permuted mean scores. They are not confidence intervals on a population mean. For draw-averaged observed scores we first average each perturbation's repeated draws, then resample perturbations. The five panels remain the cross-panel evidence units.

K.1 Magnitude-stratified permutations

A uniform permutation over the whole library need not preserve structure shared by similar perturbations. We therefore repeated both null constructions within strata. This sensitivity analysis was specified after the global results were known and before any stratified draw. Within each scored library the candidates are ordered by the Euclidean norm of their true pseudobulk response, with ties broken by the scorer's candidate order. That order is split into three contiguous groups of sizes as equal as possible. Each of the 2,000 permutations acts uniformly and independently within every stratum. A receiving perturbation obtains a prediction, and under the bundle null its support norm, from the same stratum. Scoring, rules, null constructions and permutation count are those of the global analysis, on the same 270 internal prediction sets and the 24 published model-context pairs of Appendix L. The smallest stratum in these populations has six members, allowing permutation without exclusions. The method-specific libraries of Appendix M are smaller. The strata read candidate truths, which the evaluator already holds, and define a reference distribution without altering any prediction. Batch was not used as a stratum. Only arc-h1 records batches, and each of its held-

Table K.1: **Matched-null behavior extends beyond the known-map estimator.** All entries are ℓ_1 PDS at $k = 8$ on the first support draw. Null entries are means of 2,000 permutations. Raw scores are compared with the whole-bundle null, support-norm scores with the direction-only null retaining the receiving support norm, and own-truth norms read the receiving answer under both nulls. Rounding to three decimals can make close means appear identical. Full-precision scores, both nulls, ℓ_2, cosine and intervals accompany these summaries. Estimator configurations and biological panels are unchanged from the primary evaluation.

		Raw		Support norm		Own-truth norm	
Panel	Estimator	Observed	Bundle	Observed	Direction	Observed	Null
arc-h1	Unshrunk	0.948	0.517	0.948	0.532	0.936	0.539
	EB	0.804	0.517	0.936	0.539	0.934	0.544
	Affine OT	0.850	0.517	0.912	0.537	0.904	0.546
	Projection	0.854	0.517	0.887	0.537	0.899	0.555
	Pooler	0.801	0.517	0.867	0.537	0.876	0.552
replogle	Unshrunk	0.961	0.503	0.961	0.512	0.927	0.515
	EB	0.766	0.503	0.932	0.544	0.950	0.545
	Affine OT	0.866	0.503	0.951	0.521	0.939	0.529
	Projection	0.913	0.503	0.952	0.523	0.961	0.535
	Pooler	0.864	0.503	0.916	0.528	0.946	0.544
norman	Unshrunk	0.982	0.508	0.982	0.549	0.976	0.557
	EB	0.858	0.508	0.978	0.576	0.981	0.585
	Affine OT	0.900	0.508	0.951	0.556	0.944	0.565
	Projection	0.881	0.508	0.945	0.572	0.955	0.595
	Pooler	0.739	0.508	0.910	0.582	0.928	0.606
adamson	Unshrunk	0.917	0.527	0.917	0.552	0.914	0.573
	EB	0.744	0.528	0.787	0.564	0.799	0.575
	Affine OT	0.895	0.526	0.898	0.551	0.910	0.570
	Projection	0.904	0.527	0.895	0.547	0.929	0.572
	Pooler	0.886	0.527	0.895	0.549	0.910	0.566
rpe1	Unshrunk	0.947	0.503	0.947	0.546	0.960	0.580
	EB	0.815	0.503	0.751	0.564	0.858	0.670
	Affine OT	0.861	0.502	0.876	0.549	0.908	0.602
	Projection	0.882	0.503	0.861	0.549	0.912	0.619
	Pooler	0.854	0.503	0.787	0.554	0.876	0.647

out perturbations spans 46 to 48 of its 48 batches, so batch does not assign perturbations to groups. With one stratum and the global random stream, the stratified generator reproduces the stored global nulls exactly.

Stratification raises most null means. Across the 270 internal sets the median increase is 0.036 for raw ℓ_1, 0.050 for support-norm ℓ_1 and 0.035 for own-truth ℓ_1 under the direction null, with maxima of 0.161, 0.161 and 0.149, all on adamson, the smallest library. The increase is consistent with shared direction among perturbations of similar magnitude. Under ℓ_1, no internal observation changes position relative to either null range under any rule, and counts above the direction-null upper quantile remain 267, 269 and 269 of 270 for the three rules. Four internal cases move from above the upper quantile to inside the stratified range, all at $k = 1$ on the first draw. They are affine OT on arc-h1 under raw and support-norm ℓ_2, and empirical Bayes on adamson under cosine and own-truth ℓ_2. For published models, own-truth ℓ_1 leaves the same 14 of 24 pairs outside their ranges, with no status changes. No pair changes status under own-truth ℓ_2, raw ℓ_1 or cosine either. Two pairs move inside the stratified range under other rules. One is scGPT on Adamson's first split, under raw ℓ_2 and the training-fixed ℓ_2 norm. The other is GEARS on RPE1's second split, under both training-fixed norms. Tables K.2 and K.3 give the values behind Figure 2 and Table L.1. For these cases, separation depends on the permutation reference. Neither reference is a population-level exchangeability test, and the three strata do not exclude finer magnitude structure. Appendix M applies the same stratified reference to every published method on its own library.

Table K.2: **Magnitude strata raise the null but leave the figure's comparisons unchanged.** Empirical Bayes at $k = 8$, first draw, direction-only null. Null entries give the permutation mean and the upper 97.5% quantile of 2,000 permutations, global or restricted to three strata of true response norm. Every observed score exceeds both upper quantiles.

		support-norm ℓ_1			own-truth ℓ_1		
panel	N	observed	global	stratified	observed	global	stratified
arc-h1	30	0.936	0.539/0.612	0.592/0.676	0.934	0.544/0.606	0.593/0.670
replogle	181	0.932	0.544/0.570	0.545/0.574	0.950	0.545/0.569	0.552/0.578
norman	56	0.978	0.576/0.631	0.597/0.657	0.981	0.585/0.634	0.605/0.663
adamson	18	0.787	0.564/0.651	0.658/0.719	0.799	0.575/0.654	0.652/0.722
rpe1	163	0.751	0.564/0.575	0.569/0.580	0.858	0.670/0.681	0.674/0.686

Table K.3: **Published-model own-truth scores keep their status under magnitude strata.** Entries give the stratified direction-null mean and upper 97.5% quantile for the own-truth ℓ_1 scores of Table L.1. †: observation inside the stratified central 95% range, the same ten cases as under the global null, whose means are in Table L.1. Mean is constant, so its null equals its score.

Screen / split	N	GEARS	scGPT	Linear	Mean
Adamson / 1	21	0.5636/0.6122	0.5617/0.6145	0.5451/0.6032†	0.5329/0.5329†
Adamson / 2	21	0.5613/0.6190	0.5315/0.5465†	0.5639/0.6304†	0.5261/0.5261†
K562 / 1	107	0.5854/0.6247	0.5504/0.5795	0.6179/0.6499	0.6281/0.6281†
K562 / 2	108	0.6042/0.6445	0.5707/0.6067	0.6227/0.6606	0.6373/0.6373†
RPE1 / 1	171	0.7515/0.7605	0.7218/0.7284	0.7443/0.7513	0.7490/0.7490†
RPE1 / 2	164	0.7882/0.7924†	0.6734/0.6850	0.7668/0.7743	0.7720/0.7720†

L Application to published external-model predictions

We apply the diagnostic to published GEARS, scGPT, linear-model and mean predictions from Ahlmann-Eltze et al. (2025), using both supplied partitions of Adamson, Replogle K562 essential and Replogle RPE1 essential. No model is retrained. These screens overlap the primary study. The six contexts therefore test different models and protocols, not six independent biological populations. The source training, validation and test assignments are preserved. Prediction record identifiers and publication-normalized labels are matched exactly to source parameter keys. Entries absent from the source parameter keys are counted separately and are not test perturbations. Duplicate labels matched to those keys are rejected. We align expression vectors by gene identifiers and use the 1,000 genes with highest control expression, retaining source order at ties. Each selected name must occur once in the original axis of every scored prediction, training response and control. The two K562 entries named TBCE rank 4,355 and 4,437 in control expression. Both lie outside this fixed set and are neither merged nor substituted. The candidate library is the intersection of finite held-out predictions across the four models and ground truth. The gene set stays fixed across targets, with no target-label-dependent deletion. This reproduces the PDS ranking and tie conventions, not the source publication's scalar-error evaluation or its nine-model intersection.

Predicted and measured responses are relative to the source control pseudobulk. The fixed normalization magnitude is the mean Euclidean norm of the training responses, excluding control. It reads no held-out response. Own-truth normalization uses the receiving target's measured norm and remains an oracle diagnostic. No few-shot support is constructed, and no inverse is assumed for an unknown model map. Coverage and exclusions are reported in Table L.2, and the conversion code also exports named exclusions.

For each context, 2,000 shared uniform permutations preserve pairing across models. Whole-bundle permutations move complete predictions, while direction-only permutations retain the receiving prediction's magnitude. Both use the same training-fixed norm under fixed normalization and the receiving truth norm under own-truth normalization. Observed intervals and model differences use 2,000 shared perturbation bootstraps conditional on the fixed candidate library. Null intervals are central randomization ranges, not confidence intervals. The two source partitions are reported

Table L.1: **Published-model scores and their matched nulls on common libraries.** All 24 cases of the four original methods on the common finite targets of each split (N). Entries give own-truth ℓ_1 score / direction-null mean. †: observation inside the central 95% range of 2,000 permutations (not a confidence interval). Mean is constant. Linear is the self-trained linear model. Splits reuse three biological screens. The same ten cases fall inside when permutations are restricted to magnitude strata (Appendix K.1).

Screen / split	N	GEARS	scGPT	Linear	Mean
Adamson / 1	21	$0.6327/0.5335$	$0.6349/0.5392$	$0.5760^{\dagger}/0.5472$	$0.5329^{\dagger}/0.5329$
Adamson / 2	21	$0.6304/0.5380$	$0.5442^{\dagger}/0.5310$	$0.5986^{\dagger}/0.5528$	$0.5261^{\dagger}/0.5261$
K562 / 1	107	$0.7513/0.5742$	$0.6456/0.5474$	$0.6791/0.6195$	$0.6281^{\dagger}/0.6281$
K562 / 2	108	$0.7303/0.5925$	$0.7014/0.5587$	$0.7330/0.6175$	$0.6373^{\dagger}/0.6373$
RPE1 / 1	171	$0.7737/0.7487$	$0.7331/0.7153$	$0.7624/0.7427$	$0.7490^{\dagger}/0.7490$
RPE1 / 2	164	$0.7889^{\dagger}/0.7888$	$0.6920/0.6592$	$0.7883/0.7664$	$0.7720^{\dagger}/0.7720$

separately. The analysis code evaluates identity-aligned, intentionally misaligned and constant-prediction controls alongside the observed scores. The supplied mean baseline is also checked for equality across targets.

The common libraries contain 21/21 source test perturbations in both Adamson splits, 107/272 and 108/272 in K562, and 171/384 and 164/384 in RPE1. GEARS, mean and ground truth cover all source test labels with finite scored genes. The excluded K562 and RPE1 labels are missing from scGPT predictions and have nonfinite linear-model predictions. These sets coincide within each split. No value is imputed. The results describe the common candidate libraries, not the complete source test sets. Accompanying results retain named exclusions and source-matching counts.

Training-fixed norm is not a uniform improvement. Its ℓ_1 change ranges from -0.0671 to $+0.0444$ over the 24 model/context pairs. For GEARS, the K562 first-split change is $+0.0444$ (95% paired conditional interval $[0.0091, 0.0804]$), but the RPE1 changes are -0.0667 ($[-0.0970, -0.0369]$) and -0.0671 ($[-0.0983, -0.0380]$). The K562 second-split change is -0.0040 ($[-0.0421, 0.0331]$). Corresponding ℓ_2 changes across all models range from -0.0657 to $+0.0438$.

The mean baseline is exactly constant across perturbations in all six contexts. Raw and training-fixed scores therefore attain their constant-prediction reference. Own-truth ℓ_1 instead gives it 0.5329 and 0.5261 on Adamson, 0.6281 and 0.6373 on K562, and 0.7490 and 0.7720 on RPE1. Its null distributions equal the observed score because reassigning identical source directions has no effect. On RPE1's second split, GEARS reaches 0.7889 with own-truth normalization, while the null mean is 0.7888 and its central range is $[0.7845, 0.7931]$. This comparison does not resolve directional identity, despite the high score. The comparison differs on K562's first split, where GEARS scores 0.7513 against an own-truth null mean of 0.5742. All evaluated predictions and truths have nonzero norms. These results extend the information-access diagnostic to published predictors. They establish neither a model-independent inverse nor improved biological prediction.

Table L.1 gives the common-library own-truth scores beside their null means. Figure L.1 shows all contexts. Tables L.3–L.8 give distance scores, Table L.9 gives cosine, and Tables L.10–L.11 give paired model differences.

L.1 Sensitivity to available prediction coverage

To assess sensitivity to coverage, we score each model on all its finite test predictions, retaining the genes, training norm and resampling rules above. This descriptive analysis was specified after the coverage losses were observed. Eight libraries expand, with GEARS and the mean baseline reaching all 272 K562 and 384 RPE1 targets in both splits. The other 16 reproduce the common-library results exactly. Each mean comparator uses the corresponding model's library. Between-library changes combine differences in target composition and ranking candidates. They are not paired effects.

On RPE1 split 2, GEARS still lies within its own-truth null range, scoring 0.7856 versus $[0.7818, 0.7872]$ (Table L.12). It nevertheless exceeds the same-library mean by 0.0223 ($[0.0145, 0.0302]$). Fixed-norm changes span -0.0822 to $+0.0316$. K562 split 2 changes sign

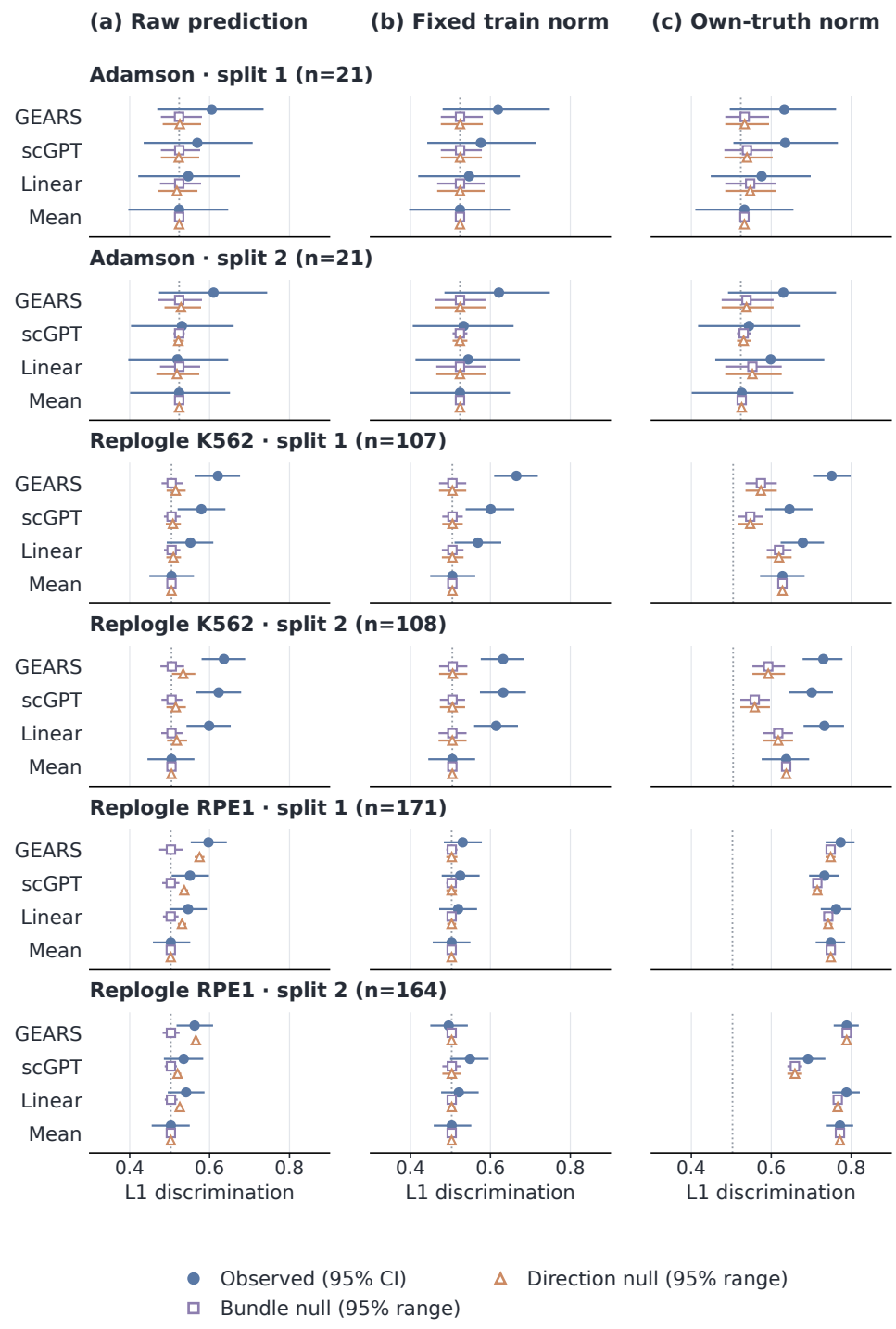


Figure L.1: **Matched controls for published predictions.** All six source contexts and four models. Columns (a)–(c) show raw ℓ_1, training-fixed norm and oracle own-truth norm scores. Filled circles give observed means with conditional 95% perturbation-bootstrap intervals. Open squares and triangles give bundle and direction nulls with central 95% permutation ranges (2,000 resamples each). Dotted lines mark $(N+1)/(2N)$, which need not apply to own-truth normalization. Linear denotes the self-trained linear model. Source splits are not independent biological replicates.

Table L.2: External prediction coverage. The two published splits of each dataset reuse biological screens and are not independent screens. The common test set is the intersection of finite predictions and truth across all four predefined models. Training counts exclude controls. The lower panel gives source-test exclusions by model or truth, with overlapping exclusions counted once in the common-set total.

Dataset	Split	Train	Source test	Common test	Excluded	Genes
Adamson	1	54	21	21	0	1000
Adamson	2	54	21	21	0	1000
Replogle K562	1	733	272	107	165	1000
Replogle K562	2	733	272	108	164	1000
Replogle RPE1	1	1035	384	171	213	1000
Replogle RPE1	2	1035	384	164	220	1000

Dataset	Split	GEARS	scGPT	Linear	Mean	Truth
Adamson	1	0	0	0	0	0
Adamson	2	0	0	0	0	0
Replogle K562	1	0	165	165	0	0
Replogle K562	2	0	164	164	0	0
Replogle RPE1	1	0	213	213	0	0
Replogle RPE1	2	0	220	220	0	0

from -0.0040 to $+0.0069$, with both intervals containing zero (Table L.13). K562 split 1 retains a positive correction effect, and both RPE1 splits retain negative effects. These coverage results support these observations within their stated scope but add no independent biological replication.

M MATCHED REFERENCES FOR EVERY PUBLISHED METHOD

We extend the published-model analysis of Appendix L to every prediction method in the same archive (Ahlmann-Eltze et al., 2025). The inclusion rule, endpoints and reporting rule were fixed before extracting any predictions from the added methods. Genes, partitions, the training-fixed norm, the official sort and the 2,000 paired permutations and bootstraps per context are those of Appendix L, and the magnitude-stratified reference is that of Appendix K.1. Each method is scored on its own finite test library, as in Appendix L.1, and compared with the published mean on the identical targets, because an intersection across all methods would discard most targets. Repeated extraction of the GEARS, scGPT, self-trained linear and mean predictions reproduces their libraries exactly. Their full-library scores and paired intervals also reproduce the earlier values with zero difference.

The archive holds 16 prediction methods per context. Two linear models, with scGPT and scFoundation gene embeddings, omit some of the 1,000 scored genes in every context. They are excluded because the protocol requires a fixed gene space. The other 13 methods and the mean are scored in all six contexts, giving 78 model–context pairs besides the mean (Table M.2). The linear models use perturbation embeddings from the training data, GEARS, the K562 or RPE1 screen or random vectors, or random gene embeddings. Nonfinite predictions leave the K562- and RPE1-embedding models 8 to 14 of the 21 Adamson targets, so their Adamson intervals are wide.

Class A requires a paired 95% bootstrap interval for model minus mean above zero and an observed score inside the central 95% direction-null range. A pair is in class B when it beats the mean and its score lies above that range, and in class C when it does not beat the mean. Table M.1 gives the counts for three ℓ_1 rules and both references, and Table M.3 the classes under raw and training-fixed ℓ_1. Under own-truth ℓ_1, 4 of the 32 pairs that beat the mean lie inside their global range and 7 inside their stratified range. The three added by strata are the K562- and RPE1-embedding linear models on Adamson libraries of 8 to 13 targets. The median share of the gain over the mean that the direction-null mean already reaches is near zero under own-truth and training-fixed ℓ_1 and 15% under raw ℓ_1, a descriptive summary, not a metric.

Table L.3: Adamson external-model ℓ_1 discrimination. Each entry is a mean followed by an interval. Observed intervals are conditional 95% paired-perturbation bootstrap intervals. Null intervals are central 95% randomization ranges, not confidence intervals. All use 2,000 resamples. Fixed train norm uses training responses only, and own-truth norm is an oracle diagnostic. Linear denotes the published self-trained linear model.

Model	Observed [95% CI]	Bundle null [95% range]	Direction null [95% range]
Split 1 ($n = 21$), raw			
GEARS	0.6054 [0.4717, 0.7324]	0.5236 [0.4807, 0.5782]	0.5255 [0.4853, 0.5760]
scGPT	0.5692 [0.4376, 0.7052]	0.5241 [0.4807, 0.5737]	0.5229 [0.4807, 0.5714]
Linear	0.5465 [0.4240, 0.6735]	0.5240 [0.4785, 0.5760]	0.5181 [0.4739, 0.5669]
Mean	0.5238 [0.3991, 0.6440]	0.5238 [0.5238, 0.5238]	0.5238 [0.5238, 0.5238]
Split 1 ($n = 21$), fixed train norm			
GEARS	0.6190 [0.4830, 0.7460]	0.5239 [0.4784, 0.5783]	0.5239 [0.4784, 0.5783]
scGPT	0.5760 [0.4444, 0.7121]	0.5241 [0.4785, 0.5760]	0.5241 [0.4785, 0.5760]
Linear	0.5465 [0.4218, 0.6713]	0.5239 [0.4694, 0.5828]	0.5239 [0.4694, 0.5828]
Mean	0.5238 [0.3991, 0.6463]	0.5238 [0.5238, 0.5238]	0.5238 [0.5238, 0.5238]
Split 1 ($n = 21$), own-truth norm			
GEARS	0.6327 [0.4988, 0.7597]	0.5335 [0.4875, 0.5918]	0.5335 [0.4875, 0.5918]
scGPT	0.6349 [0.5079, 0.7642]	0.5392 [0.4853, 0.6009]	0.5392 [0.4853, 0.6009]
Linear	0.5760 [0.4512, 0.6961]	0.5472 [0.4875, 0.6100]	0.5472 [0.4875, 0.6100]
Mean	0.5329 [0.4126, 0.6531]	0.5329 [0.5329, 0.5329]	0.5329 [0.5329, 0.5329]
Split 2 ($n = 21$), raw			
GEARS	0.6100 [0.4762, 0.7416]	0.5239 [0.4739, 0.5782]	0.5283 [0.4898, 0.5760]
scGPT	0.5306 [0.4059, 0.6576]	0.5238 [0.5125, 0.5351]	0.5218 [0.5125, 0.5329]
Linear	0.5193 [0.3990, 0.6440]	0.5241 [0.4785, 0.5737]	0.5184 [0.4694, 0.5714]
Mean	0.5238 [0.4036, 0.6485]	0.5238 [0.5238, 0.5238]	0.5238 [0.5238, 0.5238]
Split 2 ($n = 21$), fixed train norm			
GEARS	0.6213 [0.4875, 0.7461]	0.5239 [0.4648, 0.5850]	0.5239 [0.4648, 0.5850]
scGPT	0.5329 [0.4082, 0.6553]	0.5235 [0.5079, 0.5397]	0.5235 [0.5079, 0.5397]
Linear	0.5442 [0.4150, 0.6712]	0.5238 [0.4671, 0.5850]	0.5238 [0.4671, 0.5850]
Mean	0.5238 [0.4014, 0.6463]	0.5238 [0.5238, 0.5238]	0.5238 [0.5238, 0.5238]
Split 2 ($n = 21$), own-truth norm			
GEARS	0.6304 [0.4943, 0.7596]	0.5380 [0.4785, 0.6032]	0.5380 [0.4785, 0.6032]
scGPT	0.5442 [0.4195, 0.6689]	0.5310 [0.5170, 0.5465]	0.5310 [0.5170, 0.5465]
Linear	0.5986 [0.4625, 0.7302]	0.5528 [0.4875, 0.6236]	0.5528 [0.4875, 0.6236]
Mean	0.5261 [0.4036, 0.6531]	0.5261 [0.5261, 0.5261]	0.5261 [0.5261, 0.5261]

Table L.4: Adamson external-model ℓ_2 discrimination. Each entry is a mean followed by an interval. Observed intervals are conditional 95% paired-perturbation bootstrap intervals. Null intervals are central 95% randomization ranges, not confidence intervals. All use 2,000 resamples. Fixed train norm uses training responses only, and own-truth norm is an oracle diagnostic. Linear denotes the published self-trained linear model.

Model	Observed [95% CI]	Bundle null [95% range]	Direction null [95% range]
Split 1 ($n = 21$), raw			
GEARS	0.6100 [0.4739, 0.7393]	0.5237 [0.4807, 0.5782]	0.5251 [0.4853, 0.5737]
scGPT	0.5782 [0.4399, 0.7166]	0.5244 [0.4830, 0.5737]	0.5243 [0.4830, 0.5737]
Linear	0.5646 [0.4330, 0.6894]	0.5241 [0.4785, 0.5737]	0.5231 [0.4762, 0.5737]
Mean	0.5238 [0.4013, 0.6463]	0.5238 [0.5238, 0.5238]	0.5238 [0.5238, 0.5238]
Split 1 ($n = 21$), fixed train norm			
GEARS	0.6145 [0.4807, 0.7415]	0.5238 [0.4785, 0.5782]	0.5238 [0.4785, 0.5782]
scGPT	0.5782 [0.4422, 0.7166]	0.5245 [0.4807, 0.5737]	0.5245 [0.4807, 0.5737]
Linear	0.5533 [0.4240, 0.6758]	0.5238 [0.4694, 0.5805]	0.5238 [0.4694, 0.5805]
Mean	0.5238 [0.4014, 0.6417]	0.5238 [0.5238, 0.5238]	0.5238 [0.5238, 0.5238]
Split 1 ($n = 21$), own-truth norm			
GEARS	0.6304 [0.4943, 0.7574]	0.5327 [0.4853, 0.5918]	0.5327 [0.4853, 0.5918]
scGPT	0.6417 [0.5102, 0.7778]	0.5374 [0.4898, 0.5964]	0.5374 [0.4898, 0.5964]
Linear	0.5850 [0.4603, 0.7052]	0.5490 [0.4898, 0.6145]	0.5490 [0.4898, 0.6145]
Mean	0.5306 [0.4082, 0.6485]	0.5306 [0.5306, 0.5306]	0.5306 [0.5306, 0.5306]
Split 2 ($n = 21$), raw			
GEARS	0.5986 [0.4603, 0.7325]	0.5242 [0.4762, 0.5760]	0.5291 [0.4898, 0.5760]
scGPT	0.5624 [0.4286, 0.6917]	0.5239 [0.5102, 0.5397]	0.5229 [0.5102, 0.5374]
Linear	0.5306 [0.4081, 0.6576]	0.5240 [0.4785, 0.5738]	0.5194 [0.4739, 0.5646]
Mean	0.5238 [0.4014, 0.6485]	0.5238 [0.5238, 0.5238]	0.5238 [0.5238, 0.5238]
Split 2 ($n = 21$), fixed train norm			
GEARS	0.6145 [0.4785, 0.7460]	0.5242 [0.4694, 0.5850]	0.5242 [0.4694, 0.5850]
scGPT	0.5533 [0.4218, 0.6826]	0.5237 [0.5079, 0.5420]	0.5237 [0.5079, 0.5420]
Linear	0.5533 [0.4195, 0.6872]	0.5236 [0.4671, 0.5806]	0.5236 [0.4671, 0.5806]
Mean	0.5238 [0.3946, 0.6440]	0.5238 [0.5238, 0.5238]	0.5238 [0.5238, 0.5238]
Split 2 ($n = 21$), own-truth norm			
GEARS	0.6304 [0.4921, 0.7596]	0.5394 [0.4785, 0.6032]	0.5394 [0.4785, 0.6032]
scGPT	0.5737 [0.4444, 0.7007]	0.5303 [0.5125, 0.5510]	0.5303 [0.5125, 0.5510]
Linear	0.6145 [0.4739, 0.7506]	0.5574 [0.4921, 0.6281]	0.5574 [0.4921, 0.6281]
Mean	0.5261 [0.4014, 0.6440]	0.5261 [0.5261, 0.5261]	0.5261 [0.5261, 0.5261]

Table L.5: Replogle K562 external-model ℓ_1 discrimination. Each entry is a mean followed by an interval. Observed intervals are conditional 95% paired-perturbation bootstrap intervals. Null intervals are central 95% randomization ranges, not confidence intervals. All use 2,000 resamples. Fixed train norm uses training responses only, and own-truth norm is an oracle diagnostic. Linear denotes the published self-trained linear model.

Model	Observed [95% CI]	Bundle null [95% range]	Direction null [95% range]
Split 1 ($n = 107$), raw			
GEARS	0.6206 [0.5649, 0.6736]	0.5051 [0.4824, 0.5297]	0.5152 [0.4951, 0.5374]
scGPT	0.5794 [0.5227, 0.6369]	0.5049 [0.4883, 0.5245]	0.5083 [0.4931, 0.5255]
Linear	0.5521 [0.4954, 0.6066]	0.5050 [0.4885, 0.5242]	0.5095 [0.4945, 0.5258]
Mean	0.5047 [0.4518, 0.5582]	0.5047 [0.5047, 0.5047]	0.5047 [0.5047, 0.5047]
Split 1 ($n = 107$), fixed train norm			
GEARS	0.6649 [0.6120, 0.7160]	0.5048 [0.4739, 0.5368]	0.5048 [0.4739, 0.5368]
scGPT	0.6008 [0.5406, 0.6572]	0.5050 [0.4821, 0.5282]	0.5050 [0.4821, 0.5282]
Linear	0.5685 [0.5128, 0.6245]	0.5048 [0.4811, 0.5300]	0.5048 [0.4811, 0.5300]
Mean	0.5047 [0.4519, 0.5594]	0.5047 [0.5047, 0.5047]	0.5047 [0.5047, 0.5047]
Split 1 ($n = 107$), own-truth norm			
GEARS	0.7513 [0.7071, 0.7961]	0.5742 [0.5382, 0.6112]	0.5742 [0.5382, 0.6112]
scGPT	0.6456 [0.5880, 0.7006]	0.5474 [0.5195, 0.5757]	0.5474 [0.5195, 0.5757]
Linear	0.6791 [0.6262, 0.7294]	0.6195 [0.5914, 0.6483]	0.6195 [0.5914, 0.6483]
Mean	0.6281 [0.5746, 0.6808]	0.6281 [0.6281, 0.6281]	0.6281 [0.6281, 0.6281]
Split 2 ($n = 108$), raw			
GEARS	0.6359 [0.5826, 0.6865]	0.5054 [0.4795, 0.5335]	0.5338 [0.5086, 0.5616]
scGPT	0.6226 [0.5692, 0.6764]	0.5049 [0.4822, 0.5291]	0.5153 [0.4946, 0.5380]
Linear	0.5988 [0.5448, 0.6503]	0.5048 [0.4816, 0.5295]	0.5180 [0.4962, 0.5410]
Mean	0.5046 [0.4471, 0.5593]	0.5046 [0.5046, 0.5046]	0.5046 [0.5046, 0.5046]
Split 2 ($n = 108$), fixed train norm			
GEARS	0.6319 [0.5784, 0.6816]	0.5055 [0.4742, 0.5396]	0.5055 [0.4742, 0.5396]
scGPT	0.6321 [0.5764, 0.6862]	0.5051 [0.4761, 0.5340]	0.5051 [0.4761, 0.5340]
Linear	0.6143 [0.5618, 0.6662]	0.5046 [0.4727, 0.5376]	0.5046 [0.4727, 0.5376]
Mean	0.5046 [0.4470, 0.5591]	0.5046 [0.5046, 0.5046]	0.5046 [0.5046, 0.5046]
Split 2 ($n = 108$), own-truth norm			
GEARS	0.7303 [0.6812, 0.7757]	0.5925 [0.5556, 0.6319]	0.5925 [0.5556, 0.6319]
scGPT	0.7014 [0.6474, 0.7517]	0.5587 [0.5253, 0.5941]	0.5587 [0.5253, 0.5941]
Linear	0.7330 [0.6835, 0.7795]	0.6175 [0.5830, 0.6518]	0.6175 [0.5830, 0.6518]
Mean	0.6373 [0.5788, 0.6923]	0.6373 [0.6373, 0.6373]	0.6373 [0.6373, 0.6373]

Table L.6: Replogle K562 external-model ℓ_2 discrimination. Each entry is a mean followed by an interval. Observed intervals are conditional 95% paired-perturbation bootstrap intervals. Null intervals are central 95% randomization ranges, not confidence intervals. All use 2,000 resamples. Fixed train norm uses training responses only, and own-truth norm is an oracle diagnostic. Linear denotes the published self-trained linear model.

Model	Observed [95% CI]	Bundle null [95% range]	Direction null [95% range]
Split 1 ($n = 107$), raw			
GEARS	0.6290 [0.5751, 0.6803]	0.5051 [0.4822, 0.5286]	0.5162 [0.4965, 0.5378]
scGPT	0.6026 [0.5440, 0.6605]	0.5049 [0.4892, 0.5233]	0.5079 [0.4944, 0.5235]
Linear	0.5516 [0.4959, 0.6056]	0.5049 [0.4896, 0.5235]	0.5094 [0.4956, 0.5250]
Mean	0.5047 [0.4519, 0.5591]	0.5047 [0.5047, 0.5047]	0.5047 [0.5047, 0.5047]
Split 1 ($n = 107$), fixed train norm			
GEARS	0.6728 [0.6200, 0.7250]	0.5048 [0.4739, 0.5366]	0.5048 [0.4739, 0.5366]
scGPT	0.6337 [0.5755, 0.6908]	0.5051 [0.4847, 0.5277]	0.5051 [0.4847, 0.5277]
Linear	0.5713 [0.5156, 0.6275]	0.5048 [0.4808, 0.5303]	0.5048 [0.4808, 0.5303]
Mean	0.5047 [0.4510, 0.5587]	0.5047 [0.5047, 0.5047]	0.5047 [0.5047, 0.5047]
Split 1 ($n = 107$), own-truth norm			
GEARS	0.7471 [0.7032, 0.7918]	0.5727 [0.5372, 0.6098]	0.5727 [0.5372, 0.6098]
scGPT	0.6832 [0.6279, 0.7364]	0.5487 [0.5235, 0.5754]	0.5487 [0.5235, 0.5754]
Linear	0.6772 [0.6257, 0.7264]	0.6174 [0.5890, 0.6462]	0.6174 [0.5890, 0.6462]
Mean	0.6291 [0.5778, 0.6799]	0.6291 [0.6291, 0.6291]	0.6291 [0.6291, 0.6291]
Split 2 ($n = 108$), raw			
GEARS	0.6320 [0.5798, 0.6819]	0.5053 [0.4800, 0.5329]	0.5319 [0.5081, 0.5584]
scGPT	0.6610 [0.6053, 0.7159]	0.5050 [0.4834, 0.5277]	0.5133 [0.4943, 0.5333]
Linear	0.5941 [0.5402, 0.6442]	0.5049 [0.4832, 0.5283]	0.5164 [0.4961, 0.5383]
Mean	0.5046 [0.4484, 0.5581]	0.5046 [0.5046, 0.5046]	0.5046 [0.5046, 0.5046]
Split 2 ($n = 108$), fixed train norm			
GEARS	0.6301 [0.5764, 0.6806]	0.5055 [0.4764, 0.5376]	0.5055 [0.4764, 0.5376]
scGPT	0.6770 [0.6199, 0.7318]	0.5051 [0.4794, 0.5317]	0.5051 [0.4794, 0.5317]
Linear	0.6093 [0.5562, 0.6597]	0.5048 [0.4751, 0.5358]	0.5048 [0.4751, 0.5358]
Mean	0.5046 [0.4466, 0.5591]	0.5046 [0.5046, 0.5046]	0.5046 [0.5046, 0.5046]
Split 2 ($n = 108$), own-truth norm			
GEARS	0.7312 [0.6837, 0.7762]	0.5899 [0.5544, 0.6272]	0.5899 [0.5544, 0.6272]
scGPT	0.7397 [0.6841, 0.7897]	0.5547 [0.5237, 0.5870]	0.5547 [0.5237, 0.5870]
Linear	0.7287 [0.6806, 0.7729]	0.6126 [0.5787, 0.6466]	0.6126 [0.5787, 0.6466]
Mean	0.6342 [0.5782, 0.6883]	0.6342 [0.6342, 0.6342]	0.6342 [0.6342, 0.6342]

Table L.7: Replogle RPE1 external-model ℓ_1 discrimination. Each entry is a mean followed by an interval. Observed intervals are conditional 95% paired-perturbation bootstrap intervals. Null intervals are central 95% randomization ranges, not confidence intervals. All use 2,000 resamples. Fixed train norm uses training responses only, and own-truth norm is an oracle diagnostic. Linear denotes the published self-trained linear model.

Model	Observed [95% CI]	Bundle null [95% range]	Direction null [95% range]
Split 1 ($n = 171$), raw			
GEARS	0.5971 [0.5555, 0.6404]	0.5033 [0.4764, 0.5314]	0.5748 [0.5651, 0.5844]
scGPT	0.5510 [0.5079, 0.5958]	0.5029 [0.4840, 0.5217]	0.5366 [0.5298, 0.5437]
Linear	0.5461 [0.5022, 0.5905]	0.5028 [0.4859, 0.5196]	0.5309 [0.5230, 0.5389]
Mean	0.5029 [0.4607, 0.5490]	0.5029 [0.5029, 0.5029]	0.5029 [0.5029, 0.5029]
Split 1 ($n = 171$), fixed train norm			
GEARS	0.5305 [0.4860, 0.5759]	0.5032 [0.4922, 0.5147]	0.5032 [0.4922, 0.5147]
scGPT	0.5243 [0.4807, 0.5703]	0.5029 [0.4923, 0.5135]	0.5029 [0.4923, 0.5135]
Linear	0.5191 [0.4744, 0.5638]	0.5030 [0.4944, 0.5115]	0.5030 [0.4944, 0.5115]
Mean	0.5029 [0.4582, 0.5473]	0.5029 [0.5029, 0.5029]	0.5029 [0.5029, 0.5029]
Split 1 ($n = 171$), own-truth norm			
GEARS	0.7737 [0.7388, 0.8059]	0.7487 [0.7387, 0.7583]	0.7487 [0.7387, 0.7583]
scGPT	0.7331 [0.6974, 0.7682]	0.7153 [0.7073, 0.7241]	0.7153 [0.7073, 0.7241]
Linear	0.7624 [0.7266, 0.7958]	0.7427 [0.7352, 0.7507]	0.7427 [0.7352, 0.7507]
Mean	0.7490 [0.7139, 0.7824]	0.7490 [0.7490, 0.7490]	0.7490 [0.7490, 0.7490]
Split 2 ($n = 164$), raw			
GEARS	0.5625 [0.5199, 0.6060]	0.5033 [0.4851, 0.5223]	0.5655 [0.5610, 0.5701]
scGPT	0.5352 [0.4882, 0.5815]	0.5033 [0.4902, 0.5156]	0.5198 [0.5133, 0.5263]
Linear	0.5412 [0.4980, 0.5846]	0.5035 [0.4906, 0.5174]	0.5254 [0.5198, 0.5311]
Mean	0.5030 [0.4576, 0.5477]	0.5030 [0.5030, 0.5030]	0.5030 [0.5030, 0.5030]
Split 2 ($n = 164$), fixed train norm			
GEARS	0.4955 [0.4522, 0.5408]	0.5030 [0.4990, 0.5071]	0.5030 [0.4990, 0.5071]
scGPT	0.5487 [0.5016, 0.5929]	0.5034 [0.4824, 0.5234]	0.5034 [0.4824, 0.5234]
Linear	0.5208 [0.4792, 0.5679]	0.5032 [0.4968, 0.5095]	0.5032 [0.4968, 0.5095]
Mean	0.5030 [0.4608, 0.5495]	0.5030 [0.5030, 0.5030]	0.5030 [0.5030, 0.5030]
Split 2 ($n = 164$), own-truth norm			
GEARS	0.7889 [0.7589, 0.8164]	0.7888 [0.7845, 0.7931]	0.7888 [0.7845, 0.7931]
scGPT	0.6920 [0.6486, 0.7328]	0.6592 [0.6429, 0.6747]	0.6592 [0.6429, 0.6747]
Linear	0.7883 [0.7551, 0.8190]	0.7664 [0.7595, 0.7734]	0.7664 [0.7595, 0.7734]
Mean	0.7720 [0.7393, 0.8024]	0.7720 [0.7720, 0.7720]	0.7720 [0.7720, 0.7720]

Table L.8: Replogle RPE1 external-model ℓ_2 discrimination. Each entry is a mean followed by an interval. Observed intervals are conditional 95% paired-perturbation bootstrap intervals. Null intervals are central 95% randomization ranges, not confidence intervals. All use 2,000 resamples. Fixed train norm uses training responses only, and own-truth norm is an oracle diagnostic. Linear denotes the published self-trained linear model.

Model	Observed [95% CI]	Bundle null [95% range]	Direction null [95% range]
Split 1 ($n = 171$), raw			
GEARS	0.5966 [0.5542, 0.6392]	0.5033 [0.4775, 0.5299]	0.5720 [0.5616, 0.5823]
scGPT	0.5572 [0.5134, 0.5998]	0.5029 [0.4853, 0.5205]	0.5328 [0.5266, 0.5389]
Linear	0.5458 [0.5011, 0.5907]	0.5028 [0.4866, 0.5193]	0.5306 [0.5226, 0.5388]
Mean	0.5029 [0.4611, 0.5471]	0.5029 [0.5029, 0.5029]	0.5029 [0.5029, 0.5029]
Split 1 ($n = 171$), fixed train norm			
GEARS	0.5310 [0.4865, 0.5764]	0.5032 [0.4913, 0.5151]	0.5032 [0.4913, 0.5151]
scGPT	0.5338 [0.4918, 0.5786]	0.5029 [0.4935, 0.5123]	0.5029 [0.4935, 0.5123]
Linear	0.5190 [0.4740, 0.5655]	0.5029 [0.4941, 0.5118]	0.5029 [0.4941, 0.5118]
Mean	0.5029 [0.4589, 0.5475]	0.5029 [0.5029, 0.5029]	0.5029 [0.5029, 0.5029]
Split 1 ($n = 171$), own-truth norm			
GEARS	0.7663 [0.7312, 0.7995]	0.7371 [0.7264, 0.7475]	0.7371 [0.7264, 0.7475]
scGPT	0.7310 [0.6932, 0.7677]	0.7009 [0.6934, 0.7088]	0.7009 [0.6934, 0.7088]
Linear	0.7553 [0.7189, 0.7900]	0.7338 [0.7259, 0.7422]	0.7338 [0.7259, 0.7422]
Mean	0.7415 [0.7049, 0.7757]	0.7415 [0.7415, 0.7415]	0.7415 [0.7415, 0.7415]
Split 2 ($n = 164$), raw			
GEARS	0.5563 [0.5141, 0.6006]	0.5032 [0.4865, 0.5213]	0.5606 [0.5557, 0.5655]
scGPT	0.5416 [0.4939, 0.5873]	0.5033 [0.4907, 0.5164]	0.5215 [0.5155, 0.5274]
Linear	0.5416 [0.4983, 0.5842]	0.5035 [0.4911, 0.5164]	0.5265 [0.5210, 0.5325]
Mean	0.5030 [0.4570, 0.5468]	0.5030 [0.5030, 0.5030]	0.5030 [0.5030, 0.5030]
Split 2 ($n = 164$), fixed train norm			
GEARS	0.4954 [0.4521, 0.5400]	0.5030 [0.4988, 0.5073]	0.5030 [0.4988, 0.5073]
scGPT	0.5541 [0.5067, 0.5977]	0.5032 [0.4851, 0.5213]	0.5032 [0.4851, 0.5213]
Linear	0.5211 [0.4801, 0.5677]	0.5032 [0.4965, 0.5097]	0.5032 [0.4965, 0.5097]
Mean	0.5030 [0.4598, 0.5486]	0.5030 [0.5030, 0.5030]	0.5030 [0.5030, 0.5030]
Split 2 ($n = 164$), own-truth norm			
GEARS	0.7684 [0.7366, 0.7990]	0.7697 [0.7649, 0.7745]	0.7697 [0.7649, 0.7745]
scGPT	0.7298 [0.6875, 0.7683]	0.6692 [0.6545, 0.6833]	0.6692 [0.6545, 0.6833]
Linear	0.7774 [0.7436, 0.8093]	0.7534 [0.7460, 0.7610]	0.7534 [0.7460, 0.7610]
Mean	0.7614 [0.7279, 0.7935]	0.7614 [0.7614, 0.7614]	0.7614 [0.7614, 0.7614]

Table L.9: External-model cosine discrimination across all predefined contexts. Observed brackets are conditional 95% paired-perturbation bootstrap intervals. Null brackets are central 95% randomization ranges from 2,000 permutations and are not confidence intervals. Cosine depends on predicted direction, so the two matched-null constructions coincide. Linear denotes the published self-trained linear model.

Model	Observed [95% CI]	Bundle null [95% range]	Direction null [95% range]
Adamson, split 1 ($n = 21$)			
GEARS	0.6259 [0.4898, 0.7483]	0.5238 [0.4717, 0.5873]	0.5238 [0.4717, 0.5873]
scGPT	0.6327 [0.5034, 0.7665]	0.5245 [0.4671, 0.5918]	0.5245 [0.4671, 0.5918]
Linear	0.5578 [0.4354, 0.6735]	0.5243 [0.4580, 0.5964]	0.5243 [0.4580, 0.5964]
Mean	0.5238 [0.4014, 0.6395]	0.5238 [0.5238, 0.5238]	0.5238 [0.5238, 0.5238]
Adamson, split 2 ($n = 21$)			
GEARS	0.6259 [0.4966, 0.7483]	0.5247 [0.4580, 0.5964]	0.5247 [0.4580, 0.5964]
scGPT	0.6077 [0.4830, 0.7279]	0.5240 [0.4966, 0.5533]	0.5240 [0.4966, 0.5533]
Linear	0.5941 [0.4512, 0.7257]	0.5240 [0.4467, 0.6032]	0.5240 [0.4467, 0.6032]
Mean	0.5238 [0.3991, 0.6463]	0.5238 [0.5238, 0.5238]	0.5238 [0.5238, 0.5238]
Replogle K562, split 1 ($n = 107$)			
GEARS	0.6413 [0.5810, 0.6991]	0.5048 [0.4701, 0.5435]	0.5048 [0.4701, 0.5435]
scGPT	0.6889 [0.6359, 0.7394]	0.5055 [0.4734, 0.5373]	0.5055 [0.4734, 0.5373]
Linear	0.5603 [0.5034, 0.6142]	0.5050 [0.4811, 0.5294]	0.5050 [0.4811, 0.5294]
Mean	0.5047 [0.4516, 0.5551]	0.5047 [0.5047, 0.5047]	0.5047 [0.5047, 0.5047]
Replogle K562, split 2 ($n = 108$)			
GEARS	0.5967 [0.5392, 0.6538]	0.5052 [0.4757, 0.5370]	0.5052 [0.4757, 0.5370]
scGPT	0.6869 [0.6297, 0.7413]	0.5047 [0.4667, 0.5451]	0.5047 [0.4667, 0.5451]
Linear	0.5937 [0.5366, 0.6512]	0.5048 [0.4749, 0.5342]	0.5048 [0.4749, 0.5342]
Mean	0.5046 [0.4491, 0.5619]	0.5046 [0.5046, 0.5046]	0.5046 [0.5046, 0.5046]
Replogle RPE1, split 1 ($n = 171$)			
GEARS	0.5284 [0.4829, 0.5760]	0.5031 [0.4933, 0.5126]	0.5031 [0.4933, 0.5126]
scGPT	0.5284 [0.4819, 0.5751]	0.5029 [0.4973, 0.5083]	0.5029 [0.4973, 0.5083]
Linear	0.5251 [0.4795, 0.5717]	0.5030 [0.4946, 0.5117]	0.5030 [0.4946, 0.5117]
Mean	0.5029 [0.4587, 0.5473]	0.5029 [0.5029, 0.5029]	0.5029 [0.5029, 0.5029]
Replogle RPE1, split 2 ($n = 164$)			
GEARS	0.5059 [0.4625, 0.5526]	0.5030 [0.4983, 0.5077]	0.5030 [0.4983, 0.5077]
scGPT	0.5443 [0.4984, 0.5920]	0.5030 [0.4953, 0.5109]	0.5030 [0.4953, 0.5109]
Linear	0.5291 [0.4836, 0.5788]	0.5032 [0.4953, 0.5109]	0.5032 [0.4953, 0.5109]
Mean	0.5030 [0.4608, 0.5496]	0.5030 [0.5030, 0.5030]	0.5030 [0.5030, 0.5030]

Table L.10: Paired external-model differences from the published mean baseline on split 1. Each entry is model minus mean with a conditional 95% interval from 2,000 shared perturbation bootstraps. Positive values indicate larger discrimination scores. Normalization rules rescore the same model outputs. Own-truth normalization is an oracle diagnostic. Linear denotes the published self-trained linear model.

Rule	GEARS − mean	scGPT − mean	Linear − mean
Adamson ($n = 21$)			
Raw ℓ_1	0.0816 [0.0227, 0.1565]	0.0454 [−0.0091, 0.1066]	0.0227 [−0.0295, 0.0862]
Fixed train norm, ℓ_1	0.0952 [0.0294, 0.1769]	0.0522 [0.0091, 0.1066]	0.0227 [−0.0317, 0.0884]
Own-truth norm, ℓ_1	0.0998 [0.0227, 0.1905]	0.1020 [0.0226, 0.2018]	0.0431 [−0.0204, 0.1134]
Raw ℓ_2	0.0862 [0.0226, 0.1655]	0.0544 [−0.0046, 0.1270]	0.0408 [−0.0204, 0.1134]
Fixed train norm, ℓ_2	0.0907 [0.0249, 0.1701]	0.0544 [−0.0045, 0.1270]	0.0295 [−0.0317, 0.0998]
Own-truth norm, ℓ_2	0.0998 [0.0272, 0.1882]	0.1111 [0.0227, 0.2177]	0.0544 [−0.0159, 0.1361]
Cosine	0.1020 [0.0317, 0.1859]	0.1088 [0.0204, 0.2109]	0.0340 [−0.0317, 0.1043]
Replogle K562 ($n = 107$)			
Raw ℓ_1	0.1159 [0.0709, 0.1619]	0.0748 [0.0397, 0.1140]	0.0474 [0.0169, 0.0831]
Fixed train norm, ℓ_1	0.1603 [0.1130, 0.2109]	0.0961 [0.0461, 0.1488]	0.0638 [0.0273, 0.1066]
Own-truth norm, ℓ_1	0.1232 [0.0716, 0.1773]	0.0176 [−0.0414, 0.0777]	0.0510 [0.0155, 0.0935]
Raw ℓ_2	0.1243 [0.0795, 0.1699]	0.0979 [0.0616, 0.1397]	0.0469 [0.0151, 0.0836]
Fixed train norm, ℓ_2	0.1681 [0.1220, 0.2170]	0.1290 [0.0788, 0.1830]	0.0666 [0.0294, 0.1084]
Own-truth norm, ℓ_2	0.1181 [0.0674, 0.1692]	0.0542 [0.0002, 0.1123]	0.0481 [0.0093, 0.0913]
Cosine	0.1366 [0.0964, 0.1785]	0.1842 [0.1427, 0.2276]	0.0556 [0.0275, 0.0857]
Replogle RPE1 ($n = 171$)			
Raw ℓ_1	0.0942 [0.0647, 0.1251]	0.0481 [0.0255, 0.0726]	0.0432 [0.0193, 0.0700]
Fixed train norm, ℓ_1	0.0276 [0.0175, 0.0388]	0.0214 [0.0025, 0.0399]	0.0161 [0.0045, 0.0294]
Own-truth norm, ℓ_1	0.0247 [0.0155, 0.0351]	−0.0159 [−0.0320, −0.0003]	0.0134 [0.0020, 0.0267]
Raw ℓ_2	0.0937 [0.0632, 0.1245]	0.0543 [0.0297, 0.0804]	0.0429 [0.0199, 0.0685]
Fixed train norm, ℓ_2	0.0280 [0.0170, 0.0416]	0.0309 [0.0107, 0.0509]	0.0161 [0.0041, 0.0294]
Own-truth norm, ℓ_2	0.0248 [0.0144, 0.0369]	−0.0105 [−0.0297, 0.0092]	0.0139 [0.0015, 0.0286]
Cosine	0.0255 [0.0145, 0.0381]	0.0254 [0.0075, 0.0456]	0.0222 [0.0108, 0.0360]

Table L.11: Paired external-model differences from the published mean baseline on split 2. Each entry is model minus mean with a conditional 95% interval from 2,000 shared perturbation bootstraps. Positive values indicate larger discrimination scores. Normalization rules rescore the same model outputs. Own-truth normalization is an oracle diagnostic. Linear denotes the published self-trained linear model.

Rule	GEARS − mean	scGPT − mean	Linear − mean
Adamson ($n = 21$)			
Raw ℓ_1	$0.0862\,[0.0295, 0.1542]$	$0.0068\,[-0.0113, 0.0249]$	$-0.0045\,[-0.0567, 0.0476]$
Fixed train norm, ℓ_1	$0.0975\,[0.0227, 0.1837]$	$0.0091\,[-0.0159, 0.0340]$	$0.0204\,[-0.0454, 0.0930]$
Own-truth norm, ℓ_1	$0.1043\,[0.0204, 0.1973]$	$0.0181\,[-0.0045, 0.0408]$	$0.0726\,[-0.0181, 0.1746]$
Raw ℓ_2	$0.0748\,[0.0181, 0.1451]$	$0.0385\,[0.0113, 0.0680]$	$0.0068\,[-0.0295, 0.0431]$
Fixed train norm, ℓ_2	$0.0907\,[0.0181, 0.1746]$	$0.0295\,[-0.0068, 0.0703]$	$0.0295\,[-0.0295, 0.0907]$
Own-truth norm, ℓ_2	$0.1043\,[0.0295, 0.1905]$	$0.0476\,[0.0113, 0.0884]$	$0.0884\,[-0.0001, 0.1883]$
Cosine	$0.1020\,[0.0317, 0.1769]$	$0.0839\,[0.0385, 0.1315]$	$0.0703\,[-0.0272, 0.1701]$
Replogle K562 ($n = 108$)			
Raw ℓ_1	$0.1313\,[0.0841, 0.1845]$	$0.1180\,[0.0751, 0.1657]$	$0.0941\,[0.0520, 0.1417]$
Fixed train norm, ℓ_1	$0.1272\,[0.0814, 0.1821]$	$0.1275\,[0.0691, 0.1887]$	$0.1097\,[0.0676, 0.1599]$
Own-truth norm, ℓ_1	$0.0929\,[0.0484, 0.1442]$	$0.0640\,[0.0063, 0.1238]$	$0.0957\,[0.0542, 0.1400]$
Raw ℓ_2	$0.1274\,[0.0807, 0.1808]$	$0.1564\,[0.1109, 0.2059]$	$0.0895\,[0.0494, 0.1357]$
Fixed train norm, ℓ_2	$0.1254\,[0.0829, 0.1767]$	$0.1723\,[0.1155, 0.2319]$	$0.1047\,[0.0628, 0.1532]$
Own-truth norm, ℓ_2	$0.0971\,[0.0544, 0.1443]$	$0.1055\,[0.0510, 0.1620]$	$0.0945\,[0.0541, 0.1391]$
Cosine	$0.0921\,[0.0609, 0.1253]$	$0.1823\,[0.1312, 0.2345]$	$0.0891\,[0.0574, 0.1217]$
Replogle RPE1 ($n = 164$)			
Raw ℓ_1	$0.0595\,[0.0172, 0.1054]$	$0.0321\,[-0.0034, 0.0687]$	$0.0381\,[0.0227, 0.0545]$
Fixed train norm, ℓ_1	$-0.0076\,[-0.0224, 0.0053]$	$0.0457\,[0.0078, 0.0854]$	$0.0177\,[0.0086, 0.0292]$
Own-truth norm, ℓ_1	$0.0169\,[0.0038, 0.0299]$	$-0.0800\,[-0.1111, -0.0475]$	$0.0162\,[0.0054, 0.0296]$
Raw ℓ_2	$0.0533\,[0.0120, 0.0982]$	$0.0385\,[0.0049, 0.0748]$	$0.0386\,[0.0242, 0.0539]$
Fixed train norm, ℓ_2	$-0.0077\,[-0.0223, 0.0056]$	$0.0510\,[0.0183, 0.0854]$	$0.0180\,[0.0088, 0.0295]$
Own-truth norm, ℓ_2	$0.0070\,[-0.0103, 0.0228]$	$-0.0316\,[-0.0587, -0.0026]$	$0.0160\,[0.0040, 0.0307]$
Cosine	$0.0028\,[-0.0152, 0.0219]$	$0.0413\,[0.0183, 0.0657]$	$0.0261\,[0.0130, 0.0409]$

Table L.12: Sensitivity to using each model's full finite test library. All 24 predefined model–split cases use the same fixed set of 1,000 genes. n_c/n_f gives common/full target counts. Raw ℓ_1 scores on the common and full libraries are descriptive comparisons, because both the evaluated targets and ranking candidates can change. Models with different libraries must not be ranked against one another. Own-truth normalization is an oracle diagnostic. The full-library direction-null entry is its mean and central 95% range from 2,000 permutations, not a confidence interval. Linear denotes the published self-trained linear model, and splits reuse biological screens.

Model	n_c/n_f	Raw ℓ_1 Common	Raw ℓ_1 Full	Own-truth Full	Direction null [95% range]
Adamson, split 1					
GEARS	21/21	0.6054	0.6054	0.6327	0.5335 [0.4875, 0.5918]
scGPT	21/21	0.5692	0.5692	0.6349	0.5392 [0.4853, 0.6009]
Linear	21/21	0.5465	0.5465	0.5760	0.5472 [0.4875, 0.6100]
Mean	21/21	0.5238	0.5238	0.5329	0.5329 [0.5329, 0.5329]
Adamson, split 2					
GEARS	21/21	0.6100	0.6100	0.6304	0.5380 [0.4785, 0.6032]
scGPT	21/21	0.5306	0.5306	0.5442	0.5310 [0.5170, 0.5465]
Linear	21/21	0.5193	0.5193	0.5986	0.5528 [0.4875, 0.6236]
Mean	21/21	0.5238	0.5238	0.5261	0.5261 [0.5261, 0.5261]
Replogle K562, split 1					
GEARS	107/272	0.6206	0.6362	0.7516	0.5658 [0.5430, 0.5890]
scGPT	107/107	0.5794	0.5794	0.6456	0.5474 [0.5195, 0.5757]
Linear	107/107	0.5521	0.5521	0.6791	0.6195 [0.5914, 0.6483]
Mean	107/272	0.5047	0.5018	0.6191	0.6191 [0.6191, 0.6191]
Replogle K562, split 2					
GEARS	108/272	0.6359	0.6201	0.7365	0.5894 [0.5660, 0.6130]
scGPT	108/108	0.6226	0.6226	0.7014	0.5587 [0.5253, 0.5941]
Linear	108/108	0.5988	0.5988	0.7330	0.6175 [0.5830, 0.6518]
Mean	108/272	0.5046	0.5018	0.6452	0.6452 [0.6452, 0.6452]
Replogle RPE1, split 1					
GEARS	171/384	0.5971	0.6132	0.7896	0.7625 [0.7558, 0.7690]
scGPT	171/171	0.5510	0.5510	0.7331	0.7153 [0.7073, 0.7241]
Linear	171/171	0.5461	0.5461	0.7624	0.7427 [0.7352, 0.7507]
Mean	171/384	0.5029	0.5013	0.7611	0.7611 [0.7611, 0.7611]
Replogle RPE1, split 2					
GEARS	164/384	0.5625	0.5665	0.7856	0.7846 [0.7818, 0.7872]
scGPT	164/164	0.5352	0.5352	0.6920	0.6592 [0.6429, 0.6747]
Linear	164/164	0.5412	0.5412	0.7883	0.7664 [0.7595, 0.7734]
Mean	164/384	0.5030	0.5013	0.7633	0.7633 [0.7633, 0.7633]

Table L.13: Paired ℓ_1 changes on each model's full finite library. Fixed-norm changes subtract raw scores from scores normalized to the training-only fixed norm. Common-library changes are shown as point estimates. Full-library brackets are conditional 95% intervals from 2,000 paired perturbation bootstraps. The last column compares own-truth-normalized model and published mean predictions on exactly the same model-specific library. These are changes in evaluation, not newly trained prediction improvements. All positive, negative and unresolved effects are retained. Common/full differences are not paired causal effects.

Model	Fixed norm − raw Common	Fixed norm − raw Full [95% CI]	Own-truth model − mean Full [95% CI]
Adamson, split 1			
GEARS	0.0136	0.0136 [−0.0113, 0.0408]	0.0998 [0.0227, 0.1905]
scGPT	0.0068	0.0068 [0.0000, 0.0159]	0.1020 [0.0226, 0.2018]
Linear	0.0000	0.0000 [−0.0227, 0.0272]	0.0431 [−0.0204, 0.1134]
Mean	0.0000	0.0000 [−0.0227, 0.0181]	0.0000 [0.0000, 0.0000]
Adamson, split 2			
GEARS	0.0113	0.0113 [−0.0363, 0.0612]	0.1043 [0.0204, 0.1973]
scGPT	0.0023	0.0023 [−0.0181, 0.0272]	0.0181 [−0.0045, 0.0408]
Linear	0.0249	0.0249 [−0.0045, 0.0590]	0.0726 [−0.0181, 0.1746]
Mean	0.0000	0.0000 [−0.0273, 0.0317]	0.0000 [0.0000, 0.0000]
Replogle K562, split 1			
GEARS	0.0444	0.0316 [0.0129, 0.0498]	0.1325 [0.1027, 0.1662]
scGPT	0.0213	0.0213 [0.0013, 0.0433]	0.0176 [−0.0414, 0.0777]
Linear	0.0164	0.0164 [−0.0196, 0.0539]	0.0510 [0.0155, 0.0935]
Mean	0.0000	0.0000 [−0.0223, 0.0218]	0.0000 [0.0000, 0.0000]
Replogle K562, split 2			
GEARS	−0.0040	0.0069 [−0.0118, 0.0249]	0.0913 [0.0636, 0.1218]
scGPT	0.0095	0.0095 [−0.0120, 0.0312]	0.0640 [0.0063, 0.1238]
Linear	0.0155	0.0155 [−0.0234, 0.0573]	0.0957 [0.0542, 0.1400]
Mean	0.0000	0.0000 [−0.0234, 0.0233]	0.0000 [0.0000, 0.0000]
Replogle RPE1, split 1			
GEARS	−0.0667	−0.0822 [−0.1044, −0.0595]	0.0285 [0.0213, 0.0368]
scGPT	−0.0267	−0.0267 [−0.0626, 0.0096]	−0.0159 [−0.0320, −0.0003]
Linear	−0.0271	−0.0271 [−0.0542, −0.0016]	0.0134 [0.0020, 0.0267]
Mean	0.0000	0.0000 [−0.0223, 0.0225]	0.0000 [0.0000, 0.0000]
Replogle RPE1, split 2			
GEARS	−0.0671	−0.0720 [−0.0925, −0.0523]	0.0223 [0.0145, 0.0302]
scGPT	0.0135	0.0135 [−0.0173, 0.0475]	−0.0800 [−0.1111, −0.0475]
Linear	−0.0204	−0.0204 [−0.0462, 0.0068]	0.0162 [0.0054, 0.0296]
Mean	0.0000	0.0000 [−0.0198, 0.0195]	0.0000 [0.0000, 0.0000]

Table M.1: **Classification counts for every published method.** 78 model–context pairs, each on its own finite test library against the mean on the same targets. A: beats the mean inside the central 95% range of 2,000 direction permutations. B: beats the mean above that range. C: does not beat the mean. Below: beats the mean below the range. Share: median of (null mean − mean score)/(model − mean score) over classes A and B, descriptive.

Rule	Reference	Pairs	A	B	C	Below	A/(A+B)	Share
own-truth ℓ_1	global	78	4	28	46	0	12%	−1%
own-truth ℓ_1	stratified	78	7	25	46	0	22%	−1%
raw ℓ_1	global	78	0	30	47	1	0%	15%
raw ℓ_1	stratified	78	4	27	47	0	13%	15%
training-fixed ℓ_1	global	78	1	31	46	0	3%	0%
training-fixed ℓ_1	stratified	78	3	29	46	0	9%	0%

GEARS on RPE1's second split beats the mean by 0.0223 ($[0.0145, 0.0302]$), and its direction-null mean reaches 95% of that gain. The linear model with random perturbation embeddings beats the mean on RPE1's first split by 0.0024 ($[0.0003, 0.0045]$) while scoring 0.76356 inside its range $[0.75970, 0.76362]$. scBERT predicts one direction for every perturbation in every context, with a minimum pairwise cosine of 1 to nine decimals and a norm coefficient of variation below 4×10^{-8}. Its null ranges are therefore points equal to its scores. Its raw and training-fixed ℓ_1 scores equal $(N+1)/(2N)$ to machine precision, for example 0.50185 at $N = 270$, and match the mean exactly. Under own-truth normalization, which reads each receiving perturbation's measured norm, it beats the mean on both K562 splits by 0.0012 ($[0.0002, 0.0022]$) and 0.0040 ($[0.0025, 0.0057]$) and falls below it on both RPE1 splits by 0.0013. A published model can therefore beat the mean without any perturbation-specific direction.

Under raw ℓ_1 no gain over the mean lies inside its global range. GEARS on RPE1's second split beats the mean by 0.0652 ($[0.0371, 0.0944]$) yet scores 0.5665, just below its direction-null range $[0.5666, 0.5723]$, and inside the stratified one. The raw direction null keeps each receiving prediction's own norm, so a prediction can score below the reassignment of its own directions. Apart from scBERT's point ranges, three classifications depend on scores within 10^{-4} of an endpoint estimated from 2,000 permutations. These are the present case, the random-embedding model above, and scGPT on Adamson's first split under the training-fixed norm, whose score equals the upper endpoint. These classifications can change with the permutation draw. Differences between individual entries in Tables 2 and M.3 are therefore descriptive. Under the training-fixed norm only scGPT on Adamson's first split lies inside its global range. Descriptively, 7 of the 46 own-truth pairs that do not beat the mean still lie above their null range, namely Geneformer and the random-gene linear model on both K562 splits and scGPT on K562's first split and both RPE1 splits, and 15 lie below the mean with an interval excluding zero, including CPA in four contexts. These counts are not independent. Six of the 13 methods share the linear architecture, and the six contexts reuse three screens. Null ranges are randomization ranges, not confidence intervals, and overlap does not show that a model lacks biological information.

N PER-POINT INVERSE RECOVERY

Zero weights and numerical conditioning. The fraction of genes with $B_g = 0$ runs from 0.00% on replogle to 13.10% on norman and the recovery is unaffected. The residuals of 0.0 and 3×10^{-4} at these two extremes show that score recovery does not require every coordinate to be recoverable. On norman, the map sends 13.10% of coordinates to the shared response $\bar{\delta}$, yet the remaining coordinates restore the score. Only 5 of norman's 1008 perturbation-settings change rank, none by more than one. This bounds those coordinates' contribution to this score. It is not a claim that they carry no biological information.

$\mathrm{cond}(B)$, the ratio of the largest to the smallest positive weight, ranges from 14.4 to 89588, nearly four orders of magnitude. The most ill-conditioned setting, arc-h1 at $k = 1$, has exactly zero residual. In double precision, the largest residual in the table is 3.2×10^{-4}. These measurements quantify the numerical limitation of the explicit inverse.

Table M.2: **Finite test targets per method and context.** Source test is the size of each partition's test set. Dashes mark methods excluded because their predictions omit scored genes.

	Adamson		K562		RPE1	
Method	1	2	1	2	1	2
Source test	21	21	272	272	384	384
CPA	21	21	272	272	384	384
GEARS	21	21	272	272	384	384
Geneformer	21	18	271	271	382	382
scBERT	18	18	270	267	372	370
scGPT	21	21	107	108	171	164
UCE, 4 layers	21	21	107	108	171	164
UCE, 33 layers	21	21	107	108	171	164
Linear, self-trained	21	21	107	108	171	164
Linear, GEARS pert.	21	21	272	272	384	384
Linear, K562 pert.	8	14	272	272	207	216
Linear, RPE1 pert.	11	13	214	219	384	384
Linear, random pert.	21	21	272	272	384	384
Linear, random gene	21	21	107	108	171	164
Mean	21	21	272	272	384	384
Linear, scGPT gene	–	–	–	–	–	–
Linear, scFoundation gene	–	–	–	–	–	–

Table M.3: **Classes under raw and training-fixed ℓ_1.** Classes as in Table 2. L: beats the mean but lies below the direction-null range. $*$: inside the magnitude-stratified range.

	Raw ℓ_1						Training-fixed ℓ_1					
	Adamson		K562		RPE1		Adamson		K562		RPE1	
Method	1	2	1	2	1	2	1	2	1	2	1	2
GEARS	B	B	B	B	B	L*	B	B	B	B	B	C
scGPT	C	C	B	B	B	C	A	C	B	B	B	B
CPA	C	C	C	C	C	C	C	C	C	C	C	C
Geneformer	C	C	B	B	B	B	C	C	B	B	C	B
scBERT	C	C	C	C	C	C	C	C	C	C	C	C
UCE, 4 layers	C	C	C	C	C	C	C	C	C	C	C	C
UCE, 33 layers	C	C	C	C	C	C	C	C	C	C	C	C
Linear, self-trained	C	C	B	B	B	B	C	C	B	B	B	B
Linear, GEARS pert.	C	C	B	B	C	C	C	C	B	B	C	B
Linear, K562 pert.	B*	B	B	B	B	B	B*	B	B	B	B	B
Linear, RPE1 pert.	B*	B*	B	B	B	B	B*	B	B	B	B	B
Linear, random pert.	C	C	C	C	C	C	C	C	C	C	C	C
Linear, random gene	C	C	C	C	C	C	C	C	C	C	C	C

Per perturbation, not only on average. The residual above is an average of ranks, so it is quantized in steps of $1/N^2$. Norman's largest, 3.19×10^{-4}, is exactly $1/56^2$, corresponding to one net rank change in the mean. An average of zero could also hide larger moves that cancel. We therefore compare each perturbation's rank under the inverted estimator with its rank under the unshrunk baseline over all three support draws.

Table N.2: **After the inverse, 84 of 8064 perturbation-settings change rank, none by more than two.**

dataset	settings	perturbation-settings	rank changed	largest change	settings with none
arc-h1	18	540	0	0	18
replogle	18	3258	0	0	18
norman	18	1008	5	1	13
adamson	18	324	0	0	18
rpe1	18	2934	79	2	1

Table N.1: **The closed-form inverse returns the score to the unshrunk baseline in all 30 settings, from a dataset with no irrecoverable coordinates to one with 13.10% of them, and across four orders of magnitude of conditioning.** Residual is the inverted estimator minus the unshrunk baseline, with $\mathrm{cond}(B)$ ranging over the six support sizes. Per-setting rows are in Table N.3 below.

dataset	$B_g = 0$	$\mathrm{cond}(B)$	$\max \lvert\text{residual}\rvert$
replogle	0.00%	14.4 – 236.5	0.0
rpe1	0.35%	224.8 – 4729.8	2×10^{-4}
arc-h1	5.10%	2858.8 – 89588	0.0
adamson	8.95%	3156.8 – 46960	0.0
norman	13.10%	1119.3 – 35766	3×10^{-4}

Table N.3: **The inverse returns the score to the unshrunk baseline in all 30 settings, with largest residual** 3.2×10^{-4}**.** Unclipped deltas, ℓ_2.

dataset	k	B_g=0	cond(B)	PDS@1	inverted	unshrunk	residual
arc-h1	1	5.10%	89588	0.5533	0.7211	0.7211	+0.00000
arc-h1	2	5.10%	45271	0.6300	0.8222	0.8222	+0.00000
arc-h1	4	5.10%	22757	0.7333	0.8944	0.8944	+0.00000
arc-h1	8	5.10%	11409	0.8756	0.9478	0.9478	+0.00000
arc-h1	16	5.10%	5713	0.9589	0.9689	0.9689	+0.00000
arc-h1	32	5.10%	2859	0.9567	0.9667	0.9667	+0.00000
replogle	1	0.00%	236	0.5771	0.8242	0.8242	+0.00000
replogle	2	0.00%	155	0.6424	0.8959	0.8959	+0.00000
replogle	4	0.00%	92	0.7347	0.9376	0.9376	+0.00000
replogle	8	0.00%	51	0.8442	0.9692	0.9692	+0.00000
replogle	16	0.00%	27	0.9432	0.9862	0.9862	+0.00000
replogle	32	0.00%	14	0.9834	0.9936	0.9936	+0.00000
norman	1	13.10%	35766	0.6543	0.9018	0.9021	-0.00032
norman	2	13.10%	17889	0.7267	0.9426	0.9423	+0.00032
norman	4	13.10%	8946	0.8386	0.9767	0.9767	+0.00000
norman	8	13.10%	4474	0.9174	0.9911	0.9911	+0.00000
norman	16	13.10%	2238	0.9809	0.9939	0.9939	+0.00000
norman	32	13.10%	1119	0.9946	0.9971	0.9971	+0.00000
adamson	1	8.95%	46960	0.5370	0.7747	0.7747	+0.00000
adamson	2	8.95%	32437	0.5679	0.9012	0.9012	+0.00000
adamson	4	8.95%	20041	0.6883	0.9352	0.9352	+0.00000
adamson	8	8.95%	11360	0.8241	0.9136	0.9136	+0.00000
adamson	16	8.95%	6087	0.9136	0.9475	0.9475	+0.00000
adamson	32	8.95%	3157	0.9753	0.9784	0.9784	+0.00000
rpe1	1	0.35%	4730	0.6187	0.8066	0.8067	-0.00011
rpe1	2	0.35%	2868	0.6879	0.8809	0.8809	+0.00004
rpe1	4	0.35%	1605	0.7838	0.9135	0.9137	-0.00019
rpe1	8	0.35%	854	0.8705	0.9530	0.9529	+0.00004
rpe1	16	0.35%	441	0.9374	0.9732	0.9733	-0.00011
rpe1	32	0.35%	225	0.9731	0.9889	0.9890	-0.00008

Of 8064 perturbation-settings, 84 change rank, none by more than 2, and in 68 of the 90 settings no perturbation moves at all. Rank changes cancel in 12 settings, as the mean alone could conceal, but none exceeds two ranks. Three panels recover every perturbation exactly on every draw.

Penalties under an invertible map. The inverse also measures the penalty each score assigns to an invertible map on the observed data. On replogle the empirical-Bayes map has $B_g > 0$ for every gene at every support size, so it is an exact bijection of gene space and the shrunk prediction carries neither more nor less information than the unshrunk one. Whether a score should assign the two predictions equal values depends on its target. A retained-identity target requires equal values and an output-scale target does not. We do not choose between these targets. The comparison quantifies

Table O.1: **Under ℓ_2 the cellwise clip costs every estimator whose unclipped prediction can be measured, so it is not neutral across the panel of methods.** Mean over the six support sizes of PDS(unclipped) minus PDS(clipped). Positive means the clip lowers the score. The largest mean cost is for the trained neural pooler on rpe1, at $+0.0226$. The transport map clips its own output before the scoring pipeline does (Appendix D), so its 0.0000 on four of five datasets and -0.0002 on arc-h1 measure what the second clip adds, not neutrality. Because the clip costs estimators unequally, Section 6 applies the inverse to the unclipped prediction and reports the clip's effect separately. Under ℓ_1 the asymmetry is larger. On rpe1 at $k = 8$, clipping costs the empirical-Bayes map 0.040 and the unshrunk baseline 0.012, a factor of 3.3. The difference between them, 0.028, is 17% of the 0.1609 gap that setting reports.

dataset	naive k-shot mean	EB-affine	affine OT	projection	trained pooler
arc-h1	+0.0015	+0.0020	-0.0002	+0.0024	+0.0046
replogle	+0.0048	+0.0124	+0.0000	+0.0098	+0.0132
norman	+0.0028	+0.0078	+0.0000	+0.0117	+0.0115
adamson	+0.0026	+0.0031	+0.0000	+0.0015	+0.0067
rpe1	+0.0006	+0.0006	+0.0000	+0.0089	+0.0226

how much of each score's penalty a scalar rescaling removes. On the six settings where the map is a bijection, the median penalty it is charged is 0.1640 by the reported score, 0.0231 by the global quotient and 0.0088 by the per-prediction one, with worst cases 0.2534, 0.1206 and 0.0277. The reported score therefore penalizes an information-preserving map 18.6 times as much as the per-prediction quotient, and 8.9 times as much over all 30 settings, where the map is bijective only on the surviving coordinates. The comparison is not circular, because $\overline{\mathrm{PDS}}^{\star}$ quotients a one-parameter scalar family while the map is per-gene affine about a non-zero center.

O Effect of cellwise clipping

Cellwise clipping is not invertible. The inverse in Section 6 therefore acts on unclipped predictions, with clipping costs reported separately. Table O.1 reports the discrimination cost of clipping for each estimator and panel.

O.1 A common pseudobulk clipping control

To distinguish prediction changes from downstream clipping, we compare raw deltas d_i, the common transformation $\max(\mu_{\mathrm{ctrl}} + d_i, 0) - \mu_{\mathrm{ctrl}}$, and the stored result obtained by clipping predicted cells before averaging. The same coordinatewise lower floor is applied to every estimator. The comparison includes the unshrunk mean, empirical-Bayes map, projection and trained pooler at all 120 first-draw settings, plus the mean and empirical-Bayes map at 120 settings on the two additional draws. Affine OT is excluded because its stored output has already been clipped internally. These 240 prediction sets comprise all eligible cases.

Common pseudobulk clipping changes predictions in every included estimator family but leaves all ℓ_1 ranks and scores unchanged. The reason is specific to this distance and these nonnegative true pseudobulks. If a predicted coordinate $p_g < 0$ and every candidate coordinate $q_{jg} \geq 0$, replacing p_g by zero subtracts $-p_g$ from each distance $|p_g - q_{jg}|$. Summing over clipped coordinates subtracts the same constant from every candidate distance. Candidate order is consequently unchanged. This argument applies to a shared floor on pseudobulks, not to clipping cells before taking their mean.

The common floor changes some ℓ_2 and cosine rankings. Setting-mean changes range from -0.00222 to $+0.00147$ for ℓ_2 and from -0.00309 to $+0.00309$ for cosine. Empirical-Bayes scores remain unchanged under all three distances. Stored cellwise clipping has larger, estimator-dependent effects. Table O.2 compares its effect on the empirical-Bayes-minus-unshrunk ℓ_1 difference. The deficit persists under the common pseudobulk transformation and is deeper under cellwise clipping in each panel's median. This isolates sensitivity to a downstream operation without making cellwise clipping invertible or establishing which estimate is biologically more informative.

Table O.2: **A common pseudobulk floor preserves the empirical-Bayes discrimination deficit.** Entries are median-over-support-size ℓ_1 score differences, empirical Bayes minus the unshrunk mean, after averaging support draws within each perturbation. Negative values favor the unshrunk estimate. These medians describe five panels, not independent repeated experiments. Per-setting paired perturbation intervals and all four eligible estimators are computed by the supplementary analysis code.

Panel	Raw	Common pseudobulk floor	Stored cellwise clip
arc-h1	−0.18315	−0.18315	−0.18815
replogle	−0.21250	−0.21250	−0.23141
norman	−0.15689	−0.15689	−0.17815
adamson	−0.22119	−0.22119	−0.24331
rpe1	−0.16850	−0.16850	−0.18819

Prediction and candidate identities, control means and gene axes are checked before every paired comparison. We record the number of altered coordinates, the maximum numerical change and zero-vector counts, which distinguish a rank-invariant transformation from an absent intervention. The raw, common-floor and stored-cellwise scores are calculated with the same candidate ordering and official tie convention. Uncertainty uses the same paired perturbation bootstrap as Appendix U. The conditional intervals do not account for variation in the candidate library.

P CORRECTIONS THAT DO NOT USE THE KNOWN MAP

Section 6 inverts the known empirical-Bayes map. The analyses below concern corrections that do not use it, namely a per-gene map fitted from training pairs (Appendix P.1), per-gene rescaling, whitening and rank-reduction tests (Appendix P.2) and empirical-Bayes shrinkage inside the training response subspace (Appendix P.3).

P.1 LIMITS OF SHRINKAGE-MAP RECOVERY FROM TRAINING PAIRS

The inverse in Section 6 is supported for the known empirical-Bayes map. We do not evaluate inverses for the other estimators. A training-only fit did not meet the recovery criterion even for the known map. It therefore does not justify extending inverse recovery to unknown maps.

Training-only map fit. The map was fitted per gene from training perturbations, without any test label. For each training perturbation, k support cells were drawn and the target δ_q was the mean of the remaining cells. Support and target are disjoint, matching the pair used to score a held-out perturbation. The fit regresses the estimator's output on a noisy response proxy, not on the estimator's own input, the support mean. Its coefficient therefore also reflects the association between support and query. The unsuccessful recovery concerns this proxy fit, not every way of identifying an input-output map. With $\hat{\delta}_q$ the estimator's own prediction from that support,

$$\hat{B}_g \;=\; \frac{\sum_q (\hat{\delta}_{qg} - \bar{\delta}_g)(\delta_{qg} - \bar{\delta}_g)}{\sum_q (\delta_{qg} - \bar{\delta}_g)^2}.$$

Recovery criterion. Before any fit ran, we required the Spearman correlation between $\hat{B}$ and the known B, over genes with $B_g > 0$, to reach 0.9 in every one of 30 settings, five panels at six support sizes on the first support draw. Recovery of this known map was a prerequisite for evaluating fitted-map inverses for other estimators.

Table P.2: **Both per-gene corrections fail, and least squares fails because its MSE-optimal coefficient is itself a shrinkage coefficient.** rpe1, $k = 8$, two-fold cross-fitting over the 163 held-out perturbations with the score computed within fold, so every arm faces a candidate library of the same size. Per-gene least squares uses the other fold's predictions and truths. Whitening scales each side by the other fold's per-gene spread. Neither uses a perturbation's own truth, but both read held-out truths and neither is train-only. Per-gene least squares lowers the score for every arm including the unshrunk baseline, because the MSE-optimal per-gene coefficient is a shrinkage coefficient and shrinking toward the cross-perturbation mean is exactly what removes contrast between perturbations. Whitening is worse for the empirical-Bayes map. That is consistent with Proposition Q.1, which concerns the whole diagonal group and does not predict that this or any one transform must lose. The rank column counts the directions holding 99.9% of the spectral energy of the predicted deltas.

estimator	rank	PDS before correction	after whitening	after per-gene LS
naive k-shot mean	162	0.9513	0.9071	0.8461
EB-affine	161	0.8718	0.5666	0.8462
affine OT	68	0.8710	0.8661	0.8101
projection	20	0.8884	0.8888	0.8563
trained pooler	20	0.8612	0.8535	0.8571

Table P.1: **The training-only map fit does not meet the recovery criterion across settings.**

dataset	$k = 1$	2	4	8	16	32
arc-h1	0.9273	0.9540	0.9644	0.9680	0.9734	0.9722
replogle	0.9642	0.9785	0.9892	0.9926	0.9940	0.9935
norman	0.2235	0.1972	0.1714	0.1598	0.1444	0.1458
adamson	0.8711	0.9028	0.9204	0.9323	0.9368	0.9333
rpe1	0.9694	0.9780	0.9795	0.9788	0.9720	0.9609

Recovery outcome and scope. Only 23 of 30 settings reach 0.9 (Table P.1). The requirement is unmet on adamson at $k = 1$ and at every support size on norman. Adamson was evaluated first. After its failure, a second registration covered the remaining fits to localize the failure, without changing the verdict that the criterion was not met. No fit for another estimator or score recovered through a fitted map is reported. The inverse therefore retains the known-map scope stated in Section 6.

Post hoc diagnosis on norman. Checks run before the failed criterion was accepted ruled out three alternative explanations. The data partitions and prior matched those used for scoring, and slope attenuation from control-cell noise predicts far less loss than observed. With the recorded map exact on the training pairs, $\hat{B}_g/B_g$ reduces to how closely a perturbation's support-cell mean tracks its own held-out-cell mean across training perturbations, and this association is weak on norman. Its cause remains unresolved. This is a post hoc diagnostic observation.

P.2 Limits of alternative corrections

The two tested per-gene corrections do not reproduce recovery by the known map's inverse. Table P.2 compares whitening and per-gene least squares alongside each estimator's effective rank. The subsequent invariance and rank-reduction tests distinguish insensitivity to invertible transformations from sensitivity to restricted representations.

Table P.3 compares the scores under invertible transformations.

Table P.4 reports the rank-reduction test for the two score variants.

P.3 Low-rank-prior denoising retains the discrimination deficit

Denoising within the perturbation-specific subspace could, in principle, preserve discrimination better than shrinkage toward a shared response. We evaluate a composition of two closed-form estimators, applying EB-affine shrinkage within the training response subspace used by projection.

Table P.3: **A global rescaling, a per-gene rescaling and a shrinkage toward $\bar{\delta}$ are all bijections of the estimate and carry the same information.** Their score spread is the pre-specified invariance criterion used to select the reported bound for a retained-identity target. The official score, whose target is the fixed-coordinate prediction, varies fifty times as much as the scalar bound, reflecting its target, not a defect. Here $\overline{\mathrm{PDS}}^{\star}$ is maximized over 241 scales between 0.05 and 6.

score	spread across bijections	meets criterion
PDS@1	0.1311	no
$\overline{\mathrm{PDS}}^{\star}$	0.0026	yes
per-gene coordinate ascent	0.0040	yes
per-gene whitening	0.1254	no

Table P.4: **Both scores are monotone in surviving rank, and in this test only $\overline{\mathrm{PDS}}^{\star}$ stays near the chance floor at rank 2.** This motivates using the scalar quotient as a bound instead of the version with more degrees of freedom. It is not a claim that a quotient scores identity (Appendix I). Zeroing 30% of genes is not invertible yet scores 0.8834 against the bijections' 0.885. Non-invertibility alone therefore does not establish loss of discriminative information. $\overline{\mathrm{PDS}}^{\star}$ is maximized over 241 scales between 0.05 and 6.

score	at surviving rank 2	chance floor
per-gene coordinate ascent	0.6454	0.5
$\overline{\mathrm{PDS}}^{\star}$	0.5237	0.5

The low-rank-prior variant is the conjugate-Gaussian posterior mean under the prior covariance $\Sigma_a = \mathrm{diag}(\mathbf{t}) + V^\top \mathrm{diag}(\boldsymbol{\lambda}) V$. The rows of V are the top d right singular vectors of the centered training responses, each λ_j is the corresponding variance less the mean noise floor and floored at zero, and $\mathbf{t} = \max(10^{-6}, \boldsymbol{\tau}^2 - \mathrm{diag}(V^\top \mathrm{diag}(\boldsymbol{\lambda}) V))$ elementwise. The rank-d structure therefore redistributes EB-affine's per-gene prior variance rather than adding to it. It is computed by Woodbury with no $G \times G$ inverse. With $\boldsymbol{\lambda}$=0 it reduces to EB-affine, except that zero prior variances are floored at 10^{-6}. The subspace rank is the only free choice, so we fixed two values in advance and report both, d=20, the projection rank, and d=3, the implementation's default. The endpoints and interpretation criteria were registered before the run.

Under the primary protocol, the low-rank-prior variant's response-endpoint area under the curve (AUC) exceeds EB-affine's on three of five datasets at d=20 and four of five at d=3. The low-rank prior thus sometimes improves response accuracy, as expected. But on perturbation discrimination it never outperforms the unshrunk k-shot mean. The deficit runs -0.029 to -0.170 across the five datasets at d=20 and -0.052 to -0.133 at d=3, negative in all ten dataset–rank settings, closest to zero on adamson and never reaching it. Denoising within the response subspace improves response accuracy in some settings but retains the discrimination deficit. It does not resolve the trade-off.

This exploratory analysis is outside the confirmatory family. Criterion outcomes for the four pre-specified alternatives to the unshrunk reference are unchanged. Rerunning the unshrunk mean on the same partitions reproduced its primary per-perturbation official scores with zero difference in every metric. The variant and primary methods are therefore scored on identical cells.

Q INVARIANCE OF CANDIDATE RANKINGS

Table Q.1 distinguishes which distances preserve rankings under each transformation.

Discrimination is a ranking score, so invariance means the candidate order is unchanged and the distance values may move freely. Every entry was verified numerically on $G = 60$, $N = 25$. Only cosine is invariant to prediction-only scaling, the relevant transformation when shrinkage changes predictions against a fixed candidate library. Scaling both sides leaves all three rankings invariant and is irrelevant to this problem. No ranking is invariant to per-gene scaling, even when the same diagonal weight is applied to the candidate library. This was established numerically, not derived. Per-gene standardization therefore does not provide this invariance.

Table Q.1: **Only cosine preserves candidate rankings under prediction-only scalar rescaling.** Each entry compares rankings before and after transformation over 400 random draws at $G = 60$, $N = 25$. Invariance requires an unchanged ranking in all 400. The first row is relevant to shrinkage, which acts on predictions with the candidate library fixed. Orthogonal invariance is not assumed to be desirable in the gene basis.

transform applied	ℓ_1	ℓ_2	cosine
scalar, prediction only	no	no	invariant
scalar, both sides	invariant	invariant	invariant
diagonal, prediction only	no	no	no
diagonal, both sides	no	no	no
orthogonal, both sides	no	invariant	invariant

ℓ_1 is not invariant to orthogonal changes of gene coordinates, whereas ℓ_2 and cosine are. This is not presented as a flaw. The gene basis is natural for expression data, and invariance to its rotation is not obviously desirable. Stated precisely, ℓ_1 discrimination is not a function of response geometry alone, and ℓ_1 is the distance the Virtual Cell Challenge uses.

Q.1 Diagonal-group impossibility

Proposition Q.1 (Diagonal-group impossibility)**.** *Let $\mathcal{D} = \{\mathrm{diag}(b) : b \in \mathbb{R}^G_{>0}\}$ act on the prediction. The orbit of d under $\mathcal{D}$ is exactly its sign-pattern class, $\mathcal{O}(d) = \{v : \mathrm{sign}(v) = \mathrm{sign}(d)\}$. Consequently any score invariant to $\mathcal{D}$ in its first argument satisfies $S(d; \cdot) = \tilde{S}(\mathrm{sign}(d); \cdot)$.*

Proof. Positive diagonal scaling preserves signs coordinatewise, which gives one inclusion. For the other, set $b_g = v_g/d_g$ where $d_g \neq 0$, positive because the signs agree, and $b_g = 1$ where $d_g = 0$, in which case $v_g = 0$ as well. Then $v = \mathrm{diag}(b)d$. Invariance on orbits gives the second claim. A $\mathcal{D}$-invariant score must therefore treat a prediction at cosine 0.998 to the truth as equivalent to one at cosine 0.756 whose per-gene magnitudes have been scrambled, provided the signs agree. The example uses fixed seeds, with checks that the diagonal is strictly positive and the sign patterns agree.

This limits the scope of the remedy in Liu et al. (2025), without criticizing it. Cosine quotients the scalar subgroup $\{cI\} \subset \mathcal{D}$, the only distance here invariant to scaling the prediction alone (Appendix Q), and Proposition Q.1 says that quotienting by all of $\mathcal{D}$ leaves nothing but signs. It does not identify a unique largest informative subgroup, and no such claim is made, but it marks the boundary of that direction. The whitening failure in Section 6, a transformation in the $\mathcal{D}$ direction that loses 0.31 of score (Appendix P.2), is consistent with this boundary. The proposition concerns the whole group and does not by itself predict that any individual transform must lose.

R Shared-response dominance in full

The map underlying Section 7 is shown in Figure 1(a). With $\bar{\delta} = 0$ and $B = bI$ for some $b > 0$, the map contracts toward the origin. These two conditions suffice for exact recovery by one global factor and invariance of a directional distance. With $\bar{\delta} = 0$ but unequal weights the map still turns the prediction, and a four-point example with equal candidate norms loses score that no scalar recovers (Appendix T). With $\bar{\delta}$ large the map also turns the prediction toward it. On these panels B is far from uniform, but its unevenness does not order the datasets and ρ_{shared} does.

$\bar{\delta}$ is fitted on training perturbations, but the denominator of Equation (5) is measured over the held-out ones, so ρ_{shared} describes the held-out panel before any method is scored without being a train-only statistic. A train-only denominator reorders adamson above rpe1 and breaks the relation on all three draws (Appendix R.1). ρ_{shared} is invariant to k and the support draw, with bit-identical values in all 90 settings.

Recovery under fixed-norm correction. The fixed-norm implementation of Liu et al. (2025) reads the true response norm. We instead match the norm of the unshrunk support estimate, avoiding test labels. Recovery ranges from $+92.4\%$ to -48.9%, strictly ordered by ρ_{shared} on the first

Table R.1: **Shared-response dominance orders both magnitude corrections on this draw, and orders the group optimum on all three.** Medians over the settings with $|\text{gap@1}| > 0.02$, since a ratio formed on a gap of a few score quanta is not stable. The last column is the deployable form of the fixed-norm recommendation, matching each prediction to the norm of the unshrunk estimate. The expected position of arc-h1 was recorded before its recovery scores were computed. Summarized in Section 7.

dataset	ρ_{shared}	n_{test}	global rescaling	deployable fixed norm
arc-h1	0.225	30	95.3%	+92.4%
replogle	0.390	181	82.2%	+71.4%
norman	0.453	56	55.4%	+54.7%
adamson	0.791	18	41.2%	+40.0%
rpe1	0.914	163	1.3%	−48.9%

draw. Only rpe1 worsens at all six support sizes. Its own-truth form still leaves a gap, and the same true norm lifts an identity-free prediction by more than it recovers from the real prediction. We therefore credit neither effect to normalization alone and do not subtract them (Appendix S). Unlike the near-chance predictor of Liu et al. (2025), our predictors start far above chance. Norm matching nearly closes the gap at small ρ_{shared} but does not do so at large ρ_{shared}. The train-only scalar $\beta_{\text{train}} = 1/\bar{B}(k)$ likewise avoids test labels, but worsens the score in 14 of 30 settings and leaves one unchanged. Both rules' outcomes vary with ρ_{shared} (Table S.1).

Prospective evidence. Before scoring arc-h1, its $\rho_{\text{shared}} = 0.225$ predicted the largest global recovery under a fixed gap filter and summary, and observed recovery was 95.3%. Across five panels, global-recovery point estimates are strictly ordered on all three draws. Only the extremes are stable under query resampling (Appendix W). Support-norm recovery is ordered on two draws, with norman moving from third to first on the other. Repeated draws test within-panel stability, not additional independent panels. Pairing ρ_{shared} with the cosine gap gives rank correlation $+1.0$, but its within-panel perturbation-bootstrap interval includes zero (Appendix W.1), and resampling norman in target blocks narrows it only to $[0.30, 1.00]$ (Appendix W), which precludes a claim about the strength of the association. Generalization requires additional panels. Low ρ_{shared} with poor scalar recovery, or the reverse, would refute the ordering.

Metric scope. The scalar-recovery ordering is established under ℓ_2, whose exact bound follows Proposition 1. The inverse also recovers under the Virtual Cell Challenge's ℓ_1 metric, but its absorbed-fraction ordering remains unresolved. No validated exact search is available for that group's optimum. Two searches did not meet their pre-specified reference-optimum validation criteria. Their errors were as large as the effect to be resolved, so no ℓ_1 optima are reported from them (Appendix W.2).

R.1 ρ_{shared} WITHOUT THE HELD-OUT RESPONSES

The reported ρ_{shared} requires measured responses from the held-out panel. A denominator based on training perturbations defines a different descriptor, compared below. The numerator, the prior mean $\bar{\delta}$, is fitted on training perturbations alone. In contrast, the denominator represents the mean distance from each held-out response to $\bar{\delta}$.

A train-only denominator can be constructed from the empirical-Bayes prior fitted on training perturbations, whose variance $\boldsymbol{\tau}^2$ is saved for every run. Up to the factor $(n-1)/n$ of a sample variance, $\sum_g \text{Var}_q(\Delta_g^{\text{train}})$ equals $\text{mean}_q \|\Delta_q^{\text{train}} - \bar{\delta}\|^2$, and τ^2 further subtracts the sampling noise and truncates at zero. The quantity $\|\bar{\delta}\| / \sqrt{\sum_g \tau_g^2}$ is therefore a noise-corrected root-mean-square proxy for the same ratio, with the spread taken over the training perturbations. The criterion, fixed before computation, required preservation of the five-panel ordering.

Table R.2: **The train-only spread changes the dataset order.**

dataset	n_{test}	as reported	test, RMS	train, RMS
arc-h1	30	0.2252	0.2105	0.2284
replogle	181	0.3902	0.3804	0.4772
norman	56	0.4534	0.4137	0.5967
adamson	18	0.7910	0.7326	1.2379
rpe1	163	0.9139	0.8900	1.0539

The ordering fails. The train-only denominator places adamson above rpe1 (Table R.2), which breaks the monotone relation in Section 7 on all three support draws.

The held-out RMS column isolates the effect of changing the summary statistic. It replaces the mean of norms with a noise-debiased root mean square on the same held-out perturbations. All five remain strictly increasing. The reordering therefore comes from the population, not the summary statistic. Specifically, ρ_{shared} measured over the training perturbations represents a distinct quantity from ρ_{shared} measured over the held-out ones, most sharply on adamson, 0.73 against 1.24, the panel with 18 held-out perturbations against 68 training ones.

A practitioner therefore cannot derive this ρ_{shared} from their own training split. The stated recommendation remains benchmark-level reporting. ρ_{shared} uses no predictions, so it is not a function of any method, and the responses it does need are the candidate library a discrimination benchmark publishes anyway. Whoever assembles the panel can compute it once, before any method is scored.

S DEPLOYABLE CORRECTIONS AND ORACLE BOUNDS BY SETTING

Settings worsened by deployable corrections. We count these settings over the six support sizes of each panel, separately from the median recovery fractions in Section 7.

Table S.1: **Deployable corrections are worse than none more often as ρ_{shared} grows, and at every size on rpe1.**

dataset	ρ_{shared}	PDS@β_{train}	deployable fixed norm
arc-h1	0.225	0 of 6	0 of 6
replogle	0.390	0 of 6	0 of 6
norman	0.453	4 of 6	0 of 6
adamson	0.791	4 of 6	0 of 6
rpe1	0.914	6 of 6	6 of 6

Both columns are non-decreasing in ρ_{shared}. The fixed-norm column is 0, 0, 0, 0, 6 on each of the three support draws separately. The global correction worsens 14 of 30 settings and ties in one more (adamson at $k = 1$, where both scores equal 174/324). These counts complement median recovery, showing substantial signal recovery at small ρ_{shared} and worsening at large ρ_{shared} without forming ratios.

Table S.2 compares deployable corrections with oracle bounds across all six support sizes.

Oracle recovery and identity-free lift. The oracle column reads each perturbation's own true norm, and Appendix I shows that reading it lifts an identity-free prediction off the floor. The table puts the two quantities next to each other at $k = 1$. Recovery is the oracle score minus the reported score, the gap is the unshrunk baseline minus the reported score, and the lift is what the same oracle norm adds to the fitted shared response given to every perturbation.

Table S.2: **Deployable corrections and oracle bounds recover different amounts of discrimination.** The oracle column matches each prediction to the norm of its own true response and is not deployable. This is the correction evaluated by Liu et al. (2025).

dataset	k	PDS@1	PDS@β_{train}	fixed norm	oracle norm	$\overline{\text{PDS}}^{\star}_{\text{glob}}$	$\overline{\text{PDS}}^{\star}_{\text{pred}}$
arc-h1	1	0.5533	0.6944	0.7033	0.7067	0.7189	0.8978
arc-h1	2	0.6300	0.7844	0.8133	0.8144	0.8156	0.9156
arc-h1	4	0.7333	0.8844	0.8967	0.9033	0.9000	0.9511
arc-h1	8	0.8756	0.9411	0.9556	0.9567	0.9544	0.9878
arc-h1	16	0.9589	0.9656	0.9700	0.9722	0.9722	0.9811
arc-h1	32	0.9567	0.9700	0.9722	0.9722	0.9722	0.9800
replogle	1	0.5771	0.6058	0.6325	0.7687	0.7099	0.8737
replogle	2	0.6424	0.6869	0.7479	0.8611	0.8249	0.9197
replogle	4	0.7347	0.7977	0.8795	0.9330	0.9154	0.9645
replogle	8	0.8442	0.9182	0.9620	0.9721	0.9677	0.9858
replogle	16	0.9432	0.9706	0.9859	0.9887	0.9861	0.9930
replogle	32	0.9834	0.9889	0.9927	0.9936	0.9932	0.9961
norman	1	0.6543	0.6770	0.7101	0.8460	0.7446	0.9251
norman	2	0.7267	0.7545	0.8316	0.9133	0.8418	0.9601
norman	4	0.8386	0.8249	0.9225	0.9710	0.9311	0.9841
norman	8	0.9174	0.8903	0.9719	0.9885	0.9758	0.9927
norman	16	0.9809	0.9413	0.9927	0.9943	0.9930	0.9974
norman	32	0.9946	0.9627	0.9959	0.9974	0.9965	0.9984
adamson	1	0.5370	0.5370	0.6142	0.6049	0.6358	0.8457
adamson	2	0.5679	0.5957	0.7500	0.6728	0.7531	0.9043
adamson	4	0.6883	0.6327	0.7870	0.8086	0.8056	0.9198
adamson	8	0.8241	0.7377	0.8457	0.8611	0.8488	0.9414
adamson	16	0.9136	0.8426	0.9475	0.9568	0.9537	0.9846
adamson	32	0.9753	0.9012	0.9815	0.9846	0.9815	0.9938
rpe1	1	0.6187	0.5377	0.5538	0.7443	0.6195	0.8625
rpe1	2	0.6879	0.5594	0.5936	0.7898	0.6958	0.8863
rpe1	4	0.7838	0.5987	0.6870	0.8462	0.7873	0.9149
rpe1	8	0.8705	0.6775	0.8292	0.9070	0.8721	0.9497
rpe1	16	0.9374	0.8091	0.9272	0.9515	0.9377	0.9773
rpe1	32	0.9731	0.9242	0.9720	0.9771	0.9733	0.9906

Table S.3: **Oracle recovery can be smaller than its lift on an identity-free input.**

dataset	gap at $k = 1$	oracle recovery	share of the gap	lift with no identity
arc-h1	0.1678	0.1533	91.4%	0.0311
replogle	0.2471	0.1916	77.6%	0.1068
norman	0.2478	0.1916	77.3%	0.1492
adamson	0.2377	0.0679	28.6%	0.0031
rpe1	0.1880	0.1256	66.8%	0.1574

On rpe1, recovery in Table S.3 is smaller than the lift. No part can therefore be attributed to the correction independently of the norm it reads. The lift is not subtracted from recovery because it is measured on a different input. It has not been established as the share of a real prediction's recovery supplied by the true norm, so subtraction would lack a justified interpretation. The lift does not depend on k, since the identity-free prediction does not. The gap does vary. At $k = 32$, it falls below the 0.02 filter on all five panels, and the resulting share exceeds 100% on three. The table therefore uses $k = 1$, where every gap exceeds the filter many times over.

Table S.4 summarizes the absorbed fractions using the fixed gap filter.

Table S.4: **The absorbed fractions are medians over the 23 settings whose absolute reported gap exceeds 0.02.** Each gap is the empirical-Bayes estimator's score minus the unshrunk baseline's. The gap at 1 uses the reported score of both, and the gap at optimum uses each arm's own global optimum. Absorbed is $1 - |\text{gap at optimum}|/|\text{gap at 1}|$, and the per-dataset median over the rows marked yes, those with $|\text{gap at 1}| > 0.02$, gives the values reported in Section 7. These are arc-h1 95.3%, replogle 82.2%, norman 55.4%, adamson 41.2% and rpe1 1.3%.

dataset	k	base. PDS@1	base. $\overline{\text{PDS}}^{\star}_{\text{glob}}$	gap@1	gap@opt	abs. %	used
arc-h1	1	0.7211	0.7278	−0.1678	−0.0089	94.7	yes
arc-h1	2	0.8222	0.8233	−0.1922	−0.0078	96.0	yes
arc-h1	4	0.8944	0.8944	−0.1611	+0.0056	96.6	yes
arc-h1	8	0.9478	0.9500	−0.0722	+0.0044	93.8	yes
arc-h1	16	0.9689	0.9722	−0.0100	+0.0000	100.0	no
arc-h1	32	0.9667	0.9667	−0.0100	+0.0056	44.4	no
replogle	1	0.8242	0.8305	−0.2471	−0.1206	51.2	yes
replogle	2	0.8959	0.9053	−0.2534	−0.0804	68.3	yes
replogle	4	0.9376	0.9515	−0.2029	−0.0361	82.2	yes
replogle	8	0.9692	0.9778	−0.1250	−0.0101	91.9	yes
replogle	16	0.9862	0.9894	−0.0430	−0.0034	92.1	yes
replogle	32	0.9936	0.9940	−0.0102	−0.0008	91.9	no
norman	1	0.9021	0.9050	−0.2478	−0.1604	35.3	yes
norman	2	0.9423	0.9573	−0.2156	−0.1154	46.4	yes
norman	4	0.9767	0.9802	−0.1381	−0.0491	64.4	yes
norman	8	0.9911	0.9930	−0.0737	−0.0172	76.6	yes
norman	16	0.9939	0.9955	−0.0131	−0.0026	80.5	no
norman	32	0.9971	0.9981	−0.0026	−0.0016	37.5	no
adamson	1	0.7747	0.8179	−0.2377	−0.1821	23.4	yes
adamson	2	0.9012	0.9167	−0.3333	−0.1636	50.9	yes
adamson	4	0.9352	0.9506	−0.2469	−0.1451	41.2	yes
adamson	8	0.9136	0.9228	−0.0895	−0.0741	17.2	yes
adamson	16	0.9475	0.9599	−0.0340	−0.0062	81.8	yes
adamson	32	0.9784	0.9815	−0.0031	+0.0000	100.0	no
rpe1	1	0.8067	0.8068	−0.1880	−0.1874	0.3	yes
rpe1	2	0.8809	0.8818	−0.1930	−0.1860	3.6	yes
rpe1	4	0.9137	0.9143	−0.1300	−0.1270	2.3	yes
rpe1	8	0.9529	0.9535	−0.0824	−0.0813	1.3	yes
rpe1	16	0.9733	0.9738	−0.0359	−0.0361	-0.5	yes
rpe1	32	0.9890	0.9891	−0.0159	−0.0158	0.9	no

T DIRECTIONAL CHANGES FROM UNEQUAL GENE WEIGHTS

ρ_{shared} measures one way shrinkage turns a prediction, toward the shared response. A second route is not captured by ρ_{shared}. The empirical-Bayes weight B_g differs from gene to gene, so even with $\bar{\delta} = 0$ the map $\hat{\delta} \mapsto B\hat{\delta}$ rescales some coordinates more than others and changes direction. A constructed example demonstrates this distinction. Take four candidates of equal norm, $(1, 1)$, $(0, \sqrt{2})$, $(-1, -1)$ and $(0, -\sqrt{2})$, predictions equal to them, $\bar{\delta} = 0$ so that $\rho_{\text{shared}} = 0$, and $B = \text{diag}(0.1, 0.9)$. The score falls from 1.0000 to 0.8750. Equal candidate norms make every ranking invariant to a global scalar, so even the best rescaling recovers nothing. The scalar-contraction limit therefore needs B close to a multiple of the identity as well as ρ_{shared} small.

On these panels, B is far from uniform. Table T.1 measures the unevenness as the coefficient of variation of B over its positive entries at $k = 8$.

Table T.1: **Uneven gene weights do not order the observed magnitude recovery.**

dataset	ρ_{shared}	unevenness of B	absorbed
arc-h1	0.225	0.770	95.3%
replogle	0.390	0.744	82.2%
norman	0.453	0.952	55.4%
adamson	0.791	1.122	41.2%
rpe1	0.914	0.757	1.3%

Gene-weight heterogeneity is substantial on every panel, but it does not reproduce their ordering. The rank correlation of ρ_{shared} with absorbed fraction is -1.00, of unevenness with absorbed fraction -0.20, and of ρ_{shared} with unevenness $+0.20$. The two descriptors therefore do not substitute for one another. The data support a narrower account than a mechanism with one control variable. Both the shared center and unequal weights can turn a prediction, and on these five panels scalar recovery follows the route ρ_{shared} measures. A panel where unevenness dominates could break the ordering, and we have not observed one.

A direction-only descriptor does not give the same order. Systema measures how much of a panel is shared as the cosine between each perturbation's shift and the average perturbation effect, and sorts datasets by it (Viñas Torné et al., 2026). We computed that statistic (Table T.2) from the two inputs ρ_{shared} reads, the cosine of each held-out response with the fitted shared response $\bar{\delta}$, averaged over the held-out perturbations. This is a Systema-style statistic, not Systema's value. Systema weights perturbed cells, whereas this calculation weights perturbations.

Table T.2: **A direction-only descriptor does not reproduce the magnitude-recovery order.**

dataset	ρ_{shared}	mean cosine	median cosine	absorbed
arc-h1	0.225	0.153	0.143	95.3%
replogle	0.390	0.287	0.297	82.2%
norman	0.453	0.442	0.504	55.4%
adamson	0.791	0.579	0.719	41.2%
rpe1	0.914	0.445	0.516	1.3%

Cosine orders arc-h1, replogle and norman as ρ_{shared} does, but places rpe1 below adamson and within 0.003 of norman. Global rescaling, however, absorbs 1.3% of rpe1's gap, compared with 41.2% of adamson's and 55.4% of norman's. The median gives the same order, and the cosine to the held-out library mean in place of $\bar{\delta}$ moves rpe1 below norman as well. So on these panels the direction of the shared response does not carry what orders the absorbed fraction. ρ_{shared} compares the size of the shared part with the spread of the specific part, and rpe1's shared response is large against that spread without being better aligned with each response. Five panels determine one swap, making this an observation on these panels, not a test. The comparison and both possible outcomes were registered before computation.

U Separating the shared center and unequal weights

The empirical-Bayes transformation changes both the center of contraction and the relative gene weights. We vary these components independently after fitting, using

$$z(m_0, B_0) = m_0 + B_0(\hat{\delta} - m_0), \qquad m_0 \in \{0, \bar{\delta}\}, \quad B_0 \in \{\bar{B}I, B\}, \tag{U.1}$$

where $\bar{B}$ is the arithmetic mean of all diagonal entries, including zeros. The unshrunk prediction, candidate library and fitted map are identical across the four arms. No model is refitted. This gives 360 controlled-map arms at the 90 panel–support-size–draw settings. The scalar comparison preserves average contraction, not the full prior or fitted-estimator variance. These interventions identify sensitivity to two specified map components, not a unique additive cause of the original estimator's deficit.

We record the uncorrected ℓ_1, ℓ_2 and cosine scores, support-norm correction, reciprocal mean-weight rescaling, and the tie-optimistic ℓ_2 global bound. Empirical scores use the official package's

Table U.1: **The effect of unequal gene weights depends on the shared center.** Entries are the median [minimum, maximum] across six support sizes and three draws of the tie-optimistic global ℓ_2 bound difference, diagonal minus same-mean scalar weights. Interaction is the original-center contrast minus the zero-center contrast. Ranges describe repeated settings, not confidence intervals or new biological replicates.

Panel	Zero center	Original center	Interaction
arc-h1	$+0.013$ [-0.007, +0.043]	$+0.022$ [+0.002, +0.058]	$+0.008$ [-0.001, +0.023]
replogle	$+0.002$ [-0.001, +0.011]	$+0.109$ [+0.008, +0.176]	$+0.103$ [+0.008, +0.176]
norman	-0.011 [-0.055, +0.000]	$+0.078$ [+0.007, +0.130]	$+0.102$ [+0.007, +0.149]
adamson	-0.005 [-0.025, +0.015]	$+0.134$ [+0.046, +0.182]	$+0.136$ [+0.071, +0.185]
rpe1	$+0.008$ [+0.001, +0.020]	$+0.144$ [+0.078, +0.193]	$+0.136$ [+0.071, +0.189]

sorting convention. The global bound uses Equation (2), including its favorable treatment of ties, and is labeled separately. The zero-center scalar arm is an invertible contraction, so multiplication by $1/\bar{B}$ recovers the unshrunk prediction. All 90 mean weights are positive, with a minimum of 0.0237. Reconstructing the original map from its components agrees with the saved predictions to 5.0×10^{-7} or better, within their single-precision storage error. As an implementation check, a breakpoint event sweep agrees with the original global-bound enumerator to 10^{-14} on all four arms of two panels.

Table U.1 reports the diagonal-minus-scalar contrast under each center. At the original center, retaining unequal weights raises the global bound in all 90 settings, by 0.00222–0.19312. At zero center the contrast has both signs, with 52 positive, 31 negative and seven numerically zero settings. Restoring the original center lowers the bound in 89 scalar-weight settings, with one tie, and in 85 diagonal-weight settings, with four increases and one tie. The center-by-weight interaction is positive in 85 settings, negative in three and zero in two. Signs treat absolute differences at most 10^{-12} as numerical zero, below the smallest rank quantum in these panels. Figure 3c displays every setting. Here, fitted weights partly offset the shared center's effect. They are not uniformly harmful relative to scalar contraction of the same mean strength. This result limits a single-factor interpretation of the panel-level ρ_{shared} ordering.

The analysis code computes each perturbation's score in all four arms, absolute deficits and paired contrasts under every reported distance and correction. Within each panel and support size, 2,000 paired bootstrap samples resample perturbations after averaging their three draws. Intervals are conditional on the fixed candidate library and the fitted correction or selected global optimum. They do not refit the optimum or quantify generalization to a new panel. Draw-specific results are also computed. Interacting contrasts are not converted into percentages constrained to sum to the original deficit.

V SUPPORT-DRAW ROBUSTNESS

Evidence and uncertainty across five panels. The ordering by ρ_{shared} under the group optimum persists across three independent support draws on five datasets. The deployable fixed-norm form orders the datasets on two draws of the three, and on the remaining draw norman rises from third place to first (Table V.1). These repeated settings do not increase the number of independent panels. Pairing ρ_{shared} against the cosine gap gives a rank correlation of $+1.0$, but a paired bootstrap that resamples perturbations within each dataset returns a 95% interval containing zero (Appendix W.1). We therefore do not report the permutation p-value suggested by the point estimate. Adamson's 18 held-out perturbations give the widest interval. More perturbations there would improve within-panel precision, while additional panels are needed to test generalization. A dataset with small ρ_{shared} and poor global recovery, or the reverse, would break the ordering.

Prospective prediction and its registration limits. The descriptor was derived from the map and fixed before computation, not selected from a screen. Before the ordering appeared, the dated record contained only ρ_{shared} and a noise-debiased variant. Both were tested against the cosine gap, a different outcome for which a synthetic model predicted a rise followed by a fall. No panel lies on the falling part. The ordering first appeared on four datasets. Before arc-h1's scores were computed,

we predicted that arc-h1, whose $\rho_{\text{shared}} = 0.225$ is the smallest value, would show the highest global absorption, above the value then in second place. The recorded falsifier was absorption below 55.4%. The recipe, fixed at the same time, takes medians over settings with $|\text{gap@1}| > 0.02$ on unclipped deltas under ℓ_2. The measured value is 95.3%, consistent with the predicted position, whose null probability is $1/5$ if guessed. The registration had two limitations. A hand transcription recorded the second-place value as 87.1% without applying the specified filter, which gives 82.2%. Moreover, a value between 55.4% and that threshold would have broken the ordering without triggering either recorded rule.

Support redraws assess sensitivity to cell sampling within each dataset. They do not test generalization to a new dataset. The main paper's figure and medians use the first support draw. The two additional draws assess the ordering, inverse recovery across 90 settings and the single cut. Three criteria were specified before any second-draw results were generated. The ordering and inverse met their requirements. The single cut did not.

The runs repeat the protocol exactly, at support draws 2 and 7, on all five datasets at all six support sizes, for the shrunk estimator and the unshrunk baseline. That is 120 further result sets and, with the original draw, 90 settings. The rank-limited projector was not rerun because its primary results already contain per-perturbation scores for all three draws. All 120 new result sets reproduce the primary per-perturbation scores to 1.1×10^{-16}, confirming that the new predictions and the primary scores describe the same runs.

Ordering by ρ_{shared} (criterion 1). The requirement was strict ordering within draw 2 and draw 7 using the same summary fixed in advance, the median over settings with $|\text{gap@1}| > 0.02$ on unclipped deltas under ℓ_2.

Table V.1: **The ordering by ρ_{shared} holds on every draw under global rescaling, and not under the deployable fixed norm.**

		global rescaling (%)			deployable fixed norm (%)		
dataset	ρ_{shared}	draw 1	draw 2	draw 7	draw 1	draw 2	draw 7
arc-h1	0.225	95.3	89.2	93.2	+92.4	+77.1	+92.3
replogle	0.390	82.2	78.4	84.9	+71.4	+62.4	+70.8
norman	0.453	55.4	75.5	68.3	+54.7	+79.8	+52.4
adamson	0.791	41.2	53.2	46.3	+40.0	+48.6	+45.7
rpe1	0.914	1.3	−0.0	0.7	−48.9	−62.3	−57.9

Every column in the left block is strictly decreasing, meeting the criterion. The three draws add no points along the ρ_{shared} axis because ρ_{shared} is draw-invariant. They resample only the quantity being ordered. The ordering therefore persists across three support redraws at each of the five values. Absorbed fractions vary between draws, from 55.4 to 75.5 on norman and 41.2 to 53.2 on adamson. The point ordering repeats across draws but the values do not, so we do not quote a per-dataset absorbed fraction as a constant. Under query resampling only the extremes of that order are stable (Appendix W). The right block, the deployable form, is also non-monotonic on draw 2, where norman at 79.8 exceeds arc-h1 at 77.1 and inverts the top of the ordering. This comparison was exploratory and was not specified in advance. The deployable correction's ordering is therefore draw-dependent, as reported in the main text.

The transition the cut describes. On official ℓ_2 scores from clipped cells, the rank-limited projector outscores the invertible estimator in 60 of the 90 settings, all but one at small k, the exception being adamson at k=32 on one draw. Its predictions occupy 14 to 20 directions where the empirical-Bayes ones occupy 17 to 179, and the two estimators also differ in representation, so the comparison is between complete estimators and names no single cause. The mean shrinkage weight describes the projector's wins more closely than support size does. The transition falls at a different k on each dataset but at nearly the same mean shrinkage weight $\bar{B}(k)$. A single cut fitted on three datasets and locked at 0.2655 before the other two were seen is correct in 6 of 6 settings on arc-h1 and 5 of 6 on replogle. Because $\bar{B}$ is identical across draws, the cut predicts the same outcome for a dataset–support-size pair on every draw. Its correctness changes with the draw for 2 of the 30 pairs. Thus

$\bar{B}$ places three transition points on the k grid within an interval of width 0.06, without defining a draw-independent classification rule.

Single $\bar{B}$ cut (criterion 2). The requirement was at most 3 misclassifications by a single cut over all 90 settings. This limit scales the previously reported one-in-30 error rate to 90 settings and requires that the error rate not increase. The best pooled cut is 0.2844 and misclassifies 5 settings (Table V.2).

Table V.2: **The pooled cut exceeds the error limit.**

setting	$\bar{B}$		side of the cut
replogle k=16, draws 1, 2 and 7	0.3063	(3 settings)	misranked, above the cut
adamson k=32, draw 2	0.2868		misranked, above the cut
rpe1 k=8, draw 7	0.2442		not misranked, below the cut

The error count exceeds the limit. The best single threshold correctly classifies 85 of 90 settings and is conditional on the draw. The first-draw one-in-30 error rate does not generalize to the pooled analysis. The two classes now overlap on an interval of width 0.0620, against 0.0219 on the first draw alone.

$\bar{B}$ is fitted from the training responses, so it is exactly draw-invariant across all 30 dataset–support-size pairs. The range of $\bar{B}$ across the three draws is $0.000\mathrm{e}{+00}$, so the cut predicts the same outcome for a pair on every draw. Its correctness changes in 2 of those 30 pairs, adamson at k=32 and rpe1 at k=8, both of which sit inside the overlap interval. The 90 settings are not 90 independent tests. Three of the five errors arise from one persistent exception counted once per draw, replogle at k=16, the sole exception on the first draw. The pooled count remains the specified assessment across draws, and the single cut does not satisfy the error limit.

Known-map inverse (criterion 3). The requirement was 90 of 90 within 10^{-3}. The result is 90 of 90, largest residual 3.19×10^{-4}, mean -5.6×10^{-6}. The largest residual over the 90 settings equals that in the original 30 settings. The three largest are all on norman, where 13.10% of genes have $B_g = 0$. This agrees with the advance prediction because the inverse uses the known diagonal map and does not depend on which support cells were drawn.

Invariance of the fitted prior. The fitted shared response $\bar{\delta}$ and between-perturbation variance τ^2 are identical across all three draws and all six support sizes on all five datasets, with maximum absolute difference $0.000\mathrm{e}{+00}$. Because ρ_{shared} combines this fitted center with the fixed held-out candidate responses, it is also unchanged across support draws.

W RESAMPLING STABILITY OF THE SCALAR-RECOVERY ORDERING

The intervals elsewhere in the paper resample individual perturbations and do not quantify the absorbed-fraction ordering in Section 7. A pre-registered analysis therefore resampled each panel's held-out query perturbations 2,000 times with replacement, conditional on the fixed candidate library, as is every other interval in the paper. On norman the resampling unit is a target-connected component, since its 56 labels form 25 components that share atomic targets. One replicate per panel is shared by all support sizes, both arms and all three draws, because the held-out set is the same across them. Within each replicate, we recompute the denominator of ρ_{shared} over resampled responses and each setting's gap at the reported scale. Each arm's global ℓ_2 bound is recomputed by maximizing the reweighted mean score over the exact union of breakpoints in Proposition 1. We then reapply the 0.02 gap filter and recompute each panel's median absorbed fraction and the ordering. Unit weights reproduce every reported bound to 10^{-12} and all 15 reported medians exactly, and no replicate leaves a panel without a setting that passes the filter.

The order of the five ρ_{shared} values is unchanged in 89.6% of replicates. Ordering the panels by the resampled ρ_{shared}, the median absorbed fractions are strictly decreasing in 44.2%, 40.9% and 50.8% of replicates on draws 1, 2 and 7, and on all three draws together in 19.6% (Table W.1). Under the specified criterion, a proportion below one half requires describing the ordering as unstable to query

Table W.1: **The extremes of the scalar-recovery ordering are stable under query resampling and its middle is not.** Median absorbed fraction (%) with its 2.5–97.5% percentile interval over 2,000 joint resamples of the query perturbations, norman blocked by target component, each arm's global bound reoptimized. The ρ_{shared} interval comes from the same replicates.

panel	ρ_{shared}	draw 1	draw 2	draw 7
arc-h1	0.225 [0.197, 0.258]	95.3 [48.9, 97.4]	89.2 [71.2, 97.5]	93.2 [62.4, 97.7]
replogle	0.390 [0.377, 0.403]	82.2 [73.6, 89.1]	78.4 [69.4, 87.6]	84.9 [76.6, 91.6]
norman	0.453 [0.387, 0.520]	55.4 [25.7, 69.6]	75.5 [56.1, 91.7]	68.3 [39.7, 80.7]
adamson	0.791 [0.669, 0.957]	41.2 [5.7, 79.5]	53.2 [20.3, 87.0]	46.3 [14.6, 88.0]
rpe1	0.914 [0.883, 0.947]	1.3 [−1.3, 6.7]	0.0 [−2.3, 3.3]	0.7 [−0.8, 3.4]

resampling, not strict. Descriptively, with the panels in their point order, rpe1 is the lowest of the five in 97.9%, 100% and 99.8% of replicates, replogle exceeds norman in 99.9%, 61.4% and 98.2%, and the adjacent pairs arc-h1 and replogle, and norman and adamson, are ordered in 68–90% and 72–86%. The extremes therefore separate, while the order among the middle panels is not resolved by these data.

The same components change norman's other intervals moderately. For all 378 norman observed-score intervals in the matched-null analysis, the original perturbation-level interval was reproduced exactly. Blocked intervals are a median 1.21 times as wide (interquartile range 1.04 to 1.39, range 0.52 to 1.90). No conclusion in the paper rests on a norman interval bound. Rerunning the association bootstrap of Appendix W.1 with norman blocked leaves the median rank correlation at $+0.90$, narrows its 95% interval from $[0.00, 1.00]$ to $[0.30, 1.00]$ and lowers the share of replicates at exactly $+1.0$ from 23.1% to 19.4%.

W.1 Paired bootstrap for the cosine-gap association

One replicate resamples, within each dataset, its held-out perturbations with replacement, and recomputes both coordinates from that same resample. Joint resampling preserves the pairing that independent resampling of the two coordinates would break. Over 2000 replicates the rank correlation between ρ_{shared} and the cosine gap has median $+0.900$ with a 95% interval of $[0.000, 1.000]$, and exactly $+1.0$ in 23.1% of replicates. The two extreme datasets separate on both coordinates and the middle three do not. The interval is conditional on $\bar{\delta}$, which is fitted on training perturbations and held fixed, because the saved results do not carry the training deltas. It is also conditional on each panel's candidate library. Each perturbation's cosine gap comes from saved scores computed against the full held-out library. A replicate therefore reweights perturbations without rebuilding the library, rescoring or reapplying a gap filter. Perturbations are resampled one at a time, not by shared atomic target on norman. This interval is for the cosine gap. Appendix W gives the resampling interval of the absorbed fraction that Section 7 orders, and the association with norman blocked by target component.

W.2 Unresolved ordering under ℓ_1

The ordering in Section 7 is established under ℓ_2 and remains unresolved under ℓ_1, the distance used by the Virtual Cell Challenge. Two numerical searches did not meet validation criteria specified before the searches, so their outputs cannot establish the ordering.

Each attempt used Section 7's recipe with only the distance changed. Proposition 1 has no simple ℓ_1 counterpart. Under squared Euclidean distance the quadratic term cancels between two candidates and leaves one root per candidate. Under ℓ_1 each candidate difference is piecewise affine in the scalar, with a kink wherever a scaled prediction coordinate crosses a candidate coordinate, so an exact optimum requires enumerating those kinks and the roots between them. Each search therefore required validation under ℓ_2, where the exact optimum is known. No ℓ_1 result could be reported unless the search reproduced that optimum to 10^{-4}. One rank of one perturbation moves the mean score by $1/N^2$, from 3.1×10^{-3} on adamson to 3.05×10^{-5} on replogle, so this tolerance resolves a single rank on the three panels with $N \leq 56$ and admits up to two ranks on rpe1 and three on replogle. Neither search met that tolerance.

The first search, 61 scale values with four rounds of local refinement, missed the ℓ_2 optimum by up to 3.1×10^{-3} over the 90 settings. It misses at 34 of them by more than arithmetic noise, and every one of those 34 misses is an exact integer multiple of that panel's $1/N^2$, from one quantum to five, the largest being exactly $1/18^2$ on adamson. Four further settings differ by at most 2.2×10^{-16}, reflecting floating-point arithmetic, not a rank change. These discrete errors reflect missed rank transitions. The maximum is frequently attained exactly at a breakpoint, where a candidate is equidistant from the prediction and Equation (2) counts the tie in the prediction's favor. A grid does not guarantee evaluation at such isolated points. The second attempt instead compared the maximum over open intervals against the exact interior maximum. It also exceeded the tolerance. An interior maximum can occupy a short interval, which local refinement around the best sampled point can miss. It was stopped once 19 of its first 25 settings were already over the threshold, since the criterion concerns the maximum error and could no longer be met.

The unvalidated ℓ_1 medians from the first search are retained as evidence of the failed analysis. They place norman above replogle on all three draws, with inversions of 0.2, 9.2 and 5.4 points. A score error of 3.1×10^{-3} divided by gaps of 0.05 to 0.25 amounts to one to six absorbed-fraction points. The effect and error are comparable, precluding inference under the validation criterion.

Resolving the ordering requires a validated enumeration of the ℓ_1 breakpoints, the ratios Δ_{jg}/d_{ig} merged per candidate pair. A third attempt tested five versions of such an enumeration. The final version still disagreed with an independent brute-force enumeration on two of six toy panels, in opposite directions, despite reproducing Proposition 1 exactly under ℓ_2 on all six. It therefore yielded no reportable result. The failed implementation and its independent brute-force routine remain as historical source, not a runnable validation suite. These validation failures underpin the limitation in Section 8.

X CROSS-BATCH HOLDOUT

The response-side control uses a protocol different from the main paper's primary protocol. The following discrimination control uses two partitions derived from the primary protocol, retaining its held-out set, training cells, prediction controls and query cells while changing only the source of each support draw. The per-perturbation results allow the response-control summary table to be recomputed. Its exact original split specification was not retained, limiting reproduction of that run. The third-party arc-h1 input is not redistributed.

The few-shot splits draw support and query from the same experiment, so a batch effect shared by both could lift every method together. Only arc-h1 records a batch label (48 batches). The four genetic screens carry none and cannot be tested. On arc-h1 we re-score the 29 perturbations that keep at least eight cells in two batches under a batch-disjoint draw, taking support from one batch and query from the others. Table X.1 contrasts this with the standard draw on the same fixed perturbation set. The shrinkage advantage over the pooled mean shift is essentially unchanged, $+0.077/+0.052/+0.051$ at $k = 1/4/8$ under the batch-disjoint draw versus $+0.068/+0.051/+0.045$ standard, each with paired Wilcoxon $p < 0.003$. The response advantage is a property of the transport, not of support and query sharing a batch.

Batch-disjoint discrimination under the primary protocol. Discrimination retrieves candidates from the held-out set and therefore requires the primary library. We used two batch-aware partitions derived from the primary protocol, not the earlier response-control subset. They keep its 30 held-out perturbations, its training cells and so the prior, its 2000 prediction controls and each perturbation's 32 query cells, which is the candidate library, and they modify only where that perturbation's support originates. In the *free* arm the support is drawn only from batches holding none of its query cells, in the *shared* arm only from batches holding one, and the *standard* arm is the primary partition itself, whose support at $k = 8$ puts 0 to 7 of its 8 cells in the query's batches. All 30 perturbations have at least 53 cells in the first pool and 72 in the second, so none is dropped at any support size. The reported quantities, noise floor and interpretation criteria were fixed before the runs.

The three arms were evaluated at six support sizes on three draws, giving 54 scored runs on one machine. Four conditions were verified before interpreting scores. Every run met the leakage checks and used the official scorer. The candidate library and prior are identical across arms at each setting. Every recorded support is its partition's support prefix, of the right size, and obeys its arm's batch

Table X.1: **The response advantage persists under cross-batch holdout.** Evaluation on arc-h1 reports Pearson δ mean across seeds $\{1, 2, 7\}$ on the same 29 perturbations, where *std* denotes ordinary draws and *xb* denotes support and query from disjoint batches. The final two columns report the EB − mean-shift gap alongside its two-sided paired Wilcoxon p-value.

k	draw	EB-affine	affine OT	naive k-shot mean	trained pooler	Δ	p
1	std	0.230	0.145	0.163	0.187	+0.068	2e-3
	xb	0.233	0.145	0.156	0.195	+0.077	1e-3
4	std	0.327	0.240	0.277	0.264	+0.051	2e-3
	xb	0.320	0.232	0.268	0.241	+0.052	2e-3
8	std	0.395	0.302	0.350	0.316	+0.045	2e-3
	xb	0.390	0.294	0.339	0.325	+0.051	6e-4

rule. Rerunning the standard arm reproduces the primary results exactly, with 252 of 252 arrays and 1080 official per-perturbation rows identical.

Table X.2 gives the identity deficit per arm. The specified comparison contrasts the free and standard arms at every support size where the standard deficit exceeds 0.02. Its resolution floor is the larger of the two arms' ranges over the three draws. The outcome meets the criteria for a mixed result. Without batch overlap, the deficit is deeper at $k = 1$, 2 and 4, and shallower at $k = 8$, 16 and 32. Only the $k = 16$ difference exceeds the draw range, by 0.0011. Three draws on one panel give a range of 0.02 to 0.09 on differences of 0.004 to 0.034, so this panel cannot separate the two readings, and additional draws would be needed to improve within-panel resolution.

The control establishes a narrower result. The response advantage survives here too, and is larger under batch-disjoint support at every size except $k = 1$. The empirical-Bayes map outscores the unshrunk mean shift in Pearson δ by $0.082/0.080/0.067/0.055/0.042/0.028$ at $k = 1/2/4/8/16/32$ in the free arm against $0.085/0.066/0.057/0.043/0.033/0.022$ in the standard arm, each two-sided paired Wilcoxon over the 30 perturbations at $p \leq 1.4 \times 10^{-4}$. The identity deficit is present under every support pool. The exact inverse restores the unshrunk score in all 54 settings, to 0. Magnitude-correction recovery does not move with the support pool either, the global bound recovering 93.7%, 91.2% and 94.2% of the gap in the standard, free and shared arms and the deployable fixed norm 89.3%, 87.2% and 92.8%. The unshrunk estimate scores higher with support/query batch overlap at five of the six sizes, including 0.811 versus 0.798 at $k = 1$ and 0.963 versus 0.922 at $k = 8$. The difference is in the direction the confound predicts but is not large enough to shift the deficit at this resolution.

A batch-related interpretation of the identity deficit therefore remains open. Specifically, if perturbation identity is carried partly by batch structure that shrinkage removes, the deficit would overstate how much biological signal is lost, and this panel cannot say whether it does. Four of the five datasets have no batch labels, preventing a test of the confound there. This does not establish its absence.

Y Reproduction across execution environments

Table Y.1 summarizes agreement across execution environments.

Table Y.2 identifies the seven grid points with nonzero differences and the affected metrics.

The saved quantities per scored grid point are the predicted pseudobulk before and after clipping, the real pseudobulk, the predicted and real deltas, the control mean, the candidate distance matrices under all three distances, and for the empirical-Bayes estimator the diagonal B, the shared response $\bar{\delta}$, and the fitted variances. These are the pseudobulk quantities used by the score, not cell-level predictions. They require a few megabytes per setting, compared with roughly 260 GB of predicted cells per dataset. Saved results for all five panels total under a gigabyte.

Table X.2: **Batch-disjoint discrimination has a mixed outcome on arc-h1.** The identity deficit, PDS@1 of the empirical-Bayes map minus PDS@1 of the unshrunk mean shift under squared ℓ_2, median over draws $\{1, 2, 7\}$, for support drawn from all non-query cells (standard), from batches holding no query cell (free) and from batches holding one (shared). The floor is the larger of the standard and free arms' ranges over the three draws, the resolution three draws give. The comparison is free against standard wherever the standard deficit exceeds 0.02, covering every size except $k = 32$. Four of those five differences lie within the floor and the fifth exceeds it. The outcome is therefore mixed and supports neither interpretation.

k	standard	free	shared	floor	free − standard	verdict
1	-0.2311	-0.2356	-0.2589	0.0856	-0.0044	within
2	-0.2300	-0.2578	-0.2833	0.0844	-0.0278	within
4	-0.1744	-0.2089	-0.1933	0.0622	-0.0344	within
8	-0.1022	-0.0900	-0.1233	0.0411	$+0.0122$	within
16	-0.0222	-0.0033	-0.0233	0.0178	$+0.0189$	above
32	-0.0100	-0.0056	-0.0089	0.0078	$+0.0044$	not tested

Table Y.1: **Every reported discrimination score exactly reproduces the primary results at all 150 grid points.** Differences are confined to differential-expression quantities at the grid points listed below.

quantity	scope	value
grid points compared	5 datasets × 6 support sizes × 5 estimators	150
official metrics per grid point	per perturbation	28
missing-value masks identical	all grid points	150/150
$\mathrm{PDS}_{\ell_1}, \mathrm{PDS}_{\ell_2}, \mathrm{PDS}_{\cos}$ max abs diff	all grid points	0.000e+00
grid points differing on any metric		7/150

Y.1 Environment, compute and configuration

The experiments ran on three shared Linux hosts, enabling the cross-environment comparison above. Most settings were scored on two hosts, an Intel Xeon Platinum 8458P (176 cores, 1 TB RAM) under Rocky Linux 9.6 and an Intel Xeon Silver 4210R (40 cores, 126 GB RAM) under Ubuntu 20.04.6. One further host scored a single dataset. All three run Python 3.11.15. The environment specification pins the minor version, Python 3.11, and 3.11.15 is the patch these runs resolved to. It pins the scientific stack and scorer directly, but not transitive dependencies or the linear-algebra library. Exact discrimination-score agreement therefore rests on the cross-host comparison in Appendix Y, not on version pinning. The hosts are shared and oversized for this workload. These specifications describe the machines used, not minimum hardware requirements.

The scorer is the official cell-eval package, version 0.7.2. The supporting stack is NumPy 2.4.6, SciPy 1.17.1, pandas 2.3.3, scikit-learn 1.5.2, AnnData 0.12.16, Scanpy 1.11.5 and PyTorch 2.5.1.

CPU execution. The reported scores were computed on CPUs. Control-cell subsampling was the only stochastic operation in each scored run. Scores from runs on different hosts agree exactly in the reproduction table above, despite differences in visible GPU hardware.

The complete three-draw evaluation takes about six hours of wall-clock time in parallel across the three hosts. Contention from unrelated processes on the slowest host dominates this time. The same per-dataset workload took under an hour on each of the other two hosts. Every prior and pooler fit takes a few CPU-seconds, the closed-form prior a fraction of the trained pooler's. Neither is a bottleneck, and we make no efficiency claim from any of this.

Configuration. No hyperparameter search was performed for any reported result, and none of the values below was chosen post-hoc after observing outcomes. Each parameter is supplied explicitly and recorded for every scored run. Preprocessing takes the top $2{,}000$ highly variable genes, fitted on training cells only. The affine map works in a 64-dimensional control-fitted PCA latent. The distribution metrics use 50 components and 128 projection directions. The unshrunk k-shot mean and

Table Y.2: **All seven grid points with numerical differences concern differential-expression metrics, not discrimination.** Rerunning norman $k = 32$ twice on the same machine gave $0.000\mathrm{e}{+00}$ against the first run and 2.683×10^{-4} against the primary results. The estimator is therefore deterministic, with the difference arising between execution environments. The replogle entry is one unit in the last place.

dataset	estimator	k	max abs diff	metric carrying it
norman	affine OT	1	1.000e+01	the count of significant genes in the prediction
norman	affine OT	2	8.000e+00	the count of significant genes in the prediction
norman	affine OT	4	4.000e+00	the count of significant genes in the prediction
norman	affine OT	8	1.000e+00	the count of significant genes in the prediction
norman	affine OT	16	1.000e+00	the count of significant genes in the prediction
norman	affine OT	32	2.683e-04	the area under the receiver operating characteristic curve
replogle	EB-affine	2	9.714e-17	the rank correlation over significant genes

EB-affine have no tunable choices. Affine OT carries two fixed design choices, the 64-component latent and an eigenvalue floor of 10^{-4}, and projection carries its rank of 20. None was searched, so the five form a fixed set of estimators whose effect on the prediction is interpretable, not a tuned comparison. The trained pooler is scikit-learn's multilayer perceptron regressor with one hidden layer of 64 units. It uses $\alpha{=}10^{-3}$, at most 500 iterations, early stopping with patience 20 on a 10% validation split, and seed 7. The remaining library defaults are ReLU units, Adam at a learning rate of 10^{-3} and no input scaling. It is fitted inside the rank-20 train-only response subspace, with architecture $20 \to 64 \to 20$ and $2{,}644$ parameters. It is fitted at the same k it is scored at, on four k-shot draws per training perturbation, because the noise level of its input depends on k.